Repeated Measures Analysis with Discrete Data Using the SAS® System

Gordon Johnston, SAS Institute Inc., Cary, NC

Abstract

The analysis of correlated data arising from repeated measurements when the measurements are assumed to be multivariate normal has been studied extensively. In many practical problems, however, the normality assumption is not reasonable. When the responses are discrete and correlated, for example, different methodology must be used in the analysis of the data. Generalized Estimating Equations (GEEs) provide a practical method with reasonable statistical efficiency to analyze such data. This paper provides an overview of the use of GEEs in the analysis of correlated data using the SAS System. Emphasis is placed on discrete correlated data, since this is an area of great practical interest.

Introduction

GEEs were introduced by Liang and Zeger (1986) as a method of dealing with correlated data when, except for the correlation among responses, the data can be modeled as a generalized linear model. For example, correlated binary and count data in many cases can be modeled in this way.

A SAS macro, written by M.R. Karim at Johns Hopkins University is available to fit such models by solving GEEs. Work is in progress to add the capability to solve GEEs to the GENMOD procedure in SAS/STAT software. This paper provides an overview of the GEE methodology that will be implemented in the GENMOD procedure. Refer to Diggle, Liang, and Zeger (1994) and the other references at the end of this paper for details on this method.

Correlated data can arise from situations such as

- longitudinal studies, in which multiple measurements are taken on the same subject at different points in time
- clustering, where measurements are taken on subjects that share a common category or characteristic that leads to correlation. For example, incidence of pulmonary disease among family members may be correlated because of hereditary factors.

The correlation must be accounted for by analysis methods appropriate to the data. Possible consequences of analyzing correlated data as if it were independent are

- incorrect inferences concerning regression parameters due to underestimated standard errors
- inefficient estimators, that is, more mean square error in regression parameter estimators than necessary

Example of Longitudinal Data

These data, from Thall and Vail (1990), are concerned with the treatment of epileptic seizure episodes. These data were also analyzed in Diggle, Liang, and Zeger (1994). The data consists of the number of epileptic seizures in an eight-week baseline period, before any treatment, and in each of four two-week treatment periods, in which patients received either a placebo or the drug Progabide as an adjunct to other chemotherapy. A portion of the data is shown in Table 1.

Table 1. Epileptic Seizure Data

Patient ID	Treatment	Baseline	Visit1	Visit2	Visit3	Visit4
104	Placebo	11	5	3	3	3
106	Placebo	11	3	5	3	3
107	Placebo	6	2	4	0	5
.						
.						
.						
101	Progabide	76	11	14	9	8
102	Progabide	38	8	7	9	4
103	Progabide	19	0	4	3	0
.						
.						
.						

Within-subject measurements are likely to be correlated, whereas between-subject measurements are likely to be independent. The raw correlations among the counts between visits are shown in Figure 1. They indicate strong correlation in the number of seizures between the visits. The seizures data will be analyzed in later sections as count data with a specified correlation structure.

Figure 1. Raw Correlations

	Visit 1	Visit 2	Visit 3	Visit 4
Visit 1	1.00	.69	.54	.72
Visit 2		1.00	.67	.76
Visit 3			1.00	.71
Visit 4				1.00

Generalized Linear Models for Independent Data

Let $Y_i,\ i = 1, \ldots, n$ be independent measurements. Generalized linear models for independent data are characterized by

- a systematic component

  $$g(E(Y_i)) = g(\mu_i) = \mathbf{x}_i{}'\boldsymbol{\beta}$$

  where $\mu_i = E(Y_i)$, g is a link function that relates the means of the responses to the linear predictor $\mathbf{x}_i{}'\boldsymbol{\beta}$, $\mathbf{x}_i$ is a vector of independent variables for the *i*th observation, and β is a vector of regression parameters to be estimated.

- a random component: $Y_i,\ i = 1, \ldots, n$ are independent and have a probability distribution from an exponential family:

  $Y_i \sim$ exponential family:
  binomial, Poisson,
  normal, gamma,
  inverse gaussian

The exponential family assumption implies that the variance of Y_i is given by $V_i = \phi v(\mu_i)$, where v is a variance function that is determined by the specific probability distribution and ϕ is a dispersion parameter that may be known or may be estimated from the data, depending on the specific model. The variance function for the binomial and Poisson distributions are given by

- binomial: $v(\mu) = \mu(1 - \mu)$
- Poisson: $v(\mu) = \mu$

The maximum likelihood estimator of the $p \times 1$ parameter vector β is obtained by solving the estimating equations

$$\sum_{i=1}^{m} \frac{\partial \mu_i'}{\partial \boldsymbol{\beta}} v_i^{-1}(y_i - \mu_i(\beta)) = \mathbf{o}$$

for β. This is a nonlinear system of equations for β and it can be solved iteratively by the Fisher scoring or Newton-Raphson algorithm.

