University of Pennsylvania

UPenn Biostatistics Working Papers

Year 2006 *Paper* 13

Improved generalized estimating equation analysis via xtqls for implementation of quasi-least squares in Stata

Justine Shults[*] Sarah J. Ratcliffe[†]

Mary Leonard[‡]

[*]

[†]

[‡]

Improved generalized estimating equation analysis via xtqls for implementation of quasi-least squares in Stata

Justine Shults

and

Sarah J. Ratcliffe

and

Mary Leonard

Department of Biostatistics and Epidemiology
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
6th Floor Blockley Hall, 423 Guardian Drive
Philadelphia, Pennsylvania, 19104-6021, USA
Homepage: http://www.cceb.upenn.edu/faculty/?id=167
E-mail: shults@cceb.med.upenn.edu

Abstract. Quasi-least squares (QLS) is a method based on the popular generalized estimating equation (GEE) approach that is widely used for analysis of correlated cross-sectional and longitudinal data. This article summarizes the development of QLS that occurred in several manuscripts and describes its implementation with the user-written program xtqls in Stata. In addition, it demonstrates the following advantages of QLS: *(i)* QLS allows for implementation of some correlation structures that have not yet been implemented in the framework of GEE; *(ii)* QLS can be applied as an alternative to GEE if the GEE estimate is infeasible; and *(iii)* QLS is a method in the framework of GEE that uses the same estimating equation for estimation of β as GEE; as a result, implementation of QLS can involve programs already available for GEE. In particular, xtqls calls up the Stata program xtgee within an iterative approach that alternatives between updating estimates of the correlation parameter α and then using xtgee to solve the GEE estimating equation for β at the current estimate of α. The benefit of this approach is that following implementation of xtqls, all the usual post-regression estimation commands are readily available to the user. The xtqls program is available on the website for the Longitudinal Analysis for Diverse Populations project:

`http://www.cceb.upenn.edu/~sratclif/QLSproject.html.`

Keywords: correlated data; clustered data; longitudinal data; generalized estimating equations; quasi-least squares

1 Introduction

This manuscript describes the method of quasi-least squares and its implementation using the user-written program xtqls.

2 Methods

2.1 Set-up and notation

We consider the usual set-up for generalized estimating equations (GEE, Liang and Zeger, 1986), for which the data comprise correlated measurements collected on each of a group of independent clusters, or subjects. For example, consider a longitudinal study, in which serial measurements are collected on unrelated subjects at baseline and then at one and three months post-baseline. Or, consider a cross-sectional study of rats within litters, in which length and weight are measured once on all rats. In each of these studies it is reasonable to assume that measurements between the clusters (subjects or litters, respectively) are independent, but that within clusters they are correlated.

The typical goal of a GEE analysis is to relate the expected value of the outcome variable with covariates measured on each of the subjects, while adjusting for the potential correlation within the measurements on each cluster. The correlation is typically considered to be a nuisance parameter that is of secondary interest to the relationship between the outcome and covariates; however, the association can sometimes be of scientific interest. For example, in a cross-sectional study that relates the birth weight of rats with maternal feed during pregnancy, the degree of similarity of weights within litters might be important to assess.

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

A key feature of GEE is that the number of clusters should be relatively large in order for assumptions regarding the asymptotic properties of the estimators to be valid. A popular rule of thumb is that the data should contain at least 30 clusters; in general, the required sample size for a particular study will depend on the degree of correlation and the study design, as discussed in §2.4 of Diggle et al. (2002). Usually, the size of the clusters is small relative to the number of clusters; e.g. a typical longitudinal study of 30 subjects might contain 3 or 4 measurements per subject.

GEE analyses specify the relationship between the outcome and covariates measured on each subject by specifying a generalized linear model for the expected value of the outcome variable. In particular, the expected value and variance of measurement y_{ij} on subject (or cluster) i are assumed to equal $E(y_{ij}) = g^{-1}(x'_{ij}\beta) = u_{ij}$ and $Var(y_{ij}) = \phi h(u_{ij})$, 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.

Adjustment for the intra-cluster correlation of measurements is achieved by specifying a *working correlation structure* that describes the pattern of association between measurements within each cluster. The working structure for subject (or cluster) i, denoted by $Corr(Y_i) = R_i(\alpha)$, depends on a correlation parameter α that can be scalar or vector-valued. We note that α must take value in a particular region (the feasible region) in order for the correlation matrix to be positive definite. The covariance matrix of Y_i is then given by $Cov(Y_i) = \phi {A_i}^{1/2} R_i(\alpha) {A_i}^{1/2}$, where $A_i = diag(h(u_{i1}), \ldots, h(u_{in_i}))$ and ϕ is a scalar parameter that can be known or unknown.

Some useful correlation structures for analysis of correlated data include the following:

1. `The Equicorrelated (Exchangeable):` For this structure all correlations within a cluster are identical, so that $Corr(y_{ij}, y_{ik}) = \alpha$. This structure is often plausible in cross-sectional analyses, e.g. to describe the pattern of association of blood pressure among family members at baseline.The feasible region for this structure is $(-1/(n_m - 1), 1)$, where n_m represents the maximum value of n_i over $i = 1, 2, \ldots, m$.

