University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2008 *Paper* 27

”%QLS SAS Macro: A SAS macro for Analysis of Longitudinal Data Using Quasi-Least Squares”.

Hanjoo Kim* Justine Shults†

*University of Pennsylvania School of Medicine, hanjoo@mail.med.upenn.edu

†University of Pennsylvania, jshults@mail.med.upenn.edu

”%QLS SAS Macro: A SAS macro for Analysis of Longitudinal Data Using Quasi-Least Squares”.

Hanjoo Kim and Justine Shults

Abstract

Quasi-least squares (QLS) is an alternative computational approach for estimation of the correlation parameter in the framework of generalized estimating equations (GEE). QLS overcomes some limitations of GEE that were discussed in Crowder (Biometrika 82 (1995) 407-410). In addition, it allows for easier implementation of some correlation structures that are not available for GEE. We describe a user written SAS macro called %QLS, and demonstrate application of our macro using a clinical trial example for the comparison of two treatments for a common toenail infection. %QLS also computes the lower and upper boundaries of the correlation parameter for analysis of longitudinal binary data that were described by Prentice (Biometrics 44 (1988), 1033-1048). Furthermore, it displays a warning message if the Prentice constraints are violated; This warning is not provided in existing GEE software packages and other packages that were recently developed for application of QLS (in Stata, Matlab, and R). %QLS allows for analysis of normal, binary, or Poisson data with one of the following working correlation structures: the first-order autoregressive (AR(1)), equicorrelated, Markov, or tri-diagonal structures. Keywords: longitudinal data, generalized estimating equations, quasi-least squares, SAS.

%QLS SAS Macro: A SAS macro for Analysis of Longitudinal Data Using Quasi-Least Squares

Hanjoo Kim
University of Pennsylvania
School of Medicine

Justine Shults
University of Pennsylvania
School of Medicine

Abstract

Quasi-least squares (QLS) is an alternative computational approach for estimation of the correlation parameter in the framework of generalized estimating equations (GEE). QLS overcomes some limitations of GEE that were discussed in Crowder (*Biometrika* **82** (1995) 407-410). In addition, it allows for easier implementation of some correlation structures that are not available for GEE. We describe a user written SAS macro called **%QLS**, and demonstrate application of our macro using a clinical trial example for the comparison of two treatments for a common toenail infection. **%QLS** also computes the lower and upper boundaries of the correlation parameter for analysis of longitudinal binary data that were described by Prentice (*Biometrics* **44** (1988), 1033-1048). Furthermore, it displays a warning message if the Prentice constraints are violated; This warning is not provided in existing GEE software packages and other packages that were recently developed for application of QLS (in Stata, Matlab, and R). **%QLS** allows for analysis of normal, binary, or Poisson data with one of the following working correlation structures: the first-order autoregressive (AR(1)), equicorrelated, Markov, or tri-diagonal structures.

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

1. Introduction

Quasi-least squares (QLS) (Chaganty 1997; Shults & Chaganty 1998; Chaganty & Shults 1999) is a two-stage approach for fitting longitudinal data that are correlated over time using an assumed working correlation structure in the framework of generalized estimating equations (GEE) (Liang & Zeger 1986). Both QLS and GEE estimate the regression parameter β via solution of the same estimating equation which contains an additional unknown correlation parameter α. These methods alternate between updating the estimate of β, and updating the estimate of α iteratively, but they differ with respect to estimation of α; GEE uses a moment estimate of α based on the Pearson residual, while QLS solves estimating equations for α that are orthogonal to the GEE estimating equation of β.

There are primarily two important reasons to consider the implementation of QLS over GEE. Crowder (1995) pointed out that there may be no feasible estimate of α under misspecification of the working correlation structure, which could result in a failure to converge for the iterative GEE estimation procedure. On the other hand, the QLS estimates of α are guaranteed to be feasible for several structures, including the first-order autoregressive (AR(1)), equicorrelated, Markov, and tri-diagonal structures. The other main advantage of the application of QLS is that it allows for relatively straightforward implementation of more complex and/or biologically plausible correlation structures than are currently available in GEE due to difficulty in obtaining a moment estimate of α. For example, Shults (1996) and Chaganty & Shults (1999) described estimating equations for α for the Markov correlation structure that is a generalization of the AR(1) structure for measurements that are unequally spaced in time. Furthermore, QLS has been successfully implemented in multi-outcome longitudinal studies which contain multiple sources of correlations, e.g. correlations between multiple outcomes as well as within subject for each outcome over time, using Kronecker product-based correlation structures (see Shults & Morrow (2002), Chaganty & Naik (2002), Shults, Whitt & Kumanyika (2004), and Shults & Ratcliffe (2007b)). In addition, Shults, Mazurick & Landis (2006) implemented a banded Toeplitz strucure that had previously not been implemented for GEE.

There has been a considerable effort to disseminate and promote the use of QLS via the development of user written QLS programs in some of the major statistical software packages. For example, Shults et al. (2007a) developed **xtqls** using Stata's proprietary higher programming language. Xie & Shults (2008) also developed a user written QLS program called **qlspack** using R. Theses softwares estimate the regression parameter β by calling existing GEE procedures, e.g. **xtgee** in Stata and the user written package **geepack** (Yan 2002) in R, and then solving the stage one and two estimating equations for α iteratively. Lastly, Ratcliffe & Shults (2008) developed **GEEQBOX** software in Matlab for the implementation of both GEE and QLS; Prior to introduction of **GEEQBOX**, there was no software available for the implementation of GEE in Matlab. These QLS softwares can be downloaded from `www.cceb.upenn.edu/~sratclif/QLSproject.html`, and **GEEQBOX** is available for download on the web-site for the *Journal of Statistical Software* via `http://www.jstatsoft.org/v25/i14`.

