2012

Alternatives to Mixture Model Analysis of Correlated Binomial Data

N. Rao Chaganty
Old Dominion University

Roy Sabo
Virginia Commonwealth University

Yihao Deng
Indiana University - Purdue University Fort Wayne

International Scholarly Research Network
ISRN Probability and Statistics
Volume 2012, Article ID 896082, 10 pages
doi:10.5402/2012/896082

Research Article

Alternatives to Mixture Model Analysis of Correlated Binomial Data

N. Rao Chaganty,[1] Roy Sabo,[2] and Yihao Deng[3]

[1] *Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077, USA*
[2] *Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, Richmond, Virginia 23298-0032, USA*
[3] *Department of Mathematical Sciences, Indiana University-Purdue University Fort Wayne, Fort Wayne, IN 46805-1499, USA*

Correspondence should be addressed to Roy Sabo, rsabo@vcu.edu

Received 28 February 2012; Accepted 29 March 2012

Academic Editors: P. D'Urso, A. Hutt, and M. Scotto

While univariate instances of binomial data are readily handled with generalized linear models, cases of multivariate or repeated measure binomial data are complicated by the possibility of correlated responses. Likelihood-based estimation can be applied by using mixture distribution models, though this approach can present computational challenges. The logistic transformation can be used to bypass these concerns and allow for alternative estimating procedures. One popular alternative is the generalized estimating equation (GEE) method, though systematic errors can lead to infeasible correlation estimates or nonconvergence problems. Our approach is the coupling of quasileast squares (QLSs) method with a rarely used matrix factorization, which achieves a simplified estimation platform—as compared to the mixture model approach—and does not suffer from the convergence problems in GEE method. A noncontrived example is provided that shows the mechanical breakdown of GEE using several statistical software packages and highlights the usefulness of the QLS approach.

1. Introduction

Binomial data occur when observations on a given subject consist of a fixed series of Bernoulli trials, resulting in a proportional outcome. Maximum likelihood estimation is readily available in a generalized linear modeling framework when subjects consist of univariate measures (i.e., one Bernoulli or binomial trial per subject). However, estimation becomes more complicated when several Bernoulli or binomial trials are observed for each subject. In this case subject responses could be multivariate (consisting of several series of separately defined trials) or repeated measure (where the set of trials are defined similarly), and in both

Table 1: The proportion (p_{ij}) of successful outcomes for the ith subject during the jth repetition.

	Rep. 1	Rep. 2	$\cdots$	Rep. t
Sub. 1	p_{11}	p_{12}	$\cdots$	p_{1t}
Sub. 2	p_{21}	p_{22}	$\cdots$	p_{2t}
Sub. 3	p_{31}	p_{32}	$\cdots$	p_{3t}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
Sub. m	p_{m1}	p_{m2}	$\cdots$	p_{mt}

instances there is the possibility that the intrasubject responses are correlated. Here we use the term "subject" for convenience but it could be an item, store, location, plot in agriculture experiments, and so on. For example, real-life data situations where we encounter correlated proportions include (1) bankers interested in the proportion of customers making the ith type of transaction at the jth bank branch; (2) the proportion of CD deposits at the ith branch in the jth month of a year; (3) retail managers interested in the proportion of customers purchasing the ith item at the jth store; (4) marketers interested in the proportion of subjects viewing the ith advertisement type on the jth website; (5) information technology specialists interested in the proportion of students who use the ith computer program in the jth computer lab; (6) biologists interested in the proportion of hatched English sole eggs kept in solutions at different temperatures and salinity levels. We will discuss the last example later in this paper. Other examples where correlated binomial data occur are seed testing experiments described in Gilliland et al. [1].

Data layout of the aforementioned examples is presented in Table 1. In all of these examples the proportions within each row are correlated but the rows can be assumed to be independent. The within-row correlation, while complicating matters, must be accounted for in order to obtain proper variance estimates and inference for any regression parameters representing the associations between the vector of proportions and covariates. Thus, the problem is to estimate the parameters of interest within the ensemble of all parameters. In this context one could use a likelihood-based approach utilizing mixture-distribution models.

