Journal of Statistical Software

April 2008, Volume 25, Issue 14. http://www.jstatsoft.org/

GEEQBOX: A MATLAB Toolbox for Generalized Estimating Equations and Quasi-Least Squares

Sarah J. Ratcliffe
University of Pennsylvania
School of Medicine

Justine Shults
University of Pennsylvania
School of Medicine

Abstract

The **GEEQBOX** toolbox analyzes correlated data via the method of generalized estimating equations (GEE) and quasi-least squares (QLS), an approach based on GEE that overcomes some limitations of GEE that have been noted in the literature. **GEEQBOX** is currently able to handle correlated data that follows a normal, Bernoulli or Poisson distribution, and that is assumed to have an AR(1), Markov, tri-diagonal, equicorrelated, unstructured or working independence correlation structure. This toolbox is for use with MATLAB.

Keywords: correlated data, longitudinal data, generalized estimating equations, quasi-least squares, MATLAB.

1. Introduction

The method of generalized estimating equations (GEE, Liang and Zeger 1986) is widely used because it allows for straight-forward analysis of correlated outcomes that can be discrete or continuous. GEE relies on the specification of the correlation structure, which results in some limitations. For example, Crowder (1995) used simple examples to demonstrate that if the pattern in the correlations is misspecified, there may be no solution (asymptotically) to the GEE moment-based estimating equation for the correlation parameter. In practice, this can result in failure to converge in a GEE analysis. Another limitation is that relatively few correlation structures have been implemented in the major statistical software packages that implement GEE. Although a simple structure is often reasonable to describe the expected pattern of associations, expansion of GEE for implementation of more complex structures could be beneficial when the correlations are of scientific interest, or a particular pattern is biologically plausible.

The method of quasi-least squares (QLS) is a two-stage approach for estimation of the correlation parameter in the framework of GEE that overcomes some of the limitations that were just described; see Chaganty (1997) for a description of stage one of QLS for data with an equal number of observations per subject (balanced data), Shults (1996) and Shults and Chaganty (1998) for stage one for unbalanced data, and Chaganty and Shults (1999) for stage two of QLS. First, QLS can sometimes yield meaningful results when GEE fails to converge, or when the estimated correlation matrix is not positive definite for GEE. For example, Shults *et al.* (2007) demonstrated that application of a simple (tri-diagonal) correlation structure in analysis of data from a study of obesity in children with renal disease resulted in a non-positive definite estimated correlation matrix for GEE; in contrast, the estimated correlation matrix for QLS (implemented in Stata, StataCorp. 2003) was positive definite. Next, QLS allows for relatively straightforward implementation of patterned correlation structures. For example, see Shults and Morrow (2002), Shults *et al.* (2004), and Shults *et al.* (2006) for studies whose analysis benefited from QLS with structures that previously had not been implemented in the framework of GEE.

In this manuscript, we consider a study that was described in Nunez-Anton and Woodworth (1994) and Chaganty and Shults (1999). In this trial, profoundly deaf subjects were surgically implanted with one of two types of hearing aids. Tests designed to measure hearing ability were then administered to the patients at 1, 9, 18 and 30 months post-implant. Because measurements from this trial were not equally spaced in time, it would be reasonable to consider application of a correlation structure that depends on the actual temporal spacing of measurements, in addition to the usual simple patterns for the correlations that are applied in a GEE analysis. For this reason, we demonstrate implementation via QLS of the Markov structure that was described in Naik and Prabhala (2002); the Markov structure assumes that the correlation between measurements depends on the temporal spacing of measurements and declines with increasing separation in time, which are both reasonable assumptions for data from this study.

This article presents the **GEEQBOX** toolbox for analysis of correlated data with GEE and QLS using the mathematical software MATLAB (The MathWorks, Inc. 2007). The toolbox currently allows for:

- three possible data distributions,
- six assumed correlation structures,
- estimation by either GEE or QLS.

This article does not replace the user guide or statistical documentation of QLS that the reader can find on the internet at `http://www.cceb.upenn.edu/~sratclif/QLSproject.html`. Rather, it provides an overview of the features of the **GEEQBOX** toolbox and some examples of its use.

The paper is organized as follows. Section 2 provides some notation and a brief description of the methods of GEE and QLS. Section 3 describes the technical features of the **GEEQBOX** toolbox. An example data analysis is shown in Section 4 to demonstrate the implementation of **GEEQBOX**. Discussion, including conclusions and plans for continued expansion of the toolbox, are then provided in Section 5.