2. `The first-order autoregressive AR(1):` For this structure the correlation among repeated measurements on a subject will be smaller for measurements that are farther apart in terms of order of measurement, so that $Corr(y_{ij}, y_{ik}) = \alpha^{j-k}$. This structure is often reasonable in longitudinal trials with equally spaced measurements, e.g. in a depression study in which Hamilton depression scores are measured at baseline and then once weekly for six months. The feasible region for this structure is $(-1, 1)$. However, a negative value for α may be biologically implausible because it may be unreasonable to allow the intra-subject correlations to alternate in sign, e.g. for α^2 and α^3 to be positive and negative, respectively.

3. `The Markov correlation structure:` For this structure the correlation among repeated measurements on a subject will be smaller for measurements that are farther apart in terms of timing of measurement, so that $Corr(y_{ij}, y_{ik}) = \alpha^{|t_{ij} - t_{ik}|}$. This structure generalizes the AR(1) structure to allow for unequal spacing of measurements. The feasible region for this structure is $(-1, 1)$. However, as for the AR(1) structure, a negative value for α is typically not biologically plausible.

4. `The tri-diagonal correlation structure:` For this structure the correlation among measurements that are separated by one measurement occasion will be constant, so that $Corr(y_{ij}, y_{ik}) = \alpha$ for $|j - k| = 1$ and is zero otherwise. This structure is not widely applied in practice, but it is implemented in Stata's xtgee procedure and in other standard software packages that implement GEE. The feasible region for this structure is $(-1/c_m, 1/c_m)$, where $c_m = 2\sin\left(\frac{\pi[n_m - 1]}{2[n_m + 1]}\right)$ and n_m is the maximum value of n_i over $i = 1, 2, \ldots, m$; this interval is approximately $(-1/2, 1/2)$ for large n and contains $(-1/2, 1/2)$ for all n.

5. `The unstructured correlation matrix:` For this structure there is no assumed pattern to the correlations within a subject, so that $Corr(Y_{ij}, Y_{ik}) =$

α_{jk}. This structure is typically reasonable for studies with a common set of timings of measurements for all subjects. Its drawback is that the dimension of the correlation parameter will be very large for clusters of even moderate size; e.g. a study with clusters of size $n = 5$ will require estimation of $\frac{n\times(n-1)}{2} = 10$ correlation parameters.

6. `The working independent correlation matrix:` Another popular structure is the identity matrix. It's application is straightforward because it does not involve estimation of any correlation parameters. However, incorrect application of an identity structure can result in loss in efficiency in estimation of the regression parameter, especially when the true correlations are large ; e.g. see Sutradhar and Das (2000) and Wang and Carey (2003).

2.2 Brief Review of the GEE approach

GEE (Liang and Zeger, 1986) is an iterative approach that alternates between (i) updating the estimate of the regression parameter β by solving the GEE estimating equation for β and (ii) updating the estimate of the correlation parameter α. Typically, moment estimates are used for estimation of α; the XT reference manual (2005) for Stata 9.0 describes the estimates that are implemented for GEE in the xtgee command in Stata, for the following correlation structures: the equicorrelated (exchangeable), AR(1), tri-diagonal (MA(1)), identity, and unstructured. The Stata estimates differ slightly from those suggested by Liang and Zeger (1986), as also mentioned in Section 2.3. The identity matrix can also be specified in Stata 9.0, but this does not require a special algorithm, since for this structure $\alpha = 0$.

The distribution of the GEE estimate of β, $\hat{\beta}_{GEE}$, is asymptotically normal. Stata 9.0, via xtgee and related commands, provides estimates of the *model-based* and *sandwich-type* estimates of the covariance matrix of $\hat{\beta}$. 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:

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

where

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

and

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{m} Z_i(\hat{\beta})' Z_i(\hat{\beta}).$$

The *robust sandwich* covariance matrix is typically applied when there is less certainty regarding the choice of working correlation structure. However, it has been the experience of these authors that standard errors are not necessarily smallest for this structure, so that application of the sandwich covariance matrix is not always the most conservative choice. The sandwich matrix takes the following form:

$$\widehat{Cov}_R(\hat{\beta}) = W_m{}^{-1} C_m W_m{}^{-1}, \tag{2}$$

where

$$C_m = \sum_{i=1}^{m} X_i' A_i{}^{1/2} R_i^{-1}(\hat{\alpha}) Z_i(\hat{\beta}) Z_i'(\hat{\beta}) R_i^{-1}(\hat{\alpha}) A_i{}^{1/2} X_i.$$

Stata 9.0 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 in GEE analyses.

2.3 Limitations to GEE that have been noted in the literature

GEE is one of the most widely applied and heavily cited statistical methods. For example, a search by these authors (in July, 2006) for the seminal paper on GEE (Liang and Zeger, 1986) on the ISI web of knowledge web-site yielded 4026 citations. However, GEE, like all statistical approaches, has some limitations. The first limitation concerns infeasibility of the moment estimates of α. Crowder (1995) noted that if the working correlation structure is misspecified, there may be no solution (asymptotically) to a moment-based estimating equation for α. In practice, this can result in failure to converge in a GEE analysis. Shults and Chaganty (1998) demonstrated that the Liang and Zeger (1986) suggested estimates for the AR(1) structure will often take value greater than one, especially for larger values of α. (However, we note that Stata 9.0 implements an algorithm by Newton (1988) for the AR(1) structure which, judging from the experience of these authors, does not have a problem with infeasibility (estimates $\hat{\alpha} > 1$).) In Section 4.2 we consider an obesity study in renal transplanted patients for which we demonstrate that the GEE estimate of α is infeasible for the tri-diagonal structure, so that the estimated correlation matrix is not positive-definite.