In the case of binomial data the mixture model would consist of both binomial and logit-normal components. However, parameter estimation in the mixture model could experience convergence problems due to the multitude of marginal means, regression, and correlation parameters. A simplified alternative approach would be to transform the variable-specific proportions for each subject in a way that would simplify the assumed probability distribution. The logit of the proportions would transform the outcome scale from $[0,1]$ to $(-\infty,\infty)$, which could make appropriate a multivariate normal-based methodology. One such procedure could be the generalized estimating equations (GEEs) proposed by Liang and Zeger [2]. Though a popular methodology for estimating regression parameters in cases of longitudinal or repeated measure data, this procedure suffers problems estimating correlations. As will be seen in subsequent sections, the GEE method can fail to converge even for cases of continuous data, which is the case if the logit transformation is used on binomial data.

Coull and Agresti [3] discussed random effects models for logit-transformed correlated binomial data. Here we suggest the method of quasileast squares (QLSs), developed by Chaganty [4] and Chaganty and Shults [5]. While generally seen as an alternative method to solving the maximum likelihood score equation for correlation parameters in the case of Gaussian data (Sabo and Chaganty [6]), the estimation of correlation in the QLS procedure

can also be supplemented with a little-known matrix factorization that makes it distinct from the maximum likelihood method. In this sense the QLS procedure is applicable for estimating correlated continuous data, which is appropriate for logit-transformed binomial proportions.

The rest of this paper is outlined as follows. The likelihood-based mixed-model approach is discussed in Section 2, while the logit transformation of binomial data and the GEE methodology are discussed in Section 3. We briefly outline the QLS estimating procedure in Section 4, while also highlighting the matrix factorization for use in estimating correlation. A noncontrived example is given in Section 5 that shows the usefulness of the QLS approach, as well as the convergence problems experienced by several statistical software packages in implementing the GEE method. A brief conclusion follows in Section 6.

2. Maximum Likelihood Estimation Using Mixture Distribution Models

For $i = 1, \ldots, m$ subjects, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})'$ be a vector of t possibly dependent binomial random variables, where y_{ij} is the number of successes out of n_{ij} trials with success probability p_{ij} for the jth variable of the ith subject. Also assume that $\mathbf{x}_{ij} = (x_{i1}, \ldots, x_{ik})$ is the vector of k covariates corresponding to the jth variable in the ith subject, such that $\mathbf{X}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{it})'$ is the $t \times k$ matrix of all covariates for the ith subject.

The general mixture distribution model for binomial data is given by

$$f(\mathbf{y}_i) = \int_{[0,1]^t} \prod_{j=1}^{t} \binom{n_{ij}}{y_{ij}} p_{ij}^{y_{ij}} (1 - p_{ij})^{n_{ij} - y_{ij}} \; \mathbf{G}(d\mathbf{p}_i), \tag{2.1}$$

where $\mathbf{G}$ is a multivariate cumulative distribution function with support in $[0,\ 1]^t$ and $\mathbf{p}_i = (p_{i1}, \ldots, p_{it})$. Basically, we assume that $\mathbf{p}_i$ is distributed as $G(\cdot)$, and, given $\mathbf{p}_i$, the vector $\mathbf{y}_i$ consists of t independent binomial variables. Then the marginal distribution of $\mathbf{y}_i$ is given by (2.1). A popular choice for $\mathbf{G}$ is the multivariate logit-normal distribution; that is, the distribution obtained under the assumption $\text{logit}(\mathbf{p_i}) = (\log(p_{i1}/(1 - p_{i1})), \ldots, \log(p_{it}/(1 - p_{it})))$ is multivariate normal with mean $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{it})$ and covariance matrix Σ. Here $\mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$ represents the mean as a function of the covariates and a k-dimensional regression parameter vector $\boldsymbol{\beta}$. To make model (2.1) identifiable we make the common assumption that $\Sigma = \phi\mathbf{R}$, where $\mathbf{R}$ is a correlation matrix and ϕ is a scale parameter (Joe [7], page 219). This condition is necessary for model identification as the following simple example shows. Suppose $t = 2$ and that the vector $\mathbf{p}$ is multivariate logit-normal distributed with mean $\boldsymbol{\mu} = 0$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & 0.3\sigma_1\sigma_2 \\ 0.3\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. It is easy to verify that two sets of choices for the variances σ_1^2 and σ_2^2 can result in identical binary distribution for $\mathbf{y}$ as shown in Table 2.

