University of North Carolina at Chapel Hill
The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2011 *Paper* 22

ORTH: R and SAS Software for Regression Models of Correlated Binary Data Based on Orthogonalized Residuals and Alternating Logistic Regressions

Kunthel By[*] Bahjat F. Qaqish[†] John S. Preisser[‡]
Jamie Perin[**] Richard C. Zink[††]

[*]University of North Carolina at Chapel Hill
[†]University of North Carolina, Chapel Hill, qaqish@bios.unc.edu
[‡]University of North Carolina at Chapel Hill, jpreisse@bios.unc.edu
[**]Johns Hopkins Bloomberg School of Public Health, jperin@jhsph.edu
[††]SAS Institute, Cary, NC, Richard.Zink@jmp.com

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/uncbiostat/art22

Copyright ©2011 by the authors.

ORTH: R and SAS Software for Regression Models of Correlated Binary Data Based on Orthogonalized Residuals and Alternating Logistic Regressions

Kunthel By, Bahjat F. Qaqish, John S. Preisser, Jamie Perin, and Richard C. Zink

Abstract

In this article, we describe a new software for modeling correlated binary data based on orthogonalized residuals (Zink and Qaqish, 2009), a recently developed estimating equations approach that includes, as a special case, alternating logistic regressions (Carey et al., 1993). The software is flexible with respect to fitting in that the user can choose estimating equations for the association model based on alternating logistic regressions or orthogonalized residuals, the latter choice providing a non-diagonal working covariance matrix for second moment parameters providing potentially greater efficiency. Regression diagnostics based on this method are also implemented in the software. The mathematical details of the procedure are briefly reviewed and the software is applied to medical data sets.

ORTH: R and SAS Software for Regression Models of Correlated Binary Data Based on Orthogonalized Residuals and Alternating Logistic Regressions

Kunthel By[1], Bahjat F. Qaqish[1], John S. Preisser[1], Jamie Perin[2], and Richard C. Zink[3]

[1]University of North Carolina at Chapel Hill

[2]Johns Hopkins Bloomberg School of Public Health

[3]SAS Institute, Cary, NC

Abstract

In this article, we describe a new software for modeling correlated binary data based on orthogonalized residuals (Zink and Qaqish, 2009), a recently developed estimating equations approach that includes, as a special case, alternating logistic regressions (Carey et al., 1993). The software is flexible with respect to fitting in that the user can choose estimating equations for the association model based on alternating logistic regressions or orthogonalized residuals, the latter choice providing a non-diagonal working covariance matrix for second moment parameters providing potentially greater efficiency. Regression diagnostics based on this method are also implemented in the software. The mathematical details of the procedure are briefly reviewed and the software is applied to medical data sets.

Keywords: estimating equations, logistic regression, regression diagnostics, permutation invariance, association models

1 Introduction

Statistical methods for the regression analysis of correlated binary data have been around for three or four decades (Pendergast et al., 1996). However, it was only with the introduction of the generalized estimating equations (GEE) approach of Liang and Zeger (1986) that methodological breakthroughs coupled with advances in computing provided a general approach surpassing the restricted capabilities of earlier methods, particularly weighted least squares (Koch et al., 1977). For the situation where the association is not of interest, first-order GEE, also known as GEE1, provides a computationally fast approach for fitting marginal mean models under an assumed correlation structure. It is well known that if the assumed correlation structure is incorrect, parameter estimates still maintain consistency although some efficiency is lost. Extensions on the work of Liang and Zeger (1986) have allowed the active modeling of the association structure (when it is of interest); see for example Zhao and Prentice (1990); Liang et al. (1992). The computational effort expended in these methods restricted their usage to data structures characterized by small cluster sizes. Computational gains may be achieved by imposing extra conditions, but must be paid for by some degree of loss in efficiency (Prentice, 1988; Lipsitz et al.,

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

1991). Alternating logistic regressions, (Carey et al., 1993) addresses some of these concerns while in turn introducing some of its own complications (Kuk, 2004). Some of these complications were addressed by Zink and Qaqish (2009) through the use of orthogonalized residuals (ORTH). It is on this work that our software is based. We have written an R package, aptly named `orth`, and a SAS macro (also named `orth`) based on Zink's estimation algorithm, a variant of iterative reweighted least squares, and have incorporated diagnostics based on the work of Preisser and Qaqish (1996) with extensions thereof to diagnostics for marginal association models (Preisser et al., 2011). For expositional convenience, when we say ORTH, we mean the method based on orthogonalized residuals, and when we say `orth`, we mean the software package based on ORTH.

To get an appreciation of what ORTH does, it is necessary to look briefly at the background. It goes without saying that to do it justice, a modicum of math is necessary. We lay this out in the next section. Before that however, we need to introduce some notation. Let n_i denote the size of the i-th cluster and let

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} & Y_{i2} & \cdots & Y_{in_i} \end{bmatrix}^\top, \quad i = 1, \ldots, K$$

denote the vector of binary responses for cluster i where K is the number of clusters. The symbol $\boldsymbol{\mu}_i$ shall mean $\boldsymbol{\mu}_i = E[\mathbf{Y}_i]$ and the symbol $\boldsymbol{\psi}_i$ shall mean

$$\boldsymbol{\psi}_i = [\psi_{i12},\ \psi_{i13},\ \ldots,\ \psi_{i(n_i-1)n_i}]^\top,$$

where ψ_{ijk} denotes the pairwise odds ratio between the j-th and k-th responses in cluster i (Carey et al., 1993). The symbol $\boldsymbol{\beta}$, of dimension $p \times 1$, is the parameter vector associated with the marginal mean model. The symbol $\boldsymbol{\alpha}$, of dimension $q \times 1$, is the parameter vector associated with the marginal association model. For link functions g_1 and g_2, our marginal models are cast as

$$\begin{aligned} \text{Mean Model}: &\quad g_1(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta} \\ \text{Association Model}: &\quad g_2(\boldsymbol{\psi}_i) = \mathbf{Z}_i\boldsymbol{\alpha} \end{aligned}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices of dimensions $n_i \times p$ and $m_i \times q$ respectively with $m_i = \binom{n_i}{2}$.