Another limitation of GEE is that relatively few correlation structures have been implemented in the major statistical software packages that implement GEE. For example, the Markov correlation structure is a relatively simple and useful structure that has not yet been implemented for GEE (Shults and Chaganty, 1998). Stata 9.0 currently implements only five correlation structures for GEE, in addition to the identity structure and a user-specified structure that is treated as fixed in the analysis. Although a simple structure is often reasonable to describe the expected pattern of associations, expansion of GEE analyses to incorporate more complex structures can be helpful, e.g. when the association is of scientific interest, or when a more complex structure is plausible for a particular study design. See Shults and Morrow (2002), Shults, Whitt, and Kumanyika (2004), and Shults, Mazurick and Landis (2006) for discussion of studies that benefited from analysis with more complex correlation structures than are typically implemented for GEE.

2.4 Overview of QLS approach

QLS is a two-stage approach in the framework of GEE that was described for balanced data (stage one) in Chaganty (1997); unbalanced data (stage one) in Shults (1996) and Shults and Chaganty (1998); and for unbalanced data (stage two) in Chaganty and Shults (1999). Like GEE, QLS is iterative and alternates between estimation of β and of α in each stage of the procedure. For estimation of β, QLS uses the same estimating equation as GEE. However, QLS differs from GEE with respect to estimation of the correlation parameter. Rather than implement the moment estimates that are typically implemented in a GEE analysis, QLS obtains a solution to an asymptotically unbiased estimating equation for α. In stage one, QLS alternates till convergence between updating the estimates of β and solving the *stage one estimating equation* for α:

$$\frac{\partial}{\partial \alpha}\left\{\sum_{i=1}^{m} Z_i'(\beta)\left\{R_i^{-1}(\alpha)\right\} Z_i(\beta)\right\} = 0, \tag{3}$$

where $Z_i(\beta) = (z_{i1}, z_{i2}, \ldots, z_{in_i})_{n_i \times 1}$ for $z_{ij} = (y_{ij} - u_{ij})/h(u_{ij})$ is the vector of Pearson residuals on subject i.

The solution $\hat{\alpha}$ to (3) is not consistent. Stage two of QLS therefore obtains a consistent estimate $\hat{\alpha}_{QLS}$ as the solution to the *stage two estimating equation* for α:

$$\sum_{i=1}^{m} trace\left\{\frac{\partial R_i^{-1}(\delta)}{\partial \delta} R_i(\alpha)\right\}\Bigg|_{\delta=\hat{\alpha}} = 0. \tag{4}$$

Section 3.5.2 provides solutions to (3) and (4) for several working correlation structures.

The final QLS estimate $\hat{\beta}_{QLS}$ of β is then obtained by solving the GEE estimating equation for β evaluated at $\hat{\alpha}_{QLS}$. The asymptotic distribution of $\hat{\beta}_{QLS}$ is the same as the asymptotic distribution of the GEE estimate $\hat{\beta}_{GEE}$. As a result, we demonstrate in Section 4.0 that testing and construction of confidence intervals for β with QLS is easily accomplished with the xtgee procedure in Stata 9.0 that implements GEE. Please see Sun et al. (2006) for more details about the QLS approach and a comparison with other methods.

2.5 How QLS expands implementation of GEE

In this manuscript we demonstrate that QLS can be used to expand implementation of GEE. First, in Section 4 we demonstrate that QLS can be successfully applied when GEE fails to yield a feasible estimate of α. QLS might therefore be considered as an alternative approach if $\hat{\alpha}$ is infeasible, or if the GEE iterative estimation procedure fails to converge.

Next, in Section 4 we demonstrate that QLS can be used to implement a useful and relatively simple structure (the Markov) that has not yet been implemented in the framework of GEE. This demonstrates that QLS can expand application of GEE by allowing consideration of patterns of association that are more complex than those currently available for GEE, but that are biologically plausible, or reasonable for a particular study design.

However, it is important to note that failure of GEE to converge or infeasibility of $\hat{\alpha}$ may be a sign that some model assumptions are wrong. For example, Prentice (1988) noted that $\hat{\alpha}$ must satisfy additional constraints to be feasible in analyses of binary data. Shults et al (2006) demonstrated that infeasibility of $\hat{\alpha}$ for binary outcomes can be very likely when the AR(1) structure has been misspecified as equicorrelated and α is large. Failure to converge, or infeasibility of $\hat{\alpha}$ should therefore prompt careful examination of the choice of working structure.

3 The xtqls command

3.1 Syntax

The xtqls command has following syntax which is very similar to the syntax for the xtgee procedure:

xtqls *depvar* [*indepvars*] , *options*

where *depvar* is the dependent variable; *indepvars* are the covariates; and *options* are the required options that are described below, in Section 3.3.

3.2 Description

The xtqls command provides QLS estimates of the regression and correlation parameter. QLS is a method in the framework of GEE, so that implementation of xtqls might be considered whenever GEE is appropriate and especially if GEE fails to converge, or if a correlation structure not available for GEE can be implemented in QLS. QLS allows for implementation of the equicorrelated, AR(1), Markov, and tri-diagonal correlation structures.

