GEEQBOX: User's Guide Version 1.0

Sarah J. Ratcliffe and Justine Shults

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine,
6th flr Blockley Hall, 423 Guardian Drive, Philadelphia, PA, 19104-6021, USA.
Email:sratclif@cceb.med.upenn.edu

August 24, 2006

Contents

1 Introduction **1**

2 Statistical Background **2**

2.1 Notation 3

2.2 Working Correlation Structures that Can be Implemented with GEEQBOX 3

2.3 GEE Estimates of the Correlation Parameter 4

2.4 QLS Estimates of the Correlation Parameter 5

2.5 Confidence Intervals and P-values for Tests of Hypotheses Involving the Regression Parameter 7

3 Performing Analyses **8**

3.1 Data Set-up 8

3.2 Results 9

3.3 Example 9

4 Functions **15**

4.1 Reference Tables 15

4.2 Commands and Functions 17

Bibliography **23**

1

Introduction

This document is for help with the use of the MATLAB toolbox GEEQBOX. It details how to use the m-files, lists the functions available, and gives a detailed description of each function.

The GEEQBOX toolbox can be used to analyze data with correlated outcomes. Version 1.0 of the toolbox currently currently allows for:

- Three possible data distributions.
- Six assumed correlation structures.
- Estimation by either generalised estimating equations (GEE) or quasi-least squares (QLS).

The toolbox, documentation and updates are available from the website:
http://www.cceb.upenn.edu/~sratclif/QLSproject.html

Acknowledgments

Work on this software was supported by the NIH grant R01CA096885.

2

Statistical Background

This GEEQBOX toolbox is for use with correlated data. The most common use is in londitudinal data analysis, where the outcome is observed on each subject at multiple time points over the course of the study. For simplicity, we used subject to indicate the outcomes that are correlated though the correlated measurements may come from a group membership, etc. The outcome can be discrete or continuous and the data can be analyzed using a form of regression analysis for correlated data. The two methods implemented in this toolbox are generalized estimating equations and quasi-least squares.

Both GEE and QLS are iterative approaches that alternate between
1. updating the estimate of the regression parameter β by solving the GEE estimating equation for β and
2. updating the estimate of the correlation parameter α via moment estimation (GEE) or solving an unbiased estimating equation for α in two stages (QLS).

The method of generalized estimating equations (GEE) (Liang and Zeger, 1986) is extremely popular because it allows for straight-forward analysis of correlated outcomes via a marginal model. That is, the model only depends on the covariates of interest, and it does not require any distributional assumptions about the outcomes. The basic idea of GEE is an extension of a generalized linear model which incorporates the covariance matrix of the outcomes for each subject.

The method of quasi-least squares (QLS) is a two-stage approach in the framework of GEE. Chaganty (1997) described QLS for an equal number of observations per subject (balanced data) while Shults (1996) and Shults and Chaganty (1998) extended stage one for several correlation structures and unbalanced data. However, the stage one QLS estimate of the correlation parameter typically is not consistent, even if the correlation structure is correctly specified. Chaganty and Shults (1999) therefore introduced a second stage of QLS that provides a solution to an unbiased estimating equation for the correlation parameter.

GEEQBOX implements both GEE and QLS because QLS may allow for easier implementation of some complex correlation structures, e.g. the Markov correlation structure that is appropriate for longitudinal

studies with variable temporal spacing of measurements. In addition, QLS can also be considered as an alternative approach should the method of GEE fail to converge, or should GEE yield an estimate of the correlation parameter that corresponds to an estimated correlation matrix that is not positive definite (Shults and Chaganty, 1998). For more discussion of QLS that includes a comparison of stage one of QLS versus stage two and a comparison with other approaches see Sun *et al.* (2006).

2.1 Notation

For analysis of a longitudinal study, we assume that outcome measurements $Y_i = (y_{i1}, \cdots, y_{in_i})'$ and associated covariates $x'_{ij} = (x_{ij1}, \cdots, x_{ijp})$ were collected on subject i at times $T_i = (t_{i1}, \cdots, t_{in_i})'$, for $i = 1, \cdots, m$. The data are considered balanced and equally spaced when $n_i = n \; \forall \; i$ and $|t_{ij} - t_{ik}| = \gamma \; \forall \; i, j, k$, respectively. For analysis of a cross-sectional study, e.g. if one measurement is collected on each of several subjects within multiple clusters, then $Y_i = (y_{i1}, \cdots, y_{in_i})'$ represents the n_i measurements that were collected within cluster i.

The expected value and variance of measurement y_{ij} on subject i are assumed to equal $E(y_{ij}) = g^{-1}(x'_{ij}\beta) = u_{ij}$ and $Var(y_{ij}) = \phi h(u_{ij})$, respectively, where ϕ is a known or unknown scale parameter. We also let $U_i(\beta)$ represent the $n_i \times 1$ vector of expected values u_{ij} on subject i. For longitudinal and cross-sectional studies, observations are assumed to be independent if they are measured on different subjects. However, within subjects, they are assumed to be correlated, with a pattern of association that can be described by a *working correlation structure* . The working correlation structure for subject i, denoted by $Corr(Y_i) = R_i(\alpha)$, depends on a correlation parameter α that can be scalar or vector-valued. We will underline correlation parameters of dimension greater than one. The covariance matrix of Y_i is then given by $Cov(Y_i) = \phi {A_i}^{1/2} R_i(\alpha) {A_i}^{1/2}$, where $A_i = diag(h(u_{i1}), \ldots, h(u_{in_i}))$.