2 Methods

A good place to start is with the second-order GEE of Liang et al. (1992), denoted hereafter as GEE2. Define $W_{ijk} = Y_{ij}Y_{ik}$ and

$$\mathbf{W}_i = \begin{bmatrix} W_{i12} & W_{i13} & \cdots & W_{i(n_i-1)n_i} \end{bmatrix}^\top.$$

Let $\boldsymbol{\delta}_i = E[\mathbf{W}_i]$, $\mathbf{Y}_i^* = (\mathbf{Y}_i^\top,\ \mathbf{W}_i^\top)^\top$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$. The second order generalized estimating equation is

$$U_{\boldsymbol{\theta},GEE2} = \begin{pmatrix} U_{\beta,GEE2} \\ U_{\alpha,GEE2} \end{pmatrix} = \sum_{i=1}^{K} \begin{bmatrix} \mathbf{D}_i & \mathbf{0} \\ \mathbf{A}_i & \mathbf{C}_i \end{bmatrix}^\top \boldsymbol{\Sigma}_{i*}^{-1} \begin{bmatrix} \mathbf{Y}_i - \boldsymbol{\mu}_i \\ \mathbf{W}_i - \boldsymbol{\delta}_i \end{bmatrix} = \mathbf{0} \tag{1}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

where $\mathbf{\Sigma}_{i*} = \text{cov}(\mathbf{Y}_i^*)$ and

$$\mathbf{A}_i = \frac{\partial \boldsymbol{\delta}_i}{\partial \boldsymbol{\beta}}, \quad \mathbf{C}_i = \frac{\partial \boldsymbol{\delta}_i}{\partial \boldsymbol{\alpha}}, \quad \mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \ .$$

Throughout $U_{\boldsymbol{\beta}}$ denotes the estimating function for the marginal mean parameters and $U_{\boldsymbol{\alpha}}$ denotes the estimating function for the marginal association parameters. For example, $U_{\beta,GEE2}$ and $U_{\alpha,GEE2}$ correspond to the GEE2 estimating functions for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ respectively. Note that $\mathbf{\Sigma}_{i*}$ involves 3-rd and 4-th order moments. Even with restrictions on these higher order moments, the amount of computation can still be prohibitive for large cluster sizes. Furthermore, misspecification of $\mathbf{\Sigma}_{i*}$ may lead to biased estimates of $\boldsymbol{\beta}$ - even if the marginal mean model $g_1(\boldsymbol{\mu}_i)$ is correctly specified. The reason for this is that $U_{\beta,GEE2}$ is a weighted sum of $\mathbf{Y}_i - \boldsymbol{\mu}_i$ and $\mathbf{W}_i - \boldsymbol{\delta}_i$ where the weights depend upon correctly specified components of $\mathbf{\Sigma}_i^*$. Lipsitz et al. (1991) proposed a procedure to estimate $\boldsymbol{\theta}$ that provides unbiased estimates of $\boldsymbol{\beta}$, even if the model involving $\boldsymbol{\alpha}$ is misspecified. By setting $\mathbf{A}_i = \mathbf{0}$ and $\text{cov}(\mathbf{Y}_i, \mathbf{W}_i) = \mathbf{0}$ in expression (1), the resulting estimating equation for $\boldsymbol{\beta}$ is

$$U_{\beta,GEE1} = \sum_{i=1}^{K} \mathbf{D}_i^\top V_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{2}$$

where

$$\begin{aligned} V_i &= \text{diag}\{\sqrt{\sigma_{ijj}}\}\mathbf{R}_{iYY}(\boldsymbol{\alpha})\,\text{diag}\{\sqrt{\sigma_{ijj}}\}, \\ \sigma_{ijk} &= \text{cov}(Y_{ij}, Y_{ik}), \end{aligned}$$

and $\mathbf{R}_{iYY}(\boldsymbol{\alpha})$ is a working correlation matrix for $\mathbf{Y}$. Additionally specifying a working diagonal covariance matrix for $\text{cov}(\mathbf{W}_i)$ gives

$$U_{\boldsymbol{\alpha}} = \sum_{i=1}^{K} \mathbf{C}_i^\top \Big[\text{diag}(\text{var}[\mathbf{W}_i])\Big]^{-1} (\mathbf{W}_i - \boldsymbol{\delta}_i) \ . \tag{3}$$

Capitalizing on the diagonal structure of the working covariance matrix, (3) permits fast computations for large cluster sizes but sacrifices some efficiency. Prentice (1988) had earlier proposed a similar method to fit linear models to correlations among binary data.

Alternating logistic regressions (ALR) (Carey et al., 1993) takes a somewhat different approach. While keeping the estimating equation of GEE1, their estimating equation for $\boldsymbol{\alpha}$ is based on *conditional residuals* which are defined by

$$Y_{ij} - \xi_{ijk}$$

where, for $j > k$,

$$\xi_{ijk} = E[Y_{ij}|Y_{ik}] = \mu_{ij} + \frac{\sigma_{ijk}}{\sigma_{ikk}}(Y_{ik} - \mu_{ik}) \ .$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Based on $\boldsymbol{\xi}_i = [\xi_{i12},\ \xi_{i13},\ \ldots,\ \xi_{i(n_i-1)n_i},\ ]^\top$, they defined their estimating equation for $\boldsymbol{\alpha}$ as

$$U_{\alpha,ALR} = \sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{\xi}_i}{\partial \boldsymbol{\alpha}}\right]^\top \mathbf{S}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\xi}_i) = \mathbf{0} \tag{4}$$

where $\mathbf{S}_i = \text{diag}[\boldsymbol{\xi}_i(1-\boldsymbol{\xi}_i)]$, and the solution $\widehat{\boldsymbol{\alpha}}$ is invariant to permutations of the elements in $\mathbf{Y}_i$.

