Statistics and Its Interface Volume 2 (2009) 187–196

Analysis of multi-level correlated data in the framework of generalized estimating equations via xtmultcorr procedures in Stata and qls functions in Matlab

Justine Shults* and Sarah J. Ratcliffe

Many medical studies yield data with multiple sources of correlation. For example, in a study of repeated measurements collected on each eye of spouses, three sources of correlation may be present, due to the fact that measurements within the same family will be more similar if they are measured on the same eye (left versus right), within the same person (husband versus wife), or at the same measurement occasion. This article reviews an algorithm for analysis of data with two or more sources of correlation (Shults, Whitt, Kumanyika, 2004) that can be implemented using quasi-least squares, an approach in the framework of generalized estimating equations. It then describes and demonstrates implementation of this algorithm with xtmultcorr procedures in Stata and the qls functions in Matlab. The Stata and Matlab procedures are available on the website for the Longitudinal Analysis for Diverse Populations project: http://www.cceb.upenn.edu/∼sratclif/QLSproject.html.

Keywords and phrases: Cholesky decomposition, correlated data, generalized estimating equations, multi-level data, multivariate data, quasi-least squares.

1. INTRODUCTION

We consider the usual set-up for generalized estimating equations (GEE, Liang and Zeger, 1986), for which measurements are collected on multiple subjects, or clusters. Measurements from different clusters are assumed to be independent, but measurements from within the same cluster are assumed to be correlated.

The typical GEE analysis involves one source of correlation, where a source of correlation is defined as a factor that impacts the similarity (and therefore the correlation) of measurements. For example, in a longitudinal study that measures the weight of unrelated subjects, the timing of measurements will represent the one source of correlation because measurements within a subject that are collected more closely together in time should be more similar (and therefore more highly correlated) than if they are measured farther apart in time, e.g. weights at baseline and two months post-baseline should be more highly correlated than weights collected at baseline and 12 months post-baseline.

It is well known that the dependence amongst observations should be taken into account when measurements fail to be independent. For example, consider the estimation of the probability of success with its associated confidence interval. Miao and Gastwirth (2004) demonstrated that even if the data are only slightly correlated, "the coverage probabilities of the usual confidence intervals can deviate noticeably from their nominal level".

This article considers regression analysis of data with multiple sources of correlation, which are sometimes referred to as multi-level data or multivariate longitudinal data. For example, suppose the longitudinal study described above was modified to measure weights on siblings over time. In this situation, we might expect an additional source of correlation in the data, due to anticipated similarity between siblings. Or, consider a study in which grip strength is measured on both hands of elderly twins at baseline and at one month post baseline. This study might be expected to yield data with three sources of correlation, because we anticipate that two measurements within a sibling pair will be more similar if they are measured on the same twin (older versus younger), on the same hand (left versus right), or at the same measurement occasion (baseline or one month post baseline). Many other examples of multi-level correlated data are described in Goldstein (1995).

The usual goal of a GEE analysis is to explore the relationship between the expected value of the outcome variable and covariates measured on each of the subjects, while adjusting for the correlation within the measurements on each cluster. Although the regression parameter is often the primary parameter of interest, the estimated correlations can often yield interesting information. For example, if the intra-hand correlations of grip strength in the intervention to improve strength in elderly subjects are negative, this might suggest that subjects are focusing their efforts on one hand to the detriment of the other. Or, if the intra-twin correlations of birth weight are negative in an intervention to improve birth weight in twins born to high risk mothers,

*Corresponding author.

this might suggest that improvement in birth weight for one twin is occurring at the expense of the other.

Although many studies yield data with multiple sources of correlation, GEE is typically implemented for data with one source of correlation, with relatively simple correlation structures used to describe the pattern of association amongst the repeated measurements on each subject (or cluster). In this manuscript we review an algorithm for analysis of data with multiple sources of correlation (Shults et al., 2004) in the framework of GEE. We then demonstrate implementation of this algorithm with the the user-written xtmultcorr and qls programs in Stata and Matlab, respectively.

We note that the majority of our descriptions in this manuscript are for data with three sources of correlation; however, all results can easily be generalized for data with two or more than three sources of correlation.

2. METHODS

2.1 Notation

The usual notation for a typical GEE analysis (Shults, Ratcliffe, Leonard, 2007) with one source of correlation is readily modified for data with multiple sources. For example, consider a study in which repeated measurements are collected on independent subjects at baseline, and then at three and six months post-baseline. In this study we might anticipate that there is one source of correlation, due to the timing of measurements; y_{ij} and x'_{ij} might then represent the measurement of the outcome variable and vector of associated covariates that are collected on subject i at time j. Note that this notation for data with one source of correlation involves two subscripts, i that denotes the subject (or cluster) and j that denotes the value of the first (and only) source of correlation. To extend our notation to a study with three sources of correlation, we simply expand the number of subscripts to four, so that $y_{ij_1j_2j_3}$ and $x'_{ij_1j_2j_3}$ represents the value of the outcome variable and associated $p \times 1$ vector of covariates that are collected on subject (or cluster) i when the values of the first, second, and third source of correlation are j_1, j_2 and j_3, respectively.