2.2 Working Correlation Structures that Can be Implemented with GEEQBOX

GEEQBOX currently implements the following structures, with plans to implement additional structures that will be made available on the website. The `feasible region` for each structure is defined as the interval on which α will yield a positive definite correlation matrix.

- `The Equicorrelated:` This structure assumes that all pairwise correlations within a cluster are equal, so that $Corr(y_{ij}, y_{ik}) = \alpha$. This structure is plausible for cross-sectional studies, e.g. to describe the pattern of association of weights among litter-mates of baby rats. The feasible region for this structure is $(-1/(n_m - 1), 1)$, where n_m represents the maximum value of n_i over $i = 1, 2, \ldots, m$.
- `The first-order autoregressive AR(1):` This structure assumes that the correlation among repeated measurements on a subject depends on their separation in order of measurement, so that

$Corr(y_{ij}, y_{ik}) = \alpha^{j-k}$. This structure is plausible for longitudinal studies in which the collection times of measurements are equally spaced in time, e.g. in a weight loss intervention that measures weights on subjects at baseline and then at three and six months post-baseline. The feasible region for this structure is $(-1, 1)$. However, a negative value for α may be biologically implausible because it may be unreasonable to allow the intra-subject correlations to alternate in sign, e.g. for α^2 and α^3 to be positive and negative, respectively.

- `The Markov correlation structure:` This structure assumes that the correlation among repeated measurements on a subject depends on their `timing` of measurement, so that $Corr(y_{ij}, y_{ik}) = \alpha^{|t_{ij}-t_{ik}|}$. This structure generalizes the AR(1) structure to allow for unequal spacing of measurements. The feasible region for this structure is $(-1, 1)$. However, as for the AR(1) structure, a negative value for α is typically not biologically plausible. GEEQBOX therefore uses QLS to obtain an estimate of $\alpha \in (0, 1)$. We note that GEEQBOX does not implement the Markov structure for GEE because it is not straightforward to obtain a moment estimate for this structure.

- `The tri-diagonal correlation structure:` This structure assumes that the correlation among measurements on a subject is constant for measurements that are separated by one measurement occasion, so that $Corr(y_{ij}, y_{ik}) = \alpha$ for $|j-k| = 1$ and is zero otherwise. The authors are not aware of many practical applications for this structure, but it was implemented in Liang and Zeger (1986) and in most standard software packages that implement GEE. The feasible region for this structure is $\left(\frac{-1}{2\sin\frac{\pi(n_m-1)}{2(n_m+1)}}, \frac{1}{2\sin\frac{\pi(n_m-1)}{2(n_m+1)}} \right)$, where 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.

- `Unstructured:` This structure does not assume any pattern for the intra-subject correlations, so that $Corr(Y_{ij}, Y_{ik}) = \alpha_{jk}$. This structure has been implemented in QLS (Chaganty, 1997; Chaganty and Shults, 1999) but the algorithms are somewhat complex. GEEQBOX therefore implements a moment estimate using GEE.

- `Working Independent:` Another popular structure is the identity matrix. Implementation of this structure is straightforward because β can then be estimated in a non-iterative process. However, several authors have shown that incorrect application of the working independence structure can result in a serious loss in efficiency in estimation of β (e.g. Sutradhar and Das, 2000; Wang and Carey, 2004; Shults *et al.*, 2006)

2.3 GEE Estimates of the Correlation Parameter

For GEE, GEEQBOX implements the following moment estimates that are implemented in PROC GENMOD in SAS.

For the equicorrelated structure, the GEE moment estimate is given by:

$$\hat{\alpha}_{GEE-EQUI} = \frac{\sum_{i=1}^{m} \sum_{j \neq k} z_{ij} z_{ik}}{(N^* - p)\hat{\phi}_{GEE}}$$

where

$$N^* = \sum_{i=1}^{m} n_i(n_i - 1),$$

$$\hat{\phi}_{GEE} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}^2}{N - p},$$

$N = \sum_{i=1}^{m} n_i$, z_{ij} is the Pearson residual for subject i at time t_{ij} and p is the dimension of β.

For the AR(1) and tri-diagonal estimates, the GEE moment estimate is:

$$\hat{\alpha}_{GEE-TRI} = \hat{\alpha}_{GEE-AR1} = \frac{\sum_{i=1}^{m}\sum_{j=2}^{n_i} z_{ij}z_{i,j-1}}{(N^{**} - p)\hat{\phi}_{GEE}}$$

where $N^{**} = \sum_{i=1}^{m}(n_i - 1)$.

For the unstructured correlation matrix, GEEQBOX implements the following moment estimate for element j, k of the matrix:

$$R_i\,[j,k] = \frac{\sum_{i=1}^{m} z_{ij}z_{ik}}{(m - p)\hat{\phi}_{GEE}}$$

A moment based estimator has not been proposed in the literature for implementation of the more general Markov correlation for GEE, which provides motivation for implementation of QLS.

