Paper 245-2009

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

Hanjoo Kim, Forest Research Institute, Inc., Jersey City, NJ
Justine Shults, University of Pennsylvania, Philadelphia, PA

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 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.

INTRODUCTION

Quasi-least squares (QLS) is a two-stage approach for analysis of longitudinal data that is based on the popular generalized estimating equations (GEE) (Liang and Zeger, 1986). GEE and QLS are identical in their approach to estimation of the regression parameter, but differ with respect to estimation of the correlation parameter α . QLS was motivated by Dunlop (1994) and was developed in several stages: in Chaganty (1997) (stage one, for balanced and equally spaced data); Shults (1996) and Shults and Chaganty (1998) (stage one, for unbalanced and unequally spaced data); and Chaganty and Shults (1999) (stage two, for unbalanced and unequally spaced data).

QLS has several attractive features. First, Crowder (1995) constructed simple examples to demonstrate that feasible estimates of α may fail to exist when the correlation structure is not correctly specified in the analysis. In contrast, 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. In addition, QLS allows for relatively straightforward implementation of more complex structures than are currently available for GEE. For example, Shults (1996) and Shults and Chaganty (1998) implemented 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 with multiple sources of correlations, e.g. correlations between multiple outcomes as well as within subject for each outcome over time, using Kronecker product correlation structures (Shults & Morrow, 2002; Chaganty & Naik, 2002; and Shults, Whitt, & Kumanyika, 2004). (However, it is important to note that the prior manuscripts on the KP structure only considered data that were totally balanced or that were balanced within subjects.) In addition, Shults, Mazurick & Landis (2006) also implemented a banded Toeplitz structure for analysis of multiple bouts of repeated measurements.

QLS has been implemented in several major statistical software packages. For example, Shults, Ratcliffe & Leonard (2007) developed the xtqls procedure for use in Stata statistical software. Xie and Shults (2008) developed the qlspack for implementation of QLS in R software. Finally, Ratcliffe and Shults (2008) developed GEEQBOX for implementation of both GEE and QLS in Matlab; Prior to GEEQBOX, there was no software available for application of GEE in Matlab. In this manuscript, we present our user written SAS macro called %QLS version 1 for implementation of QLS in SAS version 9.1 using SAS/IML. We demonstrate our software using a randomized clinical trial reported by Debacker et al 1994). %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 software, is that it computes the so called `Prentice boundary' (Prentice, 1988) of the estimate of α for analysis of binary data.
In addition to our discussion of %QLS version 1, we note that we have recently extended %QLS for analysis of multiple outcomes that are measured on subjects over time. Our SAS procedure %QLS version 2 can be used to apply the KP structure that was proposed by Lefkopoulou M, Moore D, Ryan L. (1989) and Galecki (1994) and implemented by Roy and Khattree (2005). Unlike the previous work on the KP structure, our software can handle

unbalanced data, which result when study investigators planned for an equal number of measurements to be collected on several outcomes on each participant, but some measurements were missed on some subjects. In this manuscript we briefly describe QLS and the structures that can be applied with our software. We then demonstrate application of %QLS version 1, including 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. For additional details regarding QLS and %QLS version 1 and 2, please see Kim et al. (2008a) and Kim & Shults (2008b, 2008c).

DESCRIPTION OF QLS

For data with one level of association (e.g. within serial measurements on independent subjects or within families in a cross-sectional study), we assume that our data comprise independent vectors of measurements $Y_i' = (y_{i1}, y_{i2}, ..., y_{in_i})$ collected on cluster (or subject) i at times $t_{i1}, t_{i2}, ..., t_{in_i}$, for $i = 1, 2, ..., m.$ We do not specify the full distribution of outcomes, but assume that the mean and variance satisfy $E(y_{ij}) = u_{ij}$ and $Var(y_{ij}) = \phi h(u_{ij})$, where $\phi > 0$ is a known or unknown scale parameter and $h(\circ)$ is a known variance function. When we observe y_{ij} we also collect a vector of explanatory variables $x_{ij}' = (x_{ij1}, x_{ij2,}, ..., x_{ijp})$. We examine the association between the covariates and marginal mean of the outcome variable via the relation $u_{ij} = g^{-1}(x_{ij}'\beta)$, where $\beta \in \Re^p$ is a vector of unknown regression coefficients and $g(\circ)$ is an invertible link function. To describe the pattern of association among observations within clusters (subjects), we apply a working covariance matrix for Y_i , $\Sigma_i = A_i^{1/2}(\beta)F_i(\underline{\alpha})A_i^{1/2}(\beta).$ Here $F_i(\underline{\alpha})$ is the $n_i \times n_i$ working correlation matrix of Y_i ; $\underline{\alpha} \in \Re^q$ is a vector of parameters that fully characterizes $F_i(\underline{\alpha})$; and $A_i(\beta) = diag\left\{h(u_{i1}), h(u_{i2}), \ldots, h(u_{in_i})\right\}.$ Because the working structure $F_i(\underline{\alpha})$ may be misspecified, we assume that the true correlation structure of Y_i has form $T_i(\underline{\rho})$, where $\underline{\rho} \in \Re^q$ is the vector or parameters that characterizes $T_i(\underline{\rho})$. We define the generalized error sum of squares :