Next, it will be helpful to refer to $Y_i[a, b, c]$ as the vector of outcomes of measurements $y_{ij_aj_bj_c}$ on subject i that has been sorted first according to j_a, and then j_b, and then j_c. Our approach for analysis of multi-level data is appropriate for data that are balanced overall, so that $j_c = 1, 2, \ldots, n_c$ in $Y_i[a, b, c]$ for all i and (a, b, c) in the class of all permutations of $(1, 2, 3)$. For example, in a longitudinal study of visual acuity measured on both eyes of spouses at baseline and six months post-baseline, the study will be balanced overall if there are no missing observations; in this case, we might let spouse, eye, and time represent the first, second, and third sources of correlation, with $j_1 = 1$, 2 (1 for wife, 2 for husband), $j_2 = 1$, 2 (1 for left eye, 2 for right eye), and $j_3 = 1$, 2, 3 (number of measurement occasion) within husband/wife cluster i.

Our manuscript provides programs in Stata and Matlab for balanced data. We also note that GEE is usually applied for analyses that involve a relatively large number of small clusters (or subjects), with 30 often used as an informal lower bound for the number of clusters that are required to yield valid results. In multi-level analyses the cluster sizes can be larger than in the typical analysis, that might involve 3 or 4 measurements per cluster. For example, in the study of visual acuity described above, if the data are totally balanced, the cluster sizes will be $n_1 \times n_2 \times n_3 = 2 \times 2 \times 3 = 12$. We suspect that the correlations may play a more prominent role in multi-level studies because they typically yield data with larger cluster sizes, and hence provide more information with which to estimate the pattern of association in the data.

2.2 Specification of a working correlation structure for analysis of multi-level data in the framework of GEE

GEE analyses specify a generalized linear model to describe the relationship between the outcome and covariates measured on each subject: The expected value and variance of measurement $y_{ij_1j_2j_3}$ on subject (or cluster) i are assumed to equal $E(y_{ij_1j_2j_3}) = g^{-1}(x'_{ij_1j_2j_3}\beta) = u_{ij_1j_2j_3}$ and $Var(y_{ij_1j_2j_3}) = \phi h(u_{ij_1j_2j_3})$, respectively, where ϕ is a known or unknown scale parameter. We also let $U_i(\beta)$ represent the $w_i \times 1$ vector of expected values $u_{ij_1j_2j_3}$ on subject i, where $w_i = n_1 \times n_2 \times n_3$.

The intra-cluster correlation of measurements is then accounted for in the analysis by specifying a patterned correlation matrix R_i to describe the pattern of association of measurements within cluster i, for $i = 1, 2, \ldots, m$. Some popular structures for data with one source of correlation include the following:

1. `The Equicorrelated (Exchangeable):` This assumes equality of correlations within each cluster, so that $R_i[j, k] = \alpha$. For $n \times n$ structure R_i, R_i will be positive definite for α in $(-1/(n-1), 1)$.
2. `The first-order autoregressive AR(1):` This assumes correlations will be smaller for measurements that are farther apart in terms of measurement occasion, so that $R_i[j, k] = \alpha^{j-k}$ for α in $(-1, 1)$.
3. `The tri-diagonal correlation structure:` This structure has ones on the diagonal and α on the off-diagonal, so that $R_i[j, k] = \alpha$ for $|j - k| = 1$ and is zero otherwise. If R_i is $n \times n$, it will be positive definite for α in $(-1/c_m, 1/c_m)$, where $c_m = 2\sin\left(\frac{\pi[n-1]}{2[n+1]}\right)$. This interval is approximately $(-1/2, 1/2)$ for large n and contains $(-1/2, 1/2)$ for all n.

We will use the structures just described to construct a biologically plausible correlation structure for data with

multiple sources of correlation. (As noted earlier, we will describe all results for data with three sources of correlation, but the results can be readily generalized for two or > 3 sources.)

To construct a biologically plausible structure for data with multiple sources of correlation:

1. First, within each cluster (or subject) identify one source of correlation as the first source, another source as the second, and another as the third.
2. For each source of correlation, choose the correlation structure that would be appropriate if the source under consideration was the only source of correlation in the data. Let R_a be the correlation structure for source a; $a = 1, 2, 3$.
3. When done, construct the correlation structure for the vector of measurements $Y_i[1, 2, 3]$ as the Kronecker product of R_1, R_2, and R_3, i.e. $Corr(Y_i[1, 2, 3]) = R_1 \bigotimes R_2 \bigotimes R_3$. The covariance matrix of $Y_i[1, 2, 3]$ is then given by $Cov(Y_i[1, 2, 3]) = \phi A_i^{1/2} R_1 \bigotimes R_2 \bigotimes R_3 A_i{}^{1/2}$, where $A_i = diag(h(u_{i111}), \ldots, h(u_{in_1n_2n_3}))$; it is also easily shown that $Corr(Y_i[a, b, c]) = R_a \bigotimes R_b \bigotimes R_c$, where (a, b, c) is any permutation of $(1, 2, 3)$.

