University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2009 *Paper* 32

"Implementation of quasi-least squares With the R package qlspack"

Jichun Xie* Justine Shults†

*University of Penn, jichun@mail.med.upenn.edu

†Univeristy of Pennsylvania Department of Biostatistics, jshults@mail.med.upenn.edu

”Implementation of quasi-least squares With the R package qlspack”

Jichun Xie and Justine Shults

Abstract

Quasi-least squares (QLS) is an alternative method for estimating the correlation parameters within the framework of generalized estimating equations (GEE) that has two main advantages over the moment estimates that are typically applied for GEE: (1) It guarantees a consistent estimate of the correlation parameter and a positive definite estimated correlation matrix, for several correlation structures; and (2) It allows for easier implementation of some correlation structures that have not yet been implemented in the framework of GEE. Furthermore, because QLS is a method in the framework of GEE, existing software can be employed within the QLS algorithm for estimation of the correlation and regression parameters. In this manuscript we describe and demonstrate the user written package qlspack that allows for implementation of QLS in R software. Our package qlspack calls up the geepack package Yan (2002) and Halekoh et al. (2006) to update the estimate of the regression parameter at the current QLS estimate of the correlation parameter; hence, geepack related functions for standard error estimation can be used after implementing qlspack.

IMPLEMENTATION OF QUASI-LEAST SQUARES: With the R Package qlspack

Jichun Xie
University of Pennsylvania

Justine Shults
University of Pennsylvania

Abstract

Quasi-least squares (QLS) is an alternative method for estimating the correlation parameters within the framework of generalized estimating equations (GEE) that has two main advantages over the moment estimates that are typically applied for GEE: (1) It guarantees a consistent estimate of the correlation parameter and a positive definite estimated correlation matrix, for several correlation structures; and (2) It allows for easier implementation of some correlation structures that have not yet been implemented in the framework of GEE. Furthermore, because QLS is a method in the framework of GEE, existing software can be employed within the QLS algorithm for estimation of the correlation and regression parameters. In this manuscript we describe and demonstrate the user written package **qlspack** that allows for implementation of QLS in R software. Our package **qlspack** calls up the **geepack** package (Yan (2002) and Halekoh, Højsgaard, and Yan (2006)) to update the estimate of the regression parameter at the current QLS estimate of the correlation parameter; hence, **geepack** related functions for standard error estimation can be used after implementing **qlspack**.

Keywords: Cholesky decomposition, correlated data, generalized estimating equations, quasi-least squares, R.

1. Introduction

1.1. Example Data

In this manuscript we describe and demonstrate our user-written package **qlspack** (now available at `http://CRAN.R-project.org/`) that allows for implementation of quasi-least squares (QLS, Chaganty and Shults 1999) in R software (R Development Core Team 2008).

We demonstrate implementation of the **qlspack** package on a data set that represents repeated blood pressure measurements following induced heart attack in 43 rats (page 166 of Davis

2002). These data have the typical structure for a longitudinal GEE analysis because they involve repeated measurements on each rat and the number of rats is relatively large in comparison to the number of measurements per rat. (However, we do note that some rats have 9 measurements; in many longitudinal trials the number of measurements per subject is smaller.) Furthermore, measurements collected on different rats might be reasonably assumed to be independent, while measurements from the same rat might tend to be more similar, and thefore should be correlated.

For demonstration purposes, we will consider simple models that relate the expected blood pressure on the rats with the time of measurement and the group membership of each rat. In addition, we will also consider a binary outcome that takes value one if the rat's blood pressure exceeds the median. As is usual for a GEE analysis, we will relate the expected value of each of these outcome variables with covariates measured on each of the rats, while also adjusting for the potential correlation within the measurements on each cluster.

1.2. Set-up and notation

The set-up and notation are identical for QLS and GEE. In terms of notation, we assume that measurements $Y_i = (y_{i1}, \cdots, y_{in_i})'$ and associated covariates $x_{ij}^\top = (x_{ij1}, \cdots, x_{ijp})$ are collected on rat (subject) i at times $T_i = (t_{i1}, \cdots, t_{in_i})^\top$, for $i = 1, \cdots, m$. The data are considered balanced when $n_i = n \ \forall \ i$ and equally spaced when $|t_{ij} - t_{ij-1}| = \gamma \ \forall \ i$ and $j = 2, \ldots, n_i$. We also let $Y_i = (y_{i1}, \cdots, y_{in_i})^\top$ represent the n_i measurements that were collected within cluster i and define $N = \sum_{i=1}^{m} n_i$. The rat data are not balanced because the number of measurements per rat varies between 1 and 9 (mean = 6.8); in addition, the measurements are unequally spaced in time because they were taken at 1, 5, 10, 15, 30, 60, 120, 180, and 240 minutes after heart attack induction. Note that the rats in this study had a common set of measurement times, which are often planned for in longitudinal studies, though sometimes only approximately achieved in practice.

To conduct a QLS (or GEE) analysis of this data will involve first specifying a generalized linear model for the expected value of the outcome variable: The expected value and variance of measurement y_{ij} on rat (subject) i are then given by $E(y_{ij}) = g^{-1}(x_{ij}^\top \beta) = u_{ij}$ and $Var(y_{ij}) = \phi h(u_{ij})$, respectively, where ϕ is a known or unknown scale parameter. We also let $U_i(\beta)$ represent the $n_i \times 1$ vector of expected values u_{ij} on rat (subject) i.