If $\boldsymbol{\beta}, \mathbf{R}, \phi$ are the only parameters of interest, we can obtain maximum likelihood estimates by maximizing the likelihood $L(\boldsymbol{\beta}, \mathbf{R}, \phi) = \prod_{i=1}^{m} f(\mathbf{y}_i)$. However, if the $E(y_{ij}/n_{ij}) = p_{ij}$'s are also of interest, we can obtain estimates of these parameters using either the empirical Bayes (EB) method or the EM algorithm considering the full likelihood

$$L(\mathbf{p}_i, \boldsymbol{\beta}, \mathbf{R}, \phi) = \prod_{i=1}^{m} \prod_{j=1}^{t} \binom{n_{ij}}{y_{ij}} p_{ij}^{y_{ij}} (1 - p_{ij})^{n_{ij} - y_{ij}} h(\mathbf{p}_i, \boldsymbol{\beta}, \mathbf{R}, \phi). \tag{2.2}$$

Table 2: Identical distribution for y with two different choices for $\boldsymbol{\Sigma}$.

(y_1, y_2)	Joint probability of $\mathbf{y}$	
	$\sigma_1 = 3.0$, $\sigma_2 = 2.9$	$\sigma_1 = 3.8$, $\sigma_2 = 4.0$
$(1,1)$	0.2877	0.2877
$(1,0)$	0.2123	0.2123
$(0,1)$	0.2877	0.2877
$(0,0)$	0.2123	0.2123

Equation (2.2) is the full specification of (2.1) with covariates, regression parameters, correlation, and variance described in $h(\cdot)$, the multivariate logit-normal density function. One quickly notices that the likelihood (2.2) has parameters that increase with m, and solutions to the maximization of (2.2) will require roots of complex nonlinear equations. These considerations may make the full likelihood approach subject to computational difficulties and convergence problems. Further, such specific definitions for the components in the mixture model may affect estimator robustness.

3. Alternatives to Likelihood-Based Estimation

For reasons outlined earlier it makes sense to consider the vector of logit-transformed proportions $\hat{\mathbf{u}}_i = (\hat{u}_{i1}, \ldots, \hat{u}_{it})$, where $\hat{u}_{ij} = \text{logit}(\hat{p}_{ij}) = \log[\hat{p}_{ij}/(1-\hat{p}_{ij})]$ and $\hat{p}_{ij} = y_{ij}/n_{ij}$. Note that $\hat{\mathbf{u}}_i$ is distributed as multivariate normal with parameters $E(\hat{\mathbf{u}}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$ and $\text{Cov}(\hat{\mathbf{u}}_i) = \phi\mathbf{R}$. The focus on these normally distributed random variables, rather than the mixture-distribution-based binomial random variables, can allow us to relax distributional assumptions and utilize distribution-free methodologies for parameter estimation such as the generalized estimating equations (GEEs). This methodology is a two-stage process, in which the estimate of the regression parameter $\boldsymbol{\beta}$ is updated by a residual-based moment estimate of $\mathbf{R}$. Specifically, estimation is iterated between the two equations

$$\sum_{i=1}^{m}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)'\hat{\mathbf{R}}^{-1}(\hat{\mathbf{u}}_i - \boldsymbol{\mu}_i) = 0, \qquad \hat{\mathbf{R}} = \frac{\mathbf{Z}}{\hat{\phi}}, \tag{3.1}$$

until convergence. Here $\mathbf{Z} = \sum_{i=1}^{m}\mathbf{z}_i\mathbf{z}_i'$, $\hat{\phi} = \sum_{i=1}^{m}(\mathbf{z}_i'\mathbf{z}_i)/(mt-k)$, where $\mathbf{z}_i = \hat{\mathbf{u}}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}})$. The problem with this methodology is that the diagonal elements of $\mathbf{Z}$ are not necessarily equal to $\hat{\phi}$, implying that the diagonal elements of $\hat{\mathbf{R}} = \mathbf{Z}/\hat{\phi}$ are not necessarily unity. However, the GEE methodology, as implemented in software packages, forces the diagonal elements of $\hat{\mathbf{R}}$ to unity (i.e., it changes the values from whatever they are to 1), and thus matrix $\hat{\mathbf{R}}$ is not guaranteed to be positive definite. This can lead to (most harmlessly) convergence problems, but it can also lead to artificially deflated estimator variances for the regression parameters and is thus subject to improper or incorrect inference.

