Using SAS/GENMOD Procedure to Fit GEE Regression Models

Miin-Jye Wen        Lily Yeh
Department of Statistics    Department of Nursing
National Cheng-Kung University

Abstract

Researchers are often interested in analyzing data that results from longitudinal studies. Estimating equations for generalized linear modeling of longitudinal data have attracted a great deal of attention over the last decade. Liang and Zeger (1986) presented an approach to these problems involving generalized estimating equations (GEEs) extended from generalized linear models (GLMs) into a regression setting with correlated observations within subjects. This paper provides a brief review of the GLM and GEE methodologies, and illustrates their implementation with a home-care example using the GENMOD procedure with SAS/STAT software to solve GEE in the analysis of correlated data.

Keywords: SAS, GEE, GLM.
Introduction

The class of generalized linear models is an extension of traditional linear models that allow the mean of a population to depend on a linear predictor through a nonlinear link function and allows the probability distribution of the response to be any member of an exponential family of distributions. This family was first introduced by Nelder and Wedderburn (1972) and consists of normal error linear regression models and a nonlinear exponential, Poisson regression models, logistic and probit models for binary data, as well as many other models, such as log-linear models for the categorical data. Refer to McCullagh and Nelder (1989) for a thorough account of statistical modeling using generalized linear models. However, frequently researchers are interested in analyzing data that results from a longitudinal period or a repeated measure design with a correlation existing between the observations on a specific subject.

Correlated data can arise from longitudinal studies, in which multiple measurements are taken on the same subject at different points in time. Longitudinal data can be defined as data collected from the observations of subjects on a number of variables over time. The main advantage of a longitudinal study is its effectiveness for studying change over time. In fact, we speak of longitudinal data whenever we have observed some occurrence more than once. Longitudinal data correlation must be considered for any appropriate analysis method. Liang and Zeger (1986) formalized an approach to this problem using Generalized Estimating Equations (GEEs) to extend Generalized Linear Models (GLMs) into a regression setting with correlated observations within subjects. For details on the GEEs method, refer to Liang and Zeger (1986), Zeger and Liang (1986), and Diggle, Liang and Zeger (1994). In this paper, we briefly present a review of GLMs and GEEs methodologies, introduce a home-care example that will be implemented in the
SAS/GENMOD Procedure and provide a tutorial on how to fit this example using the SAS system.

**Methodology**

The class of generalized linear models can be described in the following:

1. $Y_1, ..., Y_n$ are $n$ independent responses that follow a probability distribution belonging to the exponential family of probability distributions with expected value $E(Y_i) = \mu_i$, $i=1,...,n$.

2. A linear predictor based on the $p$ explanatory variables $X_{i1}, ..., X_{ip}$, $i=1,...,n$ is utilized, denoted by $X_i'\beta$:

   $$X_i'\beta = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

3. The link function $g$ relates the linear predictor to the mean response:

   $$X_i'\beta = g(\mu_i) = g(E(Y_i))$$

4. Generalized linear models may have nonconstant variances $\sigma_i^2$ for the responses $Y_i$, but the variances $\sigma_i^2$ must be a function of the predictor variables through the mean response $\mu_i$.

5. Any regression model that belongs to the family of generalized linear models can be analyzed in a unified fashion, thus providing a common statistical methodology framework for different types of responses. One of the attractive properties of the generalized linear models is that it allows for linear as well as non-linear models under a single framework. McCullagh and Nelder (1989) provided further details about the estimate parameters and tests for generalized linear models and their analysis.

6. It is possible to fit models where the underlying data are normal, inverse Gaussian,
gamma, Poisson, binomial, geometry and negative binomial using a suitable choice of link functions \( g(.) \).

When the data are longitudinal types or repeated measures, there exists a correlation between the observations in generalized linear models. Liang and Zeger (1986) and Zeger and Liang (1986) introduced generalized estimating equations (GEEs) to account for correlated data. The GEE approach builds upon previous methods of variance estimation developed to protect against inappropriate assumptions about the variance. The GEE setting is described in the following steps:

1. We are not assuming that response variables \( Y_{ij} \) (i: subject & i=1,..., n, J: corresponding to a specific time & j=1,...t) is a number of the exponential family, but we are assuming that the mean and variance are characterized as in the GLM. Response variables can be continuous, counted, or binary variables.