We then adjust for the correlation amongst repeated measurements on each rat by specifying a *working correlation structure* to describe the pattern of association between measurements within each rat. This working structure for rat (subject) i will be denoted by $Corr(Y_i) = R_i(\alpha)$. The interval on which α yields a positive definite (feasible) correlation matrix will be referred to as the feasible region for $R_i(\alpha)$. The covariance matrix of Y_i is then given by $Cov(Y_i) = \phi {A_i}^{1/2} R_i(\alpha) {A_i}^{1/2}$, where $A_i = diag(h(u_{i1}), \ldots, h(u_{in_i}))$ and ϕ is a scalar parameter that can be known or unknown.

1.3. Correlation structures implemented in qlspack

The **qlspack** package allows for implementation of the following correlation structures:

1. **The Equicorrelated (Exchangeable)**: For this structure all correlations within a cluster are identical, so that $Corr(y_{ij}, y_{ik}) = \alpha$. The feasible region for this structure is $(-1/(n_m - 1), 1)$, where n_m is the maximum value of n_i over $i = 1, 2, \ldots, m$.

2. **The first-order autoregressive AR-1**: For this structure $Corr(y_{ij}, y_{ik}) = \alpha^{|j-k|}$, with feasible region $(-1, 1)$. Note that although the feasible region is $(-1, 1)$ for the AR-1 structure, a negative value for α is often biologically implausible because it yields within subject correlations that alternate in sign. For example, if $\alpha = -0.90$ then the correlation between the first and second measurements on a subject will be $(-0.90)^{|2-1|} = -0.90$, while the correlation between the first and third measurements will be $0.90^{|3-1|} = 0.81$.

3. **The Markov correlation structure**: For this structure $Corr(y_{ij}, y_{ik}) = \alpha^{|t_{ij}-t_{ik}|}$, with feasible region $(-1, 1)$. As for the AR-1 structure, a negative value for α is often biologically implausible.

4. **The tri-diagonal correlation structure**: For this structure $Corr(y_{ij}, y_{ik}) = \alpha$ for $|j-k| = 1$ and is zero otherwise. The feasible region for this structure is $(-1/c_m, 1/c_m)$, where $c_m = 2\sin\left(\frac{\pi[n_m-1]}{2[n_m+1]}\right)$ and n_m is the maximum value of n_i over $i = 1, 2, \ldots, m$; this interval is approximately $(-1/2, 1/2)$ for large n and contains $(-1/2, 1/2)$ for all n.

Note that the working independent (identity) structure is not implemented in **qlspack** because the estimates for this structure are identical for QLS and GEE; hence existing software for implementation of GEE (e.g. R **geepack**) could be used to apply the identity structure. We also note that an algorithm for implementation of an unstructured correlation matrix is provided in (Chaganty and Shults 1999); however, the moment estimate (Liang and Zeger 1986) is more straightforward to implement and can also easily be obtained in existing software for GEE.

1.4. Why use QLS?

There are two primary reasons to consider application of QLS for estimation of the correlation parameter in the framework of GEE. First, as pointed out by Crowder (1995), if the working correlation structure is misspecified in a GEE analysis, a feasible estimate of α may fail to exist. In this situation, the iterative GEE approach may fail to converge, or (if there is convergence in the estimates) $\hat{\alpha}$ may be infeasible. For example, Shults, Ratcliffe, and Leonard (2007) considered a GEE analysis for which of $\hat{\alpha}$ was infeasible for the working tri-diagonal structure, so that the estimated correlation matrix was not positive definite. As summarized in (Shults *et al.* 2007), QLS estimates of α have been proven to be always feasible, for several working correlation structures. QLS might therefore be applied when $\hat{\alpha}$ is infeasible for GEE, or when GEE fails to converge for a particular working correlation structure. (Note however, that infeasibility may be a strong indication that the working structure has been misspecified, as also discussed in Shults et al. (2009).)

Another important reason to consider application of QLS is to allow for application of a biologically plausible correlation structure that has not yet been implemented in the framework of GEE. For example, the AR-1 structure is plausible for longitudinal studies because this structure forces the correlation to decrease with increasing separation in measurement occasion, e.g. if $\alpha = 0.5$ then the correlation between the first and second measurement on each subject is 0.5, while the correlation between the first and third measurement is 0.25. Because the AR-1 structure only depends on measurement occasion, it is appropriate for studies in which the measurements are equally spaced in time, e.g. if $\alpha = 0.9$ then the correlation between *any two* consecutive measurements on a subject must be 0.9.

The Markov correlation generalizes the AR-1 structure to allow for dependence on the actual timing of measurements and is therefore appropriate for analysis of data that are unequally spaced in time. For example, if the measurements have unequal spacing in time, the Markov structure will not force the correlation between all consecutive measurements to be equal. For example, if $\alpha = 0.90$ in the rat study then the correlation between the first and second measurements on each rat (taken at 1 and 5 minutes post-heart attack induction) is $0.90^{5-1} = 0.6561$, while the correlation between the fifth and sixth measurements is $0.90^{60-30} = 0.04$ for the Markov structure (versus 0.90 for the AR-1 structure). The Markov structure is more plausible in this example because the first and second measurements on each rat are closer together in time than the fifth and sixth measurements and therefore might be expected to be more similar, and therefore more highly correlated.

Unforunately, although the Markov structure is relatively simple and biologically plausible for many longitudinal studies, it has not been implemented in the software packages that implement GEE. QLS allows for straightforward implementation of the Markov correlation structure, as we will demonstrate in this manuscript. Also 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. Future planned updates of the **qlspack** package will include implementation of additional correlation structures that have not yet been implemented in the framework of GEE.