This manuscript reflects our continued ongoing effort to develop a user written QLS software in SAS (SAS 2003) which is one of the most widely used software packages in practice, due to its versatility including varied and powerful procedures for data management and statistical analysis. In this manuscript, we present our user written SAS macro called **%QLS** for the implementation of QLS developed under SAS version 9.1 using SAS/IML, an interactive matrix language in SAS. We demonstrate **%QLS** using a randomized clinical trial reported by De Backer et al. (1996) for the comparison of two treatments for dermatophyte toe onychomycosis (DTO), a common toenail infection that is difficult to treat (see §4 for more description of the study). **%QLS** can be used for analysis of longitudinal data whose outcome follows a normal, binary, or poisson distribution, with an AR(1), equicorrelated, Markov, or tri-diagonal correlation to describe the pattern of association among the repeated measurements on each subject, or cluster.

A unique feature of **%QLS**, which is not available in current GEE packages and other user-written QLS softwares, is that it computes the so called 'Prentice boundary' (Prentice 1988) of the estimate of α for analysis of binary data (see §2.5 for more detail).

Our outline for this manuscript is as follows: §2 describes each working correlation structure implemented in **%QLS**; then it briefly describes QLS and the Prentice constraints for α. §3 describes a complete list of the parameters in **%QLS**. Finally, §4 demonstrates **%QLS** using the clinical trial data (toenail data) for testing the time-averaged difference between the two treatment groups; This includes a demonstration of an analysis that results in a violation of the Prentice constraints, in addition to an application of the Markov structure that currently is unavailable for GEE.

2. Description of quasi-least squares

In this section, we provide a brief description of QLS similar to that provided in Shults et al. (2007a), Ratcliffe & Shults (2008), and Xie & Shults (2008). For a more through treatment of the development of QLS and its related asymptotic distributional properties, see Chaganty (1997) for a description of stage one of QLS for balanced and equally spaced data; Shults & Chaganty (1998) for stage one of QLS for unbalanced and unequally spaced data; and Chaganty & Shults (1999) for the second stage of QLS. Prior to Chaganty & Shults (1999), the QLS estimate of α was in general not consistent, even if the correlation structure was correctly specified. As noted earlier, QLS is a method in the framework of GEE; For an excellent text on GEE, see Hardin & Hilbe (2002).

We consider outcome measurements $y_i = (y_{i1}, \ldots, y_{in_i})'$ with associated covariates $x_{ij} = (x_{ij1}, \ldots, x_{ijk})'$ collected at measurement occasions $j = 1, \ldots, n_i$, for each subject $i = 1, \ldots, m$. For the model specification, we assume that the mean and variance of the outcome variable satisfy $\mathrm{E}(y_{ij}) = g^{-1}(x'_{ij}\beta) = \mu_{ij}$ and $\mathrm{Var}(y_{ij}) = \phi h(\mu_{ij})$, respectively, where ϕ is known as the dispersion parameter. Unlike the generalized linear model (GLM), both GEE and QLS assume that the covariance matrix $\mathrm{Cov}(y_i) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$ where $A_i = \mathrm{diag}(h(\mu_{i1}), \ldots, h(\mu_{in_i}))$, and $R_i(\alpha)$ is known as the *working correlation matrix* that describes the pattern of association among the repeated measurements on each subject.

2.1. Working correlation structures implemented in %QLS

%QLS currently allows for application of the AR(1), equicorrelated, Markov, and tri-diagonal structures that are described as follows:

1. **The first-order autoregressive (AR(1)) structure**: This structure assumes that $\mathrm{Corr}(y_{ij}, y_{ik}) = \alpha^{|j-k|}$, with feasible region $(-1,\ 1)$; The *feasible region* for α is defined as the interval on which α yields a positive definite correlation matrix.

2. **The equicorrelated structure**: This structure assumes that $\mathrm{Corr}(y_{ij}, y_{ik}) = \alpha$. The feasible region for this structure is $(-1/(n_m - 1),\ 1)$, where n_m is the maximum value of n_i over $i = 1, 2, \ldots, m$. Note that for relatively large m, this structure may be unrealistic unless all pairwise correlations are roughly identical across all time points.

3. **The Markov correlation structure**: This structure assumes that $\mathrm{Corr}(y_{ij}, y_{ik}) = \alpha^{|t_{ij} - t_{ik}|}$, with feasible region $(-1,\ 1)$. This structure is a generalization of the AR(1) structure and useful in modeling data that are unequally spaced in time. In addition, this structure has not been implemented in any of the currently available packages that implement GEE.

4. **The tri-diagonal correlation structure**: This structure assumes that $\text{Corr}(y_{ij}, y_{ik}) = \alpha$ 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.

%QLS does not allow for application of the independent structure (identity matrix) because QLS is identical to GEE for this structure. In addition, application of the unstructured matrix is complex for QLS. In SAS, we therefore suggest application of **PROC GENMOD** with the **repeated** statement and the option **corr=ind** for application of the independent correlation structure, or **corr=un** for application of the unstructured correlation matrix for GEE.

2.2. Estimating equations and the algorithm of the quasi-least squares

QLS is a two-stage procedure for estimation of the regression parameter β and the correlation parameter α. In stage one it alternates between updating the GEE estimating equation of β and the estimating equation of α. After convergence in stage one, an updated estimate of α is obtained by solving the stage two estimating equation for α; A final estimate of β is then obtained by solving the GEE estimating equation for β. The estimating equations are as follows:

The GEE estimating equation for β:

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)' A_i^{-1/2}R_i^{-1}(\alpha)A_i^{-1/2}\left[y_i - \mu_i(\beta)\right] = 0 \tag{1}$$

The stage one estimating equation for α:

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

where $Z_i(\beta) = (z_{i1}, \ldots, z_{in_i})'$ and $z_{ij} = (y_{ij} - \mu_{ij})/h(\mu_{ij})$.