For example, consider the study of visual acuity described above for data that are balanced overall. To specify a correlation structure for the first source of correlation (spouse), we imagine that the only source of correlation is due to spouse, e.g. we have a study in which only one measurement was collected on one eye of each spouse; in this situation a reasonable structure for R_1 is 2×2 structure

$$R_1 = \begin{pmatrix} 1 & \alpha_1 \\ \alpha_1 & 1 \end{pmatrix}.$$

Similarly, a reasonable structure for the second source of correlation (eye) is given by

$$R_2 = \begin{pmatrix} 1 & \alpha_2 \\ \alpha_2 & 1 \end{pmatrix}.$$

For the third source of correlation, we might specify an AR(1) correlation structure:

$$R_3 = \begin{pmatrix} 1 & \alpha_3 & {\alpha_3}^2 \\ \alpha_3 & 1 & \alpha_3 \\ {\alpha_3}^2 & \alpha_3 & 1 \end{pmatrix}.$$

The correlation structure for $Y_i[1, 2, 3]$ is then constructed as $R_1(\alpha_1) \bigotimes R_2(\alpha_2) \bigotimes R_3(\alpha_3)$.

This Kronecker product structure has been discussed and implemented by several authors, including Galecki (1994) and Naik and Rao (2001). Roy and Khattree (2005) and Lu and Zimmerman (2005) developed tests for this structure for multivariate repeated measures data that are normally distributed. It is a popular structure for analysis of multi-level correlated data because it forces the correlation between measurements to be smaller when they have more disagreement with respect to the sources of correlation in the data, which is often biologically plausible.

For example, if $Corr(Y_i[1, 2, 3]) = R_1 \bigotimes R_2 \bigotimes R_3$, then using the definition of Kronecker product, $Corr(y_{ij_1j_2j_3}, y_{ik_1k_2k_3}) = R_1[j_1, k_1] \times R_2[j_2, k_2] \times R_3[j_3, k_3]$. If we apply this to our visual acuity example, then we see that the correlation between measurements collected on the same spouse and same eye (left versus right) at visits one and two is α_3. However, if the measurements are collected on the same spouse, but on different eyes, at visits one and two, then the correlation is $\alpha_2 \times \alpha_3$. Finally, if the measurements within families were collected on different spouses and different eyes at visits one and two, then the correlation would be $\alpha_1 \times \alpha_2 \times \alpha_3$. Therefore, if the parameters are all positive, we see that the correlation between measurements will be smaller as their degree of disagreement with respect to the sources of correlation in the data decreases.

Shults and Morrow (2002) and Chaganty and Naik (2002) implemented the Kronecker product structure for data with two sources of correlation. Shults et al. (2004) proposed a general algorithm that allows for data with 2 or more sources of correlation; this is the approach we use in this manuscript.

2.3 Overview of algorithm for implementation of the Kronecker Product Structure

The xtmultcorr and qls procedures implement the Kronecker product correlation structure using quasi-least squares (Chaganty and Shults, 1999), an approach in the framework of GEE that allows for easier implementation of some correlation structures, including the Kronecker product structure. An important advantage of GEE is that it does not require specification of the full distribution of a random variable, but only of the first two moments of the distribution. For example, for a continuous random variable, GEE does not require an assumption of normality, but instead requires specification of the first two moments of its distribution. This loosening of the requirement to specify the full distribution is important because it has long been known, e.g. see Gastwirth (1966), that departures from normality require special attention in the analysis.

To summarize briefly, QLS is a two-stage approach that in stage one alternates between updating the estimate of the regression parameter β by solving the GEE estimating equation for β (Liang and Zeger, 1986) and updating the estimate of the correlation parameter α by solving the QLS stage one estimating equation for α. (The QLS stage one estimate of α (Chaganty, 1997) has the drawback of being inconsistent, even when the working correlation structure is correctly specified.) After convergence in stage one, the (consistent) stage two estimate of α is obtained solving the QLS stage two estimating equation for α. The final QLS

estimate of β is then obtained by again solving the GEE estimating equation for β evaluated at the stage two estimate of α. For more details of the QLS approach and its implementation for data with one source of correlation in Stata, see Sun et al. (2008). Programs are also available for implementation of QLS for one source of correlation for Matlab (Ratcliffe and Shults, 2008); Stata (Shults et al., 2007); R software (Xie and Shults, 2008); and SAS (Kim and Shults, 2008).

Programs for the stage one and stage two QLS estimating equations for α for a particular correlation structure require subject id, the variable indicating timing of measurements, and the vector of Pearson residuals as arguments. The Pearson residuals, defined for data with one source of correlation, are defined for subject i as $Z_i(\beta) = (z_{i1}, z_{i2}, \ldots, z_{in_i})\prime$ for $z_{ij} = (y_{ij} - u_{ij})/h(u_{ij})$. The covariance matrix of the vector of Pearson residuals, for data with one source of correlation, is easily shown to be equal to ϕR_i, so that the covariance matrix for the vector of Pearson residuals that is obtained by stacking the vectors for all subjects, $Z = (Z_1\prime, Z_2\prime, \ldots, Z_m\prime)\prime$, is given by $I(m \times m) \bigotimes R_i$, the Kronecker product of an $m \times m$ identity matrix and R_i.