A recurrent theme is played out several times over in the above presentation, namely, that the expression for $\mathbf{W}_i$ is a function of $\mathbf{Y}_i$ and hence necessarily correlates with $\mathbf{Y}_i$. In general, the actual covariance matrix of $\mathbf{W}_i$ is not diagonal. These are the culprits responsible for efficiency loss in the methods of Prentice and Lipsitz et al. The contribution of orthogonalized residuals addresses this failing but does so in a way that avoids some of the complications of (4). Recall that the matrix $\mathbf{S}_i$ - because it is a function of $\mathbf{Y}_i$ - is stochastic and hence is not a covariance matrix in the usual sense. As such, it is not clear how one would go about introducing a non-diagonal $\mathbf{S}_i$ if the goal is efficiency improvement. A further complication caused by a stochastic $\mathbf{S}_i$ is that standard estimating equation theory cannot be applied to study the properties of $U_{\boldsymbol{\alpha}}$ (Zink and Qaqish, 2009). Lastly, the robust covariance estimator under alternating logistic regressions is not invariant to permutations of $\mathbf{Y}_i$. By casting the problem in terms of orthogonalized residuals, ORTH resolves each of these complications.

2.1 Orthogonalized Residuals

The idea that led to the development of ORTH was to minimize the correlation between $\mathbf{Y}_i$ and the residual of the association estimating equation and to approximate the covariance of this residual with a non-diagonal matrix. This is accomplished as follows. For the i-th cluster, define elements of the $m_i \times 1$ vector of orthogonalized residuals $\mathbf{Q}_i = [Q_{i12},\ Q_{i13},\ \ldots,\ Q_{i(n_i-1)n_i}]^\top$ by

$$Q_{ijk} = W_{ijk} - \left[\mu_{ijk} + b_{ijk:j}(Y_{ij} - \mu_{ij}) + b_{ijk:k}(Y_{ik} - \mu_{ik})\right] \tag{5}$$

where

$$\begin{aligned} b_{ijk:j} &= \mu_{ijk}(1-\mu_{ik})(\mu_{ik} - \mu_{ijk})/d_{ijk}, \\ b_{ijk:k} &= \mu_{ijk}(1-\mu_{ij})(\mu_{ij} - \mu_{ijk})/d_{ijk}, \\ d_{ijk} &= \sigma_{ijj}\sigma_{ikk} - \sigma_{ijk}^2\ . \end{aligned}$$

Next, $\mathbf{R}_{iQQ} = \mathsf{CORR}(\mathbf{Q}_i)$ is approximated by an exchangeable working correlation matrix

$$\mathbf{R}^*_{iQQ} := \lambda \mathbf{1}\mathbf{1}^\top + (1-\lambda) I_{m_i} \tag{6}$$

where I_r is an $r \times r$ identity matrix, $\mathbf{1}$ is an $m_i \times 1$ vector of ones and λ is the exchangeable correlation parameter to be estimated. Letting v_{ijk} denote the variance of Q_{ijk}, we approximate $\text{cov}(\mathbf{Q}_i)$ by

$$\mathbf{P}_i = \text{diag}(\sqrt{\mathbf{v}_i})\mathbf{R}^*_{iQQ}(\lambda)\,\text{diag}(\sqrt{\mathbf{v}_i}) \tag{7}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

where $\mathbf{v}_i = \{v_{ijk}\}$. Putting all of these together, ORTH's estimating equation for $\boldsymbol{\alpha}$ is

$$U_{\alpha,ORTH} = \sum_{i=1}^{K} \mathbf{C}_i^\top \mathbf{P}_i^{-1} \mathbf{Q}_i = \mathbf{0} \tag{8}$$

where

$$\mathbf{C}_i^\top = E\left[-\frac{\partial \mathbf{Q}_i^\top}{\partial \boldsymbol{\alpha}}\right] ,$$

while that for $\boldsymbol{\beta}$ is the same as (2). The correlation parameter λ is estimated by a moment estimator:

$$\hat{\lambda}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{K}\left[\left(\sum_{j<k}\frac{Q_{ijk}}{\sqrt{v_{ijk}}}\right)^2 - \sum_{j<k}\frac{Q_{ijk}^2}{v_{ijk}}\right]; \quad M = \sum_{i=1}^{K} m_i(m_i-1) .$$

Assuming that the data is missing completely at random (MCAR), it may be shown that

$$\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_{p+q}\big(\mathbf{0}, K\mathbf{L}^{-1}\boldsymbol{\Lambda}(\mathbf{L}^{-1})^\top\big)$$

where $\mathbf{L}$ and $\boldsymbol{\Lambda}$ have manageable block structures (Zink and Qaqish, 2009). By construction, $U_{\alpha,ORTH}$ and its associated robust covariance estimator for $\boldsymbol{\alpha}$, $K\mathbf{L}^{-1}\boldsymbol{\Lambda}(\mathbf{L}^{-1})^\top$, are invariant to reordering. By not estimating λ; i.e. setting it to 0, it may be shown that $U_{\alpha,ORTH} = U_{\alpha,ALR}$. Thus, $U_{\alpha,ALR}$ is a special case of $U_{\alpha,ORTH}$. This holds for all link functions g_1 and g_2. This essentially means that alternating logistic regressions can be recast in a way consistent with standard estimating equation theory. Standard approaches based on this theory may be gainfully employed if we want to study properties of alternating logistic regressions. If it so happens that $\mathbf{R}^*_{iQQ}$ is close to the true correlation matrix of $\mathbf{Q}_i$, the incorporation of λ leads to further efficiency gains.

To understand why ORTH (and ALR) improve efficiency for $\boldsymbol{\alpha}$-estimation compared to (3), let $\mathbf{R}_{iYW} = \mathsf{CORR}(\mathbf{Y}_i, \mathbf{W}_i)$, $\mathbf{R}_{iWW} = \mathsf{CORR}(\mathbf{W}_i)$, and $\mathbf{R}_{iYQ} = \mathsf{CORR}(\mathbf{Y}_i, \mathbf{Q}_i)$. It may be shown that $\mathsf{CORR}(Y_{ij}, Q_{ijk}) = \mathsf{CORR}(Y_{ik}, Q_{ijk}) = 0$ (Zink and Qaqish, 2009). This means that $(n_i - 1)$ zeros are introduced into each row of $\mathbf{R}_{iYQ}$. It turns out that the orthogonalized residuals tend to shrink the non-zero entries of $\mathbf{R}_{iYQ}$ relative to $\mathbf{R}_{iYW}$ and the off-diagonal entries of $\mathbf{R}_{iQQ}$ relative to $\mathbf{R}_{iWW}$. Thus the procedure based upon (8) is in a sense closer to GEE2 than the procedure based upon (3), leading to more efficient estimation.