2. A brief description of QLS and GEE

This section provides some notation; a description of the correlation structures that are implemented in **GEEQBOX**; and a summary of QLS and GEE. For more detail regarding QLS and GEE, please see the references that were provided in the introduction. In addition, see Hardin and Hilbe (2003) for an excellent and comprehensive text on GEE.

2.1. Notation

For analysis of a longitudinal study, we assume that measurements $Y_i = (y_{i1}, \cdots, y_{in_i})^\top$ and associated covariates $x_{ij} = (x_{ij1}, \cdots, x_{ijp})^\top$ were collected on subject i at times $T_i = (t_{i1}, \cdots, t_{in_i})^\top$, for $i = 1, \cdots, m$. The data are considered balanced and equally spaced when $n_i = n \; \forall \; i$ and $|t_{ij} - t_{ij-1}| = \gamma \; \forall \; i$, and j, respectively. For analysis of a cross-sectional study, e.g., if one measurement is collected on each of several subjects within multiple clusters, then $Y_i = (y_{i1}, \cdots, y_{in_i})^\top$ represents the n_i measurements that were collected within cluster i.

The expected value and variance of measurement y_{ij} on subject (or cluster) i are assumed to equal $E(y_{ij}) = g^{-1}(x_{ij}^\top \beta) = u_{ij}$ and $\text{Var}(y_{ij}) = \phi h(u_{ij})$, respectively, where ϕ is a known or unknown scale parameter. We also let $U_i(\beta)$ represent the $n_i \times 1$ vector of expected values u_{ij} on subject i. For longitudinal and cross-sectional studies, observations are assumed to be independent if they are measured on different subjects or clusters, respectively. However, within subjects or clusters, they are assumed to be correlated, with a pattern of association that can be described by a *working correlation structure*. The working structure for subject (or cluster) i, denoted by $\text{Corr}(Y_i) = R_i(\alpha)$, depends on a correlation parameter α that can be scalar or vector-valued. The covariance matrix of Y_i is then given by $\text{Cov}(Y_i) = \phi {A_i}^{1/2} R_i(\alpha) {A_i}^{1/2}$, where $A_i = \text{diag}(h(u_{i1}), \ldots, h(u_{in_i}))$.

Both GEE and QLS are iterative approaches that alternate between (1) updating the estimate of the regression parameter β by solving the GEE estimating equation for β and (2) updating the estimate of the correlation parameter α via moment estimation (GEE) or solving an unbiased estimating equation for α in two stages (QLS).

2.2. Working correlation structures

GEEQBOX currently implements the following structures, with plans to implement additional structures that will be made available on the web.

- Equicorrelated: This structure assumes that all pairwise correlations within a cluster are equal, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha$. This structure is plausible for cross-sectional studies, e.g., to describe the pattern of association of weights among litter-mates of baby rats.

- First-order autoregressive AR(1): This structure assumes that the correlation among repeated measurements on a subject depends on their separation in order of measurement, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha^{j-k}$. This structure is plausible for longitudinal studies in which the collection times of measurements are equally spaced in time, e.g., in a weight loss intervention that measures weights on subjects at baseline and then at three and six months post-baseline.

- Markov: This structure assumes that the correlation among repeated measurements on a subject depends on their timing of measurement, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha^{|t_{ij}-t_{ik}|}$. This structure generalizes the AR(1) structure to allow for unequal spacing of measurements. The estimate for α will be within the interval $(-1, 1)$. However, as for the AR(1) structure, a negative value for α is typically not biologically plausible. **GEEQBOX** therefore uses QLS to obtain an estimate of $\alpha \in (0, 1)$. We note that **GEEQBOX** does not implement the Markov structure for GEE because it is not straightforward to obtain a moment estimate for this structure.

- Tri-diagonal: This structure assumes that the correlation among measurements on a subject is constant for measurements that are separated by one measurement occasion, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha$ for $|j - k| = 1$ and is zero otherwise. The authors are not aware of many practical applications for this structure, but it was implemented in Liang and Zeger (1986) and in most standard software packages that implement GEE.

- Unstructured: This structure does not assume any pattern for the intra-subject correlations, so that $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk}$. This structure has been implemented in QLS (Chaganty 1997; Chaganty and Shults 1999) but the algorithms are somewhat complex. **GEEQBOX** therefore implements a moment estimate using GEE.