2. The Method of QLS

2.1. Brief Description of QLS

QLS is a two-stage computational approach for estimation of the correlation parameters within the framework of GEE. Stage one of (QLS, Chaganty 1997), for balanced and equally spaced data; Shults (1996), and Shults and Chaganty (1998), for unbalanced and unequally spaced data) alternates between updating estimates of the regression parameter β and the correlation parameter α until there is convergence in the estimates. To estimate β in stage one, QLS solves the GEE estimating equation for β (equation (6) Liang and Zeger 1986) at the current estimate of α:

GEE estimating equation for β

$$\sum_{i=1}^{m} D_i^{\top} A_i^{-1/2} R_i^{-1}(\alpha) A_i^{-1/2} \left(Y_i - U_i(\beta)\right) = 0, \tag{1}$$

where $U_i(\beta) = E(Y_i)$ and $D_i = \frac{\partial U_i}{\partial \beta}$.

The estimation approach for β is therefore identical for QLS and GEE. However, the methods differ with respect to estimation of α. While GEE typically implements moment estimates of the correlation parameters, QLS solves an unbiased estimating equation for α in the first stage of the procedure:

Stage one QLS estimating equation for α

$$\frac{\partial}{\partial \alpha}\left\{\sum_{i=1}^{m} Z_i^{\top}(\beta)\left\{R_i^{-1}(\alpha)\right\} Z_i(\beta)\right\}=0, \tag{2}$$

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

The solution $\hat{\alpha}$ to (2) is not consistent. Stage two of QLS therefore obtains a consistent estimate $\hat{\alpha}_{\mathrm{QLS}}$ as the solution to the stage two estimating equation for α (Chaganty and Shults 1999):

Stage two QLS estimating equation for α

$$\sum_{i=1}^{m} trace\left\{\frac{\partial R_i^{-1}(\delta)}{\partial \delta} R_i(\alpha)\right\}\Bigg|_{\delta=\hat{\alpha}}=0. \tag{3}$$

The final QLS estimate of β is then obtained by again solving the GEE estimating equation (1) for β, evaluated at $\hat{\alpha}_{\mathrm{QLS}}$. An advantage of QLS is that the stage one and stage two estimates of the correlation parameters will be feasible and consistent for several correlation structures when the correlation structure is correctly specified (Chaganty and Shults 1999). In section 2.3 we present solutions to the stage one and stage two estimating equations for the correlation structures that we consider in this manuscript.

The asymptotic distribution of $\hat{\beta}_{\mathrm{QLS}}$ is identical to the asymptotic distribution of $\hat{\beta}_{GEE}$. As a result, testing and confidence intervals for the regression parameter with QLS can be implemented using existing approaches for GEE, as we will demonstrate in Section 3.

We note that the current manuscript implements QLS in R software. Earlier, we implemented QLS in the `xtqls` procedure in Stata (Shults *et al.* 2007) and in the GEEQBOX package in Matlab (Ratcliffe and Shults 2006). The two manuscripts just cited contain similar descriptions of QLS and implement the same algorithms for estimation of α for QLS, in programs written initially for Stata and then translated to Matlab (The MathWorks, Inc. 2007). In addition to the manuscripts cited earlier, other papers on QLS include Chaganty and Naik (2002), Chaganty (2003), and Shi and Chaganty (2004). We also note that an important paper for the development of QLS was by Dunlop (1994) who described the link between GEE and least squares. For an excellent and thorough description of GEE, also see the text by Hardin and Hilbe (2003). Also see Sun, Shults, and Leonard (2009) for a comparison of QLS with some alternative approaches based on unbiased estimating equations.

2.2. Algorithm for QLS

The QLS procedure implements the following algorithm:

1. Obtain a starting value for $\hat{\beta}$ by assuming $\alpha=0$ and then obtaining a solution (for β) to the GEE estimating equation for β, evaluated at $\alpha=0$. (Note that solving the GEE estimating equation at $\alpha=0$ is equivalent to using linear regression, logistic regression, or Poisson regression to obtain a starting value for $\hat{\beta}$ for outcomes that are continuous, binary, or that represent counts, respectively.)

2. Alternate between the following steps until there is convergence in the estimates of β:

   (a) Update the Pearson residuals at the current estimate of β, where the j^{th} Pearson residual on subject i is given by:

   $$z_{ij} = \frac{y_{ij} - \hat{u_{ij}}}{h(\hat{u_{ij}})}.$$

   (b) Obtain an updated estimate of α solving the QLS stage one estimating equation (2) for α; this requires specification of the working correlation structure and involves the Pearson residuals and timings of measurements on each subject. As described in Section 2.3 there is an explicit solution (that is a function of the z_{ij}) for the AR-1 structure; the method of bisection is used to solve the equation for the other structures.

   (c) Update the estimate of β by solving the GEE estimating equation (1) for β for the pre-specified working correlation structure that is evaluated at the current estimate of α.

3. After convergence in stage one, update the estimate of α by obtaining the solution to the QLS stage two estimating equation (3) for α. The stage two estimate for the AR-1 structure is a simple function of the stage one estimate; otherwise, the method of bisection is used to solve the stage two estimating equation that depends on the form of the working correlation structure and involves the stage one estimate and timings of measurements on each subject.

4. Obtain the final estimate of β by solving the GEE estimating equation (1) for β for the pre-specified working correlation structure that is evaluated at the stage two estimate of α.