2.2 Diagnostics

Deletion diagnostics for these models are also available at both the cluster level and the observation level. Formulae for these diagnostics are based on one-step approximations given by Preisser and Qaqish (1996) and extensions made thereafter (Preisser et al., 2011).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Cluster-Level Diagnostics The object of cluster-level diagnostics is to locate clusters which influence (in some sense) either the values of the estimates or the predicted values. Exact techniques for assessing this influence typically entails removing the cluster from the data and refitting the model of interest. If the parameter estimates change noticeably we regard that cluster as influential. In the correlated binary data setting where we may encounter large cluster sizes or many clusters, the combination of both a marginal mean model and a marginal association model often renders the exact approach computationally impractical (Preisser and Perin, 2007). In the ensuing discussion, we give fast computational formulae based on *one-step* approximations implemented in our software.

We start with deletion diagnostics for $\boldsymbol{\beta}$ - the parameter vector for the marginal mean model. Let $\widehat{\boldsymbol{\beta}}_{(i)}$ denote the parameter estimate associated with a design matrix $\mathbf{X}_{(i)}$ in which the rows of $\mathbf{X}$ associated with cluster i are removed. Then the influence on $\widehat{\boldsymbol{\beta}}$ - denoted DFBETA - of cluster i is defined by $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}$. Analogously, the influence on $\widehat{\boldsymbol{\alpha}}$ - to be denoted DFALPHA - of cluster i is defined by $\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}_{(i)}$. Letting $DFBETA_{C_i}$ denote cluster i's influence on $\widehat{\boldsymbol{\beta}}$, the *one-step approximation* formula for $DFBETA_{C_i}$ is defined by

$$\widehat{DFBETA}_{C_i} \approx \left(\mathbf{D}^\top \mathbf{V}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}_i^\top \mathbf{V}_i^{-1} (I_{n_i} - \mathbf{H}_{i1})^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \tag{9}$$

where

$$\mathbf{H}_{i1} = \mathbf{D}_i \left(\mathbf{D}^\top \mathbf{V}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}_i^\top \mathbf{V}_i^{-1},$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1^\top & \cdots & \mathbf{D}_k^\top \end{bmatrix}^\top,$$

$$\mathbf{V} = \text{blockdiag}(\mathbf{V}_1 \quad \cdots \quad \mathbf{V}_k),$$

and $\mathbf{R}_i(\boldsymbol{\alpha})$ is a working correlation matrix. We call $\mathbf{H}_{i1}$ the leverage matrix for $\boldsymbol{\beta}$ associated with cluster i. Similarly, let $DFALPHA_{C_i}$ denote cluster i's influence on $\boldsymbol{\alpha}$. The one-step approximation formula for $DFALPHA_{C_i}$ is defined by

$$\widehat{DFALPHA}_{C_i} \approx \left(\mathbf{C}^\top \mathbf{P}^{-1} \mathbf{C}\right)^{-1} \mathbf{C}_i^\top \mathbf{P}_i^{-1} (I_{m_i} - \mathbf{H}_{i2}) \mathbf{Q}_i \tag{10}$$

where $\mathbf{Q}_i$ is defined in (5), $\mathbf{P}_i$ is defined in (7), I_{m_i} is an $m_i \times m_i$ identity matrix and

$$\mathbf{C}_i = E\left[-\frac{\partial \mathbf{Q}_i^\top}{\partial \boldsymbol{\alpha}}\right],$$

$$\mathbf{H}_{i2} = \mathbf{C}_i \left(\mathbf{C}^\top \mathbf{P}^{-1} \mathbf{C}\right)^{-1} \mathbf{C}_i^\top \mathbf{P}_i^{-1},$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1^\top & \cdots & \mathbf{C}_k^\top \end{bmatrix}^\top,$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

and

$$\mathbf{P} = \text{blockdiag}(\mathbf{P}_1 \quad \cdots \quad \mathbf{P}_k) \ .$$

We defined cluster-level Cook's distance for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\theta}$ as follows:

$$D_{\alpha,C_i} = \frac{[\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}_{(i)}]^\top \, [\text{var}(\widehat{\boldsymbol{\alpha}})]^{-1} \, [\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}_{(i)}]}{q}, \tag{11}$$

$$D_{\beta,C_i} = \frac{[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}]^\top \left[\text{var}(\widehat{\boldsymbol{\beta}})\right]^{-1} [\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}]}{p}, \tag{12}$$

$$D_{\theta,C_i} = \frac{[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}]^\top \left[\text{var}(\hat{\boldsymbol{\theta}})\right]^{-1} [\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}]}{p+q} \ . \tag{13}$$

The one-step approximations of these formulae are obtained by substituting $\widehat{DFBETA}_{C_i}$ and $\widehat{DFALPHA}_{C_i}$ as given immediately above. Robust or model-based covariance estimates of the parameter estimates may be used in (11) to (13).

Analogous to the cluster-level deletion diagnostics presented above, observation-level deletion diagnostics are available. However, there is limited utility of observation-level deletion diagnostics in the context of correlated binary data regression models. As such, we do not present their formulae here. The interested reader may consult Preisser and Qaqish (1996), Hammill and Preisser (2006), Preisser and Perin (2007), and Preisser et al. (2011) for detailed expositions.

3 Software Details

We have written an R (R Development Core Team, 2008) package and a SAS (SAS Institute Inc., 2009) macro that implement both parameter estimation for the marginal mean and marginal association models as well as the deletion diagnostics described in section 2.2. For the R package, how we specify model structures and data are based on standard R notations. Users of R who are familiar with the modeling routines `lm()` and `glm()` for linear and generalized linear models respectively should have little trouble using `orth()`. The full details of how to use the R implementation of `orth` may be seen in By et al. (2008). The SAS implementation is slightly more cumbersome to describe; its usage is best left to an example. One of the main differences between the R and the SAS implementation is how we compute diagnostics. Since R is a functional language, functions are provided to compute/extract diagnostics. This task is separate from parameter estimation. In SAS, macro options must be turned on or off. If diagnostics are desired, then this task is performed along with estimation.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

3.1 The Correlation Parameter For The Residuals

As mentioned earlier, alternating logistic regressions is a special case of orthogonalized residuals in which the correlation parameter for the residuals λ is set to 0. Thus, users are given the option of obtaining estimates based on alternating logistic regressions by not estimating λ and setting it 0. In addition, users are able to fix λ at a given value rather than estimating it. The third and final option is to estimate λ as described in section 2.1.