Implementation of an unstructured matrix is possible using QLS, but the algorithm is complex (Chaganty and Shults, 1998). For implementation of an unstructured matrix, we therefore recommend application of xtgee in Stata. In addition, the QLS and GEE procedures are identical for the identity matrix, so that use of xtgee is also recommended for implementation of an identity structure.

Future updates of the xtqls procedure are planned, to allow for implementation of additional structures with QLS.

3.3 Options

The *options* for xtqls (all required) are as follows:

i(*var1*) where *var1* is the ID variable for subjects, or clusters

t(*vart*) where *vart* is the variable that indicates the timings of observations

f(*family*) where *family* is the distribution of depvar. The following families can be implemented in xtqls:

gau (Gaussian)
bin (Bernoulli/binomial)
poi (Poisson)

c(*correlation*) where *correlation* is the correlation structure to be implemented. The following correlation structures can be implemented in xtqls:
AR 1 (AR(1))
sta 1 (tridiagonal)
exc (equicorrelated)
Markov (Markov)

vce(*vcetype*) where *vcetype* indicates the type of covariance structure for estimation of $\hat{\beta}$. The following covariance structures can be implemented in xtqls:

model (model based covariance structure)
robust (sandwich type robust sandwich covariance matrix)
jack (obtains jack-knife standard errors)
boot (obtains boot-strapped standard errors)

3.4 Relationship to the xtgee procedure

The xtqls procedure implements the xtgee procedure and has important similarities to the xtgee procedure. In particular, the syntax is as similar to the xtgee procedure as is possible. For example, the family names and names of the correlation structures (when they are available in xtgee) are identical to the names that are used in xtgee.

However, there are some differences between xtqls and xtgee: (1) Unlike xtgee which allows more flexibility in choice of link and variance functions, xtqls implements the canonical link function and corresponding variance function that is appropriate when Y_i is distributed according to an exponential family. For continuous (Gaussian) y_{ij} xtqls applies the identity link function $g^{-1}(\gamma) = \gamma$ and variance function $h(\gamma) = 1$. For binary (Bernoulli) y_{ij} xtqls applies the logistic link function $g^{-1}(\gamma) = exp(\gamma)/(1 + exp(\gamma))$ and variance function $h(\gamma) = \gamma(1 - \gamma)$. For count (Poisson) y_{ij} xtqls applies the exponential link $g^{-1}(\gamma) = exp(\gamma)$ and identity variance function $h(\gamma) = \gamma$. (2) Unlike xtgee which requires use of the force option for implementation of the AR(1) or tri-diagonal structures when the timings are unequally spaced, xtqls does not require this option for unequal timings. Rather, xtqls treats the observations as equally spaced when these two structures are specified. (3) Not all options that are available for xtgee are available for xtqls. The authors anticipate that future versions of xtqls will be more similar to xtgee than this initial version of the procedure. (4) For implementation of the tri-diagonal, equicorrelated, and tri-diagonal structures, implementation of xtqls can be noticeably slower than implementation of xtgee.

3.5 Methods and Formulas

3.5.1 The xtqls Algorithm for Estimation of the Correlation and Regression Parameters

The xtqls procedure implements the following algorithm for estimation of β and of α:

1. Obtain a starting value for $\hat{\beta}$ by assuming $\alpha = 0$ and then fitting a GEE model using xtgee in Stata, with the option corr(Ind) that indicates application of an identity working correlation structure.

2. Alternate between the following steps till convergence in the estimates of β:

   (a) Obtain updated values of the Pearson residuals at the current estimates of β and of α.

   (b) Update the estimate of α by obtaining the solution to the stage one estimating equation (3) for α.

   (c) Construct the estimated working correlation structure $R(\hat{\alpha})$ that corresponds to the updated estimate of α. For structures other than Markov, the matrix $R(\hat{\alpha})$ will be constructed for the maximum value of n_i. For example, in a study in which the maximum number of observations per subject is 4 and the working correlation structure is AR(1), $R(\hat{\alpha})$ will be a 4×4 AR(1) structure evaluated at $\hat{\alpha}$. For the Markov structure, the dimension of $R(\hat{\alpha})$ will equal the number of distinct values of the timing variable. For example, in a study in which some subjects are measured at times $(1, 2, 4)$ and all other subjects are measured at times $(1, 3, 9)$, the dimension of $R(\hat{\alpha})$ will be 5×5.

   (d) Update the estimate of β by using the xtgee procedure, with a correlation structure that is treated as fixed and equal to $R(\hat{\alpha})$.

3. After convergence in stage one, update the estimate of α by obtaining the solution to the stage two estimating equation (4) for α.

4. Construct the estimated working correlation structure $R(\hat{\alpha})$ that corresponds to the stage two estimate of α.

5. Obtain the final estimate of β by using the xtgee procedure, with a correlation structure that is treated as fixed and equal to $R(\hat{\alpha})$.