4. Quasileast Squares

The quasileast squares (QLSs) approach, on the other hand, provides an alternative estimate of $\mathbf{R}$ and does not experience the convergence problems exhibited by GEE. A further benefit

of this method is that it does not require the assumption of normality for the joint distribution of each response or among their marginal distributions. The initial step for estimation of $\mathbf{R}$ is to minimize $\mathrm{tr}(\mathbf{R}^{-1}\mathbf{Z})$ over the set of correlation matrices. Since the diagonal elements of $\mathbf{R}$ are restricted to be one, introducing a diagonal matrix of lagrange multipliers $\boldsymbol{\Lambda}$, we can verify that the point of minimum $\tilde{\mathbf{R}}$ factors the matrix $\mathbf{Z}$ as

$$\mathbf{Z} = \tilde{\mathbf{R}}\,\boldsymbol{\Lambda}\,\tilde{\mathbf{R}}. \tag{4.1}$$

Whittle [8] has shown that for a positive definite matrix $\mathbf{Z}$ the factorization (4.1) is unique. Further $\tilde{\mathbf{R}} = \boldsymbol{\Lambda}^{-1/2}(\boldsymbol{\Lambda}^{1/2}\mathbf{Z}\boldsymbol{\Lambda}^{1/2})\boldsymbol{\Lambda}^{-1/2}$, and the diagonal matrix $\boldsymbol{\Lambda}$ satisfies the fixed-point equation $\boldsymbol{\Lambda} = \mathrm{diag}\,(\boldsymbol{\Lambda}^{1/2}\mathbf{Z}\boldsymbol{\Lambda}^{1/2})^{1/2}$, which can be solved using a simple fixed-point iterative scheme (Olkin and Pratt [9], Chaganty [4]). Next, using the first step correlation estimate $\tilde{\mathbf{R}}$, we can then obtain a consistent correlation estimate as

$$\hat{\mathbf{R}} = \tilde{\mathbf{R}}\,\boldsymbol{\Delta}\,\tilde{\mathbf{R}}, \tag{4.2}$$

where $\boldsymbol{\Delta} = \mathrm{diag}(\boldsymbol{\nu})$, $\boldsymbol{\nu} = (\tilde{\mathbf{R}} \circ \tilde{\mathbf{R}})^{-1}\mathbf{e}$, $\mathbf{e}$ is a vector of ones, and $\circ$ denotes the Hadamard product. It is possible that the correlation matrix (4.2) may not be positive definite in which case Chaganty and Shults [5] have recommended the estimate

$$\hat{\mathbf{R}} = \mathrm{diag}(\mathbf{Z})^{-1/2}\mathbf{Z}\ \mathrm{diag}(\mathbf{Z})^{-1/2}. \tag{4.3}$$

See equation (3.2) in Chaganty and Shults [5]. The quasi-least squares method uses the estimate (4.2) of $\mathbf{R}$ if it is positive definite and otherwise uses (4.3), which is clearly a positive definite correlation matrix, to update the estimate of $\boldsymbol{\beta}$ until convergence. Code for fitting this model using the R statistical software is provided in the Appendix.

5. Example

We now provide an example from Alderdice and Forrester [10], who modeled the effects of salinity and temperature on the proportion of hatched English sole eggs. In this study, the number of hatched eggs was recorded at seven salinity and five temperature levels. Measurements were taken in four separate tanks for each combination of salinity and temperature, and for each tank we have recordings of the number of fish eggs and the number hatched. Thus, the tanks represent the repeated measure component for this binomial data set. The data, as given on page 6 of Lindsey [11], is reproduced in Table 3.