$$Q(\beta, \underline{\alpha}) = \sum_{i=1}^{m} Z_i'(\beta)F_i^{-1}(\underline{\alpha})Z_i(\beta) \tag{1}$$

for $Z_i(\beta) = A_i^{-1/2}(\beta)(Y_i - U_i)$ and $U_i' = (u_{i1}, u_{i2}, \ldots, u_{in_i}).$ Also define $D_i(\beta) = \partial U_i / \partial \beta.$

The QLS stage one estimate of $\underline{\alpha}$ is obtained by minimizing $Q(\beta, \underline{\alpha})$ over the feasible region for $\underline{\alpha}$, which we define as the set $\int \in \Re^q$ such that $F_i(\underline{\alpha})$ is a positive definite correlation matrix for $\underline{\alpha} \in \int$. The QLS estimating equations for β and $\underline{\alpha}$, respectively, are given by:

$$\sum_{i=1}^{m} D_i'(\beta)A_i^{-1/2}(\beta)F_i^{-1}(\underline{\alpha})Z_i(\beta) = 0 \tag{2}$$

and

$$\frac{\partial}{\partial \alpha_j}\left[\sum_{i=1}^{m} Z_i'(\beta)F_i^{-1}(\underline{\alpha})Z_i(\beta)\right] = 0 \ for \ \ j = 1, 2, \ldots, q. \tag{3}$$

The first-stage QLS estimates $\hat{\beta}_1$ and $\hat{\underline{\alpha}}$ are obtained by choosing a starting value for β , or for $\underline{\alpha}$, and alternating between solving (2) and (3) until the estimates converge. However, $\hat{\underline{\alpha}}$ is asymptotically biased. The second stage yields consistent estimates $\hat{\underline{\rho}}$ and $\hat{\beta}$ of $\underline{\rho}$ and β that are based on $\hat{\underline{\alpha}}$ and $\hat{\beta}_{1.}$ To obtain $\hat{\underline{\rho}}$ according to Chaganty and Shults (1999), we solve the following equation for $\underline{\rho}$:

$$b(\hat{\underline{\alpha}},\ \underline{\rho}) = 0, \tag{4}$$

where $b(\underline{\alpha}, \underline{\rho}) = \left[\sum_{i=1}^{m} trace \left\{ \frac{\partial F_i^{-1}(\underline{\alpha})}{\partial \alpha_j} T_i(\underline{\rho}) \right\} \right]_{q \times 1}$.

Note that QLS yields a solution to $\left[\sum_{i=1}^{m} trace \left\{ \frac{\partial F_i^{-1}(\underline{\alpha})}{\partial \alpha_j} \left(Z_i Z_i' - \phi T_i(\underline{\rho}) \right) \right\} \right]_{q \times 1} = 0$, which is the equation we obtain if (3) is modified so as to make it unbiased. QLS thus yields a solution to an unbiased estimating equation for $\underline{\rho}$ that does not depend on estimation of ϕ, which would be required were we to attempt to directly obtain a solution to an unbiased estimating equation for $\underline{\rho}$. The final estimate $\hat{\beta}$ is then obtained by solving equation (2) for β, evaluated at $\hat{\underline{\rho}}$ and what we believe to be the true correlation structure. Confidence intervals for linear functions of β can then be constructed using the following estimate of the covariance matrix:

$$C\hat{o}v\left(\hat{\beta}\right) = \left\{ \sum_{i=1}^{m} \hat{D}_i' \hat{A}_i^{-1/2} \hat{F}_i^{-1} \hat{A}_i^{-1/2} \hat{D}_i \right\}^{-1} \left\{ \sum_{i=1}^{m} \hat{D}_i' \hat{A}_i^{-1/2} \hat{F}_i^{-1} \hat{Z}_i \hat{Z}_i' \hat{F}_i^{-1} \hat{A}_i^{-1/2} \hat{D}_i \right\} \left\{ \sum_{i=1}^{m} \hat{D}_i' \hat{A}_i^{-1/2} \hat{F}_i^{-1} \hat{A}_i^{-1/2} \hat{D}_i \right\}^{-1},$$

where $\hat{D}_i$, $\hat{A}_i$, $\hat{F}_i$, and $\hat{Z}_i$ are evaluated at $\left(\hat{\underline{\rho}}, \hat{\beta}\right)$. The parameter ϕ can be estimated consistently by

$$\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} \frac{\hat{Z}_i' \hat{Z}_i}{n_i} \text{ or (with bias correction) } \hat{\phi}_b = \frac{1}{m-p} \sum_{i=1}^{m} \frac{\hat{Z}_i' \hat{F}_i^{-1} \hat{Z}_i}{n_i} . \qquad (5)$$

Note that the QLS estimating equation (2) for β is the same as the GEE estimating equation, so that QLS is a method in the framework for GEE. The QLS and GEE estimators of β have the same asymptotic covariance matrix so that, as $m \to \infty$, the asymptotic relative efficiency of the QLS estimate of β with respect to the GEE estimate is thus 1. Desmond (1997) discussed equation (2) in the context of optimal estimating equations and noted that it is optimal according to the optimality criteria of Godambe and Kale (1991, p. 14). When considering data with more than one level of association, the notation and implementation of QLS can be generalized, as in Shults and Morrow (2002, in appendix) and Kim and Shults (2008b). For a further comparison of QLS with other approaches based on unbiased estimating equations, see Sun et al (2009).

CORRELATION STRUCTURES THAT CAN BE IMPLMENTED IN %SAS

%QLS version 1 currently allows for application of the AR(1), equicorrelated, Markov, and tri-diagonal structures that can be used to describe the correlation between measurements y_{ij} and y_{ik}, that are measured on subject i at times t_{ij} and t_{ik}. **The first-order autoregressive (AR(1)) structure:** This assumes that $Corr(y_{ij}, y_{ik}) = \alpha^{|j-k|}$. The AR(1) structure is often applied for analysis of longitudinal measurements that are equally spaced in time, because it assumes that the correlation between two measurements only depends on the measurement occasion. For example, it might be appropriate for measurements collected at baseline, and then at 6 and 12 months post baseline.

The Markov structure: This structure assumes that $Corr(y_{ij}, y_{ik}) = \alpha^{|t_{ij} - t_{ik}|}$. The Markov structure generalizes the AR(1) to measurements that are unequally spaced in time. This is an extremely useful structure because balanced data are rare. For example, even if the study design planned for measurements to be collected at baseline, and then at 6 and 12 months post baseline, there could be variability (sometimes extreme) in the temporal spacing of measurements. The Markov and AR(1) structures force the correlation between measurements to decline with increasing separation in time, which is often anticipated for biological measurements.

The equicorrelated structure: This structure assumes that all pair-wise correlations on a subject or cluster are identical, so that $Corr(y_{ij}, y_{ik}) = \alpha$. This structure is often applied for analysis of clustered data, e.g. for analysis if a cross-sectional study of rats within litters, or of subjects within classrooms.

The tri-diagonal correlation structure: This structure assumes that all correlations on a cluster are zero, except for the adjacent measurements, so that $Corr(y_{ij}, y_{ij+1}) = \alpha$. This structure is one of the original structures that was implemented by Liang and Zeger (1986) and is available in the current software packages that implement GEE, e.g. in PROC GENMOD in SAS.

Kronecker product structures: We also note that %QLS version 2 allows for application of correlation structures that are formed by taking Kronecker products between the exchangeable structure and the AR1, exchangeable, Markov, or tri-diagonal structures for multiple outcomes. Please see Kim and Shults (2008b) for more discussion of the KP structures and their application. The Kronecker product structures are appropriate for data with multiple sources of correlation, e.g. for multiple measurements collected on subjects within families.