3.2 Data Set Construction

3.2.1 R Implementation

R estimation routines require both a model formula and a data set. `orth` is no different. Since we are modeling a mean model and an association model, two model formulae must be provided. The model formula for the mean must be a two-sided formula. For example, `y ~ x1 + x2 + factor(A)` is a two-sided formula for the marginal mean model. For the association model, the model formula must be one-sided; i.e., there is no left side. For example, `~ z1 + z2 + factor(B)` is a one-sided formula for the marginal association model. Two different data frames - one for the mean model and one for the association model - must be provided separately. The structure of the data frame for the mean model must be in case-record format. This means that covariates for each observation within a cluster must be stacked on top of each other. If cluster weights are used, then the weights must exist as a column in the data frame for the mean model. The most difficult part in using `orth` is the construction of the design matrix for the association model - the $\mathbf{Z}$ matrix. As far as we know, there is no algorithm for automatically generating a general $\mathbf{Z}$ matrix. Thus, this task is left to the user. Examples are given in By et al. (2008).

3.2.2 SAS Implementation

The SAS implementation of ORTH, in the form of a SAS macro available at http://www.bios.unc.edu/∼qaqish/software.htm, is very similar to the R implementation. Like R's, SAS data sets for the $\mathbf{X}$ and $\mathbf{Z}$ matrices must be created separately. A separate SAS data set of K observations for cluster weights must also be created; if cluster weights are not used, then all should be assigned a value of one to the weight variable in this data set. Both the name of the weight data set as well as the name of the weight column must be passed in as macro arguments.

Whether the R or SAS implementation is used, missing data is not permitted. The user must remove all rows with missing data.

3.3 Limitations

There are a couple of limitations in the current version of our package. First, although the theory supports general link functions g_1 and g_2, `orth` only supports

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

the logit link for the mean model and the log odds ratio link for the association model. The other limitation is a computational issue. Users of R are familiar with the fact that R has some trouble handling large data sets and for some reason, often performs very slowly. This problem applies to `orth` in the setting of large cluster sizes. In cluster-randomized trials and survey samples, it is not uncommon to encounter clusters with very large cluster sizes. In the construction of the $\mathbf{Z}$ matrix, if we let n_i denote the cluster size for the i-th cluster, then the i-th cluster contributes m_i rows to the $\mathbf{Z}$ matrix. For example, a cluster with 60 subjects contributes 1770 rows to the $\mathbf{Z}$ matrix. A study with 100 clusters of size 60 contains 177000 rows in the $\mathbf{Z}$ matrix. A noticeable slowdown will occur. An illustration of this will be seen in the examples section. With up to 20 measurements per cluster, `orth` performs in a reasonable amount of time. While the SAS implementation is subject to these very same limitations, its performance (as measured by speed) is much better than R's in the large cluster size setting.

4 Examples

In this section, we apply `orth` to two data sets. Our first example is a medical practice data set analyzed using GEE1 by Preisser and Qaqish (1996). Our second example is based on data analyzed by Fitzmaurice and Lipsitz (1995, Table 2) and later re-analyzed by Ekholm et al. (2000) using a different approach.

4.1 Medical Practice Data

From 1990 to 1991, charts of 3889 patients were randomly chosen from 57 medical practices. Each practice may be thought of as a cluster. Practices can have multiple physicians or subclusters so that patients are nested within physicians that, in turn, are nested within practices. The number of patients in each practice ranges from 19 to 197. These may be thought of as large cluster sizes. Note that the cluster with 197 patients contributes 19306 rows to the $\mathbf{Z}$ matrix. Let Y_{ij} be an indicator for the event that the j-th patient made at least one health maintenance visit to a physician in the i-th practice in the years 1990 to 1991. For the marginal mean model, we are interested in covariates that influence the probability of a health maintenance visit, $\mu_{ij} = \Pr(Y_{ij} = 1)$. The model formula for the linear predictor is

$$\begin{aligned} &1 + NBRMDS + M3 + SPECLTY + MDAGE \\ &+ MDSEX + MDFLU + PATAGE \\ &+ BLACKPAT + MALEPAT + NOINSUR \end{aligned} \tag{14}$$

where $NBRMDS$ is the number of doctors in the practice, $M3$ is the number of patients over 50 years old seen per day (centered and scaled), $SPECLTY$ is the doctor's specialty (0 if family or general practitioner and 1 for internist), $MDAGE$ is the doctor's age (centered and scaled), $MDSEX$ (0 for male and

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

1 for female) is the doctor's gender, $MDFLU$ is the doctor's flu vaccination status (1 if he/she gives flu shot and 0 if not), $PATAGE$ is the patient's age (centered and scaled), $BLACKPAT$ indicates whether a patient is black (1 if black and 0 if not), $MALEPAT$ indicates whether a patient is male (1 if male and 0 if female), and $NOINSUR$ indicates whether a patient is not insured (1 if not insured 0 if insured).

The model formula for the marginal association model is

$$1 + SAMEMD + CLSSIZE \tag{15}$$

where $SAMEMD$ indicates whether two patients j and k have the same physician and $CLSSIZE$ denotes the size of the cluster (centered and scaled). These two variables are thought to influence the pairwise odds ratio of a health maintenance visit for patients j and k in cluster i.