The stage one QLS algorithm for implementation of the Kronecker product correlation structure that is appropriate for data with > 2 sources of correlation is based on the observation that if we pre-multiply the vector of Pearson residuals $Z_i[a, b, c]$ by the Kronecker product of the square root of the inverse of R_a and the square root of the inverse of R_b and an $n_c \times n_c$ identity matrix, then we obtain the usual correlation structure for data with one source of correlation. In particular, $Corr\big(R_a^{-1/2} \bigotimes R_b^{-1/2} \bigotimes I^{-1/2}\big) Z_i[a, b, c] = \big(R_a^{-1/2} \bigotimes R_b^{-1/2} \bigotimes I(n_c \times n_c)\big)\phi\big(R_a \bigotimes R_b \bigotimes R_c\big)\big(R_a^{-1/2} \bigotimes R_b^{-1/2} \bigotimes I(n_c \times n_c)\big) = \phi I(n_a n_b \times n_a n_b) \bigotimes R_c$. To update the correlation parameters in stage one, we were therefore able to utilize an approach that alternated between sorting, pre-multiplying, and using programs written earlier for data with one source of correlation, applied to the vector of pre-multiplied Pearson residuals.

To obtain estimates in stage two, Shults et al. (2004) proved that each correlation parameter can be updated separately, e.g. the stage two estimate of α_2 is only a function of the stage one estimate of this parameter.

2.3.1 Algorithm for estimation of the correlation and regression parameters for data with multiple sources of correlation

The xtqls and qlsr procedures implements the following algorithm for estimation of β and of α_j $(j = 1, 2, 3)$ (Please see Shults et al. (2004) for more details and justification of this approach. In particular, please see Appendix B of Shults et al. (2004) for a general algorithm for $k \geq 2$ sources of correlation.):

1. Let $\hat{\alpha_j} = 0$ represent initial estimates of α_j, for $j = 1, 2, 3$. Let $old_j = \hat{\alpha_j}$ for $j = 1, 2, 3$.
2. Obtain a starting value for $\hat{\beta}$ by assuming $\alpha_j = 0$ and then obtaining a solution to the GEE estimating equation for β at $\alpha_j = 0$ for $j = 1, 2$, and 3. (Note that this is equivalent to using linear regression, logistic regression, or Poisson regression to obtain a starting value for $\hat{\beta}$, for an outcome variable that is continuous, binary, or that represents counts, respectively.)
3. For $(a, b, c) = (1, 2, 3)$, $(3, 1, 2)$, and $(2, 3, 1)$, do the following: Sort on i, j_a, j_b, and then j_c and create a new id variable $id[a, b, c]$ that takes values $1, 2, \ldots$ and that takes a different value for each distinct value of (i, j_a, j_b). These new identification variables will be used when updating the α_j in stage one of QLS.
4. Alternate between the following steps till convergence in the estimates of β:
   (a) Obtain updated values of the Pearson residuals $z_{ij_1j_2j_3}$ at the current estimate of β.
   (b) Update the estimate of the correlation parameters by alternating between the following steps until $\Delta_j \approx 0$ for $j = 1, 2, 3$:
      i. Implement the stage one updating step (given below) for $(a, b, c) = (1, 2, 3)$.
      ii. Implement the stage one updating step for $(a, b, c) = (3, 1, 2)$.
      iii. Implement the stage one updating step for $(a, b, c) = (2, 3, 1)$.

      Stage one updating step: First, sort the data with respect to i, then j_a, j_b, and j_c. Next, pre-multiply the vectors of Pearson residuals $Z_i[a, b, c]$ within each subject i by $R_a^{-1/2} \bigotimes R_b^{-1/2} \bigotimes I(n_c \times n_c)$; call the vector of pre-multiplied residuals $Z_i^*[a, b, c]$. Solve the QLS stage one estimating equation for α_c, under the assumption that the working correlation structure is $R_c(\alpha_c)$. (Note that the programs to solve this equation will require the following as arguments: $Z_i^*[a, b, c]$ (pre-multiplied Pearson residuals), $id[a, b, c]$ (identification variable), and j_c (representing the timings for each cluster indicated by different values of $id[a, b, c]$).) After obtaining an updated estimate of α_c, update the estimate of $R_i^{-1/2}(\alpha_c)$ and then obtain Δ_c, the difference between the updated estimate of α_c and the previous estimate old_c. Next, let old_c equal the updated estimate of α_c.
   (c) Construct the estimated working correlation structure $W(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3) = R_1(\hat{\alpha}_1) \bigotimes R_2(\hat{\alpha}_2) \bigotimes R_3(\hat{\alpha}_3)$ that corresponds to the updated estimates of α_j $(j = 1, 2, 3)$.
   (d) Update the estimate of β by solving the GEE estimating equation (for β) with a correlation structure that is treated as fixed and equal to $W(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$.

5. After convergence in stage one of QLS, update the estimate of α_j $(j = 1, 2, 3)$ by solving the stage two estimating equation for each parameter. Note that for each parameter this only involves specification of the working structure and the vector of timings for each cluster.
6. Construct the estimated working correlation structure $W(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3) = R_1(\hat{\alpha}_1) \bigotimes R_2(\hat{\alpha}_2) \bigotimes R_3(\hat{\alpha}_3)$ that corresponds to the stage two estimates of α_j $(j = 1, 2, 3)$.
7. Update the estimate of β by solving the GEE estimating equation (for β) with a correlation structure that is treated as fixed and equal to $W(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$.

The xtmultcorr Stata procedures implement the xtgee procedure to solve the GEE estimating equation for β in steps $4(d)$ and 7 of the algorithm. This exploits the fact that the xtgee procedure allows for solution of the GEE estimating equation at a current estimate of the correlation structure that is treated as fixed and known. An important consequence is that all the usual post regression commands in Stata are available after implementation of xtqls.

