The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

The Stata Journal (2007)
7, Number 2, pp. 147–166

Improved generalized estimating equation analysis via xtqls for quasi–least squares in Stata

Justine Shults, Sarah J. Ratcliffe, and Mary Leonard
Department of Biostatistics and Epidemiology
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
Philadelphia, PA
jshults@mail.med.upenn.edu

Abstract. Quasi–least squares (QLS) is an alternative method for estimating the correlation parameters within the framework of the generalized estimating equation (GEE) approach for analyzing correlated cross-sectional and longitudinal data. This article summarizes the development of QLS that occurred in several reports and describes its use with the user-written program `xtqls` in Stata. Also, it demonstrates the following advantages of QLS: (1) QLS allows some correlation structures that have not yet been implemented in the framework of GEE, (2) QLS can be applied as an alternative to GEE if the GEE estimate is infeasible, and (3) QLS uses the same estimating equation for estimation of β as GEE; as a result, QLS can involve programs already available for GEE. In particular, `xtqls` calls the Stata program `xtgee` within an iterative approach that alternates between updating estimates of the correlation parameter α and then using `xtgee` to solve the GEE for β at the current estimate of α. The benefit of this approach is that after `xtqls`, all the usual postregression estimation commands are readily available to the user.

Keywords: st0122, xtqls, correlated data, clustered data, longitudinal data, generalized estimating equations, quasi–least squares

1 Introduction

This article describes the method of quasi–least squares (QLS) and the user-written program `xtqls`.

2 Methods

2.1 Setup and notation

We consider the usual setup for generalized estimating equations (GEEs; Liang and Zeger 1986), for which the data comprise correlated measurements collected on each of a group of independent clusters, or subjects. Consider a longitudinal study in which serial measurements are collected on unrelated subjects at baseline and then at 1 and 3 months

postbaseline. Or consider a cross-sectional study of rats within litters in which length and weight are measured once on all rats. In both studies, assuming that measurements between the clusters (subjects or litters, respectively) are independent but that within clusters they are correlated is reasonable.

The typical goal of a GEE analysis is to relate the expected value of the outcome variable with covariates measured on each subject while adjusting for the potential correlation within the measurements on each cluster. The correlation is considered 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 feeding during pregnancy, the degree of similarity of weights within litters might be important to assess.

For notation, we assume that measurements $Y_i = (y_{i1}, \cdots, y_{in_i})'$ and associated covariates $x'_{ij} = (x_{ij1}, \ldots, x_{ijp})$ are collected on subject i at times $T_i = (t_{i1}, \ldots, t_{in_i})'$, for $i = 1, \ldots, m$. The data are considered balanced and equally spaced when $n_i = n \; \forall \; i$ and $|t_{ij} - t_{ij-1}| = \gamma \; \forall \; i$ and $j = 2, \ldots, n_i$, respectively. For analysis of a cross-sectional study, e.g., if one measurement is collected on each of several subjects within multiple clusters, then $Y_i = (y_{i1}, \ldots, y_{in_i})'$ 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 GEEs is that the number of clusters should be relatively large for assumptions regarding the asymptotic properties of the estimators to be valid. A popular rule 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 section 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 three or four 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 $\text{Var}(y_{ij}) = \phi h(u_{ij})$, 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.

Adjustment for the intracluster 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 $\text{Corr}(Y_i) = R_i(\alpha)$, depends on a correlation parameter α that can be scalar- or vector valued. α must take a value in a particular region (the feasible region) for the correlation matrix to be positive definite. 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 $A_i = \text{diag}\{h(u_{i1}), \ldots, h(u_{in_i})\}$ and ϕ is a scalar parameter that can be known or unknown.

Some useful correlation structures for analyzing correlated data include the following:

1. The equicorrelated (exchangeable). All correlations within a cluster are identical, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha$. 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)). The correlation among repeated measurements on a subject will be smaller for measurements that are farther apart in order of measurement, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha^{j-k}$. 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 6 months. The feasible region for this structure is $(-1, 1)$. However, a negative value for α may be biologically implausible because allowing the intrasubject correlations to alternate in sign, e.g., for α^2 and α^3 to be positive and negative, respectively, may be unreasonable.