2. The link function \( g \) relates the linear predictor to the mean response:

\[
X_{ij} \beta = g(\mathbb{E}(Y_{ij})) = \mu_{ij}
\]

\( \mu_{ij} \) is a \( p \times 1 \) vector of the study variables for the \( i^{th} \) subject at the \( j^{th} \) outcome. \( \beta \) consists of the \( p \) regression parameters of interest.

3. A common choice for the link function might be \( g(a) = a \) for measured data (the identity link); \( g(a) = \log(a) \) for count data (log link); \( g(a) = \log(a/1-a) \) for binary data (logit link).

4. In addition to this, we need to model the covariance structure of the correlated measures on a given subject. The \( t \times t \) covariance matrices of \( Y_i \), as

\[
V_i = \phi A_i^{1/2}R(\alpha)A_i^{1/2}
\]

are modeled where \( \phi \) is the scale or dispersion.
parameter, \( A_i \) is a \( n_i \times n_i \) diagonal matrix of the variance functions \( V(U_y) \), and \( R(\alpha) \) are the working correlation matrix of \( Y_i \) indexed by a vector of parameter \( \alpha \).

(5) A number of working correlation matrices is displayed in Table 1.

<table>
<thead>
<tr>
<th>Structure \ (SAS defines)</th>
<th>Type</th>
</tr>
</thead>
</table>
| Independent \ (IND)      | \[
1 0 \cdots 0 \\
0 1 \cdots 0 \\
\vdots \ \\
\vdots \\
0 0 \cdots 1
\] |
| Exchangeable \ (EXCH or CS) | \[
1 \alpha \cdots \alpha \\
\alpha 1 \cdots \alpha \\
\vdots \\
\vdots \\
\alpha 0 \cdots 1
\] |
| Unstructured \ (UNSTR or UN) | \[
1 \rho_{12} \cdots \rho_{1t} \\
\rho_{12} 1 \cdots \rho_{2t} \\
\vdots \\
\vdots \\
\rho_{1t} \rho_{2t} \cdots 1
\] |
| Autoregressive \ (AR or AR(1)) | \[
1 \rho \cdots \rho^{t-1} \\
\rho 1 \cdots \rho^{t-2} \\
\vdots \\
\vdots \\
\rho^{t-1} \rho^{t-2} \cdots 1
\] |
| M-dependent \ (MDEP(Number)) | \[
1 \rho_1 \cdots \rho_{t-1} \\
\rho_1 1 \cdots \rho_{t-2} \\
\vdots \\
\vdots \\
\rho_{t-1} \rho_{t-2} \cdots 1
\] |
| Fixed \ (USER or FIXED (matrix)) | \[
1 \gamma_{12} \cdots \gamma_{1t} \\
\gamma_{12} 1 \cdots \gamma_{2t} \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\gamma_{t-1} \gamma_{t-2} \cdots 1
\] |
(6) A useful property of GEE is that it provides valid inference for $\beta$, i.e., correct standard error (variance) of $\hat{\beta}$.

(7) A set of estimating equations is solved to find the value of estimator $\hat{\beta}$. An empirical variance estimator can be used to estimate $\hat{\beta}$. This variance estimator is also referenced to a “robust” estimator. Another available variance estimate is the model-based estimate, which is consistent when both the mean model and the covariance model are correctly specified. Since in general the analysis will not know the correct covariance structure, the empirical variance estimate will be preferred when the number of clusters is large. When the number of clusters is small, (say, $n<20$), the model-based variance estimator may have better properties.

**Example using SAS system**

Before analyzing longitudinal data or repeated measurements, the researcher must first confirm the following questions to fit a GEE:

(1) Which is the appropriate family of distributions for the data?

(2) What link function is an appropriate choice?

(3) What is a reasonable structure for the correlation within subjects?

The researcher must specify both the distribution family and working correlation matrix. To conduct the software procedure, we analyzed data from a study on home care clients from Yeh et al. (2000). Subjects included 75 clients and 4 two months observations. Within-subject measurements are likely to be correlated, whereas
between-subject measurements are likely to be independent. A goal of this study was to explore the albumin and hemoglobin (Hb) of patients who were newly admitted to home care agencies. It is of interest to relate the albumin and hemoglobin use in the 10 settings to home care client characteristics. The primary outcomes were albumin and hemoglobin use in ten settings: age, gender, number of care needs (bad), intake route (ccb), care location fixed or not (ccbf), caregiver fixed or not (cca), registration for service from discharge (ccaf), 2 months (t2), 4 months (t3), and 6 months periods (t4). Note: The first visit within one month was defined as the reference data.