The Matlab programs implement an approach for updating the estimate of β that is based on the inverse of the Cholesky decomposition of the working correlation structure; this approach is described in Shults (1996) and Shults and Chaganty (1998).

3. THE XTMULTCORR COMMANDS IN STATA 9.0

3.1 Syntax

The xtmultcorr commands have the following syntax which is similar to the syntax for the xtgee procedure, expect that the command requires specification of additional variables for each source of correlation and options to indicate the correlation structure that will be implemented for each source of correlation

For data with 2 sources of correlation, the xtmultcorr2 procedure is used.

xtmultcorr2 *depvar* [*indepvars*], *options*

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

For data with 3 or 4 sources of correlation, the xtmultcorr3 and xtmultcorr4 procedures are used, respectively. The syntax will be the same as for xtmultcorr2, but the options will differ, as described in the following section.

3.2 Description

The xtmultcorr2, xtmultcorr3, and xtmultcorr4 commands provide QLS estimates of the regression and correlation parameters for data with 2, 3, and 4 sources of correlation, respectively. This is done by implementation of a Kronecker product correlation structure of 2, 3, or 4 correlation structures (for 2,3, or 4 sources of correlation, respectively), each of which would be appropriate for data with one source of correlation. The correlation structures can be chosen from among the equicorrelated, AR(1), Markov, and tri-diagonal correlation structures.

3.3 Options

The *options* for xtmultcorr2 (all required) are as follows:

1. i(*var1*) where *var1* is the ID variable for subjects, or clusters
2. l1(*var2*) where *var2* is the variable that indicates the value of source one of correlation
3. l2(*var3*) where *var3* is the variable that indicates the value of source two of correlation
4. c1(*corr1*) where *corr1* is the correlation structure for source one of correlation
5. c2(*corr2*) where *corr2* is the correlation structure for source two of correlation. The following correlation structures can be implemented in the xtmultcorr programs:
   (a) AR 1 (AR(1))
   (b) sta 1 (tridiagonal)
   (c) exc (equicorrelated)
6. f(*family*) where *family* is the distribution of depvar. The following families can be implemented in the xtmultcorr programs:
   (a) gau (Gaussian)
   (b) bin (Bernoulli/binomial)
   (c) poi (Poisson)
7. vce(*vcetype*) where *vcetype* indicates the type of covariance structure for estimation of $\hat{\beta}$. The following covariance structures can be implemented in the xtmultcorr programs:
   (a) model (model based covariance structure)
   (b) robust (sandwich type robust sandwich covariance matrix)
   (c) jack (obtains jack-knife standard errors)
   (d) boot (obtains boot-strapped standard errors)

The *options* for xtmultcorr3 (all required) are as follows:

1. i(*var1*) where *var1* is the ID variable for subjects, or clusters
2. l1(*var2*) where *var2* is the variable that indicates the value of source one of correlation
3. l2(*var3*) where *var3* is the variable that indicates the value of source two of correlation
4. l3(*var4*) where *var4* is the variable that indicates the value of source three of correlation
5. c1(*corr1*) where *corr1* is the correlation structure for source one of correlation

6. c2(*corr2*) where *corr2* is the correlation structure for source two of correlation
7. c3(*corr3*) where *corr3* is the correlation structure for source three of correlation
8. f(*family*) where *family* is the distribution of depvar
9. vce(*vcetype*) where *vcetype* indicates the type of covariance structure for estimation of $\hat{\beta}$

The *options* for xtmultcorr4 (all required) are as follows:

1. i(*var1*) where *var1* is the ID variable for subjects, or clusters
2. l1(*var2*) where *var2* is the variable that indicates the value of source one of correlation
3. l2(*var3*) where *var3* is the variable that indicates the value of source two of correlation
4. l3(*var4*) where *var4* is the variable that indicates the value of source three of correlation
5. l4(*var5*) where *var5* is the variable that indicates the value of source four of correlation
6. c1(*corr1*) where *corr1* is the correlation structure for source one of correlation
7. c2(*corr2*) where *corr2* is the correlation structure for source two of correlation
8. c3(*corr3*) where *corr3* is the correlation structure for source three of correlation
9. c4(*corr4*) where *corr4* is the correlation structure for source four of correlation
10. f(*family*) where *family* is the distribution of depvar
11. vce(*vcetype*) where *vcetype* indicates the type of covariance structure for estimation of $\hat{\beta}$

3.4 Relationship to the xtgee procedure

The xtmultcorr procedure implements the xtgee procedure and has important similarities to the xtgee procedure that are the same as the similarities between xtqls and xtgee; please see section 3.4 of Shults, Ratcliffe, Leonard (2007) for more details.

3.5 Saved results

The saved results for the xtmultcorr programs are the same as those for the xtgee procedure in Stata. For example, typing xtcorr after implementing xtmultcorr2 in an analysis will display the estimated correlation matrix.

4. EXAMPLES IN STATA

Here we demonstrate implementation of the xtmultcorr commands in Stata.

4.1 Data and variables

We will use the data set

`example_multlevel.dta`

that is available on http://www.cceb.upenn.edu/~sratclif/QLSproject.html. This is a data set that is based on Table 3.7 (p. 65) of Davis (2002). The data are from a study (Weissfeld and Kshirsagar, 1992) in which three methods of suctioning an endotracheal tube were applied in random order to 25 patients in an intensive care unit. The outcome variable was oxygen saturation, that was measured on each patient at baseline, first suctioning pass, second suctioning pass, third suctioning pass, and five minutes after suctioning.

