Paper SAS166-2015

Weighted Methods for Analyzing Missing Data with the GEE Procedure

Guixian Lin and Robert N. Rodriguez, SAS Institute Inc.

Abstract

Missing observations caused by dropouts or skipped visits present a problem in studies of longitudinal data. When the analysis is restricted to complete cases and the missing data depend on previous responses, the generalized estimating equation (GEE) approach, which is commonly used when population-averaged effects are of primary interest, can lead to biased parameter estimates.

The new GEE procedure in SAS/STAT® 13.2 implements a weighted GEE method, which provides consistent parameter estimates when the dropout mechanism is correctly specified. When none of the data are missing, the method is identical to the usual GEE approach, which is available in the GENMOD procedure. This paper reviews the concepts and statistical methods for weighted GEE and illustrates them with an example.

Introduction

Missing data frequently occur in longitudinal studies, where missing observations can be caused by dropouts or skipped visits. To draw valid inferences when data are missing, you can use different approaches, such as maximum likelihood, multiple imputation, fully Bayesian analysis, and inverse probability weighting (Little and Rubin 2002; National Research Council 2010). The GEE procedure, introduced in SAS/STAT 13.2, provides a weighted generalized estimating equations (GEE) method for analyzing longitudinal data that have missing observations. This approach extends the usual generalized estimating equations approach (Liang and Zeger 1986).

When none of the data are missing, the weighted GEE method is identical to the usual GEE method, which is available in the GENMOD procedure. The standard GEE method is valid if the data are missing completely at random (MCAR), but it can lead to biased results if the data are missing at random (MAR). The GEE procedure implements the inverse probability-weighted method to account for dropouts under the MAR assumption (Robins and Rotnitzky 1995; Preisser, Lohman, and Rathouz 2002).

The following sections introduce the weighted GEE method and provide a clinical trials example to illustrate how the use of PROC GEE to analyze longitudinal data with dropouts.

Generalized Estimating Equations Method

Longitudinal studies are frequently used in applied fields such as public health, medical research, and social science. Multiple measurements are taken on the same subject over time, enabling you to discover changes in the response over time and the relationship of changes to covariates (Fitzmaurice, Laird, and Ware 2011).

The marginal model is commonly used in analyzing longitudinal data when the population-averaged effects are of interest. To estimate the regression parameters in the marginal model, Liang and Zeger (1986) proposed the generalized estimating equations method, which is widely used.

Suppose y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$, represents the jth response on the ith subject with a vector of covariates x_{ij}. There are n_i measurements on subject i, and the maximum number of measurements per subject is T.

Suppose the responses on the ith subject are $\mathbf{Y}_i = [y_{i1}, \ldots, y_{in_i}]'$ with corresponding means $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{in_i}]'$. For generalized linear models, the marginal mean μ_{ij} of the response y_{ij} is related to a linear predictor through a link function $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$, and the variance of y_{ij} depends on the mean through a variance function $v(\mu_{ij})$.

An estimate of the parameter $\boldsymbol{\beta}$ in the marginal model can be obtained by solving the generalized estimating equations,

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}'_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$

where $\mathbf{V}_i$, the working covariance matrix of $\mathbf{Y}_i$, is specified through the working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$,

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

Here, $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix whose jth diagonal element is $v(\mu_{ij})$, which is the value of the variance function at μ_{ij}. If $\mathbf{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\mathbf{Y}_i$, then $\mathbf{V}_i$ is the true covariance matrix of $\mathbf{Y}_i$.

Only the mean and the covariance of $\mathbf{Y}_i$ are required in the GEE method; a full specification of the joint distribution of the correlated responses is not needed. This is particularly convenient because the joint distribution for noncontinuous responses involves high-order associations and is complicated to specify. Moreover, the regression parameter estimates are consistent even when the working covariance is incorrectly specified. Because of this, the GEE method is popular in situations where the marginal effect is of interest and the responses are not continuous. However, the GEE approach can lead to biased estimates when missing responses depend on previous responses. The weighted GEE method described in the following section can provide unbiased estimates.

Weighted Generalized Estimating Equations Method

In longitudinal studies, response measurements are often missing because of skipped visits or dropouts. Suppose r_{ij} is the indicator that the response y_{ij} is observed, where $r_{ij} = 1$ if y_{ij} is observed and 0, otherwise. Missing data patterns can be classified into two types: dropout and intermittent. A dropout occurs when an individual skips a particular visit and doesn't return for the subsequent visits. That is, if $r_{ij} = 0$, then $r_{ik} = 0$ for all $k > j$. Otherwise, the missing data pattern is intermittent. Intermittent patterns can be quite complicated; only dropout patterns are considered here.