The stage two estimating equation for α (evaluated at the stage one estimate $\hat{\alpha}$):

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

Note that the stage one and two estimating equations (2) and (3) involve the first derivative of the inverse of $R_i(\alpha)$, which may not be easily obtained for some correlation structures, e.g. the tri-diagonal structure. However, it can be easily shown that

$$\frac{\partial R_i^{-1}(\alpha)}{\partial\alpha} = -R_i^{-1}(\alpha)\frac{\partial R_i(\alpha)}{\partial\alpha}R_i^{-1}(\alpha) \tag{4}$$

where $\frac{\partial R_i(\delta)}{\partial \delta}$ is computed by taking the derivative of each element in $R_i(\alpha)$. For example, for the tri-diagonal structure, $\frac{\partial R_i(\delta)}{\partial \delta}$ is an $n_i \times n_i$ matrix with ones on the off-diagonal and zeros elsewhere, i.e. the (j,k)th element of $\frac{\partial R_i(\alpha)}{\partial \alpha}$ is 1 if $|j-k|=1$ and is 0 otherwise.

%QLS involves application of the following algorithm to obtain the estimates of β and α:

1. Fit a generalized linear model with an appropriate link and variance function using **PROC GENMOD** in SAS to obtain an initial estimate for β.
2. In stage one of QLS, repeat the following steps until a pre-specified convergence criterion is met for estimating β and α in stage one.
   (i) Compute the Pearson residuals at the current estimate of β, where the jth Pearson residual on subject i is given by $z_{ij} = (y_{ij} - \hat{u}_{ij})/h(\hat{u}_{ij})$.
   (ii) Compute the estimate of α by solving the QLS stage one estimating equation (2) for α using a pre-specified working correlation structure.
   (iii) Update the estimate of β by solving the GEE estimating equation (1) for β at the current estimate of α.
3. After convergence in the estimates of β and α in stage one, obtain an updated estimate of α by solving the stage two estimating equation (3) for α; Then obtain a final estimate of β by solving the GEE estimating for β that is evaluated at the stage two estimate of α.

2.3. Algebraic expressions of the estimating equations of β and α

At each iteration, **%QLS** uses the Newton-Raphson method to compute the current estimate β^* using the previous estimates β and α as follows:

$$\beta^* = \beta + \left\{ \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)' A_i^{-1/2} R_i^{-1}(\alpha) A_i^{-1/2} \left(\frac{\partial \mu_i}{\partial \beta} \right)' \right\}^{-1} \times \left\{ \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)' A_i^{-1/2} R_i^{-1}(\alpha) A_i^{-1/2} [y_i - \mu_i(\beta)] \right\}$$

Although (4) can be used to solve for stages one and two estimating equations (2) and (3), explicit expressions for α are preferred since their application is computationally more efficient. Except for the tri-diagonal structure, **%QLS** implements explicit expressions for α that were obtained by solving the stage one and two estimating equations (2) and (3) for α, for particular working structures.

For the AR(1) structure with unbalanced data, Shults & Chaganty (1998) proved that the feasible stage one estimate $\hat{\alpha}_{\text{A-ONE}}$ can be expressed as:

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

while Chaganty & Shults (1999) showed that the stage two estimate $\hat{\alpha}_{\text{A-TWO}}$ is given by

$$\hat{\alpha}_{\text{A-TWO}} = \frac{2\hat{\alpha}_{\text{A-ONE}}}{1+\hat{\alpha}^2_{\text{A-ONE}}}.$$

For the equicorrelated structure with unbalanced data, Shults (1996) showed that the stage one estimating equation of α has a unique feasible solution $\hat{\alpha}_{\text{E-ONE}}$ given by

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

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

$$\sum_{i:n_i>1} \frac{n_i\,(n_i-1)\,\hat{\alpha}_{\text{E-ONE}}\,(\hat{\alpha}_{\text{E-ONE}}\,(n_i-2)+2)}{(1+\hat{\alpha}(n_i-1))^2} \Big/ \sum_{i:n_i>1} \frac{n_i\,(n_i-1)\,\left(1+\hat{\alpha}^2_{\text{E-ONE}}(n_i-1)\right)}{(1+\hat{\alpha}_{\text{E-ONE}}(n_i-1))^2}.$$

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

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

where $e_{ij} = |t_{ij} - t_{i,j-1}|$. The stage two estimating equation of the Markov structure (Chaganty & Shults 1999) is then given by

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

2.4. Estimates of the covariance matrix

%QLS provides two types of estimated covariance matrices of the estimated regression parameter, the *model-based* and *robust sandwich-based* estimates. The robust sandwich-based estimate of the covariance matrix is the default matrix; It is often preferred when there is any uncertainty in the choice of working correlation structure. However, the standard errors are not necessarily smaller for the sandwich covariance matrix. Therefore, application of both the model based and sandwich based covariance matrices might be considered in an analysis.

%QLS computes the model-based covariance matrix as follows:

$$\widehat{\text{Cov}}_{\text{M}}(\hat{\beta}) = \hat{\phi}\sum_{i=1}^{n} X_i' A_i^{1/2} R_i^{-1}(\hat{\alpha}) A_i^{1/2} X_i.$$

where $\hat{\phi}$ is the estimate of the dispersion parameter with or without the bias correction given by

$$\hat{\phi}_{\text{BC}} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{Z_i(\hat{\beta})' Z_i(\hat{\beta})}{n_i}, \text{ or } \hat{\phi}_{\text{B}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i(\hat{\beta})' Z_i(\hat{\beta})}{n_i}$$

respectively where p is the dimension of the regression parameters β. By default, **%QLS** provides the bias corrected estimate of the dispersion parameter $\hat{\phi}_{\text{BC}}$. For the robust sandwich-based estimate of the covariance matrix, **%QLS** computes

$$\widehat{\text{Cov}}_{\text{R}}(\hat{\beta}) = W_n^{-1} \left\{ \sum_{i=1}^{n} X_i' A_i^{1/2} R_i^{-1}(\hat{\alpha}) Z_i(\hat{\beta}) Z_i'(\hat{\beta}) R_i^{-1}(\hat{\alpha}) A_i^{1/2} X_i \right\} W_n^{-1}.$$

where

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

%QLS also computes the $(1-\alpha)100\%$ confidence interval, and a p-value for testing each individual regression parameter $\beta_j = 0$, based on either the model-based or the robust sandwich-based estimate of the covariance matrix of $\widehat{\beta}$.