The goal of the analysis is to study the dependence of the proportion of eggs hatched on the temperature and salinity. After calculating $\hat{u}_{ij} = \mathrm{logit}(y_{ij}/n_{ij})$, where y_{ij} is the number of eggs hatched out of the total n_{ij} in the jth tank at the ith combination of temperature and

Table 3: Number of hatched and total eggs of English sole at different salinity and temperature levels in sea water.

		Tank 1		Tank 2		Tank 3		Tank 4	
Temp.	Salinity	Hatch	Total	Hatch	Total	Hatch	Total	Hatch	Total
15	4	236	666	203	724	183	764	212	723
15	8	600	656	697	747	615	746	641	703
15	12	407	566	343	603	365	560	302	394
25	4	203	717	177	782	155	852	138	590
25	8	591	621	564	640	714	754	532	570
25	12	475	622	465	645	506	608	415	532
35	4	1	738	3	655	10	742	3	763
35	8	526	616	419	467	410	484	374	606
35	12	272	362	352	478	392	590	382	459
10	10	303	681	329	710	262	611	301	700
10	6	277	757	234	681	263	647	287	801
40	10	387	450	389	553	388	564	318	604
40	6	276	662	247	542	248	527	149	591
20	10	351	391	559	650	527	603	476	548
20	6	585	643	620	671	437	497	667	771
30	10	447	491	462	530	475	545	499	556
30	10	522	573	615	680	539	581	517	561
30	6	563	666	600	704	562	656	615	723

salinity, the Shapiro-Wilk test for normality was performed on the transformed responses for each replicate. The results (P-values < 0.05) indicate a departure from normality, so that maximum likelihood methods for continuous data are not applicable. The data was analyzed using GEE in several statistical software packages using an unstructured working correlation matrix to account for the correlation between the four replications of the solution in the four tanks. The results using PROC GENMOD in SAS version 9.2, gee.fit module in TIBCO Spotfire S+ version 8.2, and xtgee procedure in STATA version 11, are shown in parts (i), (ii), and (iii) of Table 4. The warning message from PROC GENMOD read "WARNING: Iteration limit exceeded." Here we see that in each case the estimates failed to converge. The 0.999 correlation estimates in part (i) represent model breakdown in that programmers often use this value to indicate nonconvergence.

The warning message from TIBCO Spotfire S+ software read (sic) "Warning messages: 1: at convergence at the correlation estimate 1 is outside of the range $[-1, 1]$ in cgeefit (gee.model) 2: correlation matrix is not full rank, $2 < 4$ in: cgeefit (gee.model)." Note that the correlation between the first and second tanks in part (ii) is greater than one, clearly violating the most liberal of correlation boundaries. The warning message from xtgee in STATA version 11 read "convergence not achieved." Also, the fourth eigenvalue in part (iii) is negative, indicating that the estimated correlation matrix is not positive definite. The results of the QLS analysis are given in part (iv) of Table 4, which show that the estimated correlation matrix is positive definite.

Table 4: Analysis of English sole eggs data: (i) GEE parameter estimates and working correlation matrix using the SAS system GENMOD procedure. (ii) GEE parameter estimates and working correlation matrix using the TIBCO Spotfire S+. (iii) GEE parameter estimates and eigenvalues of the working correlation matrix using STATA version 11. (iv) QLS parameter estimates and correlation estimates using an R program given in the Appendix.

(i) GEE estimates using SAS GENMOD procedure								Working correlation		
Parm.	Est.	SE	95% C.I.		Z	Pr > |Z|	1.000	**0.999**	**0.999**	**0.999**
Int.	−1.983	1.149	−4.235	0.270	−1.73	0.085		1.000	**0.999**	**0.999**
Sal.	−0.017	0.038	−0.091	0.058	−0.43	0.664			1.000	**0.999**
Temp.	0.378	0.163	0.059	0.697	2.32	0.020				1.000
(ii) GEE estimates using TIBCO Spotfire S+								Working correlation		
Parm.	Est.	SE	95% C.I.		Z	Pr > |Z|	1.000	**1.049**	0.890	0.978
Int.	−2.096	1.562	−5.157	0.965	−1.34	0.180		1.000	0.820	0.874
Sal.	−0.010	0.040	−0.089	0.069	−0.25	0.799			1.000	0.754
Temp.	0.379	0.141	0.103	0.656	2.69	0.007				1.000
(iii) GEE estimates using STATA 11								Eigenvalues		
Parm.	Est.	SE	95% C.I.		Z	Pr > |Z|	3.664	0.259	0.114	**−0.038**
Int.	−2.096	1.562	−5.157	0.965	−1.34	0.180				
Sal.	−0.010	0.040	−0.089	0.069	−0.25	0.799				
Temp.	0.379	0.141	0.103	0.656	2.69	0.007				
(iv) QLS estimates using R 2.14.1								Working correlation		
Parm.	Est.	SE	95% C.I.		Z	Pr > |Z|	1.000	0.968	0.920	0.940
Int.	−1.936	1.623	−5.117	1.244	−1.19	0.233		1.000	0.931	0.912
Sal.	−0.018	0.042	−0.100	0.064	−0.42	0.672			1.000	0.874
Temp.	0.370	0.147	0.083	0.657	2.53	0.012				1.000