Modeling Correlation

Generalized Estimating Equations

Let $Y_{ij},\ j = 1, \ldots, n_i, i = 1, \ldots, K$ represent the *j*th measurement on the *i*th subject. There are n_i measurments on subject i and $\sum_{i=1}^{K} n_i$ total measurements.

Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurements on the *i*th subject be $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]'$ with corresponding vector of means $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{in_i}]'$ and let $\mathbf{V}_i$ be an estimate of the covariance matrix of $\mathbf{Y}_i$. The Generalized Estimating Equation for estimating β is an extension of the independence estimating equation to correlated data and is given by

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{o}$$

Working Correlations

Let $\mathbf{R}_i(\boldsymbol{\alpha})$ be an $n_i \times n_i$ "working" correlation matrix that is fully specified by the vector of parameters $\boldsymbol{\alpha}$. The covariance matrix of $\mathbf{Y}_i$ is modeled as

$$\mathbf{V}_i = \phi \mathbf{A}^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}^{\frac{1}{2}}$$

where $\mathbf{A}$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the *j*th diagonal element. If $\mathbf{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\mathbf{Y}_i$, then $\mathbf{V}_i$ will be the true covariance matrix of $\mathbf{Y}_i$.

The working correlation matrix is not usually known and must be estimated. It is estimated in the iterative fitting process using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

There are several specific choices of the form of working correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$ commonly used to model the correlation matrix of $\mathbf{Y}_i$. A few of the choices are shown here. Refer to Liang and Zeger (1986) for additional choices. The dimension of the vector $\boldsymbol{\alpha}$, which is treated as a nuisance parameter, and the form of the estimator of α are different for each choice.

- $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}_0$, a fixed correlation matrix. For $\mathbf{R}_0 = \mathbf{I}$, the identity matrix, the GEE reduces to the independence estimating equation.
- m-dependent:

$$Corr(Y_{ij}, Y_{i,j+t}) = \begin{cases} \alpha_t & t = 1, 2, \ldots, m \\ 0 & t > m \end{cases}$$

- Exchangeable: $Corr(Y_{ij}, Y_{ik}) = \alpha,\ j \neq k$
- Unstructured: $Corr(Y_{ij}, Y_{ik}) = \alpha_{jk}$

Fitting Algorithm

The following is an algorithm for fitting the specified model using GEEs.

- Compute an initial estimate of β, for example with an ordinary generalized linear model assuming independence.
- Compute the working correlations $\mathbf{R}_i(\boldsymbol{\alpha})$.
- Compute an estimate of the covariance:

$$\mathbf{V}_i = \phi \mathbf{A}^{\frac{1}{2}} \hat{\mathbf{R}}_i(\boldsymbol{\alpha}) \mathbf{A}^{\frac{1}{2}}$$

- Update β:

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r -$$

$$[\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}]^{-1} [\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)]$$

- Iterate until convergence.

Properties of GEEs

The GEE method has some desirable statistical properties that make it an attractive method for dealing with correlated data.

- GEEs reduce to independence estimating equations for $n_i = 1$.
- GEEs are the maximum likelihood score equation for multivariate Gaussian data.
- $\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow N(0, \mathbf{M}(\phi))$ if the mean model is correct even if $\mathbf{V}_i$ is incorrectly specified, where
  - $$\mathbf{M}(\phi) = \mathbf{I}_0^{-1} \mathbf{I}_1 \mathbf{I}_0^{-1}$$
  - $$\mathbf{I}_0 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$
  - $$\mathbf{I}_1 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} Cov(\mathbf{Y}_i) \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

The third property listed above means that you don't have to specify the working correlation matrix correctly in order to have a consistent estimator of the regression parameters. Choosing the working correlation closer to the true correlation increases the statistical efficiency of the regression parameter estimator, so you should specify the working correlation as accurately as possible based on knowledge of the measurement process.

Estimating the Covariance of $\hat{\boldsymbol{\beta}}$

The *model-based* estimator of $Cov(\hat{\boldsymbol{\beta}})$ is given by

$$Cov_M(\hat{\boldsymbol{\beta}}) = \mathbf{I}_0^{-1}$$

This is the GEE equivalent of the inverse of the Fisher information matrix that is often used in generalized linear models as an estimator of the covariance estimate of the maximum likelihood estimator of $\boldsymbol{\beta}$. It is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ if the mean model and the working correlation matrix are correctly specified.