3. The Markov correlation structure. The correlation among repeated measurements on a subject will be smaller for measurements that are farther apart in 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 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 tridiagonal correlation structure. The correlation among measurements that are separated by one measurement occasion will be constant, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha$ for $|j - k| = 1$ and is zero otherwise. This structure is not widely applied in practice, but it is implemented in Stata's `xtgee` command and in other standard software packages that implement GEE. The feasible region for this structure is $(-1/c_m, 1/c_m)$, where $c_m = 2\sin(\pi[n_m - 1]/2[n_m + 1])$ 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. The within-subject correlations have no assumed pattern, so that $\text{Corr}(y_{ij}, y_{ik}) = \alpha_{jk}$. 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 large for clusters of even moderate size; e.g., a study with clusters of size $n = 5$ will require estimation of $\{n \times (n - 1)\}/2 = 10$ correlation parameters.

6. The working independent correlation matrix. Applying the identity matrix is straightforward because it does not involve estimation of any correlation parameters. However, incorrect application of an identity structure can cause 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 Review of GEE

GEE is widely used because it extends generalized linear models to correlated data; for a thorough discussion of GEE, see Hardin and Hilbe (2003). In the following, we refer to GEE as the iterative approach for estimation developed by Liang and Zeger (1986) that alternates between (1) updating the estimate of the regression parameter β by solving the GEE for β and (2) updating the estimate of the correlation parameter α. Typically, moment estimates are used for estimation of α; StataCorp (2005) describes the estimates that are implemented for GEE in the `xtgee` command in Stata 9.2 for the following correlation structures: the equicorrelated (exchangeable), AR(1), tridiagonal (first-order moving average), 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 9.2, but doing so does not require a special algorithm, since for this structure $\alpha = 0$.

The distribution of the GEE estimate of β, $\widehat{\beta}_{\text{GEE}}$, is asymptotically normal. Stata 9.2, via `xtgee` and related commands, provides estimates of the *model-based* and *sandwich-type* estimates of the covariance matrix of $\widehat{\beta}$. The model-based estimate of the covariance matrix is appropriate when the user is confident that the correlation structure has been correctly specified. It has the following form:

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

where

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

and

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

The robust sandwich covariance matrix is typically applied when there is less certainty regarding the choice of working correlation structure. However, we have found that robust standard errors are not necessarily larger than their model-based counterparts, so that the sandwich covariance matrix is not always the most conservative choice. The sandwich matrix takes the following form:

$$\widehat{\text{Cov}}_R(\widehat{\beta}) = {W_m}^{-1} C_m {W_m}^{-1}$$

where

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

Stata 9.2 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 of GEE

GEE is one of the most widely applied and heavily cited statistical methods. For example, a search (in June 2007) for the seminal paper on GEE, Liang and Zeger (1986), on the ISI Web of Knowledge web site yielded 4,369 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 a value greater than one, especially for larger values of α. (However, Stata 9.2 implements an algorithm by Newton (1988) for the AR(1) structure which, judging from the experience of these authors, has no problem with infeasibility [estimates $\widehat{\alpha} > 1$].) In section 4.2, we consider an obesity study in renal transplant patients for which we demonstrate that the GEE estimate of α is infeasible for the tridiagonal structure, so that the estimated correlation matrix is not positive definite.

Another limitation of GEE is that relatively few correlation structures exist in the major statistical software packages that use GEE. For example, the Markov correlation structure is a relatively simple and useful structure that is not yet available for GEE (Shults and Chaganty 1998). Stata 9.2 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, expanding 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

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). See Sun, Shults, and Leonard (2006) for more details about the QLS approach and for a comparison with other methods.

GEE uses an iterative approach for estimation that alternates between updating $\widehat{\beta}$ by solving the GEE for β and updating $\widehat{\alpha}$ with a consistent estimate for α. QLS is a two-stage computational approach within this framework that updates $\widehat{\alpha}$ in stage one with an estimate that minimizes an objective function, the generalized error sum of squares (Chaganty and Shults 1999). In stage one, QLS alternates until convergence between updating the estimates of β and solving the *stage-one estimating equation* for α:

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

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

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

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

Section 3.5 provides solutions to (1) and (2) for several working correlation structures.