2.4 QLS Estimates of the Correlation Parameter

While GEE typically uses moment estimates for α, QLS estimates α by obtaining a solution to an unbiased estimating equation in two stages (see Sun *et al.*, 2006, for more details). In stage one, QLS alternates between updating the estimates of β and solving the *stage one estimating equation* for α until convergence.

$$\frac{\partial}{\partial\alpha}\left\{\sum_{i=1}^{m} Z_i'(\beta)\left\{R_i^{-1}(\alpha)\right\}Z_i(\beta)\right\} = 0 \tag{2.1}$$

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

The solution $\hat{\alpha}$ to (2.1) is not consistent. Stage two of QLS therefore obtains a consistent estimate $\hat{\alpha}_{QLS}$ as the solution to the *stage two estimating equation* for α.

$$\sum_{i=1}^{m} trace\left\{\frac{\partial R_i^{-1}(\delta)}{\partial\delta}R_i(\alpha)\right\}\Bigg|_{\delta=\hat{\alpha}} = 0 \tag{2.2}$$

The final QLS estimate $\hat{\beta}_{QLS}$ of β is then obtained by solving the GEE estimating equation for β evaluated at $\hat{\alpha}_{QLS}$. For estimating equations that do not have a unique solution, GEEQBOX uses the bisection method to obtain a solution in the feasible region for α.

For the AR(1) structure and for unbalanced data, Shults and Chaganty (1998) proved that the feasible stage one estimate $\hat{\alpha}$ can be expressed as:

$$\hat{\alpha}_{QONE} = \frac{\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}^2 + z_{ij-1}^2) - \sqrt{\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}^2 + z_{ij-1}^2)\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}^2 - z_{ij-1}^2)}}{2\sum_{i=1}^{m}\sum_{j=2}^{n_i} z_{ij}z_{ij-1}} \tag{2.3}$$

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

$$\hat{\alpha}_{QLS-AR1} = \frac{2\hat{\alpha}_{QONE}}{1+\hat{\alpha}_{QONE}^2}. \tag{2.4}$$

For the Markov structure and unbalanced data, Shults and Chaganty (1998) provided the QLS stage one estimating equation for α:

$$\sum_{i=1}^{m}\sum_{j=2}^{n_i}\frac{e_{ij}\alpha^{e_{ij}}\left[\alpha^{2e_{ij}}z_{ij}z_{i,j-1} - \alpha^{e_{ij}}\left(z_{ij}^2 + z_{i,j-1}^2\right) + z_{ij}z_{i,j-1}\right]}{1-\alpha^{2e_{ij}}} = 0 \tag{2.5}$$

where $e_{ij} = |t_{ij} - t_{i,j-1}|$. Note that GEEQBOX requires that $e_{ij} \geq 1\ \forall\, i$ and j. The stage two estimating equation for the Markov structure (Chaganty and Shults, 1999) is given by:

$$\sum_{i=1}^{m}\sum_{j=2}^{n_i}\left.\frac{2e_{ij}\delta^{2e_{ij}-1} - \alpha^{e_ij}e_{ij}\left[\delta^{e_{ij}-1} + \delta^{3e_{ij}-1}\right]}{1-\delta^{2e_{ij}}}\right|_{\delta=\hat{\alpha}} = 0 \tag{2.6}$$

For the equicorrelated structure and for unbalanced data, Shults and Chaganty (1998) proved that there will be a unique feasible solution to the following stage one estimating equation for α:

$$\sum_{i:n_i>1} Z_i'\,Z_i - \sum_{i:n_i>1}\frac{1+\alpha^2(n_i-1)}{(1+\alpha(n_i-1))^2}\,(Z_i'(\beta)\,e_i)^2 = 0 \tag{2.7}$$

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

$$\sum_{i:n_i>1}\frac{n_i\,(n_i-1)\,\widehat{\alpha}\,(\widehat{\alpha}\,(n_i-2)+2)}{(1+\widehat{\alpha}(n_i-1))^2} \Big/ \sum_{i:n_i>1}\frac{n_i\,(n_i-1)\,\left(1+\widehat{\alpha}^2(n_i-1)\right)}{(1+\widehat{\alpha}(n_i-1))^2} \tag{2.8}$$

For the tri-diagonal structure and unbalanced data, GEEQBOX obtains solutions to the stage one and two estimating equations (2.1) and (2.2) for the tri-diagonal structure by first constructing the tri-diagonal matrix $R_i(\hat{\alpha})$. Next, to evaluate

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

GEEQBOX implements the following expression:

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

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

2.5 Confidence Intervals and P-values for Tests of Hypotheses Involving the Regression Parameter

The asymptotic distribution of the QLS estimate $\hat{\beta}_{QLS}$ is the same as the asymptotic distribution of the GEE estimate $\hat{\beta}_{GEE}$. GEEQBOX therefore provides both model based and sandwich-based estimates of the covariance matrix of $\hat{\beta}$ (Liang and Zeger, 1986). The covariance matrix depends on the scalar parameter ϕ; GEEQBOX implements the estimate provided in Chaganty and Shults (1999). The *model based* estimate of the covariance matrix is appropriate when the user has a high degree of confidence that the correlation structure has been correctly specified. It has the following form:

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

where

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

and $\hat{\phi} = \min\left\{\hat{\phi_p}, \hat{\phi_c}\right\}$, for

$$\hat{\phi_p} = \frac{1}{m}\sum_{i=1}^{m} \frac{Z_i(\hat{\beta})' Z_i(\hat{\beta})}{n_i} \text{ and } \hat{\phi_c} = \frac{1}{m}\sum_{i=1}^{m} \frac{Z_i(\hat{\beta})' R_i^{-1}(\hat{\alpha}) Z_i(\hat{\beta})}{n_i}.$$

The *robust sandwich* covariance matrix has the following form:

$$\widehat{Cov}_R(\hat{\beta}) = {W_m}^{-1} \left\{ \sum_{i=1}^{m} X_i' {A_i}^{1/2} R_i^{-1}(\hat{\alpha}) Z_i(\hat{\beta}) Z_i'(\hat{\beta}) R_i^{-1}(\hat{\alpha}) {A_i}^{1/2} X_i \right\} {W_m}^{-1}. \tag{2.9}$$

GEEQBOX provides estimated standard errors, 95% confidence intervals, and p-values for the tests $\beta_j = 0$ that are based on both the *model* and *sandwich* covariance matrices.

3

Performing Analyses

There are two main functions, `gee` and `qls`, that can be used to analyze the correlated data. These functions take the same inputs and produce the same outputs. This chapter explains how the data must be set-up for analysis, what results are obtained and give an example of their use.

3.1 Data Set-up

Before using one of the analysis functions, the data must be organized such that each outcome measurement (Y) and associated id's, time points (t), and covariates (X) are listed on a separate row. That is,

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{nn_n} \end{bmatrix} \qquad \text{id} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ \vdots \\ 2 \\ \vdots \\ n \\ \vdots \\ n \end{bmatrix} \qquad t = \begin{bmatrix} t_{11} \\ t_{12} \\ \vdots \\ t_{1n_1} \\ t_{21} \\ \vdots \\ t_{2n_2} \\ \vdots \\ t_{n1} \\ \vdots \\ t_{nn_n} \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n_1} \\ x_{21} \\ \vdots \\ x_{2n_2} \\ \vdots \\ x_{n1} \\ \vdots \\ x_{nn_n} \end{bmatrix}$$

$$i = 1, \ldots, n, \quad j = 1, \ldots, n_i$$

where y_{ij} is the j-th measurement on the i-th subject, t_{ij} is the associated time point and x_{ij} is the row vector of covariates at time t_{ij}.

In addition, as shown above, the measurements must be sorted so that all measurements from the same subject (id) are listed on consecutive rows. If the id vector consisted of $[1\ 1\ 2\ 2\ 2\ 1\ 1]^T$, then the program would count this as 3 subjects since there are 3 changes in id numbers. However, the id's do not have to be consecutive numbers. For example, id=$[12\ 12\ 12\ 10\ 10\ 10\ 99\ 99]^T$ would produce the same results as id=$[1\ 1\ 1\ 2\ 2\ 2\ 3\ 3]^T$.

The matrix of covariates, X, should be set-up so that each column contains a separate covariate. At present, there should be no missing data in X. A constant term is not included by default in the programs. Thus, in order to include a constant in the model, a column of ones must be included as a covariate in X. This column of ones should be the *final* column of X. The programs will default to calling the beta estimate by the associated column number of X. However, this can be over-written with the optional `varnames` input. Varnames should be a cell row array that defines the names to be used instead of 1, 2, etc. For example, if X contains the covariates, in order, time, age and a constant term, then varnames could be defined as `varnames={'time','age','constant'}`.

3.2 Results

Each function produces the same printed results and output variables. The printed results consist of the initial values used by the algorithms, the estimated covariance parameter (α), scale parameter (ϕ), and the covariate parameter estimates (β). In addition, the standard errors, corresponding z-values, p-values and 95% confidence intervals for each β_j are also given. Two versions of these values are presented; the one based on the robust covariance matrix and the one based on the model based covariance matrix. The model based results should be used when the specified working correlation matrix is known to be correct; otherwise, the robust results should be used.

Three variables are produced as outputs to the functions. These are the estimated β's, α, and a structured variable results that contains the entire printed results from the robust estimations in the cell variable results.robust and the model based estimation in the cell variable results.model.

3.3 Example

We use the data provided in Table 3 of Nunez-Anton and Woodworth (1994) of illustrate the use of the functions. This data is contained in the files audio.dat (ascii file) and audio.mat (Matlab data file), which can be obtained from the web site.

The data consists of the following variables: subject id; group (=0 for A; =1 for B); month of measurement; percent; and high (=0 for percent¡50; =1 for percent$\geq$ 50). The variable percent represents the "percent correct scores on a sentence test administered under audition-only conditions to groups of subjects wearing two different cochlear implants, referred to here as A and B. The electrode array was surgically implanted 5 to 6 weeks prior to being electrically connected to the external speech processor.