6. Discussion

The logit transformation was originally applied on mortality rates in univariate bioassays (Berkson, [12]), though the idea also generalizes nicely into the cases of correlated repeated-measure, longitudinal, or multivariate binomial data discussed here. Doing so allows the data analyst to bypass complicated, parametrically saturated mixture distributions and utilize methods for correlated continuous data. Interestingly, even after the logit transformation is applied, the GEE method still experiences convergence difficulties and problems with correlation parameter estimation. Potential causes for these problems are explained in Section 3. The QLS method, on the other hand, does not experience these difficulties and handles the simultaneous estimation of both regression and correlation parameters with relative ease. This was made possible by incorporating the little-known and rarely used matrix factorization given in (4.1).

Note that the probit transformation had an earlier origin and similar function to the logit transformation (Bliss, [13]) and can also be used in place of the logit transformation shown here. However, likelihood estimation of correlated binomial data using a latent multivariate distribution has already been established for the probit link function (Ashford and Sowden, [14]) and has been compared favorably to the GEE method when analyzed on real data (Sabo and Chaganty [15]).

```
##############################################################################
# R (ver 2.14.1) program to compute QLS estimates for the mixture model #
##############################################################################

# Function to estimate the correlation matrix
# between the repeated measurements

correlation.est <- function(residuals, tol=1e-10)
 {
     t <- ncol(residuals)
     Z <- t(residuals)%*%residuals
     # start the decomposition algorithm with an identity matrix
     Lambda0 <- diag(t)
     ev <- eigen(Z)
     Lambdak <- diag(diag(ev$vec%*%diag(sqrt(ev$val))%*%t(ev$vec)))
     Diff <- diag(Lambdak -Lambda0)
     while(sum(Diff^2) > tol)
     {
        Lambda0 <- Lambdak
        ev <- eigen(sqrt(Lambda0)%*% Z %*%sqrt(Lambda0))
        M <- ev$vec%*% diag(sqrt(ev$val)) %*%t(ev$vec)
        Lambdak <- diag(diag(M))
        Diff <- diag(Lambdak - Lambda0)
     }
     Rtilde <- solve(sqrt(Lambdak))%*% M %*%solve(sqrt(Lambdak))
     Rhat <- Rtilde%*%diag(as.vector(solve(Rtilde*Rtilde)%*%rep(1,t)))%*
     %Rtilde
     ev <- eigen(Rhat)
     if (any(ev$val<0))
        Rhat <- solve(sqrt(diag(diag(Z))))%*% Z %*% solve
        (sqrt(diag(diag(Z)))) return(Rhat)
 }

# Function to calculate the regression parameter beta.

regression.est <- function(x, y, t, Rhat)
 {

     mt <- nrow(x)
     Sigma <- solve(kronecker(diag(mt/t), Rhat))
     XRinvX <- t(x)%*%Sigma%*%x
     XRinvY <- t(x)%*%Sigma%*%y
     betahat <- solve(XRinvX)%*%XRinvY
     return(betahat)
 }
# The main program starts here

d <- read.table("c:/hatch-eggs.txt", header=TRUE)
proportion <- d$Hatch/d$Total
y <- log(proportion/(1-proportion))
x <- model.matrix(~Salinity+Temperature, data=d)
tol <- 1e-10
t <- length(d$ID)/length(unique(d$ID))
mt <- nrow(x)
m <- mt/t
k <- ncol(x)
Rhatinit <- diag(t)
betahat <- regression.est(x, y, t, Rhatinit)
```