2.5. Prentice boundary of the estimate of α for analyzing binary data

Consider longitudinal binary measurements $y_{i1}, \ldots, y_{in_i}$ with expected values $E(y_{ij}) = \Pr(y_{ij} = 1) = p_{ij}$, with $q_{ij} = \Pr(y_{ij} = 0) = 1 - p_{ij}$, and the correlation between measurements y_{ij} and y_{ik} denoted by $\text{Corr}(y_{ij}, y_{ik})$. An important feature of correlated binary data is that the p_{ij}, q_{ij} and $\text{Corr}(y_{ij}, y_{ik})$ completely determine the bivariate distribution of y_{ij} and y_{ik} because the pair-wise probabilities $\text{pr}(y_{ij}, y_{ik})$ can be expressed as

$$\Pr(y_{ij}, y_{ik}) = p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} p_{ik}^{y_{ik}} q_{ik}^{1-y_{ik}} \left\{ 1 + \text{Corr}(y_{ij}, y_{ik}) \frac{(y_{ij} - p_{ij})(y_{ik} - p_{ik})}{(p_{ij}\, p_{ik}\, q_{ij}\, q_{ik})^{1/2}} \right\}. \tag{5}$$

Prentice (1988) pointed out that the probabilities in (5) will be non-negative, i.e. $\Pr(y_{ij}, y_{ik}) \geq 0$, only if the correlations satisfy the following constraints that depend on the marginal means:

$$-\max_{j \neq k} \left\{ \sqrt{\frac{p_{ij}p_{ik}}{q_{ij}q_{ik}}}, \sqrt{\frac{q_{ij}q_{ik}}{p_{ij}p_{ik}}} \right\} \leq \text{Corr}(y_{ij}, y_{ik}) \leq \min_{j \neq k} \left\{ \sqrt{\frac{p_{ij}q_{ik}}{q_{ij}p_{ik}}}, \sqrt{\frac{q_{ij}p_{ik}}{p_{ij}q_{ik}}} \right\} \tag{6}$$

for all $i = 1, \ldots, n$.

A main problem with violation of the Prentice constraints is that this leads to a lack of theoretical justification for the existence of a valid joint probability mass function, given the estimated parameter values of β and α. However, as Monlenberghs & Verbeke (2006) pointed out, the estimate of the working correlation is "merely a device to provide consistent and asymptotically normal point estimates for the marginal regression parameters which should not the made a part of formal inference." Rochon (1998) also noted that violation of the Prentice constraints appears to cause no difficulty in practice, although the potential for violation should be taken into account in the design phase of a study. For example, when conducting sample size calculations, values of α that satisfy the Prentice constraints should be considered. Shults et al. (2008) also suggest that a severe violation of bounds might be used to remove a particular working structure from a list of candidate working structures.

As noted in §1, the existing GEE and user-written QLS softwares (in Stata, Matlab, and R) do not provide the estimated Prentice constraints (6) for analysis of binary data. **%QLS**, on the other hand, computes the estimated Prentice constraints for each correlation structure,

and alert users to a potential problem by providing a warning message if the estimate of α does not fall within the interval.

3. List of parameters in the macro

A complete list of the parameters in **%QLS** is as follows:

```
%QLS(data=,
     y=,
     x=,
     id=,
     time=,
     link=,
     corr=,
     robust=,
     dispersion=,
     alpha=,
     initialout=,
     stage1out=,
     stage2out=,
     cmatrix=,
     reference=,
     converge=,
     maxiter=)
```

where

data is the name of the data set in the usual longitudinal data format to be read in **PROC GENMOD**. The data set must not contain any missing values.
y is the outcome variable.
x are the predictors (covariates) in the regression model.
id is the ID variable; **time** is the time variable.
link equals 1 for the identity link; 2 for the logit link; and 3 for the log link (default is 1).
corr equals 1 for the AR(1); 2 for the Equicorrelated; 3 for the Markov; 4 for the Tri-diagonal (default is 1).
robust equals 1 for robust sandwich-based standard errors; 2 for model-based standard errors (default is 1).
dispersion equals 1 for bias not corrected; 0 for bias-corrected (default is 1).
alpha is the significance level to be used in testing each regression coefficient (default is 0.05).
initialout equals 1 creates a SAS permanent data set in the current work space for the initial output; 0 otherwise (default is 0).
stage1out equals 1 creates a SAS permanent data set in the current work space for the stage 1 output; 0 otherwise (default is 0).
stage2out equals 1 creates a SAS permanent data set in the current work space for the

stage 2 output; 0 otherwise (default is 0).
cmatrix equals 1 creates a SAS permanent data set in the current work space for the stage 2 correlation matrix; 0 otherwise (default is 0).
reference equals 1 prints out the reference information; 0 otherwise (default is 0).
convergence is the convergence criterion for estimation of β and of α (default is 0.0001).
maxiter is the maximum number of allowable iterations for estimation of β and α (default is 100).

Note that many of the parameters have default values, so that they do not have to be specified. **%QLS** assumes the usual longitudinal data format to be read in **PROC GENMOD** without any missing observation contained in the data. If there are missing observations in the data that are coded as missing, these must be deleted prior to implementation of **%QLS**; This is equivalent to assuming that the observations are 'Missing Completely At Random' (MCAR), as in the usual GEE analysis implemented by **PROC GENMOD** with the **repeated** statement.

4. Clinical trial example

De Backer et al. (1996) reported a 12-week, randomized, double-blind, multi-center comparative trial for the comparison between the standard oral drug (terbinafine 250mg daily) and the experimental oral drug (theritraconazole 200mg daily) in the treatment of a common toenail infection called dermatophyte toe onychomycosis (DTO) which affects more than 2% of the British population (Roberts 1992). The data was also described in Monlenberghs & Verbeke (2006), and can be downloaded from `www.cceb.upenn.edu/~sratclif/QLSproject.html`.