Subjects were profoundly, bilaterally deaf, thus preconnection baseline values for the sentence test were all zero. At the time of the analysis reported here, data were available for 23 subjects in group A and 21 subjects in group B, with measurements scheduled at 1, 9, 18, 30 months after connection."(Nunez-Anton and Woodworth, 1994)

```
>> load('audio.dat');
>> id = audio(:,1);
>> group = audio(:,2);
>> month = audio(:,3);
>> percent = audio(:,4);
>> high = audio(:,5);
>> cons = ones(136,1);
```

In the worked examples presented here $\beta = (\beta_1, \beta_2, \beta_3)$ where β_1 is the regression coefficient associated with month of measurement; β_2 is the regression coefficient associated with group (group = 0 for A; group = 1 for B); and β_3 is the regression coefficient associated with cons, the constant that takes value one. There are two outcomes to be considered: the continuous outcome percent, and the binary outcome high. Thus, the X matrix of interest is set as

```
>> X=[month group cons];
```

and the associated variable names for displaying the results are

```
>> varnames={'month','group','cons'};
```

For the continuous outcome percent, we first estimate the parameters using QLS and assuming a normal distribution for Y (identity link) and an Markov correlation structure. The distribution/link can be set using the 5th input (Family), and the working correlation structure using the 6th input (CorrStruct). The normal distribution is set using the option 'n', while the Markov assumption is specified via 'markov'. These inputs are not case sensitive; thus 'N', 'Markov', 'MARKOV', etc. would all have be acceptable. Thus, the command to perform the analysis, and resulting output, would be:

```
>>[betahat,alphahat,results] = qls(id,percent,month,X,'n','markov',varnames);
Normal distribution family assumed
Markov Correlation structure assumed
Initial estimate of beta = [0.830658     -9.50759      28.2862]

Stage 1 estimate of alpha = 0.94555
Stage 1 estimate of beta = [0.966584     -10.8228      26.2529]

Stage 2 estimate of alpha = 0.98623
Stage 2 estimate of beta = [1.05273     -12.0745      25.3058]

QLS estimate of scale parameter = 720.8796
```

```
Estimates based on ROBUST covariance matrix
  Columns 1 through 5

          ''            ''            ''           ''              ''
    'Variable'    'Beta'        'Std.Error'    'z value'    'p-value'
    'month'       [  1.0527]    [   0.1399]    [ 7.5231]    [5.3513e-014]
    'group'       [-12.0745]    [   7.3972]    [-1.6323]    [     0.1026]
    'constant'    [ 25.3058]    [   5.1899]    [ 4.8760]    [1.0827e-006]

  Columns 6 through 7

    '95% CI'             ''
    'low lim'     'up lim'
    [  0.7785]    [ 1.3270]
    [-26.5727]    [ 2.4237]
    [ 15.1338]    [35.4778]

Estimates based on MODEL based covariance matrix
  Columns 1 through 5

          ''            ''            ''           ''              ''
    'Variable'    'Beta'        'Std.Error'    'z value'    'p-value'
    'month'       [  1.0527]    [   0.1388]    [ 7.5848]    [3.3307e-014]
    'group'       [-12.0745]    [   7.6073]    [-1.5872]    [     0.1125]
    'constant'    [ 25.3058]    [   5.5109]    [ 4.5919]    [4.3917e-006]

  Columns 6 through 7

    '95% CI'             ''
    'low lim'     'up lim'
    [  0.7807]    [ 1.3248]
    [-26.9845]    [ 2.8355]
    [ 14.5046]    [36.1070]
```

Similarly, using GEEs and assuming an equicorrelated covariance structure, we would obtain:

```
>> [betahat,alphahat,results] = gee(id,percent,month,X,'n','equi',varnames);
Normal distribution family assumed
Equicorrelated structure assumed
```

```
GEE estimate of alpha = 0.89733
GEE estimate of beta = [0.909073     -11.9411      28.4943]

GEE estimate of scale parameter = 718.7803

Estimates based on ROBUST covariance matrix
  Columns 1 through 5

           ''            ''            ''           ''               ''
    'Variable'    'Beta'        'Std.Error'    'z value'    'p-value'
    'month'       [  0.9091]    [   0.1288]    [ 7.0602]    [1.6627e-012]
    'group'       [-11.9411]    [   7.5661]    [-1.5782]    [     0.1145]
    'constant'    [ 28.4943]    [   5.3765]    [ 5.2998]    [1.1596e-007]

  Columns 6 through 7

    '95% CI'             ''
    'low lim'     'up lim'
    [  0.6567]    [ 1.1614]
    [-26.7705]    [ 2.8883]
    [ 17.9565]    [39.0322]

Estimates based on MODEL based covariance matrix
  Columns 1 through 5

           ''            ''            ''           ''               ''
    'Variable'    'Beta'        'Std.Error'    'z value'    'p-value'
    'month'       [  0.9091]    [   0.0825]    [11.0201]    [          0]
    'group'       [-11.9411]    [   7.8167]    [-1.5276]    [     0.1266]
    'constant'    [ 28.4943]    [   5.4790]    [ 5.2007]    [1.9859e-007]

  Columns 6 through 7

    '95% CI'             ''
    'low lim'     'up lim'
    [  0.7474]    [ 1.0708]
    [-27.2616]    [ 3.3794]
    [ 17.7557]    [39.2330]
```