The algorithm in **qlspack** uses the **geepack** R function to solve the GEE estimating equation in steps 2 (c) and 4 for β, at the current estimates of the correlation parameter. It is based on an algorithm described by Shults *et al.* (2007). We also note that Hardin and Hilbe (2003) demonstrates a similar algorithm, but with a moment estimate for α, for a correlation structure that is currently unsupported for GEE. Prior to the widespread availability of software packages that allowed for solution of the GEE estimating equation for β at fixed values of the correlation structure, Shults (1996) and Shults and Chaganty (1998) solved the GEE estimating equation using an approach based on the Cholesky decomposition of the inverse of the working correlation structure. This approach is implemented in Ratcliffe and Shults (2006).

2.3. Stage One and Stage Two Estimates of α

The QLS procedure in **qlspack** obtains provides solutions to the stage one (2) and stage two (3) estimating equations for several working correlation structures. For estimating equations that do not have an explicit solution, **qlspack** uses the bisection method to obtain a solution in the feasible region for α.

For the AR(1) structure and for unbalanced data, Shults and Chaganty (1998) proved that

the feasible stage one estimate $\hat{\alpha}$ can be expressed as:

$$\hat{\alpha}_{\text{QONE}} = \frac{\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}+z_{ij-1})^2 - \sqrt{\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}+z_{ij-1})^2 \sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}-z_{ij-1})^2}}{2\sum_{i=1}^{m}\sum_{j=2}^{n_i} z_{ij}z_{ij-1}}, \tag{4}$$

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

$$\hat{\alpha}_{\text{QLS-AR1}} = \frac{2\hat{\alpha}_{\text{QONE}}}{1+\hat{\alpha}^2_{\text{QONE}}}. \tag{5}$$

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

$$\sum_{i=1}^{m}\sum_{j=2}^{n_i} \frac{e_{ij}\alpha^{e_{ij}}\left[\alpha^{2e_{ij}} z_{ij}z_{i,j-1} - \alpha^{e_{ij}}\left(z_{ij}^2 + z_{i,j-1}^2\right) + z_{ij}z_{i,j-1}\right]}{(1-\alpha^{2e_{ij}})^2} = 0, \tag{6}$$

where $e_{ij} = |t_{ij} - t_{i,j-1}|$. Note that **qlspack** requires that $e_{ij} \geq 1 \; \forall \, i$ and j.

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

$$\sum_{i=1}^{m}\sum_{j=2}^{n_i} \frac{2e_{ij}\delta^{2e_{ij}-1} - \alpha^{e_i j}e_{ij}\left[\delta^{e_{ij}-1} + \delta^{3e_{ij}-1}\right]}{(1-\delta^{2e_{ij}})^2}\Bigg|_{\delta=\hat{\alpha}} = 0. \tag{7}$$

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

$$\sum_{i:n_i>1} Z_i^\top Z_i - \sum_{i:n_i>1} \frac{1+\alpha^2(n_i-1)}{(1+\alpha(n_i-1))^2}\,(Z_i^\top(\beta)\, e_i)^2 = 0, \tag{8}$$

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

$$\sum_{i:n_i>1} \frac{n_i\,(n_i-1)\,\widehat{\alpha}\,(\widehat{\alpha}\,(n_i-2)+2)}{\left(1+\widehat{\alpha}(n_i-1)\right)^2} \Big/ \sum_{i:n_i>1} \frac{n_i\,(n_i-1)\,\left(1+\widehat{\alpha}^2(n_i-1)\right)}{\left(1+\widehat{\alpha}(n_i-1)\right)^2}. \tag{9}$$

For the tri-diagonal structure and unbalanced data, Shults (1996) proved that there will always be a feasible solution to the stage one estimating equation for α. **qlspack** obtains solutions to the stage one and two estimating equations (2) and (3) for the tri-diagonal structure by first constructing the tri-diagonal matrix $R_i(\hat{\alpha})$ and then using the R function `solve` to obtain $R_i^{-1}(\hat{\alpha})$. Next, to evaluate

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

qlspack implements the following expression:

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

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

3. Examples

3.1. Rat Data

Here we demonstrate implementation of **qlspack** for analysis of the rat data that were described in the Introduction.

The data are available in Table 6.11 on page 166 of Davis (2002) who provides an excellent description of methods for analysis of data with repeated measurements. For convenience, they are also printed in the Appendix and are available as the text file `rat.txt` on the web-site `http://www.cceb.upenn.edu/~sratclif/QLSproject.html`. This file contains the data displayed in the appendix with some additional variables that were created for the demonstration analyses:

Columns in `rat.txt`

id2 id time group bp group1 group2 group3 group4 highbp

The description of each column is as follows:

- `id2`: the id variable for each rat that is provided in Davis (2002)
- `id`: a new id variable that takes value 1,2,..43 after sorting on id and group
- `time`: the timing of each measurement
- `group`: the group variable that takes value 1, 2, 3, or 4
- `bp`: the blood pressure value
- `group1`: indicator variables for group 1 that takes value one for rats in group 1 and that takes value 0 otherwise
- `group2`: indicator variables for group 2 that takes value one for rats in group 2 and that takes value 0 otherwise
- `group3`: indicator variables for group 3 that takes value one for rats in group 3 and that takes value 0 otherwise
- `group4`: indicator variables for group 4 that takes value one for rats in group 4 and that takes value 0 otherwise
- `highbp`: a variable that takes value 1 if the rat's blood pressure is at least 100.