An important feature of this algorithm is its use of the xtgee procedure to update $\hat{\beta}$. As a result, as we demonstrate in Section 4, all the usual post regression commands in Stata are available after implementation of xtqls. We note that this algorithm was described in a presentation by the first author at the Stata 2004 User's Group Meetings in Boston, MA that is available on the following web-site:

`http://repec.org/nasug2004/Shults_Stata_2004.ppt.`

3.5.2 Stage One and Stage Two estimates of α

The xtqls procedure obtains provides solutions to the stage one (3) and stage two (4) estimating equations for several working correlation structures. For estimating equations that do not have a unique solution, xtqls 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:

$$\hat{\alpha}_{QONE} = \frac{\sum\limits_{i=1}^{m}\sum\limits_{j=2}^{n_i}(z_{ij}^2 + z_{ij-1}^2) - \sqrt{\sum\limits_{i=1}^{m}\sum\limits_{j=2}^{n_i}(z_{ij}^2 + z_{ij-1}^2)\sum\limits_{i=1}^{m}\sum\limits_{j=2}^{n_i}(z_{ij}^2 - z_{ij-1}^2)}}{2\sum\limits_{i=1}^{m}\sum\limits_{j=2}^{n_i} z_{ij}z_{ij-1}}, \tag{5}$$

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

$$\hat{\alpha}_{QLS-AR1} = \frac{2\hat{\alpha}_{QONE}}{1+\hat{\alpha}_{QONE}^2}. \tag{6}$$

For the Markov structure and unbalanced data, Shults (1996) obtained the QLS stage one estimating equation for α:

$$\sum_{i=1}^{m}\sum_{j=2}^{n_i}\frac{e_{ij}\alpha^{e_{ij}}\left[\alpha^{2e_{ij}}z_{ij}z_{i,j-1} - \alpha^{e_{ij}}\left(z_{ij}^2 + z_{i,j-1}^2\right) + z_{ij}z_{i,j-1}\right] = 0}{1-\alpha^{2e_{ij}}}, \tag{7}$$

where $e_{ij} = |t_{ij} - t_{i,j-1}|$. Note that xtqls 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:

$$\left.\sum_{i=1}^{m}\sum_{j=2}^{n_i}\frac{2e_{ij}\delta^{2e_{ij}-1} - \alpha^{e_{ij}}e_{ij}\left[\delta^{e_{ij}-1} + \delta^{3e_{ij}-1}\right]}{1-\delta^{2e_{ij}}}\right|_{\delta=\hat{\alpha}} = 0. \tag{8}$$

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:

$$\sum_{i:n_i>1} Z_i' Z_i - \sum_{i:n_i>1}\frac{1+\alpha^2(n_i-1)}{(1+\alpha(n_i-1))^2}\,(Z_i'(\beta)\,e_i)^2 = 0, \tag{9}$$

where I_{n_i} is the identity matrix and e_i is a $n_i \times 1$ column vector of ones. Shults and Morrow (C.3,2002) obtained the stage two estimate $\hat{\alpha}_{QLS-EQC}$:

$$\sum_{i:n_i>1}\frac{n_i\,(n_i-1)\,\widehat{\alpha}\,(\widehat{\alpha}\,(n_i-2)+2)}{(1+\widehat{\alpha}(n_i-1))^2} \Big/ \sum_{i:n_i>1}\frac{n_i\,(n_i-1)\,\left(1+\widehat{\alpha}^2(n_i-1)\right)}{(1+\widehat{\alpha}(n_i-1))^2}. \tag{10}$$

For the tri-diagonal structure and unbalanced data, Shults (1996) proved that there will always be a feasible solution to the stage one estimating equation for α. Xtqls obtains solutions to the stage one and two estimating equations (3) and (4) for the tri-diagonal structure by first constructing the tri-diagonal matrix $R_i(\hat{\alpha})$ and then using the Stata function **syminv** to obtain $R_i^{-1}(\hat{\alpha})$. Next, to evaluate

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

xtqls implements the following expression:

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

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.

3.6 Saved Results

The saved results for xtqls are the same as those for the xtgee procedure in Stata. For example, typing xtcorr will display the estimated correlation matrix.

4 Examples

Here we demonstrate implementation of xtqls command in Stata.

4.1 Data and variables

We will use the data set

```
random_small.dta
```

that is available on http://www.cceb.upenn.edu/∼sratclif/QLSproject.html . This contains data from a study of obesity in children following renal transplant that was conducted at the Children's Hospital in the University of Pennsylvania. To facilitate sharing of this data for the purpose of demonstrating the xtqls procedure, 10 percent of observations were dropped prior to saving the data set

```
random_small.dta.
```

(This was done by generating the variable *random* with the uniform command, sorting on the variable *random*, and then dropping all observations corresponding to *random* ≤ 0.1.)

A description of the data is as follows:

```
Contains data from random_small.dta
  obs:          531
 vars:            5                          20 Aug 2006 09:56
 size:       12,744 (98.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
```

```
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g                  subject id
month           float  %9.0g                  month of measurement
bmiz            float  %9.0g                  BMI z-score
basebmiz        float  %9.0g                  BMI z-score at baseline
obese           float  %9.0g                  1 if subject is obese/ 0 if not
                                                obese
-------------------------------------------------------------------------------
Sorted by:  id  month
```

For the examples we consider here, we will regress BMI z-score and obesity status (yes/no) on baseline BMI z-score and on month of measurement. We will demonstrate implementation of the robust sandwich-based covariance matrix and also of the model based covariance matrix.