To model the binary outcome high, we need to use the Bernoulli distribution assumption. This is done by setting the 'Family' input to 'b'. The QLS estimates under a tr-diagonal correlation assumption are

shown below.

```
>> [betahat,alphahat,results] = qls(id,high,month,X,'b','tri',varnames);
Bernoulli distribution family assumed
Tri-diagonal Correlation structure assumed
Initial estimate of beta = [0.049843    -0.81037     -1.0891]

Stage 1 estimate of alpha = 0.31269
Stage 1 estimate of beta = [0.053484    -0.79452     -1.1604]

Stage 2 estimate of alpha = 0.51664
Stage 2 estimate of beta = [0.056485    -0.76032     -1.2516]

QLS estimate of scale parameter = 1

Estimates based on ROBUST covariance matrix
  Columns 1 through 5

           ,,           ,,             ,,           ,,           ,,
    'Variable'    'Beta'       'Std.Error'    'z value'    'p-value'
    'month'       [ 0.0565]    [   0.0183]    [ 3.0923]    [ 0.0020]
    'group'       [-0.7603]    [   0.6172]    [-1.2318]    [ 0.2180]
    'constant'    [-1.2516]    [   0.4182]    [-2.9928]    [ 0.0028]

  Columns 6 through 7

    '95% CI'            ''
    'low lim'    'up lim'
    [ 0.0207]    [ 0.0923]
    [-1.9701]    [ 0.4494]
    [-2.0712]    [-0.4319]

Estimates based on MODEL based covariance matrix
  Columns 1 through 5

           ,,           ,,             ,,           ,,           ,,
    'Variable'    'Beta'       'Std.Error'    'z value'    'p-value'
    'month'       [ 0.0565]    [   0.0193]    [ 2.9196]    [ 0.0035]
    'group'       [-0.7603]    [   0.5018]    [-1.5151]    [ 0.1297]
    'constant'    [-1.2516]    [   0.4133]    [-3.0283]    [ 0.0025]

  Columns 6 through 7
```

```
'95% CI'            ''
'low lim'    'up lim'
[ 0.0186]    [ 0.0944]
[-1.7439]    [ 0.2233]
[-2.0616]    [-0.4415]
```

4

Functions

The detailed description for each function is organized as follows:

- name of function
- a statement of purpose
- a synopsis of the function's syntax
- a description of what the function does

and, for selected functions,

- examples
- a detailed description of the mathematical procedure/equations comprising the function.

4.1 Reference Tables

The GEEQBOX toolbox contains the following functions:

	Top Level Function
`qls`	Analyzes correlated data via QLS.
`gee`	Analyzes correlated data via GEEs.

	Called Functions (sub-routines)
`counti`	Counts the number of times each unique element occurs in a vector.
`fequi1`	Computes the stage 1 QLS estimates for the equicorrelated structure assumption.
`fmark1`	Computes the stage 1 QLS estimates for the Markov correlation structure assumption.
`fmark2`	Computes the stage 2 QLS estimates for the Markov correlation structure assumption.
`ftri1`	Computes the stage 1 QLS estimates for the tri-diagonal correlation structure assumption.
`ftri2`	Computes the stage 2 QLS estimates for the tri-diagonal correlation structure assumption.
`ftri2a`	Computes the stage 2 QLS estimates for the tri-diagonal correlation structure assumption.

4.2 Commands and Functions

counti

Purpose Counts the number of times each unique element occurs in a vector

Syntax `[ni,n] = counti(x);`

Description `[ni,n] = counti(x)` for a vector of non-unique elements, `x`, returns the number of times each unique element is contained in `x` in `ni`. The number of unique elements is optionally returned in `n`.

Note: the non-unique elements of `x` must be grouped together and `ni` will contain the number of each non-unique element in the order in which they appear in `x`.

---

fequi1

Purpose Computes the stage 1 QLS estimates for the equicorrelated structure assumption.

Syntax `fa = fequi1(x,p1,p2,p3);`

Description `fa = fequi1(x,p1,p2,p3);` is used to find the optimize `x` in the stage 1 QLS function assuming an equicorrelated structure and given known parameters `p1`, `p2`, and `p3`. That is, it is called by `qls` when the equicorrelated option has been specified.

---

fmark1

Purpose Computes the stage 1 QLS estimates for the Markov correlation structure assumption.

Syntax `fa = fmark1(x,p1,p2,p3,p4);`

Description `fa = fmark1(x,p1,p2,p3,p4);` is used to find the optimize `x` in the stage 1 QLS function assuming a Markov correlation structure and given known parameters `p1`, to `p4`. That is, it is called by `qls` when the Markov option has been specified.

See Also `fmark2`

---

fmark2

Purpose Computes the stage 2 QLS estimates for the Markov correlation structure assumption.

Syntax `fa = fmark2(x,p1,p2);`

Description `fa = fmark1(x,p1,p2);` is used to find the optimize `x` in the stage 2 QLS function assuming a Markov correlation structure and given known parameters `p1`, and `p2`. That is, it is called by `qls` when the Markov option has been specified.

See Also `fmark1`

---

ftri1

Purpose Computes the stage 1 QLS estimates for the Tri-diagonal correlation structure assumption.

Syntax `fa = ftri1(x,p1,p2,p3,p4);`

Description `fa = ftri1(x,p1,p2,p3,p4);` is used to find the optimize `x` in the stage 1 QLS function assuming a tri-diagonal correlation structure and given known parameters `p1`, to `p4`. That is, it is called by `qls` when the Tri-diagonal option has been specified.

See Also `ftri2 ftri2a`

---

ftri2

Purpose Computes the stage 2 QLS estimates for the Tri-diagonal correlation structure assumption.