Note that Table 6.11 on page 166 of Davis (2002) displays duplicate ids for the rats, e.g. `id2` takes value 1 for the first rat within each group. For this reason, we needed to create a new id (`id`) that takes distinct values for different rats. The variable `id` should be used in the analysis. Other variables described above that are not included in the Appendix include `group1`, `group2`, `group3`, `group4`, and `highbp`.

Because the blood pressure measurements are not equally spaced in time, the Markov correlation structure is biologically plausible for their analysis. However, we will also demonstrate analyses that implement the AR-1, equicorrelated, and tri-diagonal correlation structures.

To demonstrate the **qlspack** package, we will fit a simple regression model for expected blood pressure that includes time and indicator variables for groups 2, 3, and 4. In addition, we will consider a binary outcome that takes value 1 if the blood pressure is at least 100 (the median blood pressure value).

3.2. Syntax for qlspack Package

The syntax for the `qls` function in **qlspack** is very similar to the syntax of the **geepack** R function. The following command will yield output that includes estimates of the regression and correlation coefficients and standard errors for the regression coefficients:

```
R> qlsfit.ar1<- qls(formula, data, id, family = "gaussian", time,
 + correlation = "ar1", std.err = san.se)
```

The argument `formula` indicates the regression model; data is the name of the R data set. We need to list all the variables we would like to include in the model in the `formula` argument.

The argument `id` is the variable name for the subject id.

The argument `family` is the name of the two parameter exponential family that is selected. Possible values for `family` include `gaussian`, `binomial`, and `poisson`. The link and variance functions that correspond to each of these family choices include: (`gaussian`) the identity link function $g^{-1}(\gamma) = \gamma$ and variance function $h(\gamma) = 1$; (*binomial*) the logistic link function $g^{-1}(\gamma) = \exp(\gamma)/(1 + \exp(\gamma))$ and variance function $h(\gamma) = \gamma(1 - \gamma)$; (`poisson`) the exponential link $g^{-1}(\gamma) = \exp(\gamma)$ and identity variance function $h(\gamma) = \gamma$. The default value for `family` is `gaussian`.

The argument `time` is the name of the timing variable. This argument is useful only when the argument `correlation` takes the value `markov`. The default value for `time` is `"NA"`.

The argument `correlation` is the name of the correlation structure. Possible values for `correlation` include `ar1` (AR-1 structure);`exchangeable` (equicorrelated structure); `markov` (Markov structure); and `tri-diagonal` (tri-diagonal structure).

The argument `std.err` is the name of the type of standard errors (for $\hat{\beta}$) that will be applied in the analysis. Possible values for *std.err* include `san.se` (sandwich robust estimate); `jack` (approximate jacknife estimate); `j1s` (1-step jacknife variance estimate); and `fij` (fully iterated jacknife variance estimate). The default value is `san.se`.

In general, the data for implementation of `qls` should have the same structure as for implementation of **geepack**. The id for different clusters should be different, but need not to be consecutive.

For detailed information like standard error estimates for $\hat{\beta}$ and corresponding p-values, we can use

```
R> summary(qlsfit.ar1)
```

It is also important to note that the **geepack** package must be installed prior to use of **qlspack**.

3.3. QLS Analysis of Blood pressure for Several Correlation Structures

Fitting the regression model with the Markov structure is achieved by first copying the `rat.txt` data into the current working directory and installing the **geepack** and **qlspack** packages. Next we can set our working directory in R and open the `rat.txt` data set. Note also that we will implement the default sandwich covariance matrix for estimation of $Cov(\hat{\beta})$.)

```
R> data(rat)
```

Next, let us fit the Markov correlation structure in a QLS analysis that regresses blood-pressure on time and indicator variables for groups 2, 3, and 4:

```
R> qlsfit.mkv <- qls(formula = bp ~ time + group2 + group3 + group4, data =
 + rat, id = rat$id, time = rat$time, family = "gaussian",
 + correlation = "markov")
R> summary(qlsfit.mkv)
```

Please note that in the `formula` argument, we must list all the variables we need to include in the model separately. We cannot use the form like `formula=bp~time+as.factor(group)`.

```
Call: qls(formula = bp ~ time + group2 + group3 + group4, data = rat,
    id = rat$id, family = "gaussian", time = rat$time,
    correlation = "markov")

Coefficients:
                 Estimate    Std.err         Wald       p(>W)
(Intercept) 102.404408615 3.50183923 855.15511945 0.000000000
time          0.001588161 0.01330833   0.01424106 0.905009229
group2        2.121321466 4.41045303   0.23133778 0.630533651
group3       -9.767634073 5.54060793   3.10787773 0.077914423
group4      -21.812644901 8.45134651   6.66138971 0.009852405

Estimated Scale Parameters:
            Estimate  Std.err
(Intercept)  396.275 64.44642

Correlation: Structure = markov  Link = identity

Estimated Correlation Parameters:
   Estimate Std.err
1 0.9428284       0
Number of clusters:   43   Maximum cluster size: 9
```

The estimated correlation for the Markov structure is 0.9428.