A total of 189 patients were randomized to each treatment group and followed over 12 weeks, with measurements taken at baseline, and at months 1, 2, 3, 6, 9, and 12. The primary outcome measure was the severity of the toe nail infection, that was defined as 1 if the infection was severe, and 0 otherwise. For the purpose of demonstration, we first consider a simple logistic regression model for comparison of the time-averaged treatment difference between the standard treatment group and the experimental treatment group.

The toenail data, **toenail.txt**, contains a total of 4 variables: **time**, **treatment**, **y**, and **id** where **time** is the time variable, **treatment** is the treatment indicator (1 for the standard arm, and 0 otherwise), **y** is the outcome variable (1 if the infection is severe, and 0 otherwise), and **id** is the ID number assigned to each patient. Let y_{ij} follow the Bernoulli distribution with $\text{pr}(y_{ij} = 1) = p_{ij}$ such that

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1(treatment_i) \tag{7}$$

where $treatment_i$ is the treatment indicator that equals 1 if the ith subject is assigned to the standard drug and 0 otherwise. One advantage of implementing the model in (7) is that the upper limit of the Prentice constraint for α is always equal to 1. In general, any model that involves covariates that do not vary within clusters will have an upper Prentice boundary (for α) of 1, due the fact that the estimated probabilities p_{ij} will not vary within subjects (clusters) when only cluster level covariates are considered in the model.

Before we demonstrate **%QLS** to fit the model (7), first we assume that the data, **toenail.txt**, is read into the current SAS workspace, e.g.

```
data toenail;
 infile "D:\toenail.txt";
 input time treatment y id;
run;
```

where we assume that the **toenail.txt** is stored in **D** directory.

4.1. Example of application of the AR(1) correlation structure

The following codes can be used to analyze the toenail data using the QLS regression model (7) with the AR1 structure:

```
%QLS(data=toenail, y=y, x=treatment, id=id, time=time, link=2, corr=1);
```

The estimated standard errors are the robust sandwich-based estimates that are set by default. The outputs from the code are as follows:

```
                Quasi-Least Squares SAS Macro Version 1.0

            Regression Analysis using Quasi-Least Squares (QLS)

                        QLS Model Information


                Variance Function              : Binomial
                Link Function                  : Logit
                Dependent Variable             : Y
                Correlation Structure          : AR(1)

                Number of Observation Read     :      1907
                Number of Clusters             :       294
                Maximum Cluster Size           :         7
                Minimum Cluster Size           :         1
                Correlation Matrix Dimension   :         7
                Number of Distinct Time Points :         7

                    TIME   0   1   2   3   6   9  12

                Number of Events               :       408
                Number of Trials               :      1907


                        Analysis of Initial Parameter Estimates

  Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

   INTERCEPT -1.217433 0.0778205    15.64   0.0000 -1.369958 -1.064908
```

```
    TREATMENT -0.168861 0.1118004     1.51   0.1309 -0.387985  0.050264

%QLS is modeling the probability that Y=1

--------------------------------page 1/3---------------------------------

Correlation  converged after   1 iterations  ( tolerance =         0 )
Reg. coeffi. converged after   2 iterations  ( tolerance = 0.0000605 )


               Analysis of Stage 1 QLS Parameter Estimates

    Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

    INTERCEPT -1.200902 0.1409324    -8.52   0.0000 -1.477125  -0.92468
    TREATMENT -0.169826 0.1971473    -0.86   0.3890 -0.556228 0.2165757


               Stage 1  Correlation Parameter  Estimate
                                     0.4423849

               Dispersion Parameter Estimate at Stage 1
                                                 1

                   Stage 1 Working Correlation Matrix

        1.0000   0.4424   0.1957   0.0866   0.0383   0.0169   0.0075
        0.4424   1.0000   0.4424   0.1957   0.0866   0.0383   0.0169
        0.1957   0.4424   1.0000   0.4424   0.1957   0.0866   0.0383
        0.0866   0.1957   0.4424   1.0000   0.4424   0.1957   0.0866
        0.0383   0.0866   0.1957   0.4424   1.0000   0.4424   0.1957
        0.0169   0.0383   0.0866   0.1957   0.4424   1.0000   0.4424
        0.0075   0.0169   0.0383   0.0866   0.1957   0.4424   1.0000

--------------------------------page 2/3---------------------------------

Correlation  converged after   1 iterations  ( tolerance =         0 )
Reg. coeffi. converged after   3 iterations  ( tolerance = 4.0938E-9 )


               Analysis of Stage 2 QLS Parameter Estimates

    Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

    INTERCEPT -1.178475 0.1392601    -8.46   0.0000 -1.451419  -0.90553
    TREATMENT -0.170937 0.1938719    -0.88   0.3779 -0.550919 0.2090446
```

```
                    Prentice Boundary
                    -.259393 1.000000

              Stage 2  Correlation Parameter  Estimate
                                           0.7399569

              Dispersion Parameter Estimate at Stage 2
                                                    1

                  Stage 2 Working Correlation Matrix

      1.0000   0.7400   0.5475   0.4052   0.2998   0.2218   0.1641
      0.7400   1.0000   0.7400   0.5475   0.4052   0.2998   0.2218
      0.5475   0.7400   1.0000   0.7400   0.5475   0.4052   0.2998
      0.4052   0.5475   0.7400   1.0000   0.7400   0.5475   0.4052
      0.2998   0.4052   0.5475   0.7400   1.0000   0.7400   0.5475
      0.2218   0.2998   0.4052   0.5475   0.7400   1.0000   0.7400
      0.1641   0.2218   0.2998   0.4052   0.5475   0.7400   1.0000

--------------------------------page 3/3--------------------------------
```

The output of **%QLS** contains the model information followed by the estimates of the stage one and two estimates of β and of α. As noted earlier, the upper limit of the Prentice interval is equal to 1 in the above output. From the stage two output, the p-value corresponding to the time-averaged treatment effect is equal to 0.38, which suggests that there is no significant time-averaged treatment difference between the standard drug versus the experimental drug.