Syntax `fa = ftri2(x,p1,p2,p3,p4,p5);`

Description `fa = ftri2(x,p1,p2,p3,p4,p5);` is used to find the optimize `x` in the stage 2 QLS function assuming a tri-diagonal correlation structure and given known parameters `p1`, to `p5`. That is, it is called by `qls` when the Tri-diagonal option has been specified. It is called when an absolute zero exists within the feasible region for the estimate.

See Also `ftri1 ftri2a`

---

ftri2a

Purpose Computes the stage 2 QLS estimates for the Tri-diagonal correlation structure assumption.

Syntax `fa = ftri2a(x,p1,p2,p3,p4,p5);`

Description `fa = ftri2a(x,p1,p2,p3,p4,p5);` is used to find the optimize `x` in the stage 2 QLS function assuming a tri-diagonal correlation structure and given known parameters `p1`, to `p5`. That is, it is called by `qls` when the Tri-diagonal option has been specified. It is called when an absolute zero does not exist within the feasible region for the estimate. Instead, it finds the value of `x` closest to zero within the feasible region.

See Also `ftri1 ftri2a`

---

gee

Purpose Analyses correlated data via GEE.

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

Description `[bhat,alpha,results] = gee(id,y,t,X);` analyses correlated data via GEEs where `y` is a $N \times 1$ vector containing the repeated measures data, measured at corresponding times `t`, on subjects `id`, regressed on fixed covariates `X`. In order for a constant to be included in the model, the final column of `X` could be a column of ones. For all these variables, the number of rows should be equal to

$$N = \sum_{i=1}^{n} n_i$$

This model assumes the data follows a normal distribution with a Markov correlation structure. `[bhat,alpha,results] = gee(id,y,t,X,Family,CorrStruct);` can be used to change these assumptions. `Family` specifies the assumed distribution of the data. It can take values:

- `1` or `n` for a Normal distribution (default);
- `2` or `b` for a Bernoulli distribution; or
- `3` or `p` for a Poisson distribution.

`CorrStruct` specifies the assumed correlation structure. It can take values:

- `1` or `ar1` for an AR(1) structure;
- `2` or `markov` for a Markov structure (default);
- `3` or `equi` for an Equicorrelated structure;
- `4` or `tri` for a Tri-diagonal correlation structure;
- `5` or `un` for an Unstructured correlation matrix; or
- `6` or `ind` for a working Independent correlation structure.

Further, `varnames` can be used to specify the variables names to be used in the output.

This is a cell row array which corresponds to the variables entered in `X`. Additionally, `maxit` specifies the maximum number of iterations (default is 100), and `tol` the tolerance used for determining convergence (default is 1e-5).

`[bhat,alpha,results]` contains the parameter estimates for the fixed effects `X` in `bhat`, and the estimated correlation parameters in `alpha`. `results` is a structured array with two elements, `robust` and `model` which contains the entire printed results / estimates under the robust and model based covariance matrices, respectively.

Examples To fit the regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

where `id` and `time` contains the i's and time points associated with each row. Under a normal distribution and Markov working correlation structure, the model would be fit with the commands

```
X = [X1,X2,ones(length(Y,1))];
varnames = {'X1','X2','constant'};
[bhat,alpha,results] = gee(id,Y,time,X,'n','markov',varnames);
```

The results from fitting this model to some data would be:

```
Normal distribution family assumed
AR(1) Correlation structure assumed

GEE estimate of alpha = 0.85796
GEE estimate of beta = [0.67155  -0.0037861      0.8628]

GEE estimate of scale parameter = 0.61616

Estimates based on ROBUST covariance matrix
  Columns 1 through 5

           ''           ''             ''           ''           ''
    'Variable'    'Beta'       'Std.Error'    'z value'    'p-value'
    [       1]    [ 0.6715]    [   0.0572]    [11.7442]    [      0]
    [       2]    [-0.0038]    [   0.0024]    [-1.5532]    [ 0.1204]
    [       3]    [ 0.8628]    [   0.0586]    [14.7258]    [      0]

  Columns 6 through 7

    '95% CI'                ''
    'low lim'    'up lim'
    [ 0.5595]    [     0.7836]
    [-0.0086]    [9.9147e-004]
```

```
    [ 0.7480]    [     0.9776]

Estimates based on MODEL based covariance matrix
  Columns 1 through 5

          ''          ''          ''          ''          ''
    'Variable'    'Beta'      'Std.Error'    'z value'    'p-value'
    [       1]    [ 0.6715]    [   0.0510]    [13.1793]    [      0]
    [       2]    [-0.0038]    [   0.0018]    [-2.0659]    [ 0.0388]
    [       3]    [ 0.8628]    [   0.0747]    [11.5564]    [      0]

  Columns 6 through 7

    '95% CI'                ''
    'low lim'    'up lim'
    [ 0.5717]    [      0.7714]
    [-0.0074]    [-1.9423e-004]
    [ 0.7165]    [      1.0091]
```

---

qls

Purpose Analyses correlated data via QLS.