Next, to fit the equicorrelated structure (with abbreviated output) we use:

```
R> qlsfit.exch <- qls(formula = bp ~ time + group2 + group3 + group4, data =
 + rat, id = rat$id, time = rat$time, family = "gaussian",
 + correlation = "exchangeable")
R> summary(qlsfit.exch)

Call: qls(formula = bp ~ time + group2 + group3 + group4, data = rat, id =
rat$id, family = "gaussian", time = rat$time, correlation =
"exchangeable")

Coefficients:
  (Intercept)          time        group2        group3        group4
 1.001260e+02 -9.154816e-04  2.537508e+00 -1.294372e+01 -1.883122e+01

Degrees of Freedom: 291 Total (i.e. Null);  286 Residual

Scale Link:                    identity
Estimated Scale Parameters:  [1] 400.6342

Correlation:  Structure = exchangeable  Link = identity
Estimated Correlation Parameters:
[1] 0.704539

Number of clusters:   43   Maximum cluster size: 9
```

The estimated correlation for the equicorrelated structure is 0.7045.

Next, to fit the AR-1 structure we use:

```
R> qlsfit.ar1 <- qls(formula = bp ~ time + group2 + group3 + group4, data =
 + rat, id = rat$id, time = rat$time, family = "gaussian",
 + correlation = "ar1")
R> summary(qlsfit.ar1)

Call: qls(formula = bp ~ time + group2 + group3 + group4, data = rat, id =
rat$id, family = "gaussian", time = rat$time, correlation = "ar1")

Coefficients:
  (Intercept)          time        group2        group3        group4
 1.012666e+02 -9.380585e-04  2.507180e+00 -1.313150e+01 -2.083418e+01

Degrees of Freedom: 291 Total (i.e. Null);  286 Residual

Scale Link:                    identity
Estimated Scale Parameters:  [1] 399.1677
```

```
Correlation:  Structure = ar1  Link = identity
Estimated Correlation Parameters:
  alpha:1
0.7762648

Number of clusters:   43   Maximum cluster size: 9
```

The estimated value for α is 0.7763.

Next, for the tri-diagonal structure, we use:

```
R> qlsfit.tri <- qls(formula = bp ~ time + group2 + group3 + group4, data =
 + rat, id = rat$id, time = rat$time, family = "gaussian",
 + correlation = "tridiagonal")
R> summary(qlsfit.tri)

Call: qls(formula = bp ~ time + group2 + group3 + group4, data = rat, id =
rat$id, family = "gaussian", time = rat$time, correlation =
"tridiagonal")

Coefficients:
 (Intercept)         time       group2       group3       group4
 97.91145037   0.01966653  12.73875053 -10.69637510 -16.16715911

Degrees of Freedom: 291 Total (i.e. Null);  286 Residual

Scale Link:                    identity
Estimated Scale Parameters:  [1] 420.5853

Correlation:  Structure = tridiagonal  Link = identity
Estimated Correlation Parameters:
  alpha:1
0.5223492

Number of clusters:   43   Maximum cluster size: 9
```

The estimated value of α was 0.5223 for the tri-diagonal structure.

3.4. Comparison of GEE and QLS for Analysis of Blood Pressure

To compare the results in GEE versus QLS, we used **geepack** as described in Halekoh *et al.* (2006) to implement the above models in GEE. For example, here we demonstrate implementation of **geepack** for the AR-1 structure, which yields results that differ slightly from those for QLS:

```
R> geeglm( highbp ~ time + group2 + group3 + group4, family =
 + gaussian, data=rat, id=rat$id, corstr="ar1")
```

```
Call: geeglm(formula = bp ~ time + group2 + group3 + group4, family =
gaussian, data = rat, id = rat$id, corstr = "ar1")

Coefficients:
  (Intercept)          time        group2        group3    group4
101.292169249  -0.004733925   3.105246676 -13.992058591  -21.285179098


Degrees of Freedom: 291 Total (i.e. Null);  286 Residual

Scale Link:                    identity Estimated Scale Parameters:
[1] 402.9202

Correlation:  Structure = ar1  Link = identity Estimated Correlation
Parameters:
    alpha
0.8696674

Number of clusters:   43   Maximum cluster size: 9
```

The equicorrelated and tri-diagonal structures could be implemented by changing the value of `corstr` in the above command. Please (see Halekoh *et al.* 2006) for more details regarding implementation of the R **geepack** for implementation of GEE in R.

Table 1 displays the estimation results for QLS versus GEE, for the sandwich based covariance matrix. Note that the Markov structure is not available for analysis in **geepack** and is therefore not included for GEE.

Table 2 displays the QLS and GEE estimates for α for each of the correlation structures displayed above.

It is interesting to note that for this example, the feasible region for the tri-diagonal structure is $(-0.526, 0.526)$. The GEE moment estimate for α is 0.742 which exceeds the upper limit for the feasible region for α. The application of GEE with moment estimates therefore yields some non positive definite estimated correlation matrices for the tri-diagonal structure, which can be checked by obtaining the eigenvalues for this structure and demonstrating that some eigenvalues are negative. For example, the estimated correlation matrix for rats who were measured at all measurement occasions is a 9×9 tri-diagonal correlation structure; if $\alpha = 0.742$ then the eigenvalues for this structure are 2.412, 2.201, 1.872, 1.459, 1, 0.541, 0.128, -0.201, and -0.411. In contrast to the GEE moment estimate, the QLS estimate for α (0.522) is inside the feasible region for α and corresponds to a 9×9 tri-diagonal structure with positive eigenvalues 1.992, 1.844, 1.614, 1.323, 1, 0.677, 0.386, 0.155, and 0.007. However, as discussed in **?** an infeasible estimate for α might be a sign that the correlation structure has been misspecified. The infeasibility of the moment estimate for the tri-diagonal structure, coupled with the fact that it is not biologically plausible for this analysis, might therefore be used to exclude this structure from consideration as a working correlation structure in this analysis.