The final QLS estimate $\widehat{\beta}_{\text{QLS}}$ of β is then obtained by solving the GEE for β evaluated at $\widehat{\alpha}_{\text{QLS}}$. The asymptotic distribution of $\widehat{\beta}_{\text{QLS}}$ is the same as the asymptotic distribution of the GEE estimate $\widehat{\beta}_{\text{GEE}}$. As a result, we demonstrate in section 4 that testing and construction of confidence intervals for β with QLS is easily accomplished with `xtgee` in Stata 9.2, which uses GEE.

2.5 How QLS expands GEE

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

Next, in section 4 we demonstrate that QLS can apply a useful and relatively simple structure (the Markov) that has not yet been implemented in the framework of GEE. We thus show 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, failure of GEE to converge or infeasibility of $\widehat{\alpha}$ may be a sign that some model assumptions are wrong. For example, Prentice (1988) noted that $\widehat{\alpha}$ must satisfy additional constraints to be feasible in analyses of binary data. Shults, Sun, and Amsterdam (2006) demonstrated that infeasibility of $\widehat{\alpha}$ for binary outcomes can be likely when the AR(1) structure has been misspecified as equicorrelated and α is large. Failure to converge, or infeasibility of $\widehat{\alpha}$, should therefore prompt careful examination of the choice of working structure.

3 The xtqls command

3.1 Syntax

`xtqls` has the following syntax, which is similar to the `xtgee` syntax:

`xtqls` *depvar* [*indepvars*] [*if*] [*in*] [*weight*], `i(`var_i`) t(`var_t`) f(`*family*`)`
`c(`*correlation*`) vce(`*vcetype*`)`

where *depvar* is the dependent variable, *indepvars* are the covariates, and *options* are the required options that we describe in section 3.3.

3.2 Description

`xtqls` provides QLS estimates of the regression and correlation parameter. QLS is a method in the framework of GEE, so that `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 the equicorrelated, AR(1), Markov, and tridiagonal correlation structures.

Using an unstructured matrix is possible with QLS, but the algorithm is complex (Chaganty and Shults 1999). For an unstructured matrix, we therefore recommend `xtgee` in Stata. The QLS and GEE procedures are also identical for the identity matrix, so that use of `xtgee` is recommended for an identity structure.

Future updates of `xtqls` are planned to allow more structures with QLS.

3.3 Options

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

`i(`var_i`)` specifies the ID variable for subjects or clusters.

`t(`var_t`)` specifies the variable for timings of observations.

`f(`*family*`)` specifies the distribution of *depvar*. *family* is one of the following:

`gau`	Gaussian
`bin`	Bernoulli/binomial
`poi`	Poisson

`c(`*correlation*`)` specifies that the correlation structure be used. *correlation* is one of the following:

`AR 1`	AR(1)
`sta 1`	tridiagonal
`exc`	equicorrelated
`Markov`	Markov

`vce(`*vcetype*`)` specifies the type of covariance structure for estimation of $\widehat{\beta}$. *vcetype* is one of the following:

`model`	model-based covariance structure
`robust`	sandwich-type robust sandwich covariance matrix
`jack`	obtains jackknife standard errors
`boot`	obtains bootstrapped standard errors

3.4 Relationship to xtgee

`xtqls` both uses and is similar to `xtgee`. In particular, the syntax is as similar to that of `xtgee` as 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` uses 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 the `force` option for the AR(1) or tridiagonal 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`. We expect future versions of `xtqls` to be more similar to `xtgee` than this initial version. (4) For the tridiagonal, equicorrelated, and tridiagonal structures, `xtqls` can be noticeably slower than `xtgee`.

3.5 Methods and formulas

The xtqls algorithm for estimation of the correlation and regression parameters

`xtqls` uses following algorithm to estimate β and α.

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

2. Alternate between the following steps until 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 estimate (1) for α.

   c. Construct the estimated working correlation structure $R(\widehat{\alpha})$ that corresponds to the updated estimate of α. For structures other than Markov, the matrix $R(\widehat{\alpha})$ 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), $R(\widehat{\alpha})$ will be a 4×4 AR(1) structure evaluated at $\widehat{\alpha}$. For the Markov structure, the dimension of $R(\widehat{\alpha})$ 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 $R(\widehat{\alpha})$ will be 5×5.

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

3. After convergence in stage one, update the estimate of α by obtaining the solution to the stage-two estimate (2) for α.