We note that %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.

PRENTICE CONSTRAINTS FOR LONGITUDINAL BINARY DATA

Prentice (1988) described additional constraints for the correlation parameter, in addition to the usual restrictions required for the correlation matrix to be positive definite. The Prentice constraints are due to the fact that the GEE analysis of binary outcomes provides estimates of the marginal probabilities that the measurements y_{ij} take value 1 (or zero); If we consider any pair of measurements y_{ij} and y_{ik} on a subject, the GEE analysis provides estimates of the correlation between the outcomes, in addition to estimates of the marginal probabilities. The bivariate distribution of y_{ij} and y_{ik} can then be expressed as a function of the estimated marginal probabilities and correlation. Prentice (1988) showed that the correlations must satisfy certain constraints, in order for the bivariate distributions to be valid, i.e. for all the pair wise probabilities to be non-negative. Please see Prentice (1988) and Kim and Shults (2008b) for more details.

Prior software for implementation of GEE did not check for a potential violation of the Prentice constraints, although it is well known that violation of these bounds may be problematic for GEE (Rochon, 1988). However, %QLS does check the boundary conditions, and issues a warning if they are not satisfied. (We note that there is some current discussion in the literature regarding the implications of a violation of bounds. For example, Shults et al. (2009) suggest that a violation could be helpful in assessing the correctness of the choice of working correlation structure.)

PARAMETERS IN %QLS

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 of 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.

CLINICAL TRIAL EXAMPLE

It is important to note that the following example is also provided in Kim and Shults (2008a): De Backer (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). The data was also described in Monleberghs & Verbeke (2006), and can be downloaded from the web-site for the LADP project: 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 $P(y_{ij}=1)=p_{ij}$ such that

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right)=\beta_0+\beta_1 treatment_{ij} \qquad (6)$$

where $treatment_{ij}$ is the treatment indicator that equals 1 if the i^{th} subject is assigned to the standard drug and 0 otherwise. One advantage of implementing (6) is that the upper limit of the Prentice constraint for α will be 1 for this model. In general, any model that involves cluster level covariates will have an upper Prentice boundary of 1, due the fact that the estimated bivariate probabilities do not vary within subjects (clusters) when only cluster level covariates are considered in the model.

Before we demonstrate %QLS to fit the model , we must first read the data, toenail.txt, 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.

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 (6) 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

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

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
```

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 treatments.

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. The following code can be used to fit model (6) with the Markov correlation structure:

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

To save space, we omit the stage one and initial outputs and only present the stage two output.

```
            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
```

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.

EXAMPLE OF APPLICATION OF THE EQUICORRELATED AND TRI-DIAGONAL CORRELATION 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 (6) 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);
```

EXAMPLE OF APPLICATION 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:

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 treatment_{ij} + \beta_2 time + \beta_3 time \times treatment \qquad (6)$$

where $treatment_{ij}$ is the treatment indicator that equals 1 if the i^{th} subject is assigned to the standard drug and 0 otherwise; $time_{ij}$ represents the time of the measurement collected on subject i at the j^{th} measurement occasion, and $math \times treatment$ is the treatment by time interaction. To fit the model in (6) 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 (6) with the AR(1) structure, we use

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

To save space, here we provide the stage two output for the AR(1) structure.

```
            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.
```

From the stage two output, the estimated stage two $\hat{\alpha}$ 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 (2009).

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 stage two output.

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

The stag two output for this analysis is as follows:

```
            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
```

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.

CONCLUSION

%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.

In this manuscript we focused on the application of %QLS version 1. However, as noted earlier, we have also developed %QLS version 2, that allows for implementation of the Kronecker product correlation structures. Please see Kim et al. (2008a) and Kim & Shults (2008b, 2008c) for additional details regarding %QLS version 1 and 2. Further updates of %QLS will be made to allow for implementation of other correlation structures that are currently unavailable for GEE, including the familial correlation structure that is described in Gleseer (1992).

REFERENCES

- Chaganty, N. R. (1997) An alternative approach to the analysis of longitudinal data via generalized estimating equations. Journal of Statistical Planning and Inference. 63: 39--54.
- Chaganty, N. R. & Naik, D. (2002) Analysis of multivariate longitudinal data using quasi-least squares. Journal of Statistical Planning and Inference. 103: 421-436.
- Chaganty, N.R. & 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.
- De Backer, M., De Keyser, P., De Vroey, C. & Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day-a double-blind comparative trial. British Journal of Dermatology: 134: 16-17.