3.5. QLS analysis of a Binary Outcome in the Rat Data

Note also that QLS analyses can be conducted for binary and count outcomes, as well. For

	$\hat{\beta}_{GEE}$	se_{GEE}	p_{GEE}		$\hat{\beta}_{\text{QLS}}$	se_{QLS}	p_{QLS}
MARKOV							
Constant					102.404	3.502	<.001
Time					0.002	0.013	0.905
Group2					2.121	4.410	0.631
Group3					−9.768	5.541	0.078
Group4					−21.813	8.451	0.010
EQUI							
Constant	100.23	5.776	<.001		100.13	5.918	<.001
Time	−0.0004	0.011	0.969		−0.0009	0.011	0.933
Group2	2.369	6.883	0.731		2.538	7.032	0.718
Group3	−12.842	8.101	0.113		−12.944	8.269	0.118
Group4	−18.853	9.703	0.052		−18.831	9.817	0.055
TRI							
Constant	101.72	3.972	<.001		97.911	4.886	<.001
Time	0.0076	0.017	0.652		0.020	0.030	0.519
Group2	−0.952	5.181	0.854		12.739	6.759	0.059
Group3	−10.531	5.645	0.062		−10.696	6.799	0.116
Group4	−19.836	8.398	0.018		−16.167	7.277	0.026
AR-1							
Constant	101.29	5.543	<.001		101.266	3.962	<.001
Time	−0.0047	0.011	0.662		−0.0009	0.011	0.934
Group2	3.105	6.508	0.633		2.507	6.058	0.679
Group3	−13.992	7.537	0.063		−13.132	7.003	0.061
Group4	−21.285	9.321	0.022		−20.834	9.057	0.021

Table 1: Estimation results for QLS versus GEE.

example, the regression model for the binary outcome highbp and Markov structure is fit using the following command:

```
R> qls( highbp ~ time + group2 + group3 + group4, data = rat, id =
 + rat$id, time = rat$time, family = binomial, correlation="markov")

Call: qls(formula = highbp ~ time + group2 + group3 + group4, data=rat,
    id = rat$id, family = binomial, time = rat$time, correlation = "markov")

Coefficients:
  (Intercept)          time        group2        group3        group4
 0.5623337877  0.0009521582  0.1683637564 -0.5914757350 -1.4953002081

Degrees of Freedom: 291 Total (i.e. Null);  286 Residual

Scale Link:                identity
Estimated Scale Parameters: [1] 1.061967
```

	Markov	TRI	EQUI	AR-1
GEE		0.742	0.623	0.870
QLS	0.943	0.522	0.705	0.776

Table 2: QLS and GEE estimate for different correlation structures.

```
Correlation:  Structure = markov  Link = identity
Estimated Correlation Parameters:
  alpha:1
0.9539716

Number of clusters:   43   Maximum cluster size: 9
```

The estimated value of α was 0.9539 for the Markov structure.

Similarly, the AR-1, exchangeable and tri-diagonal structures will be implemented by replacing `correlation="markov"` in the above command by `correlation="ar1"`, `correlation="exchangeable"` and `correlation="tri-diagonal"`, respectively.

3.6. Relationship to R package geepack

qlspack updates the QLS estimates for β by using the `geeglm` function in **geepack** to solve the GEE estimating equation (for β) at the current QLS estimates of the correlation parameter. As a result, the functions which can be applied to the return value of `geeglm` can also be applied to that of **qlspack**. Because **geepack** provides estimates of the estimated covariance matrix of β based on several approaches (robust sandwich covariance, fully iterated jacknife, 1-step jacknife, and approximate jacknife) **qlspack** therefore also allows for application of these estimates. As we have seen, the default approach is application of the robust sandwich covariance.

4. Discussion

We implemented QLS using the user-written package **qlspack** for R software. This allowed for application of the Markov correlation structure that previously was not offered as a potential working correlation structure in the major software packages for GEE. In addition, it allowed for feasible estimation with a tri-diagonal correlation structure when the moment estimate was infeasible (and therefore yielded some estimated correlation matrices that were not positive definite). Future updates of **qlspack** are planned, to incorporate additional correlation structures that have not yet been implemented in the framework of GEE and to allow for analysis of multi-level correlated data.

5. Acknowledgments

Work on this manuscript was supported by the NIH funded grant R01CA096885 "Longitudinal Analysis for Diverse Populations". The authors are grateful to Dr. Jun Yan and Dr. Søren

Højsgaard, for being very responsive to our queries about **geepack** in R.