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

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

This algorithm uses `xtgee` to update $\widehat{\beta}$. As we demonstrate in section 4, all the usual postestimation commands in Stata are available after `xtqls`. This algorithm was described in a presentation by the first author at the Stata 2004 Users Group meeting in Boston, which is available at http://repec.org/nasug2004/Shults_Stata_2004.ppt. Hardin and Hilbe (2003, 73–77) demonstrate a similar algorithm, but with a moment estimate for α, for a correlation structure that is currently unsupported for GEE.

Stage-one and stage-two estimates of α

`xtqls` gives solutions to the stage-one (1) and stage-two (2) estimating equations for several working correlation structures. For estimating equations that do not have an explicit solution, `xtqls` uses bisection 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 $\widehat{\alpha}$ can be expressed as

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

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

$$\widehat{\alpha}_{\text{QLS-AR1}} = \frac{2\widehat{\alpha}_{\text{QONE}}}{1 + \widehat{\alpha}_{\text{QONE}}^2}$$

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

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

where $e_{ij} = |t_{ij} - t_{i,j-1}|$. `xtqls` requires that e_{ij} be $\geq 1\ \forall\, i$ and j.

The stage-two estimating equation for the Markov structure (Chaganty and Shults 1999) is given by

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

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:

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

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

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

For the tridiagonal 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 (1) and (2) for the tridiagonal

structure by first constructing the tridiagonal matrix $R_i(\widehat{\alpha})$ and then using the Stata function `syminv()` to obtain $R_i^{-1}(\widehat{\alpha})$. Next, to evaluate

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

`xtqls` uses the following expression:

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

where $\{\partial R_i(\delta)\}\partial\delta$ 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 $\{\partial R_i(\delta)\}\partial\delta$ 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 `xtgee` in Stata. For example, typing `xtcorr` will display the estimated correlation matrix.

4 Examples

Here we demonstrate `xtqls` in Stata.

4.1 Data and variables

We will use the dataset `random_small.dta`, which is available at http://www.cceb.upenn.edu/∼sratclif/QLSproject.html. These data are from a study of obesity in children after renal transplant that was conducted at the Children's Hospital of Philadelphia. To facilitate sharing of these data for demonstrating `xtqls`, we dropped 10% of the observations before saving the dataset `random_small.dta`. (We did so by generating the variable `random` with the uniform command, sorting on the variable `random`, and then dropping all observations corresponding to `random` ≤ 0.1.)

(*Continued on next page*)

```
. use random_small
. describe
Contains data from random_small.dta
  obs:           531
 vars:             5                          20 Aug 2006 09:56
 size:        12,744 (99.9% of memory free)
```

variable name	storage type	display format	value 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 our examples, we will regress body mass index (BMI) z-score and obesity status (yes–no) on baseline BMI z-score and on month of measurement. We will demonstrate the robust sandwich–based covariance matrix and the model-based covariance matrix.

4.2 Example with infeasible GEE moment estimate

If we regress BMI on time and baseline BMI, then the feasible region (set of values on which α is positive definite) for the tridiagonal structure is $(-0.51764, 0.51764)$. We first use this structure with Stata's `xtgee` command, using 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 = .00021742
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)
```

bmiz	Coef.	Semi-robust 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

```
working correlation matrix not positive definite
convergence not achieved
r(430);
```

Above, `xtgee` required the option `force` because the timing variable `month` is not equally spaced on all subjects. (If this option were not supplied, then we would have received a warning that the observations were not equally spaced, in which case 97 subjects would have been omitted from estimation. Our analysis would then be based on only three subjects.)

Stata warned us that the estimated correlation matrix is not positive definite. We can see that this is indeed the case when 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  0.8262  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  0.8262
r11  0.0000  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 $\widehat{\alpha}_{\text{GEE}} = 0.8262$, which per the Stata warning is outside the feasible region $(-0.51764, 0.51764)$ for the tridiagonal structure.

Next we will use the tridiagonal structure with `xtqls`, using the sandwich-based covariance matrix. Doing so does not require the option `force`; `xtqls` will treat the timings as equally spaced for the tridiagonal and AR(1) structures. (This example will take considerably longer to run than did `xtgee` for the tridiagonal structure.)