The estimator

$$\mathbf{M} = \mathbf{I}_0^{-1} \mathbf{I}_1 \mathbf{I}_0^{-1}$$

is called the *empirical*, or *robust*, estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$. It has the property of being a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, even if the working correlation matrix is misspecified, that is, if $Cov(\mathbf{Y}_i) \neq \mathbf{V}_i$. In computing $\mathbf{M}$, $\boldsymbol{\beta}$ and ϕ are replaced by estimates, and $Cov(\mathbf{Y}_i)$ is replaced by an estimate, such as

$$(\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})'(\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$$

Progabide Example

Return to the epileptic seizure data, and model the data as the log-linear model with $v(\mu) = \mu$ (the Poisson variance function) and

$$\begin{aligned} \log(E(Y_{ij})) &= \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \\ & \quad x_{i1}x_{i2}\beta_3 + \log(t_{ij}) \end{aligned}$$

where

- Y_{ij}: number of epilectic seizures in interval j
- t_{ij}: length of interval j
- $x_{i1} = \begin{cases} 1: & \text{weeks 8-16} \\ 0: & \text{weeks 0-8} \end{cases}$

- $x_{i2} = \begin{cases} 1: & \text{progabide group} \\ 0: & \text{placebo group} \end{cases}$

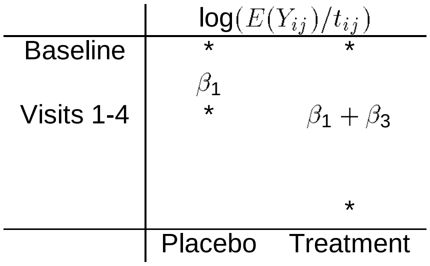
The correlations between the counts are modeled as $r_{ij} = \alpha, i \neq j$ (exchangeable correlations). For comparison, the correlations are also modeled as independent (identity correlation matrix). In this model, the regression parameters have the interpretation in terms of the log seizure rate shown in Figure 2.

Figure 2. Interpretation of Regression Parameters

Treatment	Visit	log$(E(Y_{ij})/t_{ij})$
Placebo	Baseline	β_0
	1-4	$\beta_0 + \beta_1$
Progabide	Baseline	$\beta_0 + \beta_2$
	1-4	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

As indicated schematically in Figure 3, the difference between the log seizure rates in the pretreatment (baseline) period and the treatment periods is β_1 for the placebo group and $\beta_1 + \beta_3$ for the Progabide group. A value of $\beta_3 < 0$ would indicate an effective reduction in the seizure rate.

Figure 3. Interpretation of Model



The results of fitting the model using the SAS macro are shown in Table 2. The parameter estimates are nearly identical, but the standard errors for the independence case are underestimated. The coefficient of the interaction term, β_3, is highly significant under the independence model and marginally significant with the exchangeable correlations model.

Table 2. Results of Model Fitting

Variable	Correlation Structure	Coef.	Std. Error	Coef./S.E.
Intercept	Exchangeable	1.35	.16	8.56
	Independent	1.35	.03	39.52
Visit (x_1)	Exchangeable	.11	.12	.95
	Independent	.11	.05	2.36
Treat (x_2)	Exchangeable	-.11	.19	-.56
	Independent	-.11	.05	-2.22
$x_1 * x_2$	Exchangeable	-.30	.17	-1.76
	Independent	-.30	.07	-4.32

The fitted exchangeable correlation matrix is shown in Figure 4.

Figure 4. Exchangeable Correlations

	Exchangeable Correlations			
	Visit 1	Visit 2	Visit 3	Visit 4
Visit 1	1.00	.60	.60	.60
Visit 2		1.00	.60	.60
Visit 3			1.00	.60
Visit 4				1.00

The empirical (robust) and model-based estimates of the regression parameter covariance matrix are shown in Output 1. The fit working correlation matrix is also shown. The two covariance estimates are similar, indicating an adequate correlation model.

Output 1. Covariance Estimates - Exchangeable Correlation

```
Working Correlation:
        1 0.5983035 0.5983035 0.5983035 0.5983035
0.5983035         1 0.5983035 0.5983035 0.5983035
0.5983035 0.5983035         1 0.5983035 0.5983035
0.5983035 0.5983035 0.5983035         1 0.5983035
0.5983035 0.5983035 0.5983035 0.5983035         1

Variance estimate (naive):

          INTERCPT        X1       TRT     X1TRT

INTERCPT 0.0120616 0.0015936 -0.012062 -0.001594
X1        0.118764 0.0149267 -0.001594 -0.014927
TRT      -0.700173 -0.083155 0.0246033 0.0055616
X1TRT    -0.075566  -0.63627 0.1846554 0.0368707

Variance estimate (robust):

          INTERCPT        X1       TRT     X1TRT

INTERCPT 0.0247613 -0.001152 -0.024761 0.0011518
X1       -0.063047 0.0134791 0.0011518 -0.013479
TRT      -0.812488 0.0512247 0.0375093 -0.002999
X1TRT    0.0427552 -0.678151 -0.090451 0.0293096

NOTE: Covariances are above diagonal and
      correlations are below diagonal.
```

For comparison, a correlation model that is less adequate was fit. The 2-dependent working correlation, with correlations for lags three and four set to zero, was fit. The results are shown in Output 2. The empirical (robust) and model-based covariances are much more dissimilar than for the exchangeable case, indicating a less adequate correlation model.