4.2. Example of application of the Markov correlation structure

Here we demonstrate application of the Markov correlation structure, which is currently unavailable for GEE. This is important because the toenail data is unequally spaced in time, e.g. the variable time in this data set indicates the visit number and takes value in $\{0, 1, 2, 3, 6, 9, 12\}$ for each subject. Therefore, the Markov correlation structure would be preferable for the analysis of this trial. The following code can be used to fit the model (7) with the Markov correlation structure:

```
%QLS(data=toenail, y=y, x=treatment, id=id, time=time, link=2, corr=3);
```

Here we omit the stage one output, and present the initial and stage two outputs.

```
                Quasi-Least Squares SAS Macro Version 1.0

          Regression Analysis using Quasi-Least Squares (QLS)

                        QLS Model Information
```

```
                Variance Function              : Binomial
                Link Function                  : Logit
                Dependent Variable             : Y
                Correlation Structure          : Markov

                Number of Observation Read     :      1907
                Number of Clusters             :       294
                Maximum Cluster Size           :         7
                Minimum Cluster Size           :         1
                Correlation Matrix Dimension   :         7
                Number of Distinct Time Points :         7

                    TIME   0   1   2   3   6   9  12

                Number of Events               :       408
                Number of Trials               :      1907


                    Analysis of Initial Parameter Estimates

    Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

    INTERCEPT -1.217433 0.0778205    15.64   0.0000 -1.369958 -1.064908
    TREATMENT -0.168861 0.1118004     1.51   0.1309 -0.387985  0.050264

%QLS is modeling the probability that Y=1

--------------------------------page 1/3---------------------------------

Correlation  converged after  27 iterations  ( tolerance = 0.0000556 )
Reg. coeffi. converged after   3 iterations  ( tolerance = 6.8226E-7 )


                Analysis of Stage 2 QLS Parameter Estimates

    Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

    INTERCEPT -1.345997  0.141525    -9.51   0.0000 -1.623381 -1.068613
    TREATMENT -0.204626 0.1969105    -1.04   0.2987 -0.590564 0.1813111


                          Prentice Boundary
                          -.212116 1.000000

                Stage 2  Correlation Parameter  Estimate
                                               0.7942784
```

```
                Dispersion Parameter Estimate at Stage 2
                                                    1

                    Stage 2 Working Correlation Matrix

      1.0000   0.7943   0.6309   0.5011   0.2511   0.1258   0.0630
      0.7943   1.0000   0.7943   0.6309   0.3161   0.1584   0.0794
      0.6309   0.7943   1.0000   0.7943   0.3980   0.1994   0.0999
      0.5011   0.6309   0.7943   1.0000   0.5011   0.2511   0.1258
      0.2511   0.3161   0.3980   0.5011   1.0000   0.5011   0.2511
      0.1258   0.1584   0.1994   0.2511   0.5011   1.0000   0.5011
      0.0630   0.0794   0.0999   0.1258   0.2511   0.5011   1.0000

--------------------------------page 3/3--------------------------------
```

The results are similar to those for the AR(1) structure, with an estimated α in stage two ($\hat{\alpha} = 0.79$) versus $\hat{\alpha} = 0.74$ for the AR(1) structure. Further, the p-value with respect to the time-averaged treatment effect is 0.30; hence the same conclusion follows as with the AR(1) structure.

4.3. Example of application of the equicorrelated and tri-diagonal structures

Although the equicorrelated and tri-diagonal structures may not be best candidate correlation structures for the toenail data, we include the implementation of theses structures for the purpose of demonstration. Here we only present the codes for fitting the model (7) with the equicorrelated and tri-diagonal correlation structures, but omit their outputs. For the equicorrelated correlation structure, we have

```
%QLS(data=toenail, y=y, x=treatment, id=id, time=time, link=2, corr=2);
```

For the tri-diagonal correlation structure, we have

```
%QLS(data=toenail, y=y, x=treatment, id=id, time=time, link=2, corr=4);
```

4.4. Example of violation of the Prentice boundary

Here we briefly demonstrate violation of the Prentice constraints using the toenail data. Consider the following model for testing the treatment effect over time:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1(treatment_i) + \beta_2(time_{ij}) + \beta_3(treatment_i \times time_{ij}) \qquad (8)$$

where $treatment_i$ is the treatment indicator that equals 1 if the ith subject is assigned to the standard drug and 0 otherwise, $time_{ij}$ represents the time of the measurement collected on subject i at the jth measurement occasion, and $treatment_i \times time_{ij}$ is the treatment by time interaction.

To fit the model in (8) using **%QLS**, a new variable corresponding to the interaction term must be created first, e.g.

```
data toenail;
 infile "D:\toenail.txt";
 input time treatment y id;
 interaction=treatment*time;
run;
```

To fit the model (8) with the AR(1) structure, we use

```
%QLS(data=toenail, y=y, x=treatment time interaction, id=idnum, time=time,
     link=2, corr=1);
```