4.2 An example where the GEE moment estimate is infeasible

If we regress BMI on time and baseline BMI then the feasible region (set of values on which α is positive definite) for the tri-diagonal structure is $(-0.51764, 0.51764)$. We first implement this structure using using Stata's xtgee procedure, with the sandwich-based covariance matrix:

```
. xtgee bmiz base month, i(id) t(month) f(gau) vce(robust) c(sta 1) force
```

```
 Iteration 1: tolerance = .0290889
 Iteration 2: tolerance = .0021742
 Iteration 3: tolerance = 1.180e-06
 Iteration 4:tolerance = 6.366e-09

GEE population-averaged model                 Number of obs      = 531
Group and time vars:          id month        Number of groups = 100
Link:                         identity        Obs per group: min = 2
Family:                       Gaussian                       avg = 5.3
Correlation:                  stationary(1)                  max = 11
                                              Wald chi2(2)    = 104.78
Scale parameter: .6754737                     Prob > chi2     = 0.0000

                                     (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semi-robust
        bmiz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    basebmiz |   .6350395   .0625064    10.16   0.000      .5125293    .7575498
       month |  -.0023181   .0033438    -0.69   0.488     -.0088719    .0042357
       _cons |   .9186753   .0619903    14.82   0.000     .7971766    1.040174
------------------------------------------------------------------------------
```

Note that above xtgee required the use of the option force because the timing variable month is not equally spaced on all subjects. Next, we will display the estimated correlation matrix:

.xtcorr

```
Estimated within-id correlation matrix R:

         c1       c2       c3       c4       c5       c6       c7       c8       c9
 r1  1.0000
 r2  0.8262   1.0000
 r3  0.0000   0.8262   1.0000
 r4  0.0000   0.0000   0.8262   1.0000
 r5  0.0000   0.0000   0.0000   0.8262   1.0000
 r6  0.0000   0.0000   0.0000   0.0000   0.8262   1.0000
 r7  0.0000   0.0000   0.0000   0.0000   0.0000   0.8262   1.0000
 r8  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.8262   1.0000
 r9  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.8262   1.0000
r10  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
r11  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000


        c10     c11
r10   1.0000
r11   0.8262 1.0000
```

The estimate $\hat{\alpha}_{GEE} = 0.8262$ which is outside the feasible region $(-0.51764, 0.51764)$ for the tri-diagonal structure.

Next, we will implement the tri-diagonal structure using the xtqls procedure with the sandwich-based covariance matrix. Note that this does not require the option force; as mentioned earlier, xtqls will treat the timings as equally spaced for the tri-diagonal and AR(1) structures.

. xtqls bmiz basebmi month, i(id) t(month) f(gau) vce(robust) c(sta 1)

```
Iteration 1: tolerance = .09658071
Iteration 2: tolerance =2.737e-16

GEE population-averaged model               Number of obs      = 531
```

```
Group and time vars:       id __00000S      Number of groups = 100
Link:                      identity         Obs per group: min = 2
Family:                    Gaussian                        avg = 5.3
Correlation:               fixed (specified)               max = 11
                                            Wald chi2(2)    = 94.09
Scale parameter:           .8811255         Prob > chi2     = 0.0000

                                     (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semi-robust
        bmiz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    basebmiz |   .6224297   .0738585     8.43   0.000     .4776697    .7671897
       month |   .0178934   .0036415     4.91   0.000     .0107561    .0250306
       _cons |   .7849147   .0760118    10.33   0.000     .6359344     .933895
------------------------------------------------------------------------------
```

As noted earlier, xtqls implements the xtgee procedure for a fixed correlation matrix. Therefore all usual post regression commands are available after implementation of xtqls. For example, if we use the xtcorr command to provide the estimated correlation matrix, we see that $\hat{\alpha}_{QLS} = 0.5176$ so that the estimated correlation parameter is within (but just barely) the feasible region for α.

.xtcorr

```
Estimated within-id correlation matrix R:

         c1       c2       c3       c4       c5       c6       c7       c8       c9
 r1  1.0000
 r2  0.5176   1.0000
 r3  0.0000   0.5176   1.0000
 r4  0.0000   0.0000   0.5176   1.0000
 r5  0.0000   0.0000   0.0000   0.5176   1.0000
 r6  0.0000   0.0000   0.0000   0.0000   0.5176   1.0000
 r7  0.0000   0.0000   0.0000   0.0000   0.0000   0.5176   1.0000
 r8  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.5176   1.0000
 r9  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.5176   1.0000
r10  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
r11  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000

        c10      c11
r10  1.0000
r11  0.5176   1.0000
```

When other structures were implemented, we observed that the variable month was only significant for the tri-diagonal structure and QLS. However, as discussed in Shults et al. (2006) infeasibility of $\hat{\alpha}_{GEE}$ or $\hat{\alpha}_{QLS}$ might be an indication that the correlation structure has not been correctly specified. Given that the tri-diagonal structure is not biologically plausible for this analysis, coupled with the fact that $\hat{\alpha}_{GEE}$ was infeasible for GEE, we would therefore be more inclined to accept the results of an analysis based on a more biologically plausible structure such as the Markov. We demonstrate implementation of the Markov structure in the next section.