- Working Independent: Another popular structure is the identity matrix. Implementation of this structure is straightforward because β can then be estimated in a non-iterative process. However, several authors have shown that incorrect application of the working independence structure can result in a serious loss in efficiency in estimation of β (e.g., Sutradhar and Das 2000; Wang and Carey 2004; Shults *et al.* 2006)

2.3. GEE estimates of the correlation parameter

For GEE, **GEEQBOX** implements the following moment estimates that are implemented in `PROC GENMOD` in SAS (SAS Institute Inc. 2003).

For the equicorrelated structure, the GEE moment estimate is given by:

$$\hat{\alpha}_{\text{GEE}-\text{EQUI}} = \frac{\sum_{i=1}^{m} \sum_{j \neq k} z_{ij} z_{ik}}{(N^* - p)\hat{\phi}_{\text{GEE}}}$$

where

$$N^* = \sum_{i=1}^{m} n_i(n_i - 1),$$

$$\hat{\phi}_{\text{GEE}} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ij}^2}{N - p},$$

$N = \sum_{i=1}^{m} n_i$, z_{ij} is the Pearson residual for subject i at time t_{ij} and p is the dimension of β.

For the AR(1) and tri-diagonal estimates, the GEE moment estimate is:

$$\hat{\alpha}_{\text{GEE}-\text{TRI}} = \hat{\alpha}_{\text{GEE}-\text{AR1}} = \frac{\sum_{i=1}^{m} \sum_{j=2}^{n_i} z_{ij} z_{i,j-1}}{(N^{**} - p)\hat{\phi}_{\text{GEE}}}$$

where $N^{**} = \sum_{i=1}^{m}(n_i - 1)$.

For the unstructured correlation matrix, **GEEQBOX** implements the following moment estimate for element j, k of the matrix:

$$R_i\left[j,k\right] = \frac{\sum_{i=1}^{m} z_{ij} z_{ik}}{(m-p)\hat{\phi}_{\text{GEE}}}$$

A moment based estimator has not been proposed in the literature for implementation of the more general *Markov correlation* for GEE, which provides motivation for implementation of QLS.

2.4. QLS estimates of the correlation parameter

While GEE typically uses moment estimates for α, QLS estimates α by obtaining a solution to an unbiased estimating equation in two stages (see Sun *et al.* 2006, for more details). In stage one, QLS alternates between updating the estimates of β and solving the *stage one estimating equation* for α until convergence.

$$\frac{\partial}{\partial \alpha}\left\{\sum_{i=1}^{m} Z_i^{\top}(\beta)\left\{R_i^{-1}(\alpha)\right\} Z_i(\beta)\right\} = 0 \tag{1}$$

where $Z_i(\beta) = \left(z_{i1}, z_{i2}, \ldots, z_{in_i}\right)_{n_i \times 1}$ is the vector of Pearson residuals on subject i.

The solution $\hat{\alpha}$ to (1) is not consistent. Stage two of QLS therefore obtains a consistent estimate $\hat{\alpha}_{\text{QLS}}$ as the solution to the *stage two estimating equation* for α.

$$\sum_{i=1}^{m} trace\left\{\frac{\partial R_i^{-1}(\delta)}{\partial \delta} R_i(\alpha)\right\}\Bigg|_{\delta=\hat{\alpha}} = 0 \tag{2}$$

The final QLS estimate $\hat{\beta}_{\text{QLS}}$ of β is then obtained by solving the GEE estimating equation for β evaluated at $\hat{\alpha}_{\text{QLS}}$. For estimating equations that do not have a unique solution, **GEEQBOX** uses the bisection method to obtain a solution in the feasible region for α.

For the AR(1) structure and for unbalanced data, Shults and Chaganty (1998) proved that the feasible stage one estimate $\hat{\alpha}$ can be expressed as:

$$\hat{\alpha}_{\text{QONE}} = \frac{\sum_{i=1}^{m}\sum_{j=2}^{n_i}\left(z_{ij}^2 + z_{ij-1}^2\right) - \sqrt{\sum_{i=1}^{m}\sum_{j=2}^{n_i}\left(z_{ij} + z_{ij-1}\right)^2 \sum_{i=1}^{m}\sum_{j=2}^{n_i}\left(z_{ij} - z_{ij-1}\right)^2}}{2\sum_{i=1}^{m}\sum_{j=2}^{n_i} z_{ij} z_{ij-1}} \tag{3}$$

while the stage two estimate $\hat{\alpha}_{\text{QLS-AR1}}$ (Chaganty and Shults 1999) is given by

$$\hat{\alpha}_{\text{QLS-AR1}} = \frac{2\hat{\alpha}_{\text{QONE}}}{1 + \hat{\alpha}_{\text{QONE}}^2}. \tag{4}$$