Here we provide the initial and stage two outputs for the AR(1) structure.

```
                Quasi-Least Squares SAS Macro Version 1.0

            Regression Analysis using Quasi-Least Squares (QLS)

                        QLS Model Information


                    Variance Function              : Binomial
                    Link Function                  : Logit
                    Dependent Variable             : Y
                    Correlation Structure          : AR(1)

                    Number of Observation Read     :      1907
                    Number of Clusters             :       294
                    Maximum Cluster Size           :         7
                    Minimum Cluster Size           :         1
                    Correlation Matrix Dimension   :         7
                    Number of Distinct Time Points :         7

                        TIME   0   1   2   3   6   9  12

                    Number of Events               :       408
                    Number of Trials               :      1907


                        Analysis of Initial Parameter Estimates

  Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

  INTERCEPT   -0.557058 0.1090393     5.11   0.0000 -0.770771 -0.343345
  TREATMENT   0.0235766 0.1564805     0.15   0.8802 -0.283119 0.3302727
  TIME        -0.17693 0.0245578     7.20   0.0000 -0.225062 -0.128797
```

```
    INTERACTION -0.077976 0.0394371     1.98   0.0480 -0.155271  -0.00068


%QLS is modeling the probability that Y=1

--------------------------------page 1/3--------------------------------

Correlation  converged after   1 iterations  ( tolerance =         0 )
Reg. coeffi. converged after   5 iterations  ( tolerance = 0.0000709 )


               Analysis of Stage 2 QLS Parameter Estimates

    Parameter  Estimate Stand Err        Z   Pr>|Z| [95% Con. Interval]

    INTERCEPT   -0.649358 0.1702276    -3.81   0.0001 -0.982998 -0.315718
    TREATMENT   0.1213252 0.2510978     0.48   0.6290 -0.370817 0.6134679
    TIME        -0.141402 0.0285992    -4.94   0.0000 -0.197455 -0.085348
    INTERACTION -0.120551 0.0555837    -2.17   0.0301 -0.229493 -0.011609


                          Prentice Boundary
                          -.037683 .3076519

               Stage 2  Correlation Parameter  Estimate
                                             0.7054869

               Dispersion Parameter Estimate at Stage 2
                                                     1

                   Stage 2 Working Correlation Matrix

        1.0000   0.7055   0.4977   0.3511   0.2477   0.1748   0.1233
        0.7055   1.0000   0.7055   0.4977   0.3511   0.2477   0.1748
        0.4977   0.7055   1.0000   0.7055   0.4977   0.3511   0.2477
        0.3511   0.4977   0.7055   1.0000   0.7055   0.4977   0.3511
        0.2477   0.3511   0.4977   0.7055   1.0000   0.7055   0.4977
        0.1748   0.2477   0.3511   0.4977   0.7055   1.0000   0.7055
        0.1233   0.1748   0.2477   0.3511   0.4977   0.7055   1.0000

    Warning!  Correlation parameter estimate is not within the boundary.
    The existence of a multivariate binary distribution is questionable.

--------------------------------page 3/3--------------------------------
```

From the stage two output, the estimated stage two α is 0.71, which exceeds the upper limit (0.31) of the Prentice constraints. It is also important to note that although the results are

not shown here, application of GEE for the AR(1) structure would also result in a severe violation of the Prentice constraints.

The above results suggest something different than the time-averaged model, which is that there is a difference in the likelihood of high severity between the two treatment conditions. However, graphical displays (not shown) suggest that the assumption of linearity in the logit is not appropriate for these data. For an extensive discussion of approaches for assessment of the linearity in the logit assumption, see Hilbe (2008).

For demonstration purposes, we now present a model that did not violate the linearity in the logit assumption, and that also did not result in a violation of the Prentice bounds for α. This model contains indicator variables for the second (1 month), third (2 month), fifth (6 month), and seventh (12 month) measurements on each subject; an indicator variable for the standard treatment; and a visit seven (12 month) by treatment indicator variable (all other treatment by visit indicator variables did not differ significantly from zero). The corresponding data set, **toenail2.txt**, can be also downloaded from www.cceb.upenn.edu/~sratclif/QLSproject.html. Here we present the code, and the initial and stage two outputs.

```
%qls(data=toenail2, y=y, x=time2 time3 time5 time7 treatment time7_trt,
     id=id, time=time, link=2, corr=1);
```

```
            Quasi-Least Squares SAS Macro Version 1.0

        Regression Analysis using Quasi-Least Squares (QLS)

                  QLS Model Information


            Variance Function              : Binomial
            Link Function                  : Logit
            Dependent Variable             : Y
            Correlation Structure          : AR(1)

            Number of Observation Read     :      1907
            Number of Clusters             :       294
            Maximum Cluster Size           :         7
            Minimum Cluster Size           :         1
            Correlation Matrix Dimension   :         7
            Number of Distinct Time Points :         7

               TIME   0   1   2   3   6   9  12

            Number of Events               :       408
            Number of Trials               :      1907

                Analysis of Initial Parameter Estimates
```

Parameter	Estimate	Stand Err	Z	Pr>|Z|	[95% Con.	Interval]
INTERCEPT	-1.143183	0.1019955	11.21	0.0000	-1.343091	-0.943276
TIME2	0.5386061	0.1502601	3.58	0.0003	0.2441018	0.8331104
TIME3	0.3537537	0.1548007	2.29	0.0223	0.0503499	0.6571575
TIME5	-1.179593	0.2379959	4.96	0.0000	-1.646056	-0.713129
TIME7	-0.996883	0.3003914	3.32	0.0009	-1.585639	-0.408127
TREATMENT	-0.144662	0.1182858	1.22	0.2213	-0.376498	0.0871737
TIME7_TRT	-0.751825	0.5181603	1.45	0.1468	-1.767401	0.2637501

%QLS is modeling the probability that Y=1

----------page 1/3----------

Correlation converged after 1 iterations (tolerance = 0)
Reg. coeffi. converged after 8 iterations (tolerance = 0.0000839)

Analysis of Stage 2 QLS Parameter Estimates

Parameter	Estimate	Stand Err	Z	Pr>|Z|	[95% Con.	Interval]
INTERCEPT	-1.140052	0.1428558	-7.98	0.0000	-1.420044	-0.86006
TIME2	0.1103149	0.0699347	1.58	0.1147	-0.026755	0.2473844
TIME3	0.1702156	0.0794198	2.14	0.0321	0.0145556	0.3258757
TIME5	-0.48106	0.0954562	-5.04	0.0000	-0.66815	-0.293969
TIME7	-0.236975	0.1379387	-1.72	0.0858	-0.50733	0.0333794
TREATMENT	-0.093147	0.2003516	-0.46	0.6420	-0.485829	0.2995349
TIME7_TRT	-0.318485	0.1752016	-1.82	0.0691	-0.661874	0.0249039