Output 2. Covariance Estimates - 2-Dependent Correlation

```
Working Correlation:
        1 0.6419198  0.579265         0         0
0.6419198         1 0.6419198  0.579265         0
 0.579265 0.6419198         1 0.6419198  0.579265
        0  0.579265 0.6419198         1 0.6419198
        0         0  0.579265 0.6419198         1

Variance estimate (naive):

          INTERCPT        X1       TRT     X1TRT

INTERCPT 0.0051212 0.0039807 -0.005121 -0.003981
X1       0.7898409 0.0049598 -0.003981  -0.00496
TRT      -0.696673 -0.550261 0.0105514 0.0121299
X1TRT    -0.382889 -0.484767 0.8128322 0.0211058

Variance estimate (robust):

          INTERCPT        X1       TRT     X1TRT

INTERCPT 0.0397344 -0.031138 -0.039734 0.0311383
X1       -0.565107  0.076412 0.0311383 -0.076412
TRT      -0.803623 0.4541329 0.0615265 -0.043132
X1TRT    0.3754255 -0.664345 -0.417909 0.1731308

NOTE: Covariances are above diagonal and
      correlations are below diagonal.
```

Modeling Odds Ratios for Binary Data

Diggle, Liang, and Zeger (1994) point out that modeling association among binary responses with correlation has a disadvantage, and they propose using the odds ratio instead. For binary data, the correlation between the *j*th and *k*th response is, by definition,

$$Corr(Y_{ij}, Y_{ik}) = \frac{Pr(Y_{ij} = 1, Y_{ik} = 1) - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})}}$$

The joint probability in the numerator satisfies the following bounds, by elementary properties of probability, since $\mu_{ij} = Pr(Y_{ij} = 1)$:

$$\max(0, \mu_{ij} + \mu_{ik} - 1) \leq Pr(Y_{ij} = 1, Y_{ik} = 1) \leq$$

$$\min(\mu_{ij}, \mu_{ik})$$

The correlation, therefore, is constrained to be within limits that depend in a complicated way on the means of the data.

The odds ratio, defined as

$$OR(Y_{ij}, Y_{ik}) =$$

$$\frac{Pr(Y_{ij} = 1, Y_{ik} = 1)Pr(Y_{ij} = 0, Y_{ik} = 0)}{Pr(Y_{ij} = 1, Y_{ik} = 0)Pr(Y_{ij} = 0, Y_{ik} = 1)}$$

is not constrained by the means and is preferred by many workers to correlations for binary data. Carey, Zeger, and Diggle (1993) propose an algorithm for fitting the log odds ratio as

$$\log(OR(Y_{ij}, Y_{ik})) = \mathbf{z}'_{ijk}\boldsymbol{\alpha}$$

where $\mathbf{z}'_{ijk}$ is a vector of covariates and $\boldsymbol{\alpha}$ is a vector of association parameters to be estimated. The mean is modeled with a regression model just as it is when you use correlations to model association. This implementation of GEE is called alternating logistic regression (ALR), and it uses a GEE similar to the one used to model correlations to estimate the mean regression parameters $\boldsymbol{\beta}$ alternating with a logistic regression to estimate the association parameters $\boldsymbol{\alpha}$.

The previous method treated correlation as a nuisance parameter, which must be taken into account but is not of scientific interest. The ALR method is useful if the association is a scientific focus of the analysis, since a detailed model for the association is fitted.

Conclusion

Generalized Estimating Equations provide a practical method with good statistical properties to model data that exhibit association but cannot be modeled as multivariate normal. A SAS macro is available to use GEEs to model associations among non-normal data using correlations. Work is in progress to add the capability to the SAS/STAT procedure GENMOD to use GEEs to model correlations and odds ratios.

References

Carey, V., Zeger, S.L., and Diggle, P. (1993), “Modelling Multivariate Binary Data with Alternating Logistic Regressions” *Biometrika*, 517-526

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford: Oxford Science

Liang, K.-Y. and Zeger, S.L. (1986), “Longitudinal Data Analysis Using Generalized Linear Models” *Biometrika*, 13-22

Thall, P.F. and Vail, S.C. (1990), “Some Covariance Models for Longitudinal Count Data with Overdispersion” *Biometrics*, 657-671

Zeger, S.L. and Liang, K.-Y. (1986), “Longitudinal Data Analysis for Discrete and Continuous Outcomes” *Biometrics*, 121-130