For the Markov structure and unbalanced data, Shults and Chaganty (1998) provided the QLS stage one estimating equation for α:

$$\sum_{i=1}^{m}\sum_{j=2}^{n_i} \frac{e_{ij}\alpha^{e_{ij}}\left[\alpha^{2e_{ij}} z_{ij} z_{i,j-1} - \alpha^{e_{ij}}\left(z_{ij}^2 + z_{i,j-1}^2\right) + z_{ij} z_{i,j-1}\right] = 0}{\left(1 - \alpha^{2e_{ij}}\right)^2} \tag{5}$$

where $e_{ij} = |t_{ij} - t_{i,j-1}|$. Note that **GEEQBOX** requires that $e_{ij} \geq 1 \; \forall \, i$ and j. The stage two estimating equation for the Markov structure (Chaganty and Shults 1999) is given by:

$$\sum_{i=1}^{m} \sum_{j=2}^{n_i} \frac{2e_{ij}\delta^{2e_{ij}-1} - \alpha^{e_i j} e_{ij} \left[\delta^{e_{ij}-1} + \delta^{3e_{ij}-1}\right]}{\left(1 - \delta^{2e_{ij}}\right)^2} \Bigg|_{\delta=\hat{\alpha}} = 0 \tag{6}$$

For the equicorrelated structure and for unbalanced data, Shults (1996) proved that there will be a unique feasible solution to the following stage one estimating equation for α:

$$\sum_{i:n_i>1} Z_i^{\top} Z_i - \sum_{i:n_i>1} \frac{1 + \alpha^2(n_i - 1)}{(1 + \alpha(n_i - 1))^2} \left(Z_i^{\top}(\beta)\, e_i\right)^2 = 0 \tag{7}$$

where I_{n_i} is the identity matrix and e_i is a $n_i \times 1$ column vector of ones. Shults and Morrow (2002) obtained the stage two estimate $\hat{\alpha}_{\text{QLS}-\text{EQC}}$:

$$\sum_{i:n_i>1} \frac{n_i\,(n_i-1)\,\widehat{\alpha}\,\left(\widehat{\alpha}\,(n_i-2)+2\right)}{\left(1+\widehat{\alpha}(n_i-1)\right)^2} \Big/ \sum_{i:n_i>1} \frac{n_i\,(n_i-1)\,\left(1+\widehat{\alpha}^2(n_i-1)\right)}{\left(1+\widehat{\alpha}(n_i-1)\right)^2} \tag{8}$$

For the tri-diagonal structure and unbalanced data, **GEEQBOX** obtains solutions to the stage one and two estimating equations (1) and (2) for the tri-diagonal structure by first constructing the tri-diagonal matrix $R_i(\hat{\alpha})$. Next, to evaluate

$$\frac{\partial R_i^{-1}(\delta)}{\partial \delta}\Bigg|_{\delta=\hat{\alpha}}$$

GEEQBOX implements the following expression:

$$\frac{\partial R_i^{-1}(\delta)}{\partial \delta}\Bigg|_{\delta=\hat{\alpha}} = -R_i^{-1}(\hat{\alpha}) \left.\frac{\partial R_i(\delta)}{\partial \delta}\right|_{\delta=\hat{\alpha}} R_i^{-1}(\hat{\alpha})$$

where $\frac{\partial R_i(\delta)}{\partial \delta}$ is an $n_i \times n_i$ matrix with ones on the off-diagonal and zero elsewhere, i.e., the $(j,k)^{th}$ element of $\frac{\partial R_i(\delta)}{\partial \delta}$ is 1 if $|j-k| = 1$ and is 0 otherwise.

2.5. Testing hypotheses involving the regression parameter

The asymptotic distribution of the QLS estimate $\hat{\beta}_{\text{QLS}}$ is the same as the asymptotic distribution of the GEE estimate $\hat{\beta}_{\text{GEE}}$. **GEEQBOX** therefore provides both model-based and sandwich-based estimates of the covariance matrix of $\hat{\beta}$ (Liang and Zeger 1986). The covariance matrix depends on the scalar parameter ϕ; **GEEQBOX** implements the estimate provided in Chaganty and Shults (1999). The *model-based* estimate of the covariance matrix is appropriate when the user has a high degree of confidence that the correlation structure has been correctly specified. It has the following form:

$$\widehat{\text{Cov}}_M(\hat{\beta}) = \hat{\phi} {W_m}^{-1},$$

where

$$W_m = \sum_{i=1}^{m} X_i^{\top} {A_i}^{1/2} R_i^{-1}(\hat{\alpha}) {A_i}^{1/2} X_i$$

and $\hat{\phi} = \min\left\{\hat{\phi}_p, \hat{\phi}_c\right\}$, for

$$\hat{\phi_p} = \frac{1}{m}\sum_{i=1}^{m}\frac{Z_i(\hat{\beta})^\top Z_i(\hat{\beta})}{n_i} \text{ and } \hat{\phi_c} = \frac{1}{m}\sum_{i=1}^{m}\frac{Z_i(\hat{\beta})^\top R_i^{-1}(\hat{\alpha}) Z_i(\hat{\beta})}{n_i}.$$

The *robust sandwich* covariance matrix has the following form:

$$\widehat{\mathrm{Cov}}_R(\hat{\beta}) = {W_m}^{-1}\left\{\sum_{i=1}^{m} X_i^\top {A_i}^{1/2} R_i^{-1}(\hat{\alpha}) Z_i(\hat{\beta}) Z_i^\top(\hat{\beta}) R_i^{-1}(\hat{\alpha}) {A_i}^{1/2} X_i\right\} {W_m}^{-1}. \quad (9)$$

GEEQBOX provides estimated standard errors, 95% confidence intervals, and p values for the tests $\beta_j = 0$ that are based on both the *model* and *sandwich* covariance matrices.

3. Some technical features of the GEEQBOX toolbox

The development of the toolbox began in 2005, and is for use with MATLAB. It consists of two main functions, both of which can be called like any standard function in the MATLAB environment.

- `gee` function: calculates estimates using GEE.
- `qls` function: calculates estimate using QLS.

Both functions require the same inputs and produce the same layout of results. The required inputs for both functions are an $N \times 1$ vector of repeated measures outcomes, plus corresponding vectors of subject id's, measurement times, and a matrix of fixed effects.

3.1. Data representation

Both main functions require the same inputs: `id`, `y`, `t`, and `X`, in that order. For example, using the `gee` function the command would be `gee(id, y, t, X)`.

Each row of `X` should contain the observation or covariates associated with a single time point t_{ij}. The vectors `id` and `y` should contain the associated unique numerical identifier for the subject and outcome, respectively. Thus, these four inputs should have N (total number of observations across all subjects) rows.

In addition, the measurements must be sorted so that all measurements from the same subject (`id`) are listed on consecutive rows. If `id=[1 1 2 2 2 1 1]'`, then the program would count this as 3 subjects since there are 3 changes in id numbers. However, the id's do not have to be consecutive numbers. For example, `id=[12 12 12 10 10 10 99 99]'` would produce the same results as `id=[1 1 1 2 2 2 3 3]'`.

The matrix of covariates, `X`, should be set-up so that each column contains a separate covariate. At present, there should be no missing data in `X`. A constant term is not included by default in the programs. Thus, in order to include a constant in the model, a column of ones must be included as a covariate in `X`. This column of ones should be the *final* column of `X`. The programs will default to calling the beta estimate by the associated column number of `X`.

The functions also have a number of optional inputs to control the assumed distribution of the data and correction structure, as well as naming the fixed effects in the output and controlling the convergence tolerance and maximum number of iterations. The distribution can be specified by a single number of letter in the `family` input variable. The default distribution is a normal (`n` or `1`) distribution, but Bernoulli (`b` or `2`) or Poisson (`p` or `3`) may also be specified. The correlation structure is specified in the `corr` variable. The default correlation structure is AR(1) (`ar1` or `1`) for `gee` and Markov (`markov` or `2`) for `qls`. Any of the other correlation structures described in Section 2.2 can also be optionally chosen, and avilable options are listed in the help file for each function. Thus to specify a Poisson distribution with an equicorrelated (equi) correlation structure, we would use the command `gee(id, y, t, X, 'p', 'equi')`.

The default output display uses column numbers to label the variables in X. The third optional input `varnames` can be used to overwrite these display names. The `varnames` variable is structured variable with each item being a string variable. For example, the three columns of X could be labeled A, B, and C with the commands:

```
> varnames = $\{$'A', 'B', 'C'$\}$;
> gee(id, y, t, X, 'p', 'equi', varnames);
```

3.2. Output

Each function produces the same printed results and output variables. The printed results consist of the initial values used by the algorithms, the estimated covariance parameter (α), scale parameter (ϕ), and the covariate parameter estimates (β). In addition, the standard errors, corresponding z-values, p values and 95% confidence intervals for each β_j are also produced. Two versions of these values are presented; the one based on the robust covariance matrix and the one based on the model-based covariance matrix. The model-based results should be used when the specified working correlation matrix is known to be correct; otherwise, the robust results should be used.