Prentice Boundary
-.173521 .7220668

Stage 2 Correlation Parameter Estimate
0.7161348

Dispersion Parameter Estimate at Stage 2
1

Stage 2 Working Correlation Matrix

1.0000	0.7161	0.5128	0.3673	0.2630	0.1884	0.1349
0.7161	1.0000	0.7161	0.5128	0.3673	0.2630	0.1884
0.5128	0.7161	1.0000	0.7161	0.5128	0.3673	0.2630
0.3673	0.5128	0.7161	1.0000	0.7161	0.5128	0.3673

```
0.2630   0.3673   0.5128   0.7161   1.0000   0.7161   0.5128
0.1884   0.2630   0.3673   0.5128   0.7161   1.0000   0.7161
0.1349   0.1884   0.2630   0.3673   0.5128   0.7161   1.0000

--------------------------------page 3/3--------------------------------
```

For the above model, the Prentice constraints are not violated. In addition, the results seem more in agreement with the time-averaged model, which also did not identify a significant difference between the two treatment conditions with respect to severity of toenail infection.

5. Discussion

%QLS can fit a model to longitudinal data using the method of quasi-least squares, and can consider data which follows the normal, binary, or Poisson distribution with the AR(1), Markov, equicorrelated, and tri-diagonal structures. The syntax and the output of **%QLS** are similar to the existing GEE procedures in SAS, i.e. **PROC GENMOD** with the **repeated** statement, that would be familiar to SAS users. **%QLS** assumes that there are no missing observations in the dataset; hence any observations that are coded as missing should be deleted prior to the implementation of the macro. As noted earlier, this is equivalent to assuming that the data is missing completely at random (MCAR), which is a typical assumption in a GEE analysis.

Further updates of **%QLS** will be made to allow for implementation of other structures that are not currently available for GEE, including the familial structure, and Kronecker product-based structures which account for multiple sources of correlation, e.g. multi-outcome longitudinal data for which the sources of correlations come from within subjects over time, and between multiple outcomes. It will also be of interest to develop and compare existing methods for selection of a working correlation structure and for assessment of goodness of fit of QLS (and GEE) models.

Acknowledgement

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

References

Chaganty, N. R. (1997) An alternative approach to the analysis of longitudinal data via generalized estimating equations. *J. Statist. Plan. Inf.* **63**: 39–54.

Chaganty, N. R. & Naik, D. (2002) Analysis of multivariate longitudinal data using quasi-least squares. *J. Statist. Plan. Inf.* **103**: 421-436.

Chaganty, N.R. & Shults, J. (1999) On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *J. Statist. Plan. Inf.* **76**: 127-144.

Crowder, M. (1995) On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **8**2, 407-410.

De Backer, M., De Keyser, P., De Vroey, C. & Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day-a dounble-blind comparative trial. *British Journal of Dermatology* **134**, 16-17.

Hardin, J.W. & Hilbe, J.M. (2002). *Generalized estimating equations*. Florida: Chapman & Hall/CRC Press.

Hilbe, J.M. (2008). *Logistic Regression Models*. Chapman & Hall/CRC Press.

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

Monleberghs, G. & Verbeke, G. (2006). *Models for Discrete Longitudinal Data.* New York: Springer-Verlag.

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

Ratcliffe, S. & Shults, J. (2008) **GEEQBOX**: A MATLAB Toolbox for Generalized Estimating Equations and Quasi-Least Squares. *J. Statist Software* **25**, Issue 14.

Roberts, D.T. (1992) Prevalence of dermatophyte onychomycosis in the United Kingdon: Results of an omnibus survey. *British Journal of Dermatology* **134 Suppl. 39**, 23-27.

Rochon, J. (1998) Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine* **17**, 1643-1658.

SAS Release 9. (2003) *SAS Institute Inc.*, Cary, North Carolina

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. R. (1998) Analysis of serially correlated data using quasi-least squares. *Biometrics* **54**, 1622–1630.

Shults, J., Mazurick, C.A. & Landis, J.R. (2006) Analysis of repeated bouts of measurements in the framework of generalized esquation 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**, 521–530.

Shults, J. & and Radcliffe, S. (2007) Analysis of multi-level correlated data in the framework of generalized estimating equations via xtmultcorr procedures in Stata and qls functions in Matlab. *UPenn Biostatistics Working Papers*. Working Paper 15. http://biostats.bepress.com/upennbiostat/papers/art15

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

Shults, J., Wenguang, S., Tu, X., Kim, H., Amsterdam, J., Hilbe, J., and Ten-Have T. (2008) A Comparison of Several Approaches for Choosing Between Working Correlation Structures in Generalized Estimating Equation Analysis of Longitudinal Binary Data. *Under Review*

Shults, J., Whitt, C.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.

Xie, J. & Shults, J. (2008) Implementation of quasi-least squares with the R package **qlspack**. *J. Statist Software* In Press.

Yan, J. (2002) **geepack**: yet another package for generalized estimating equations. *R News* 2002; **7**, 12–14.

Affiliation:

Hanjoo Kim
PhD Candidate
Department of Biostatistics and Epidemiology
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
423 Guardian Drive, 501 Blockley Hall Philadelphia, PA 19104-6021, U.S.A.
Telephone: +1/215/573/8950
E-mail: hanjoo@mail.med.upenn.edu

Justine Shults
Associate Professor
Department of Biostatistics and Epidemiology
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
423 Guardian Drive, 610 Blockley Hall Philadelphia, PA 19104-6021, U.S.A.
Telephone: +1/215/573/6526
E-mail: jshults@mail.med.upenn.edu
URL: http://www.cceb.upenn.edu/~sratclif/QLSproject.html