Let's open the data set in Stata and describe the variables: *Please note that the output may be slightly edited, in order to improve its appearance in the manuscript.*

```
. use example_multlevel, clear

. de Contains data from example_multlevel.dta
  obs:           375
 vars:             6
 size:        19,500 (98.1% of memory free)
-------------------------------------------------
              storage  display
variable name   type   format   variable label
-------------------------------------------------
id          double %10.0g   subject id
time        double %10.0g   measurement occasion
type        double %10.0g   method of suctioning
o2          double %10.0g   oxygen saturation
family      double %10.0g   artificial family variable
high        double %10.0g   1 if o2>96; 0 otherwise
-------------------------------------------------
Sorted by:  id  type  time
```

Let's next check the number of subjects:

```
. summ id

Variable |  Obs   Mean    Std. Dev.   Min        Max
---------+-----------------------------------------
      id |  375   13.68   7.603405     1         26

. qui tab id
. di _result(2) 25
```

There are 25 subjects.

Next, let's tabulate time and type of measurement:

```
. tab type

  method of |
 suctioning |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        125       33.33       33.33
          2 |        125       33.33       66.67
          3 |        125       33.33      100.00
------------+-----------------------------------
      Total |        375      100.00

. tab time
```

```
measurement |
   occasion |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         75       20.00       20.00
          2 |         75       20.00       40.00
          3 |         75       20.00       60.00
          4 |         75       20.00       80.00
          5 |         75       20.00      100.00
------------+-----------------------------------
      Total |        375      100.00
```

It appears that type and time are balanced overall, although this will be checked in our programs. Next, let's summarize family:

```
. summ family
Variable |   Obs    Mean    Std. Dev.  Min    Max
---------+-------------------------------------------
  family |   375   6.76    3.619023   1      13
```

The variable family is an artificial grouping (not in the original data set) that was created to demonstrate analysis of data with 3 sources of correlation. There are 12 families of size two and one family (that contains the subject with id = 26) of size one.

For the examples we consider here, we will regress a binary outcome variable on time. The binary outcome, high, takes value one if $o_2 > 96$ and takes value zero otherwise. We consider a simple logistic regression model for the mean because our goal is to demonstrate our xtmultcorr programs. We will demonstrate implementation of the robust sandwich-based covariance matrix and also of the model based covariance matrix.

4.2 Adjustment for three sources of correlation

Here we will account for three sources of correlation and will consider an analysis of the binary outcome (high).

In this analysis we will take the artificial family groupings into account, so that the measurements will be clustered according to family. Our data are not balanced with respect to family because we have 13 families of size two and one family (for id = 26) of size one. Because our programs assume the data are balanced overall, we will first drop the subject with id = 26, in order to have 12 equally sized families:

```
. drop if id==26
```

We now have 12 families of size 2. If we consider the family groupings, we will have 3 sources of correlation, due to the fact that measurements within a family may be more similar if they are collected on the same subject, according to the same type of method, or at the same time. We will identify subject, type, and time as the first, second, and third sources of correlation. (However, note that the ordering of subject, type, and time was arbitrary, e.g. we could have identified type, time, and subject as the first, second, and third sources, respectively). We specify the equicorrelated structures for sources one and two (id and type, respectively) and the AR(1) structure for source three (time). Because the variable high is binary, we specify that the family is Bernoulli via the option f(bin 1). In addition, we specify a robust sandwich based covariance matrix for estimation of the covariance matrix of the regression parameter:

```
. xtmultcorr3 high time, i(family) l1(id) l2(type)/*
*/ l3(time) c1(exc) c2(exc) c3(AR 1) f(bin 1) vce(robust)

 Data are not balanced within subjects with respect to
 sources of correlation.
```

The variable *id* must take value $1, 2, \ldots, n_1$ for each subject i, *after sorting on type and time*. We therefore need to create a new identification variable that can be used in the xtmultcorr3 program:

```
. sort family id
.qui by family id: gen id2=1 if _n==1
.qui by family: replace id2 = sum(id2)

. xtmultcorr3 high time, i(family) l1(id2) l2(type) /*
*/ l3(time) c1(exc) c2(exc) c3(AR 1) f(bin 1) vce(robust)

Estimated correlation associated with level one:

symmetric __000003[2,2]
           c1         c2
r1          1
r2  .04860953          1

Estimated correlation associated with level two:

symmetric __000004[3,3]
          c1        c2        c3
r1         1
r2  .1518791         1
r3  .1518791  .1518791         1

Estimated correlation associated with level three:

symmetric __000005[5,5]
           c1         c2         c3         c4         c5
r1          1
r2  .64306572          1
r3  .41353352  .64306572 1
r4  .26592923  .41353352  .64306572          1
r5  .17100997  .26592923  .41353352  .64306572          1


Iteration 1: tolerance = .00844386
Iteration 2: tolerance = .00004461
Iteration 3: tolerance = 7.387e-09

GEE population-averaged model       Number of obs      = 360
Group and time vars:family __00001U Number of groups   = 12
Link:                    logit      Obs per group: min = 30
Family:                 binomial                   avg = 30
Correlation: fixed (specified)                     max = 30
                          Wald-chi2(1)         =      0.31
Scale parameter:        1      Prob > chi2        = 0.5759

               (Std. Err. adjusted for clustering on family )
------------------------------------------------------------------
      |          Semi-robust
 high |   Coef.   Std. Err.    z  P>|z|  [95% Conf. Interval]
------+-----------------------------------------------------------
 time | .0308239 .0551098   0.56 0.576   -.0771892  .1388371
_cons |-.3893841 .3379246  -1.15 0.249   -1.051704  .2729359
------------------------------------------------------------------
```

As we might expect (based on the fact that the family grouping was artificial) the estimated correlation associated with family was very small, with $\hat{\alpha}_1 = .04860953$.