Three variables are produced as outputs to the functions. These are the estimated β's, α, and a structured variable that contains the entire printed results from the robust estimations in the cell variable `results.robust` and the model-based estimation in the cell variable `results.model`.

4. Example

Here we present results obtained using **GEEQBOX** applied to data provided in Table 3 of Nunez-Anton and Woodworth (1994). This data set contains the following variables: subject id; group (A or B); month of measurement; and percentage. The variable percent represents the percent correct scores on a sentence test administered under audition-only conditions to groups of subjects wearing two different cochlear implants, referred to here as A and B. The electrode array was surgically implanted 5 to 6 weeks prior to being electrically connected to the external speech processor. Subjects were profoundly, bilaterally deaf, thus preconnection baseline values for the sentence test were all zero. At the time of the analysis reported here, data were available for 23 subjects in group A and 21 subjects in group B, with measurements scheduled at 1, 9, 18, 30 months after connection."(Nunez-Anton and Woodworth 1994)

In the worked examples presented here $\beta = (\beta_1, \beta_2, \beta_3)$ where β_1 is the regression coefficient associated with month of measurement; β_2 is the regression coefficient associated with group (group = 0 for A; group = 1 for B); and β_3 is the regression coefficient associated with cons, the constant that takes value one. The worked examples below are for the continuous outcome percentage as well as a binary variable, high, that takes value 1 for values of percentage ≥ 50, and takes value 0 otherwise. We note that our goal is not to present a complete analysis, but rather to demonstrate implementation of our toolbox and to highlight some of its features.

The data is contained in the files `audio.dat` (ASCII file) or `audio.mat` (MATLAB data file). The `X` matrix of interest is set as

```
> X = [month group cons];
```

with associated variable names for displaying the results:

```
> varnames = {'month', 'group', 'cons'};
```

For the continuous outcome percent, we can fit the model

$$\text{percent} = \beta X + \varepsilon$$

assuming a normal distribution for `Y` (*identity link*) and an *equicorrelated correlation* structure via QLS or GEE using the respective commands:

```
> [betahat, alphahat, results] = qls(id, percent, month, X, 'n', 'equi',
 varnames);
> [betahat, alphahat, results] = gee(id, percent, month, X, 'n', 'equi',
 varnames);
```

The resulting estimates of β with 95% confidence-intervals (low and up lim) are given in Table 1.

To demonstrate the sensitivity of results to choice of working correlation structure, we next fit the Markov correlation structure that is appropriate when the correlation between measurements declines with increasing separation in time. As discussed in the introduction, this structure is not readily applicable for GEE. We therefore implement the Markov structure using QLS via the following command:

```
> [betahat,alphahat,results] = qls(id, percent, month, X, 'n', 'markov',
 varnames);
```

The resulting estimates of β with 95% confidence-intervals (low and up lim) are given in Table 2.

We note that in this particular example, the results are not sensitive to the choice of equicorrelated versus Markov structure.

For the binary outcome, high (= 1 if percentage ≥ 50; = 0 otherwise) is modeled using an *equicorrelated* correlation structure and a *Bernoulli link* function. The associated commands for obtaining the QLS and GEE estimates are:

Method: QLS		Covariance matrix: robust				
	estimate	std. error	z value	p value	low lim	up lim
month	.90717187	.12838767	7.0658803	1.596e-12	.65553667	1.1588071
group	-11.887807	7.5787575	-1.5685693	.11674832	-26.741898	2.9662851
cons	28.493216	5.3826517	5.2935277	1.200e-07	17.943412	39.043019

Method: QLS		Covariance matrix: model-based				
	estimate	std. error	z value	p value	low lim	up lim
month	.90717187	.102903	8.8157964	0	.70548571	1.108858
group	-11.887807	7.6561579	-1.5527118	.12049201	-26.8936	3.1179872
cons	28.493216	5.4145422	5.2623499	1.422e-07	17.880908	39.105523