Syntax

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

Description `[bhat,alpha,results] = qls(id,y,t,X);` analyses correlated data via QLS where `y` is a $N \times 1$ vector containing the repeated measures data, measured at corresponding times `t`, on subjects `id`, regressed on fixed covariates `X`. In order for a constant to be included in the model, the final column of `X` could be a column of ones. For all these variables, the number of rows should be equal to

$$N = \sum_{i=1}^{n} n_i$$

This model assumes the data follows a normal distribution with a Markov correlation structure. `[bhat,alpha,results] = qls(id,y,t,X,Family,CorrStruct);` can be used to change these assumptions. `Family` specifies the assumed distribution of the data.

It can take values:

`1` or `n` for a Normal distribution (default);
`2` or `b` for a Bernoulli distribution; or
`3` or `p` for a Poisson distribution.

`CorrStruct` specifies the assumed correlation structure. It can take values:

`1` or `ar1` for an AR(1) structure;
`2` or `markov` for a Markov structure (default);
`3` or `equi` for an Equicorrelated structure;
`4` or `tri` for a Tri-diagonal correlation structure;
`5` or `un` for an Unstructured correlation matrix; or
`6` or `ind` for a working Independent correlation structure.

Further, `varnames` can be used to specify the variables names to be used in the output. This is a cell row array which corresponds to the variables entered in `X`. Additionally, `maxit` specifies the maximum number of iterations (default is 100), and `tol` the tolerance used for determining convergence (default is 1e-5).

`[bhat,alpha,results]` contains the parameter estimates for the fixed effects `X` in `bhat`, and the estimated correlation parameters in `alpha`. `results` is a structured array with two elements, `robust` and `model` which contains the entire printed results / estimates under the robust and model based covariance matrices, respectively.

Examples To fit the regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

where `id` and `time` contains the i's and time points associated with each row. Under a normal distribution and Markov working correlation structure, the model would be fit with the commands

```
X = [X1,X2,ones(length(Y,1))];
varnames = {'X1','X2','constant'};
[bhat,alpha,results] = qls(id,Y,time,X,'n','markov',varnames);
```

The results from fitting this model to some data would be:

```
Normal distribution family assumed
Markov Correlation structure assumed
Initial estimate of beta = [0.62771   -0.002418     0.97107]

Stage 1 estimate of alpha = 0.88122
Stage 1 estimate of beta = [0.61824 -0.00027181     0.87907]

Stage 2 estimate of alpha = 0.96007
Stage 2 estimate of beta = [0.63961   0.0014781     0.78748]
```

```
QLS estimate of scale parameter = 0.60842

Estimates based on ROBUST covariance matrix
  Columns 1 through 5

            ''          ''             ''           ''           ''
    'Variable'     'Beta'      'Std.Error'    'z value'    'p-value'
    'X1'           [0.6396]    [   0.0667]    [ 9.5896]    [      0]
    'X2'           [0.0015]    [   0.0030]    [ 0.4916]    [ 0.6230]
    'Constant'     [0.7875]    [   0.0613]    [12.8423]    [      0]

  Columns 6 through 7

    '95% CI'           ''
    'low lim'    'up lim'
    [ 0.5089]    [0.7703]
    [-0.0044]    [0.0074]
    [ 0.6673]    [0.9077]

Estimates based on MODEL based covariance matrix
  Columns 1 through 5

            ''          ''             ''           ''           ''
    'Variable'     'Beta'      'Std.Error'    'z value'    'p-value'
    'X1'           [0.6396]    [   0.0457]    [13.9809]    [      0]
    'X1'           [0.0015]    [   0.0022]    [ 0.6773]    [ 0.4982]
    'Constant'     [0.7875]    [   0.0758]    [10.3936]    [      0]

  Columns 6 through 7

    '95% CI'           ''
    'low lim'    'up lim'
    [ 0.5499]    [0.7293]
    [-0.0028]    [0.0058]
    [ 0.6390]    [0.9360]
```

Bibliography

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

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.

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

Nunez-Anton, V. and Woodworth, G. (1994). Analysis of longitudinal data with unequally spaced observations and time- dependent correlated errors. *Biometrics*, **50**(2), 445–456.

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

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

Shults, J. and Morrow, A. (2002). Use of quasi-least squares to adjust for two levels of correlation. *Biometrics*, **58**, 521–30.

Shults, J., Mazurick, C., and Landis, J. (2006). Analysis of repeated bouts of measurements in the framework of generalize estimating equations. *Statistics in Medicine*. (in press).

Sun, W., Shults, J., and Leonard, M. (2006). Use of unbiased estimating equations to estimate correlation in generalized estimating equation analysis of longitudinal trials. Technical Report Working Paper 4, UPenn Biostatistics Working Papers.

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

Wang, Y. and Carey, V. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association*, **99**, 845–852.

Index

`counti`, 16, **17**

`fequi1`, 16, **17**
`fmark1`, 16, **17**
`fmark2`, 16, **17**
`ftri1`, 16, **18**
`ftri2`, 16, **18**
`ftri2a`, 16, **18**

`gee`, 8, 15, **19**
`GEEQBOX` toolbox, 1–7, 15
Generalized estimating equations (GEE), 2, 4

`qls`, 8, 15, 17–19, **21**
Quasi-least squares (QLS), 2, 5

Working Correlation Structure, 3
    AR(1), 3, 5, 6
    Equicorrelated, 3, 4, 6
    Independent, 4
    Markov, 4–6
    Tri-diagonal, 4–6
    Unstructured, 4, 5