Parameter estimates for models (14) and (15) based on alternating logistic regressions defined with orthogonalized residuals ($\lambda = 0$) are presented in the first three columns of Table 1. Doctors' age as well as the patients' age, ethnicity, gender, and insurance status have a statistically significant influence in the probability of at least one health maintenance visit. From the association parameters, we see that the odds ratio of a health maintenance visit between two

Table 1: *Parameter estimates for mean and association models based on the medical practice data. The data come from a North Carolina Early Cancer Detection Program at the Lineberger Cancer Center. The symbol * denotes $p < 0.05$*

	ALR[1]			ORTH ($\hat{\lambda} = 0.0134$)		
Effects	Est.	S. E.	χ^2	Est.	S. E.	χ^2
Mean Parms.						
(Intercept)	−0.1061	0.1726	0.38	−0.1652	0.1675	0.97
NBRMDS	−0.0344	0.0388	0.78	−0.0098	0.0458	0.05
M3	0.2488	0.1701	2.14	0.2065	0.1664	1.54
SPECLTY	−0.0781	0.2492	0.10	−0.0226	0.2378	0.01
MDAGE	−0.2642	0.0641	16.96*	−0.2440	0.0665	13.47*
MDSEX	0.4235	0.2625	2.60	0.4540	0.2561	3.14
MDFLU	−0.0721	0.0988	0.53	−0.0869	0.0975	0.79
PATAGE	−0.0966	0.0339	8.13*	−0.0967	0.0336	8.27*
BLACKPAT	−0.3948	0.1226	10.36*	−0.3910	0.1214	10.38*
MALEPAT	−0.4110	0.0653	39.56*	−0.4061	0.0653	38.65*
NOINSUR	−0.4158	0.1190	12.21*	−0.4182	0.1193	12.29*
Association Parms.						
(Intercept)	0.5384	0.1727	9.71*	0.3791	0.1497	6.42*
SAMEMD	0.2898	0.1125	6.64*	0.3531	0.0938	14.18*
CLSSIZE	−0.1788	0.0620	8.33*	−0.0477	0.2864	0.03

1 Defined with orthogonalized residuals ($\lambda = 0$)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

patients with the same doctor is 1.34 times more than the odds ratio between two patients with different doctors.

The last three columns of Table 1 presents estimates based on ORTH (λ estimated). The estimates for the mean parameters are similar to those obtained under alternating logistic regressions. However, the association parameters are noticeably different from alternating logistic regressions. The standard errors for the association parameters have changed as well. Under alternating logistic regressions, `CLSSIZE` is significant but under ORTH, it is not.

Under alternating logistic regressions, we considered Cook's distance for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using the robust covariance. Based on the one-step approximation formulae, the top two clusters with the largest Cook's distance for $\boldsymbol{\alpha}$ are clusters 34 and 52 with respective values 0.1433 and 0.1465. Cluster 5 has the largest Cook's distance for $\boldsymbol{\beta}$ with value 0.1247. It turns out that the 5-th cluster also has the largest overall Cook's distance with value 0.0997. Figure 1 presents a two dimensional plot of $DFALPHA_C$ for both the intercept and `sameMD`. The 34-th cluster seems to have the largest influence on both the intercept parameter and the parameter for `sameMD`. The 15-th cluster has the largest effect on the intercept parameter but not on that for `sameMD`.

4.2 Arthritis Clinical Trial

The data for our next example is from a clinical trial for the effects of auranofin on arthritic symptoms (self-assessed) over time. Investigators were chiefly interested in the effects of treatment (auranofin) on the probability of a good self-assessment (the response). Subjects were measured at baseline (week 0), week 1, week 5, week 9, and week 13. However, subjects are only randomized to treatment after week 1; no treatment was given prior to this. Missing data is a prominent feature of this data set. We will not concern ourselves with issues related to missing data. Rather, we assume missing completely at random.

For their marginal mean model, Fitzmaurice and Lipsitz (1995) used the model formula

$$1 + TIME + GENDER + AGE + TREATMENT$$

where AGE denotes the subject's baseline age, $TREATMENT$ is an indicator for auranofin (1 if subject is assigned to auranofin and 0 if the subject is assigned to placebo), and $TIME$ denotes a linear trend for measurement occasion (at 0, 1, 5, 9, and 13 weeks). For their association model, they assumed that the pairwise odds ratio of a good assessment follows an autoregressive structure:

$$\psi_{ijk} = \alpha^{1/|k-j|} \ .$$

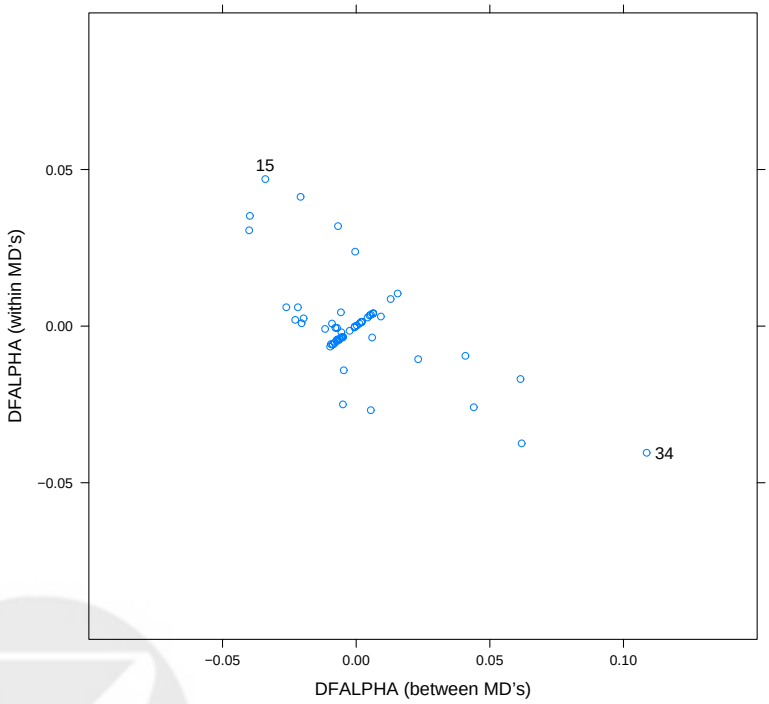
This leads to an association model of the form

$$\log \psi_{ijk} = \frac{1}{|k-j|} \log \alpha \ ,$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Figure 1: *Two-dimensional plot of* $DFALPHA_C$ *for the Intercept term (between MD's) and the sameMD term*



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

where $|k-j|$ denotes the number of weeks that elapsed between times k and j. Using our notation, we may write this as

$$\log \psi_{ijk} = \tilde{\alpha} z_{ijk},$$

where $z_{ijk} = 1/|k-j|$ and $\tilde{\alpha}$ is $\log \alpha$. The appendix provides both R and SAS code for applying alternating logistic regressions. Parameter estimates based