Method: GEE		Covariance matrix: robust				
	estimate	std. error	z value	p value	low lim	up lim
month	.90907324	.12876037	7.0601941	1.663e-12	.65670754	1.1614389
group	-11.941123	7.5661458	-1.5782306	.11451265	-26.770496	2.8882506
cons	28.494344	5.376535	5.2997599	1.160e-07	17.956529	39.032159

Method: GEE		Covariance matrix: model-based				
	estimate	std. error	z value	p value	low lim	up lim
month	.90907324	.08249241	11.020083	0	.74739108	1.0707554
group	-11.941123	7.816718	-1.5276389	.12660221	-27.261608	3.3793631
cons	28.494344	5.4789932	5.2006533	1.986e-07	17.755715	39.232974

Table 1: Estimates of β and associated 95% confidence-intervals (low and up lim) for the percent outcome assuming an equicorrelatd correlation structure.

Method: QLS		Covariance matrix: robust				
	estimate	std. error	z value	p value	low lim	up lim
month	1.0527	0.1399	7.5231	5.3513e-014	0.7785	1.3270
group	-12.0745	7.3972	-1.6323	0.1026	-26.5727	2.4237
cons	25.3058	5.1899	4.8760	1.0827e-006	15.1338	35.4778

Method: QLS		Covariance matrix: model-based				
	estimate	std. error	z value	p value	low lim	up lim
month	1.0527	0.1388	7.5848	3.3307e-014	0.7807	1.3248
group	-12.0745	7.6073	-1.5872	0.1125	-26.9845	2.8355
cons	25.3058	5.5109	4.5919	4.3917e-006	14.5046	36.1070

Table 2: Estimates of β and associated 95% confidence-intervals (low and up lim) for the percent outcome assuming a Markov correlation structure.

Method: QLS Covariance matrix: robust

	estimate	std. error	z value	p value	low lim	up lim
month	.05516138	.01580866	3.4893129	.00048426	.02417696	.08614579
group	-.97889627	.60696593	-1.6127697	.10679455	-2.1685276	.2107351
cons	-1.0439871	.41404722	-2.5214205	.01168821	-1.8555048	-.2324695

Method: QLS Covariance matrix: model-based

	estimate	std. error	z value	p value	low lim	up lim
month	.05516138	.0135889	4.0592958	.00004922	.02852762	.08179514
group	-.97889627	.59701443	-1.6396526	.10107742	-2.149023	.19123051
cons	-1.0439871	.41774854	-2.4990803	.01245161	-1.8627592	-.22521506

Method: GEE Covariance matrix: robust

	estimate	std. error	z value	p value	low lim	up lim
month	.05512979	.01595601	3.4551101	.00055007	.02385657	.086403
group	-.97956447	.60867271	-1.6093451	.10754089	-2.1725411	.21341212
cons	-1.0301008	.40730368	-2.529073	.01143642	-1.8284013	-.2318002

Method: GEE Covariance matrix: model-based

	estimate	std. error	z value	p value	low lim	up lim
month	.05512979	.01286563	4.2850434	.00001827	.02991361	.08034596
group	-.97956447	.60502254	-1.6190545	.10543554	-2.1653869	.20625793
cons	-1.0301008	.41823048	-2.4629978	.01377808	-1.8498174	-.21038408

Table 3: Estimates of β and associated 95% confidence-intervals (low and up lim) for the high outcome assuming an equicorrelatd correlation structure.

```
> [betahat, alphahat, results] = qls(id, high, month, X, 'b', 'equi',
 varnames);
> [betahat, alphahat, results] = gee(id, high, month, X, 'b', 'equi',
 varnames);
```

The resulting estimates of β with 95% confidence-intervals (low and up lim) are given in Table 3.

5. Discussion

The **GEEQBOX** MATLAB toolbox can be used to analyze correlated data via either the method of generalized estimating equations (GEE) or quasi-least squares (QLS). As demonstrated in Section 4, QLS allowed for implementation of the Markov structure, with similar results for the Markov and equicorrelated structures. If the results had differed, findings based on application of the Markov structure might be preferred because an assumption of equal correlations for all measurements within a subject (which is required for application of the equicorrelated structure) is very strong and might not be reasonable for test result data, i.e., we might anticipate that test results from two examinations that occur closely together

in time will be more similar, and therefore more highly correlated, than examinations that are taken further apart in time. In general, implementation of QLS will allow for consideration of correlation structures, including the Markov, that previously have not been available for GEE. Careful comparison of analysis results between several structures might strengthen confidence in a strong finding, e.g., if a results persists across several structures, or might result in the need to choose the most plausible structure, e.g., if the results are not consistent across structures.