A. Table 6.11 of Davis (2002)

		Minutes after Ligation								
Grp	ID	1	5	10	15	30	60	120	180	240
1	1	112.5	100.5	102.5	102.5	107.5	107.5	95.0	102.5	100.5
1	2	92.5	102.5	105.0	100.0	110.0	117.5	97.5	102.5	112.5
1	3	132.5	125.0	115.0	112.5	110.0	110.0	127.5	.	.
1	4	110.0	110.0	.	.	.	.	.	.	.
1	5	122.5	127.5	.	.	.	.	.	.	.
1	6	102.5	107.5	107.5	102.5	90.0	112.5	107.5	110.0	112.5
1	7	42.5	42.5	.	.	.	.	.	.	.
1	8	107.5	80.0	.	.	.	.	.	.	.
1	9	110	130	115	105	112.5	110	115	102.5	92.5
1	10	97.5	97.5	80	82.5	82.5	102.5	100	95	95
1	11	90	70	85	85	92.5	97.5	107.5	97.5	90
2	1	115	115	107.5	107.5	112.5	107.5	112.5	107.5	107.5
2	2	120	.	.	.	.	.	.	.	.
2	3	125	125	120	120	117.5	125	122.5	120	120
2	4	95	90	95	90	100	107.5	100	100	92.5
2	5	97.5	70	.	.	.	.	.	.	.
2	6	87.5	65.5	85	90	105	90	85	87.5	100
2	7	90	87.5	97.5	95	100	95	102.5	.	.
2	8	97.5	92.5	57.5	55	90	97.5	110	115	105
2	9	107.5	107.5	145	110	105	105	112.5	.	.
2	10	102.5	130	85	80	127.5	97.5	117.5	102.5	127.5
3	1	107.5	107.5	102.5	102.5	102.5	97.5	98.5	102.5	92.5
3	2	67.5	20	.	.	.	.	.	.	.
3	3	97.5	108.5	94.5	102.5	102.5	107.5	117.5	112.5	.
3	4	105	105	.	.	.	.	.	.	.
3	5	85	60	.	.	.	.	.	.	.
3	6	100	105	105	105	110	110	115	107.5	105
3	7	95	95	90	100	100	100	95	90	100
3	8	85	92.5	92.5	92.5	90	110	100	102.5	87.5
3	9	82.5	77.5	75	65.5	65	72.5	72.5	67.5	67.5
3	10	92.5	75	40	35	.	.	.	.	.
3	11	62.5	75	115	110	100	100	.	.	.
4	1	70	67.5	67.5	77.5	77.5	77.5	72.5	65	55
4	2	45	37.5	45	45	47.5	45	50	45	50
4	3	52.5	22.5	90	65	60	65.5	52.5	47.5	57.5
4	4	100	100	100	100	97.5	92.5	.	.	.
4	5	47.5	30	.	.	.	.	.	.	.
4	6	102.5	90	.	.	.	.	.	.	.
4	7	115	110	100	110	105	105	105	105	105
4	8	97.5	97.5	97.5	105	95	92.5	92.5	92.5	92.5
4	9	95	125	130	125	115	117.5	110	105	102.5
4	10	72.5	87.5	65	57.5	92.5	82.5	57.5	50	50
4	11	105	105	105	105	102.5	100	95	92.5	87.5

References

Chaganty N (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 (2003). "Analysis of growth curves with patterned correlation matrices using quasi-least squares." *Journal of Statistical Planning and Inference*, **117**, 123–139.

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

Chaganty N, Shults J (1999). "On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter." *Journal of Statistical Planning and Inference*, **76**, 127–144.

Crowder M (1995). "On the use of a working correlation matrix in using generalised linear models for repeated measures." *Biometrika*, **82**, 407–410.

Davis C (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag Inc.

Dunlop D (1994). "Regression for Longitudinal Data: A Bridge from Least Squares Regression." *The American Statistician*, **48**, 299–303.

Halekoh U, Højsgaard S, Yan J (2006). "The R Package **geepack** for Generalized Estimating Equations." *Journal of Statistical Software*, **15**(2), 1–11.

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

Liang K, Zeger S (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, **73**, 13–22.

Ratcliffe S, Shults J (2006). "**GEEQBOX**: A MATLAB toolbox for implementation of quasi-least squares and generalized estimating equations." *Journal of Statistical Software*, **25(14)**, 1–13.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shi G, Chaganty N (2004). "Application of quasi-least squares to analyze replicated autoregressive time series regression models." *Journal of Applied Statistics*, **31**, 1147–1156.

Shults J (1996). *The analysis of unbalanced and unequally spaced longitudinal data using quasi-least squares*. Ph.D. thesis, Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia.

Shults J, Chaganty N (1998). "Analysis of Serially Correlated Data Using Quasi-Least Squares." *Biometrics*, **54**(4), 1622–1630.

Shults J, Mazurick C, Landis J (2006). "Analysis of repeated bouts of measurements in the framework of generalized estimating equations." *Statistics in Medicine*, **25**(23), 4114–4128.

Shults J, Morrow A (2002). "Use of Quasi-Least Squares to Adjust for Two Levels of Correlation." *Biometrics*, **58**, 521–Ű530.

Shults J, Ratcliffe S, Leonard M (2007). "Improved generalized estimating equation analysis via xtqls for implementation of quasi-least squares in Stata." *Stata Journal*, **7**(2), 147–166.

Shults J, Whitt M, Kumanyika S (2004). "Analysis of data with multiple sources of correlation in the framework of generalized estimating equations." *Statistics in Medicine*, **23**(20), 3209–3226.

Sun W, Shults J, Leonard M (2009). "A Note on the Use of Unbiased Estimating Equations to Estimate Correlation in Analysis of Longitudinal Trials." *Biometrical Journal*, **51 (1)**, 5–18.

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

Yan J (2002). "**geepack**: yet another package for generalized estimating equations." *R News*, **2**, 12–14.

Affiliation:

Jichun Xie (PhD candidate) & Justine Shults (Associate Professor)
Department of Biostatistics and Epidemiology
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
Philadelphia, PA 19104
E-mail for Jichun Xie: jichun@mail.med.upenn.edu
E-mail for Justine Shults: jshults@mail.med.upenn.edu
URL: http://www.cceb.upenn.edu/~sratclif/QLSproject.html