Table 2: *Estimates under ALR as a special case of orth ($\lambda = 0$)*

Parameter	Estimate	Std Error	χ^2	p value
Mean Parameters				
Intercept	1.0934	1.3781	0.63	0.4275
TIME	−0.0272	0.0296	0.84	0.3594
GENDER (male)	0.5956	0.4785	1.55	0.2132
AGE	−0.0154	0.0247	0.39	0.5337
TREATMENT (auranofin)	1.4572	0.4509	10.44	0.0012
Association Parameters				
Distance (z_{ijk})	3.6841	1.0000	13.57	0.0002

on alternating logistic regressions using orthogonalized residuals are presented in Table 2. The marginal mean parameter estimates and standard errors are exactly the same between alternating logistic regressions as a special case of ORTH and the alternating logistic regressions based on conditional residuals (not shown) as implemented in SAS (SAS Institute Inc., 2009). For the association parameter, ALR estimates using ORTH are the same as ALR estimates using conditional residuals but their standard errors are different. From the methods section, the reader may recall that under ORTH, the estimate of the variance of the association parameters is invariant to re-ordering of the data whereas under alternating logistic regressions, as devised by Carey et al. (1993), it is not. What SAS does is to first compute the sandwich estimator under the original data structure after which it reverses the order of the data and recomputes the sandwich estimator which creates an "invariance" to ordering, though it is an approximation. What we see from SAS's `PROC GENMOD` is the average of those two numbers.

5 Concluding Remarks

We have presented a software for analyzing correlated binary data based on orthogonalized residuals. This method is in its infancy and further studies are needed. But at the very least, our software permits estimation of alternating logistic regressions under a fundamentally different approach than that practiced by SAS's `PROC GENMOD`. Choosing alternating logistic regressions based on orthogonalized residuals versus that based on conditional residuals (Carey et al.,

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

1993) has implications on the computed standard errors. We know that the sandwich estimates for the standard errors given by our software are invariant to permutations. Standard errors based on SAS's implementation of alternating logistic regressions are an average of two runs of the estimation algorithm creating "invariance" to the dataset ordering –though still an approximation. It is not clear to us that this is the correct thing to do. Furthermore, the software allows us to estimate the correlation of the residuals by an exchangeable correlation parameter. Whether this improves upon the standard errors remains unanswered. If the true correlation of the residuals indeed follows the exchangeable structure, then estimates under ORTH are more efficient than ALR. If the exchangeable structure is incorrect, then it is not clear whether our estimates are worse or better than ALR. These issues should be in the minds of the user at all times.

References

By, K., Qaqish, B.F., and Preisser, J.S. *orth: Multivariate logistic regression using orthogonalized residuals*, 2008. URL `http://cran.r-project.org`. R package version 1.5.

Carey, V., Zeger, S.L., and Diggle, P. Modelling multivariate binary data with alternating logistic regression. *Biometrika*, 80:517–526, 1993.

Ekholm, A., McDonald, J.W., and Smith, P.W. Association models for a multivariate binary response. *Biometrics*, 56:712 – 718, 2000.

Fitzmaurice, G.M. and Lipsitz, S.R. A model for binary time series data with serial odds ratio patterns. *Applied Statistics*, 44:51 – 61, 1995.

Hammill, B. G. and Preisser, J. S. A SAS/IML software program for GEE regression diagnostics. *Computational Statistics And Data Analysis*, 51:1197 – 1212, 2006.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33:133 – 158, 1977.

Kuk, A.Y.C. Permutation invariance of alternating logistic regressions for multivariate binary data. *Biometrika*, 91:758 – 761, 2004.

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

Liang, K.Y., Zeger, S.L., and Qaqish, B.F. Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, B*, 54:3–40, 1992.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, 78:153–160, 1991.

CORRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

Pendergast, J. F., Gange, J. S., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. A survey of methods for analyzing clustered binary response data. *International Statistical Review*, 64:89 – 118, 1996.

Preisser, J. S. and Perin, J. Deletion diagnostics for marginal mean and correlation model parameters in estimating equations. *Statistics and Computing*, 17:381 – 393, 2007.

Preisser, J. S., By, K., Perin, J., and Qaqish, B. F. Deletion diagnostics for alternating logistic regressions. The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series Working Paper 21, 2011. URL `http://biostats.bepress.com/uncbiostat/papers/art21`.

Preisser, J.S. and Qaqish, B.F. Deletion diagnostics for generalized estimating equations. *Biometrika*, 83:551–562, 1996.

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

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

SAS Institute Inc. *SAS/STAT® Software: Version 9.2*. SAS Institute Inc., Carey, NC, 2009.

Zhao, L. P. and Prentice, R. L. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77:642 – 648, 1990.

Zink, R.C. and Qaqish, B.F. Correlated binary regression using orthogonalized residuals. COBRA Preprint Series 51, 2009. URL `http://biostats.bepress.com/cobra/ps/art51`.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

A Analysis Of Arthritis Clinical Trial Data

A.1 R Code

The code to analyze the arthritis clinical trials data (Fitzmaurice and Lipsitz, 1995) using R follows.

```
setwd("C:/Documents and Settings/Desktop")

# This is longitudinal data where each subject is measured on five occasions
# which we label as weeks 0, 1, 5, 9, and 13.  There are several missing
# values.  Subjects are randomized to treatment after week 1.  This means that
# at weeks 0 and 1, no treatment was given.

arth = read.table("fitz.csv", h=T, sep=",")


# Removes records with missing responses.

arth2 = na.omit(arth)

# Column names : gender, age, patient, y, time, trt
#
#  gender : 1 = male, 0 = female
#  age (in years)
#  patient (subject id)
#  y :  1 indicates good self-assessment and 0 a bad self-assessment
#  time (in weeks): 0 for baseline, 1, 5, 9, 13
#  trt (treatment) :  1 for auranofin and 0 for no treatment


# Creating Z matrix associated with continuous time AR(1) log-odds ratio

n <- as.vector( table( as.factor(arth2$patient) ) )
last <- cumsum(n);
first <- last - n + 1;
z.arth <- NULL
for ( i in 1:length(n) )
{
    n.i   <- n[i]
    id.i  <- arth2$patient[ first[i] ]
    occ.i <- arth2$time[ first[i]:last[i] ]

    if (n.i == 1) { z.i <- cbind(0,0) }
    else
    {
        id.i <- rep(id.i, choose(n.i, 2))
        z.i1 <- rep(NA, choose(n.i, 2) )
        l <- 1
        for( j in seq(1, n.i - 1) )
        {
```

Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

```
            for( k in seq(j+1, n.i) )
            {
                z.i1[l] <- 1 / ( abs(occ.i[k] - occ.i[j]) )
                l <- l+1
            }
        }
        z.i <- cbind(id.i, z.i1)
    }

    z.arth <- rbind(z.arth, z.i)
}

## Remove any rows of zeros.  These correspond to cluster size 1 which do  ##
## not exists in the association matrix.  Clusters of size 1 provide no    ##
## information on the association.                                         ##

z.arth <- data.frame(z.arth)
names(z.arth) <- c("patient", "distance")
z.arth <- z.arth[(z.arth$patient != 0), ]


## invokes the orth package (assuming that it is installed)
library(orth)

## ALR: odds-ratio follows continuous time AR(1) structure.
## Model from Fitzmaurice et al 1995.
## Mean Model  : 1 + GENDER + age + TREATMENT + week
## Assoc Model : distance
## where distance denotes the reciprocal of the spacing of two measurements.

fitz.1 <- orth(y ~ time + factor(gender) + age + factor(trt),
             data=arth2, formula.z = ~ -1 + distance,
             dataz = z.arth, id=patient, estLam=F, tol=0.00001)


summary(fitz.1)


## The following is an output from the ORTH procedure.

Class:  summary.orth

Summary values based on robust covariance.  Those interested in
model-based covariance may use the 'SUMMARY()' method on this
summary.orth class.

Marginal Mean Parameters:
                Estimate    Std. Error    Chi Square      Pr(>Chi)
(Intercept)   1.09338601    1.37808222     0.6295015   0.427538230
```

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

```
time            -0.02717259    0.02964747     0.8400146   0.359392593
factor(gender)1  0.59564765    0.47848133     1.5497044   0.213179081
age             -0.01536315    0.02468207     0.3874334   0.533651587
factor(trt)1     1.45721185    0.45092111    10.4434560   0.001230847


Association Parameters (log-odds):
           Estimate    Std. Error    Chi Square      Pr(>Chi)
distance   3.684099      1.024464      12.93209  0.0003229965
```

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

A.2 SAS Code

Below is an example of SAS code used to analyze the arthritis data. Depending on the reader's experience, he or she may prefer to create the data set for the **Z** matrix in another way. In fact, `PROC IML` may be used to create the **Z** data set in exactly the same manner as the R code.

```
filename INF  "APSTAT.DAT";

** Reads in ASCII file containing the data           **;
** Note y_j = 0 denotes good self assessment and     **;
** y_j = 1 denotes bad self assessment.              **;

** Gender : 1 if male                                **;
** age    : in years                                 **;
** treat  : treatment (1 if auranofin, 0 if placebo) **;

data temp;
  infile inf;
  input gender age trt  y1-y5;
  patient = _n_;  /* Create patient ID */
run;

%let y = y;
%let x = int time gender age treat;
%let z = distance;

data XY (keep=patient &y &x)  Z (keep = patient &z);
  set temp;
  int = 1; /* intercept */
  array y_[*]  y1-y5;
  array t_[*]  t1-t5 (0 1 5 9 13);

  /* Creates the design matrix for the mean model */
  /* as well as the response.                     */
  do j = 1 to 5;
    y = y_[j];
    if (y ^= .) then do
      y = ^y;           * model prob(good);
      time = t_[j];
      treat = trt * (j>2);
      output XY;
    end;
  end;

  /* Creates the design matrix for the association model. */
  do j = 1 to 4;
    if (y_[j] ^= .) then do k = (j+1) to 5;
      if (y_[k] ^= .) then do;
        distance = 1 / abs(t_[j]-t_[k]);
```

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

```
        output Z;
      end;
    end;
  end;

run;

** Create weight matrix      **;
** Each cluster has weight 1 **;
data weight (keep=w);
  set temp;
  w=1;
run;


** Invokes the macro **;

%include "ORTH.macro";

** Performs ALR using ORTH **;
%ORTH(xydata=xy, yvar=&y, xvar=&x, id=patient, zdata=z, zvar=&z,
         wdata=weight, wvar=w, maxiter=20, epsilon=0.0000001, estlamb = NO,
         CLSOUT=clsout, OBSOUT=obsout, monitor=no, IBETA=, IALPHA=);

quit; * Stops IML from continuing;

proc print data = clsout; title2 "clsout"; run;
proc print data = obsout; title2 "obsout"; run;


/* The following is output printed by the ORTH procedure                */

Marginal Mean Parameter Estimates with Model-based (Naive) Standard Errors

                      VAR         PARM  N_STDERR

                      INT    1.0933862 1.0812585
                      TIME   -0.027173 0.0365192
                      GENDER 0.5956477 0.4353876
                      AGE    -0.015363  0.018969
                      TREAT  1.4572118 0.4726857


                      ORTHRES Macro, Version 1.0
                  Method of Orthogonalized Residuals
                Richard Conrad Zink & Bahjat F. Qaqish
                            (c) 2003

                ********** RESULTS **********
```

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uncbiostat/art22

```
Number of Clusters:        51
Maximum Cluster Size:       5
Minimum Cluster Size:       2
Number of Iterations:       9
      Outcome Variable: y


Note: Robust Standard Errors are Presented


                                LAMBDA

Note: Lambda is fixed at          0 .


      Marginal Mean Parameter Estimates

VAR         PARM    STDERR     CHISQ    PVALUE

INT    1.0933862 1.3780824 0.6295015 0.4275382
TIME   -0.027173 0.0296475 0.8400149 0.3593925
GENDER 0.5956477 0.4784813  1.549705  0.213179
AGE    -0.015363 0.0246821 0.3874335 0.5336516
TREAT  1.4572118  0.450921  10.44346 0.0012308


   Marginal Odds Ratio Parameter Estimates

VAR           PARM    STDERR     CHISQ    PVALUE

DISTANCE 3.6840989 1.0244654 12.932068  0.000323
```

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press