We fit a normal model using an AR(1) working correlation structure. The GEE methodology for the modeling of correlated response data has been incorporated into the GENMOD procedure of the SAS/STAT software. This example using the GENMOD procedure illustrates GEE strategy implementation with the new REPEATED statement, available in SAS/STAT software. GENMOD procedures include classical linear models with normal errors, logistic and probit models for binary data and log-linear models for multinomial data. The REPEATED statement specifies the covariance structure of multivariate responses for GEE model fitting in the GENMOD procedure. The TYPE option from the REPEATED statement specifies the structure of the working correlation matrix used to model the correlation of the responses from the subjects. A number of working correlation matrices are displayed in Table 1. Table 2 displays the syntax needed to specify this model for the SAS packages.

**Table 2. SAS Syntax to fit GEE’s for home care model**

```sas
Data homecare;
  Input in albumin Hb period age sex bad ccb ccbf cca ccaf;
  t2=0.0; t3=0.0; t4=0.0;
  If period=2 then t2=1;
  If period=3 then t3=1;
```
If period=4 then t4=1;
Cards;
1 3.3 11.2 1 77 2 56 1 3 1 3
1 3.3 11.2 2 77 2 56 1 3 1 3
1 3.4 8.3 3 77 2 56 1 3 1 3
1 3.4 11.2 4 77 2 56 1 3 1 3
2 3.4 12.3 1 77 2 52 2 1 2 1
2 3.7 13.1 2 77 2 52 2 1 2 1
2 3.7 13.1 3 77 2 52 2 1 2 1
3 2.9 11.4 1 61 1 90 2 1 1 1
3 3 11.2 2 61 1 90 2 1 1 1
3 3 11.2 3 61 1 90 2 1 1 1
;
;
;
proc genmod data = homecare;
class in;
  model albumin = age sex bad ccb ccbf cca ccaf t2 t3 t4/
    dist = normal   link = identity;
repeated subject = in / type =AR;
run;

Table 2 specifies SUBJECT=in which identifies the clustering variables. The TYPE = AR option specifies an autoregressive working correlation structure.

Table 3 displays the parameter estimate tables and includes the parameter estimates, standard error, 95% confidence intervals, Z-scores, and p-values for the parameters estimates.

Table 3. Analysis of GEE parameter estimates for the response variable albumin

<table>
<thead>
<tr>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Of GEE Parameter Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Standard Error Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8
The GEE regression coefficients in Table 3, $\beta$, have the same interpretation as the coefficients from a classical regression analysis. Similarly, replacing the response variable albumin using Hb from Table 2 produces similar results in Table 3 as in Table 4.

**Table 4. Analysis of GEE parameter estimates for the response variable Hb**

| Parameter | Estimate | Error | 95% Confidence Limits | Z Pr > |Z| |
|-----------|----------|-------|------------------------|--------|-----|
| Intercept | 13.3664  | 1.3405| 10.7391 15.9936        | 9.97   | <.0001 |
| age       | -0.0462  | 0.0133| -0.0722 -0.0202        | -3.48  | 0.0005 |
| sex       | -0.1298  | 0.3979| -0.9096 0.6500         | -0.33  | 0.7442 |
| bad       | -0.0001  | 0.0059| -0.0116 0.014          | -0.02  | 0.9821 |
| ccb       | 0.4441   | 0.3722| -0.2854 1.1736         | 1.19   | 0.2328 |
| ccbf      | 0.1480   | 0.2613| -0.3641 0.6600         | 0.57   | 0.5711 |
| cca       | 0.6899   | 0.9168| -1.1069 2.4868         | 0.75   | 0.4517 |
| ccaf      | -0.0059  | 0.3233| -0.6396 0.6277         | -0.02  | 0.9854 |
| T2        | 0.1307   | 0.1609| -0.1848 0.4461         | 0.81   | 0.4169 |
Discussion

In this paper, we reviewed the methodologies underlying GLMs and GEEs and fit models for a home care example using the GENMOD procedure in the SAS/STAT software. Because SAS is a popular and powerful statistical software package, we illustrated the GEE methodology using the SAS system.

There are other software packages that can run the GEE program. For instance, GEE (a SAS macro for longitudinal data analysis), written by Karim and Zeger (1988) and Groemping (1993), Stata, Suddan, and S-plus can fit such models by solving GEEs. In general, GEE’s are well supported by these software packages and are straightforward to use. Refer to Horton and Lipsitz (1999) for a detailed review of the software that fits GEE regression models.

This scheme would also be applicable to period data in business or data in social sciences with repeated measures over time.
References