- Desmond, A. F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. Journal of Statistical Planning and Inference 60, 77-121.
- Dunlop, D. Regression for Longitudinal Data: A Bridge from Least Squares Regression. (1994). The American Statistician. 48: 1994
- A.T. Galecki, General class of covariance structures for two or more repeated factors in longitudinal data analysis, Communications in Statistics-Theory and Methods 22 (1994), pp. 3105–3120.
- Gleseer, L. (1992). A note on the analysis of familial data. *Biometrika* 79, 412-415.
- Godambe, V.P. and B.K. Kale (1991). Estimating functions: an overview. In: V.P. Godambe, Ed., Estimating Functions, Oxford Univ. Press, Oxford, 3-20.
- Hardin, J.M. & Hilbe, J.M. (2002). Generalized estimating equations. Florida: Chapman & Hall/CRC Press.
- Hilbe, J.M. (2008). Logistic Regression Models. Chapman & Hall/CRC Press.
- Kim H., Shults J., Patterson S., & Goldberg-Alberts, R. Analysis of Adverse Events in Drug Safety: A Multivariate Approach Using Stratified Quasi-least Squares" (2008a). UPenn Biostatistics Working Papers*.* Working Paper 29. http://biostats.bepress.com/upennbiostat/papers/art29.
- Kim H. & Shults J. %QLS SAS Macro: A SAS macro for Analysis of Longitudinal Data Using Quasi-Least Squares. (2008b). UPenn Biostatistics Working Papers. Working Paper 27. http://biostats.bepress.com/upennbiostat/papers/art27
- Kim, H. & Shults J. Analysis of unbalanced multi-outcome longitudinal data using quasi-least squares in SAS. (2008c). UPenn Biostatistics Working Papers*.* Working Paper 30. http://biostats.bepress.com/upennbiostat/papers/art30
- Lefkopoulou M, Moore D, Ryan L. The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. Journal of the American Statistical Association 1989; 84(407): 810-815
- 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 & Shults (2008) Ratcliffe, S. & Shults, J. (2008) GEEQBOX: A MATLAB Toolbox for Generalized Estimating Equations and Quasi-Least Squares. Journal of Statistical software, 25, Issue 14, 1-14.
- Roberts, D.T. (1992) Prevalence of dermatophyte onychomycosis in the United Kingdom: 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.
- Roy, A., Khattree, R. (2005). Testing the hypothesis of a Kronecker product covariance matrix in multivariate repeated measures data. SAS Users Group International, Proceedings of the Statistics and Data Analysis Section. Paper 199-30: 1-11.
- 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 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 Ratcliffe, 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. This paper is also in press at Statistics and Its Interface
- 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. (2009) 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.
- Sun, W., Shults, J., and Leonard, M. (2009). Use of unbiased estimating equations to estimate correlation in GEE analysis of longitudinal trials. Biometrical Journal, to appear.
- Xie, J. & Shults, J. (2008) Implementation of quasi-least squares with the R package qlspack. Journal of Statistical Software, to appear.

ACKNOWLEDGMENTS

This research was partially supported by the NIH funded R01 The “Longitudinal Analysis for Diverse Populations” (LADP) project R01CA096885.

RECOMMENDED READING

Please see our web-site of the Longitudinal Analysis for Diverse Populations Project for more details regarding our research on QLS: www.cceb.upenn.edu/~sratclif/QLSproject.html.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Hanjoo Kim, PhD
Enterprise: Forest Research Institute, Inc.
Address: Harborside Financial Center Plaza V
City, State ZIP: Jersey City, NJ 07311
Work Phone: 201-427-8356
Fax: 201-427-8496
E-mail: han-joo.kim@frx.com
Web: www.frx.com

Name: Justine Shults, PhD
Enterprise: University of Pennsylvania School of Medicine
Address: 423 Guardian Drive
City, State ZIP: Philadelphia PA 19104-6021
Work Phone: 215-573-6526
E-mail: jshults@mail.med.upenn.edu
Web: http://www.cceb.med.upenn.edu/faculty/?id=167