We also note that because the xtmultcorr programs make use of the xtgee program in Stata, all the usual post-regression estimation commands are available after analysis. For example, if we use the xtcorr command in Stata we can print the estimated correlation matrix that is the Kronecker product of the estimated working structures for each source of correlation. Below, the first 5 lines of the correlation matrix are displayed:

```
. xtcorr

Estimated within-family correlation matrix R:

       c1      c2      c3      c4      c5
r1:r1  1.0000
r1:r2  0.6431  1.0000
r1:r3  0.4135  0.6431  1.0000
r1:r4  0.2659  0.4135  0.6431  1.0000
r1:r5  0.1710  0.2659  0.4135  0.6431  1.0000
```

To conduct an analysis for data with 2 sources of correlation, please note that we would use the xtmultcorr2 procedure in Stata, with corresponding commands for 2 sources of correlation. For example, for 2 sources of correlation, we would only need to specify the 2 variables that correspond to each source of correlation and the 2 types of correlation structure for each of the two sources of correlation.

5. USING MATLAB

The above analyses can also be performed in Matlab using the qls functions. Here we present a brief outline of the commands and results. Further information can be obtained using the help function, and in the user guide available online.

5.1 Syntax

The qls function has the following syntax:

```
[bhat,alpha,results] =
    qls(id,y,t,X,Family,CorrStruct,varnames,tol,maxit)
```

where id is the unique subject/cluster identification variable, y is the dependent variable, t is the variable(s) indicating the sources of correlation, and X are the covariates. The remaining parameters are optional and are used to specify the assumed distribution of the data (Family, default=Gaussian), assumed correlation structure for each source of correlation (CorrStruct, default=AR(1)), variable names for X used in result displays (varnames), and convergence tolerance (tol) and maximum number of iterations (maxit) used in the qls algorithm. When multiple sources of correlation need to be accounted for, then the qls2 (for 2 sources of correlation) and qls3 (3 sources) functions should be used. These functions have the same syntax as the qls function, except t and CorrStruct have multiple inputs.

The qls functions return three result variables; bhat contains the estimated covariate parameters, alpha contains the estimated correlation parameters, and results contains the final results displayed (estimates, standard errors, p-values, 95% confidence intervals) using both the model based and sandwich type robust covariance matrices.

5.2 Example for three sources of correlation

Here we present the same results that were presented in the previous section, but in Matlab. We will use the

```
example_multlevel
```

data set.

```
>> load example_multlevel;
>> whos
  Name            Size              Bytes  Class

  X             375x2               6000  double array
  family        375x1               3000  double array
  high          375x1               3000  double array
  id            375x1               3000  double array
  o2            375x1               3000  double array
  time          375x1               3000  double array
  type          375x1               3000  double array
  varnames        1x2                144  cell array

Grand total is 3014 elements using 24144 bytes
>>
```

To consider an analysis with 3 sources of correlation, we first need to drop subject with id = 26, and create the id2 variable to ensure balanced data.

```
>> id = id(1:360,1);
>> type = type(1:360,1);
>> time = time(1:360,1);
>> o2 = o2(1:360,1);
>> family = family(1:360,1);
>> high = high(1:360,1);
>> X = [time ones(360,1)];
>> id2 = kron(ones(12,1),[ones(15,1); 2*ones(15,1)]);
```

Next, we can analyze the binary outcome high, with family as the clustering variable and subject, type and time as our three sources of correlation:

```
>>[betah,alphah,results]=
qls3(family,high,[id2 type time],X,'b',{'3','3','1'},varnames);
Bernoulli distribution family assumed
Initial estimate of beta = [0.039611     -0.3873]
AR(1) Correlation structure assumed for Level 3
Equicorrelated structure assumed for Level 1
Equicorrelated structure assumed for Level 2

Stage 1 estimate of alpha1 = 0.024319
Stage 1 estimate of alpha2 = 0.074032
Stage 1 estimate of alpha3 = 0.36418
Stage 1 estimate of beta = [0.034523    -0.38469]
```

```
Stage 2 estimate of alpha1 = 0.048609
Stage 2 estimate of alpha2 = 0.15188
Stage 2 estimate of alpha3 = 0.64307
Stage 2 estimate of beta = [0.030824    -0.38938]

Estimated correlation associated with level one:
    1.0000    0.0486
    0.0486    1.0000

Estimated correlation associated with level two:
    1.0000    0.1519    0.1519
    0.1519    1.0000    0.1519
    0.1519    0.1519    1.0000

Estimated correlation associated with level three:
    1.0000    0.6431    0.4135    0.2659    0.1710
    0.6431    1.0000    0.6431    0.4135    0.2659
    0.4135    0.6431    1.0000    0.6431    0.4135
    0.2659    0.4135    0.6431    1.0000    0.6431
    0.1710    0.2659    0.4135    0.6431    1.0000


QLS estimate of scale parameter = 1

Estimates based on ROBUST covariance matrix
'Variable' 'Beta'    'Std.Error' 'z value' 'p-value'
'time'     [ 0.0308] [   0.0527] [ 0.5849]  [ 0.5586]
'constant' [-0.3894] [   0.3231] [-1.2053]  [ 0.2281]

                '    '95% CI'           ''
'Variable' 'Beta'    'low lim'  'up lim'
'time'     [ 0.0308]  [-0.0725]  [0.1341]
'constant' [-0.3894]  [-1.0226]  [0.2438]


Estimates based on MODEL based covariance matrix
'Variable' 'Beta'    'Std.Error' 'z value' 'p-value'
'time'     [ 0.0308] [   0.0895] [ 0.3445] [ 0.7304]
'constant' [-0.3894] [   0.3383] [-1.1510] [ 0.2497]

                '    '95% CI'           ''
'Variable' 'Beta'   'low lim'   'up lim'
'time'     [ 0.0308] [-0.1445]  [0.2062]
'constant' [-0.3894] [-1.0524]  [0.2737]
```