```
. xtqls bmiz basebmi month, i(id) t(month) f(gau) vce(robust) c(sta 1)
Iteration 1: tolerance = .09658071
Iteration 2: tolerance = 0
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)
```

bmiz	Coef.	Semi-robust 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

`xtqls` uses `xtgee` for a fixed correlation matrix. Therefore, all the usual postregression commands are available after `xtqls`. For example, if we use `xtcorr` to provide the estimated correlation matrix, we see that $\widehat{\alpha}_{\text{QLS}} = 0.5176$, 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  0.5176
r11  0.0000  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
```

For other structures, we observed that the variable `month` was significant only for the tridiagonal structure and QLS. However, as discussed in Shults, Sun, and Amsterdam (2006), infeasibility of $\widehat{\alpha}_{\text{GEE}}$ or $\widehat{\alpha}_{\text{QLS}}$ might indicate that the correlation structure has not been correctly specified. Given that the tridiagonal structure is not biologically plausible for this analysis, and that $\widehat{\alpha}_{\text{GEE}}$ 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 the Markov structure in the next section.

4.3 The Markov structure

Now let us 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	

The Markov structure is appropriate for this analysis, because this structure accounts for the variability of spacing of measurements. We next use the Markov structure with `xtqls`. Here we show the 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 = 0
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

(*Continued on next page*)

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  0.5975
r11  0.0170  0.0185  0.0210  0.0272  0.0455  0.0762  0.1275  0.2133  0.3570
        c10     c11
r10  1.0000
r11  0.5975  1.0000
```

The within-subject correlations are high for this analysis.

4.4 AR(1) and equicorrelated structure with QLS

Let us next consider the outcome of obesity (1 = obese; 0 = not obese) and use the AR(1) and equicorrelated correlation structures with `xtqls`, using 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  0.6987
r11  0.0277  0.0397  0.0568  0.0813  0.1164  0.1666  0.2384  0.3411  0.4882
        c10     c11
r10  1.0000
r11  0.6987  1.0000
```

If we had implemented the AR(1) structure by using `xtgee`, then 97 subjects would have been dropped from the analysis because of unequal spacing of measurements. Or we could have used the option `force`, which would have treated all observations as equally spaced. (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 use 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

Next let us 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  0.5065
r11  0.5065  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

We have used QLS with the user-written `xtqls` command in Stata. This command allows correlation structures such as the Markov that have not yet been used in the framework of GEE. QLS may also provide a feasible estimate when the GEE estimate is infeasible or if GEE fails to converge. `xtqls` calls `xtgee`, and therefore all the usual postregression estimation commands are available after `xtqls`. Future updates of `xtqls` will use more correlation structures, including the banded Toeplitz and other structures that are appropriate for data with multiple levels of correlation.

6 Acknowledgment

Work on this report was supported by the NIH-funded grant R01CA096885 "Longitudinal Analysis for Diverse Populations".

7 References

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

Chaganty, N. R., and J. Shults. 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 generalized linear models for repeated measures. *Biometrika* 82: 407–410.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger. 2002. *Analysis of Longitudinal Data*. 2nd ed. Oxford: Oxford University Press.

Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Liang, K.-Y., and S. L. Zeger. 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 covariate 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 N. R. Chaganty. 1998. Analysis of serially correlated data using quasi-least squares. *Biometrics* 54: 1622–1630.

Shults, J., C. A. Mazurick, and J. R. Landis. 2006. Analysis of repeated bouts of measurements in the framework of generalized estimating equations. *Statistics in Medicine* 25: 4114–4128.

Shults, J., and A. Morrow. 2002. Use of quasi-least squares to adjust for two levels of correlation. *Biometrics* 58: 521–530.

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

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

StataCorp. 2005. *Stata 9 Longitudinal/Panel-Data Reference Manual*. College Station, TX: Stata Press.

Sun, W., J. Shults, and M. Leonard. 2006. Use of unbiased estimating quations to estimate correlation in generalized estimating equation analysis of longitudinal trials. http://www.biostatsresearch.com/upennbiostat/papers/art4.

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

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

About the authors

Justine Shults, Ph.D., and Sarah Ratcliffe, Ph.D., are assistant professors, and Mary Leonard, M.D., M.S.C.E, is an associate professor in the Center for Clinical Epidemiology, Department of Epidemiology and Biostatistics, University of Pennsylvania School of Medicine.