4.3 Implementation of the Markov structure

Next, let's examine the spacing of measurements in this study. First, create a variable called lag that represents the spacing of measurements with respect to time:

```
. qui sort id month

. qui by id: gen lag = month - month[_n-1] if _n>1
```

Next, if we tabulate the variable lag we see that the spacing between measurements varies between 2 and 36 months.

```
. tab lag

        lag |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         81       18.79       18.79
          3 |         77       17.87       36.66
          5 |          6        1.39       38.05
          6 |         76       17.63       55.68
          9 |         11        2.55       58.24
         11 |          2        0.46       58.70
         12 |        160       37.12       95.82
         18 |          4        0.93       96.75
         24 |         12        2.78       99.54
         36 |          2        0.46      100.00
------------+-----------------------------------
      Total |        431      100.00
```

Application of the Markov structure is appropriate for this analysis, because as discussed earlier, this structure takes the variability of spacing of measurements into account. We next implement the Markov structure using the xtqls procedure. Here we implement that model based covariance matrix that is appropriate under the assumption that the correlation matrix has been correctly specified:

. xtqls bmiz basebmi month, i(id) t(month) f(gau) vce(model) c(Markov)

```
Iteration 1: tolerance = .08135458
Iteration 2: tolerance =6.117e-17

GEE population-averaged model              Number of obs      = 531
Group and time vars:        id month         Number of groups = 100
Link:                       identity       Obs per group: min = 2
Family:                     Gaussian                      avg = 5.3
Correlation: fixed (specified)                            max = 11
                                               Wald chi2(2)= 179.83
Scale parameter:                  .6887826      Prob > chi2 = 0.0000

------------------------------------------------------------------------------
        bmiz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    basebmiz |   .6438863   .0483796    13.31   0.000      .549064    .7387087
       month |   .0008804    .002362     0.37   0.709    -.003749    .0055099
       _cons |   .8149975    .080812    10.09   0.000     .6566088    .9733861
------------------------------------------------------------------------------
```

Next, we display the estimated correlation matrix:

. xtcorr

```
Estimated within-id correlation matrix R:

        c1       c2       c3       c4       c5       c6       c7       c8       c9
 r1  1.0000
 r2  0.9177   1.0000
 r3  0.8069   0.8792   1.0000
 r4  0.6237   0.6796   0.7730   1.0000
 r5  0.3727   0.4061   0.4619   0.5975   1.0000
 r6  0.2227   0.2426   0.2760   0.3570   0.5975   1.0000
 r7  0.1330   0.1450   0.1649   0.2133   0.3570   0.5975   1.0000
 r8  0.0795   0.0866   0.0985   0.1275   0.2133   0.3570   0.5975   1.0000
 r9  0.0475   0.0518   0.0589   0.0762   0.1275   0.2133   0.3570   0.5975   1.0000
r10  0.0284   0.0309   0.0352   0.0455   0.0762   0.1275   0.2133   0.3570
r11  0.0170   0.0185   0.0210   0.0272   0.0455   0.0762   0.1275 0.2133

        c10      c11
```

```
r10  1.0000
r11  0.5975  1.0000
```

It is interesting to note that the within subject correlations are quite high for this analysis.

4.4 Implementation of the AR(1) and equicorrelated structure with QLS

Let's next consider the outcome of obesity (1 = obese; 0 = not-obese) and implement the AR(1) and equicorrelated correlation structures with xtqls and the model based covariance matrix.

```
. xtqls obese basebmi month, i(id) t(month) f(bin 1) vce(model) c(AR 1)
```

```
Iteration 1: tolerance = .09449318
Iteration 2: tolerance = .0025892
Iteration 3: tolerance = .00016702
Iteration 4: tolerance =.00001117
Iteration 5: tolerance = 7.837e-07

GEE population-averaged model                Number of obs      = 531
Group and time vars:       id __00000S       Number of groups = 100
Link:                            logit       Obs per group: min = 2
Family:                       binomial                      avg = 5.3
Correlation: fixed (specified)                              max = 11
                                             Wald chi2(2)     = 35.66
Scale parameter: 1                                 Prob > chi2=0.0000

------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    basebmiz |   1.260941   .2115425     5.96   0.000     .8463256    1.675557
       month |   .0015922   .0067496     0.24   0.814    -.0116368    .0148212
       _cons |  -1.401252   .2658318    -5.27   0.000    -1.922272   -.8802308
------------------------------------------------------------------------------
```

The estimated correlation matrix for the AR(1) structure is then given by:

```
. xtcorr
```

```
Estimated within-id correlation matrix R:
```

```
         c1      c2      c3      c4      c5      c6      c7      c8      c9
 r1  1.0000
 r2  0.6987  1.0000
 r3  0.4882  0.6987  1.0000
 r4  0.3411  0.4882  0.6987  1.0000
 r5  0.2384  0.3411  0.4882  0.6987  1.0000
 r6  0.1666  0.2384  0.3411  0.4882  0.6987  1.0000
 r7  0.1164  0.1666  0.2384  0.3411  0.4882  0.6987  1.0000
 r8  0.0813  0.1164  0.1666  0.2384  0.3411  0.4882  0.6987  1.0000
 r9  0.0568  0.0813  0.1164  0.1666  0.2384  0.3411  0.4882  0.6987  1.0000
r10  0.0397  0.0568  0.0813  0.1164  0.1666  0.2384  0.3411  0.4882
r11  0.0277  0.0397  0.0568  0.0813  0.1164  0.1666  0.23840 0.3411

        c10     c11
r10  1.0000
r11  0.6987  1.0000
```

Note that if we had implemented the AR(1) structure using xtgee then 97 subjects would have been dropped from the analysis, due to unequal spacing of measurements. Or, we could have used the option force, which would have treated all observations as equally spaced. (As mentioned earlier, implementation of the AR(1) structure with xtqls will not require the force option because it will automatically treat the observations as equally spaced for the AR(1) structure.)