Future and ongoing work of these authors will include implementation of correlation structures that have not previously been implemented in the framework of GEE. Related research will involve the development and comparison of methods for choosing between several correlation structures, which will be especially important for studies in which the findings differ according to choice of correlation structure. Future versions of the toolbox will be made available via the web site. These will also incorporate new correlation structures, add utility functions for manipulating and displaying the clustered data, methods for analyzing nested correlation structures and a window based "point-and-click" way of running the functions.

Acknowledgments

Work on this manuscript was supported by the NIH grant R01CA096885.

References

Chaganty NR (1997). "An Alternative Approach to the Analysis of Longitudinal Data via Generalized Estimating Equations." *Journal of Statistical Planning and Inference*, **63**, 39–54.

Chaganty NR, Shults J (1999). "On Eliminating the Asymptotic Bias in the Quasi-Least Squares Estimate of the Correlation Parameter." *Journal of Statistical Planning and Inference*, **76**, 127–144.

Crowder M (1995). "On the Use of a Working Correlation Matrix in Using Generalised Linear Models for Repeated Measures." *Biometrika*, **82**, 407–410.

Hardin J, Hilbe J (2003). *Generalized Estimating Equations*. Chapman and Hall/CRC, USA.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**, 13–22.

Naik D, Prabhala S (2002). "Prediction in Growth Curve Models Under Markov Covariance Structure." *Journal of Applied Statistical Science*, **11**, 245–254.

Nunez-Anton V, Woodworth G (1994). "Analysis of Longitudinal Data with Unequally Spaced Observations and Time-dependent Correlated Errors." *Biometrics*, **50**(2), 445–456.

SAS Institute Inc (2003). *The SAS System, Version 9.1*. Cary, NC. URL http://www.sas.com/.

Shults J (1996). *The Analysis of Unbalanced and Unequally Spaced Longitudinal Data Using Quasi-Least Squares*. Ph.D. thesis, Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia.

Shults J, Chaganty NR (1998). "Analysis of Serially Correlated Data Using Quasi-Least Squares." *Biometrics*, **54**, 1622–1630.

Shults J, Mazurick CA, Landis JR (2006). "Analysis of Repeated Bouts of Measurements in the Framework of Generalize Estimating Equations." *Statistics in Medicine*, **25**, 4114–4128.

Shults J, Morrow AL (2002). "Use of Quasi-Least Squares to Adjust for Two Levels of Correlation." *Biometrics*, **58**, 521–30.

Shults J, Ratcliffe SJ, Leonard M (2007). "Improved Generalized Estimating Equation Analysis via `xtqls` for Quasi-Least Squares in Stata." *Stata Journal*, **7**, 147–166.

Shults J, Whitt MC, Kumanyika S (2004). "Analysis of Data with Multiple Sources of Correlation in the Framework of Generalized Estimating Equations." *Statistics in Medicine*, **23**, 3209–3226.

StataCorp (2003). *Stata Statistical Software: Release 8*. StataCorp LP, College Station, TX. URL http://www.stata.com/.

Sun W, Shults J, Leonard M (2006). "Use of Unbiased Estimating Equations to Estimate Correlation in Generalized Estimating Equation Analysis of Longitudinal Trials." *Technical Report Working Paper 4*, University of Pennsylvania Biostatistics Working Papers.

Sutradhar BC, Das K (2000). "On the Accuracy of Efficiency of Estimating Equation Approach." *Biometrics*, **56**, 622–625.

The MathWorks, Inc (2007). *MATLAB – The Language of Technical Computing, Version 7.5*. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

Wang YG, Carey VJ (2004). "Unbiased Estimating Equations from Working Correlation Models for Irregularly Timed Repeated Measures." *Journal of the American Statistical Association*, **99**, 845–852.

Affiliation:

Sarah J. Ratcliffe
Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine
6th flr Blockley Hall, 423 Guardian Drive
Philadelphia, PA, 19104-6021, United States of America
Telephone: +1/215/573-7398
E-mail: sratclif@upenn.edu
URL: http://www.cceb.upenn.edu/~sratclif/QLSproject.html

Justine Shults
Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine
6th flr Blockley Hall, 423 Guardian Drive
Philadelphia, PA, 19104-6021, United States of America
Telephone: +1/215/573-6526
E-mail: jshults@mail.med.upenn.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 25, Issue 14 *Submitted:* 2007-04-13
April 2008 *Accepted:* 2008-03-28