Algorithm 1: Continued.

```
residuals <- matrix(y-x%*%betahat, ncol=t, byrow=TRUE)
Rhat <- correlation.est(residuals)
betanew <- regression.est(x, y, t, Rhat)

while(sum((betanew-betahat)^2)>tol)
{
     betahat <- betanew
     residuals <- matrix(y-x%*%betahat, ncol=t, byrow=TRUE)
     Rhat <- correlation.est(residuals)
     betanew <- regression.est(x, y, t, Rhat)
}

# Calculate the scale parameter

residuals <- matrix(y-x%*%betahat, ncol=t, byrow=TRUE)
Z <- t(residuals)%*% residuals
Rhat <- correlation.est(residuals)
scale <- sum(diag(solve(Rhat)%*%Z))/(mt-k)

# Calculate model based standard errors and z-scores for betas

Sigma <- solve(kronecker(diag(m), Rhat))
Covbeta <- scale*solve(t(x)%*%Sigma%*%x)
stderrbeta <- sqrt(diag(Covbeta))
zstat <- betanew/stderrbeta

# Prepare and print the output

output <- cbind(betanew, stderrbeta, zstat, 2*(1-pnorm(abs(zstat))))
colnames(output) <- c("Estimate", "Std. Error", "z value", "P-value")
list(scale = scale, Rhat=Rhat, beta = output)
```

Algorithm 1

Appendix

For more details see Algorithm 1.

References

[1] D. Gilliland, O. Schabenberger, and H. Liu, "Intercluster correlations for binomial data: an application to seed testing," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 7, no. 1, pp. 95–106, 2002.

[2] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[3] B. A. Coull and A. Agresti, "Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution," *Biometrics*, vol. 56, no. 1, pp. 73–80, 2000.

[4] N. R. Chaganty, "An alternative approach to the analysis of longitudinal data via generalized estimating equations," *Journal of Statistical Planning and Inference*, vol. 63, no. 1, pp. 39–54, 1997.

[5] N. R. Chaganty and J. Shults, "On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter," *Journal of Statistical Planning and Inference*, vol. 76, no. 1-2, pp. 145–161, 1999.

[6] R. T. Sabo and N. R. Chaganty, "Estimation methods for an autoregressive familial correlation structure," *Communications in Statistics. Theory and Methods*, vol. 39, no. 6, pp. 973–991, 2010.

[7] H. Joe, *Multivariate Models and Dependence Concepts*, vol. 73, Chapman & Hall, London, UK, 1997.

[8] P. Whittle, "A multivariate generalization of Tchebichev's inequality," *The Quarterly Journal of Mathematics*, vol. 9, pp. 232–240, 1958.

[9] I. Olkin and J. W. Pratt, "A multivariate Tchebycheff inequality," *Annals of Mathematical Statistics*, vol. 29, pp. 226–234, 1958.

[10] D. F. Alderdice and C. R. Forrester, "Some effects of salinity and temperature on early development and survival of the English sole (Parophrys vetulus)," *Journal of the Fisheries and Research Boad of Canada*, vol. 25, pp. 495–521, 1968.

[11] J. K. Lindsey, "Likelihood analyses and tests for binary data," *Journal of Applied Statistics*, vol. 24, no. 1, pp. 1–16, 1975.

[12] J. Berkson, "Application to the logistic function to bio-assay," *Journal of the American Statistical Association*, vol. 39, pp. 357–365, 1944.

[13] C. I. Bliss, "The calculation of the dosage mortality curve," *Annals of Applied Biology*, vol. 22, pp. 134–167, 1935.

[14] J. R. Ashford and R. R. Sowden, "Multi-variate probit analysis," *Biometrics*, vol. 26, no. 3, pp. 535–546, 1970.

[15] R. T. Sabo and N. R. Chaganty, "What can go wrong when ignoring correlation bounds in the use of generalized estimating equations," *Statistics in Medicine*, vol. 29, no. 24, pp. 2501–2507, 2010.