Next, we will implement the equicorrelated correlation structure, when the outcome is obesity and with the model based covariance matrix:

```
. xtqls obese basebmi month, i(id) t(month) f(bin 1) vce(model) c(exc)
```

```
Iteration 1: tolerance = .05135684
Iteration 2: tolerance =.03097018
Iteration 3: tolerance = .00177796
Iteration 4: tolerance =.00006787
Iteration 5: tolerance = 8.058e-06
Iteration 6: tolerance =9.287e-07

GEE population-averaged model               Number of obs      = 531
Group and time vars:       id __00000S      Number of groups = 100
Link:                      logit            Obs per group: min = 2
Family: binomial                                           avg = 5.3
Correlation: fixed (specified)                             max = 11
                                                Wald chi2(2) = 35.38
Scale parameter:              1                 Prob > chi2  =0.0000
```

```
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    basebmiz |    1.37291   .2329508     5.89   0.000     .9163352    1.829486
       month |  -.0059594   .0045567    -1.31   0.191    -.0148904    .0029716
       _cons |  -1.334806   .2574099    -5.19   0.000    -1.839321   -.8302924
------------------------------------------------------------------------------
```

Let's next display the estimated correlation matrix.

```
   . xtcorr
```

```
Estimated within-id correlation matrix R:

         c1      c2      c3      c4      c5      c6      c7      c8      c9
 r1  1.0000
 r2  0.5065  1.0000
 r3  0.5065  0.5065  1.0000
 r4  0.5065  0.5065  0.5065  1.0000
 r5  0.5065  0.5065  0.5065  0.5065  1.0000
 r6  0.5065  0.5065  0.5065  0.5065  0.5065  1.0000
 r7  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  1.0000
 r8  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  1.0000
 r9  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  1.0000
r10  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065
r11  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065  0.5065 0.5065

        c10     c11
r10  1.0000
r11  0.5065  1.0000
```

5 Discussion

In this paper we have implemented quasi-least squares using the user-written xtqls procedure in Stata. As we demonstrated, this allows for implementation of correlation structures such as the Markov that have not yet been implemented in the framework of GEE. In addition, may provide a feasible estimate when the GEE estimate is infeasible, or GEE fails to converge. The xtqls procedure calls up the xtgee procedure and therefore all the usual post-regression estimation commands are available after implementation of xtqls. Future updates of xtqls will implement additional correlation structures, including the banded

Toeplitz and other structures that are appropriate for data with multiple levels of correlation.

6 Acknowledgments

Work on this manuscript was supported by the NIH funded grant R01CA096885 “Longitudinal Analysis for Diverse Populations”.

References

Chaganty, N. R. 1997. An alternative approach to the analysis of longitudinal data via generalized estimating equations. 1997 *Journal of Statistical Planning and Inference* 63: 39-54.

Chaganty, N.R. and 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.

Diggle, P.J., Heagerty, P., Liang, K.Y., and Zeger, S.L. 2002. *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Liang, K.-Y. and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13-22.

Newton, H.J. 1988. *TIMESLAB: A time series analysis laboratory.* Belmont, CA: Brooks/Cole.

Prentice, R.L. 1988. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44: 1033-1048.

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. and Chaganty, N.R. 1998. Analysis of serially correlated data using quasi-least squares. *Biometrics* 54: 1622-1630.

Shults, J. Mazurick, C.A. and Landis, J.R. Analysis of repeated bouts of measurements in the framework of generalized estimating equations. *Statistics in Medicine, in press.*

Shults, J. and Morrow, A. 2002. Use of Quasi-Least Squares to Adjust for Two Levels of Correlation. *Biometrics,* 58: 521530.

Shults, J. Sun, W. and Amsterdam, J. 2006. On the violation of bounds for the correlation in generalized estimating equation analysis of binary data from longitudinal trials. *http://www.biostatsresearch.com/upennbiostat/papers/art8.*

Shults, J., Whitt, M.C. and Kumanyika, S. 2004. Analysis of data with multiple sources of correlation in the framework of generalize estimating equations. *Statistics in Medicine* 23(20): 3209-3226.

Stata XT manual for Longitudinal/Panel Data. 2005. Stata Press: College Station, Texas.

Sun, W., Shults, J., and Leonard, M., 2006. Use of Unbiased Estimating Equations to Estimate Correlation in Generalized Estimating Equation Analysis of Longitudinal Trials. *http://www.biostatsresearch.com/upennbiostat/papers/art4*.

Sutradhar, B.C. and Das, K. 2000. On the accuracy of efficiency of estimating equation approach. *Biometrics* 56: 622-625.

Wang, Y.G. and Carey, V.J. Working correlation misspecification, estimation and covariate design: implications for generalized estimating equation performance. Biometrika 2003; 90: 29-41.

7 About the Authors

Justine Shults, Ph.D. and Sarah Ratcliffe, Ph.D. are Assistant Professors of Biostatistics and Mary Leonard, M.D., M.S.C.E, is an Associate Professor of Epidemiology and Pediatric Nephrology in the Center for Clinical Epidemiology and Biostatistics, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine. In addition, Mary Leonard, M.D., M.S.C.E is a faculty member in the Department of Pediatrics, University of Pennsylvania School of Medicine.