As for the Stata procedures, for analysis of data with 2 sources of correlation, we would apply the qls2 procedure, with appropriate options for data with 2 sources of correlation.

6. SUMMARY

In this paper we have implemented quasi-least squares for analysis of data with multiple sources of correlation with the user-written xtmultcorr procedures in Stata and the qls procedures in Matlab. It is important to note that the algorithms implemented in xtmultcorr and qls in Stata versus Matlab are technically equivalent, so that choosing between these procedures will depend on the personal preference of the user that may be based on convenience and or familiarity with a particular package. Our software is available for free download on the web-site http://www.cceb.upenn.edu/∼sratclif/QLSproject.html. Currently, efforts are underway to extend our approach for unbalanced data and to extend our SAS macro (Kim and Shults, 2008) for implementation of QLS to data with multiple sources of correlation.

ACKNOWLEDGMENTS

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

Received 1 September 2008

REFERENCES

[1] 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. MR1474184

[2] 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. MR1673345

[3] Chaganty, N. R. and Naik, D. (2002). Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference* **103** 421–436. MR1897004

[4] Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag New York, Inc. MR1883764

[5] Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics – Theory and Methods* **23** 3105–3120.

[6] Gastwirth, J. L. (1966). On Robust Procedures. *Journal of the American Statistical Association* **61** 929–948. MR0205397

[7] Goldstein, H. (1995). *Multilevel statistical models. 2nd ed.* London: Edward Arnold.

[8] Kim, H. and Shults, J. (2008). %QLS SAS Macro: A SAS macro for Analysis of Longitudinal Data Using Quasi-Least Squares. *UPenn Biostatistics Working Papers. Working Paper 27*. http://biostats.bepress.com/upennbiostat/papers/art27

[9] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[10] Lu, N. and Zimmerman, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & Probability Letters* **73** 449–457. MR2187860

[11] Miao, W. and Gastwirth, J. L. (2004). The Effect of Dependence on Confidence Intervals for a Population Proportion. *The American Statistician* **58**(2) 124–130. MR2061196

[12] Naik, D. N. and Rao, S. S. (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *Journal of Applied Statistics* **28** 91–105. MR1834425

[13] Ratcliffe, S. and Shults, J. (2008). GEEQBOX: A Matlab toolbox for generalized estimating equations and quasi-least squares. *Journal of Statisical Software* **25**(14) 1–14.

[14] Roy, A. and Khattree, R. (2005). Testing the hypothesis of a Kroneckar product covariance matrix in multivariate repeated measures data. *SAS Users Group International, Proceedings of the Statistics and Data Analysis Section*. **Paper 199-30** 1–11.

[15] 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.

[16] Shults, J. and Chaganty, N. R. (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics* **54** 1622–1630.

[17] Shults, J. and Morrow, A. (2002). Use of Quasi-Least Squares to Adjust for Two Levels of Correlation. *Biometrics* **58** 521–530. MR1925549

[18] Shults, J., Ratcliffe, Sarah J., and Leonard, M. (2007). Improved generalized estimating equation analysis via xtqls for implementation of quasi-least squares in Stata. *Stata Journal* **7** 147–166.

[19] 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.

[20] Sun, W., Shults, J., and Leonard, M. (2008). A note on the use of unbiased Estimating Equations to Estimate Correlation in Generalized Estimating Equation Analysis of Longitudinal Trials. *Biometrical Journal* **in press**. *Earlier version is posted as a technical report:* http://www.biostatsresearch.com/upennbiostat/papers/art4.

[21] Weissfeld, L. A. and Kshirsagar, A. M. (1992). A modified growth curve model and its application to clinical studies. *Australian Journal of Statistics* **34** 161–168. MR1193769

[22] Xie, J. and Shults, J. (2008). Implementation of quasi-least squares with the R package qlspack. *Journal of Statisical Software* **in press**.

Justine Shults
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
E-mail address: jshults@mail.med.upenn.edu

Sarah J. Ratcliffe
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
E-mail address: sratclif@mail.med.upenn.edu