The mechanism for missingness can be described by a statistical model for the probability of observing a missing value. Making the right assumption about the mechanism is crucial to methods that handle missing data correctly. In general, missingness mechanisms are classified into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin 1976).

- The data are said to be MCAR if the probability of a missing response is independent of its past, current, and future responses conditional on the covariates. That is, $P(r_{ij} = 0|\mathbf{Y}_i, \mathbf{X}_i) = P(r_{ij} = 0|\mathbf{X}_i)$.
- The data are said to be MAR if the probability of a missing response is independent of its current and future responses conditional on the observed past responses and the covariates. That is,

  $$P(r_{ij} = 0|r_{ij-1} = 1, X_i, Y_i) = P(r_{ij} = 0|r_{ij-1} = 1, X_i, y_{i1}, \ldots, y_{ij-1})$$

  MAR is a weaker assumption than MCAR. In fact, MCAR is a special case of MAR.
- The data are said to be MNAR if the probability of a missing response depends on the unobserved responses. MNAR is the most general and the most problematic missing data scenario.

The GEE procedure implements two weighted methods, observation-specific and subject-specific, for estimating the regression parameter $\boldsymbol{\beta}$ when dropouts occur. Both provide consistent estimates if the data are MAR.

Observation-Specific Weighted GEE Method

Suppose w_{ij} is the weight for y_{ij}, and is defined as the inverse probability of observing y_{ij}. In other words, $w_{ij} = P(r_{ij} = 1|X_i, Y_i)^{-1}$. Suppose $\mathbf{W}_i$ is a $T \times T$ diagonal matrix whose jth diagonal is $r_{ij} w_{ij}$. Consider the following weighted generalized estimating equations (Robins and Rotnitzky 1995; Preisser, Lohman, and Rathouz 2002):

$$\mathbf{S}_{ow}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$

Unlike the standard GEE, the weighted estimating equations are unbiased when the observations are appropriately weighted, leading to consistent estimates of $\boldsymbol{\beta}$.

The weights w_{ij} are generally unknown in practice, and they are estimated with a logistic regression model under the MAR assumption. Specifically, suppose $\lambda_{ij} = P(r_{ij} = 1 | r_{ij-1} = 1, X_i, Y_i)$ denotes the probability of observing the response y_{ij} given its observed previous responses.

Under the MAR assumption,

$$\lambda_{ij} = P(r_{ij} = 1 | r_{ij-1} = 1, X_i, Y_i) = P(r_{ij} = 1 | r_{ij-1} = 1, X_i, y_{i1}, \ldots, y_{ij-1})$$

Using the observed data, λ_{ij} can be predicted from a logistic regression model,

$$\text{logit}\{\lambda_{ij}\} = z_{ij}\boldsymbol{\alpha}$$

where the z_{ij} are predictors that usually include the covariates x_{ij}, the past responses, and indicators for visit times. The dropout process implies that the estimated probability of observing y_{ij} can be expressed as a cumulative product of conditional probabilities:

$$\hat{P}(r_{ij} = 1 | X_i, Y_i) = \lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{ij}(\hat{\boldsymbol{\alpha}})$$

With the estimated weights $\hat{w}_{ij} = \hat{P}(r_{ij} = 1 | X_i, Y_i)^{-1}$, the regression parameter $\boldsymbol{\beta}$ is estimated by solving the equation for $\mathbf{S}_{ow}(\boldsymbol{\beta})$. The algorithm for doing this is similar to the algorithm for solving the standard GEE.

The following algorithm fits marginal models by using the observation-specific weighted GEE method when the dropout process is missing at random:

1. Fit a logistic regression with data (r_{ij}, z_{ij}) to obtain an estimate of $\boldsymbol{\alpha}$ and estimate the weights

   $$\hat{w}_{ij} = \hat{P}(r_{ij} = 1 | X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{ij}(\hat{\boldsymbol{\alpha}})]^{-1}$$

   where $\lambda_{ij}(\hat{\boldsymbol{\alpha}})$ is the predicted probability obtained from the logistic regression.

2. Compute an initial estimate of $\boldsymbol{\beta}$ by using an ordinary generalized linear model, assuming independence of the responses.

3. Compute the working correlation matrix $\mathbf{R}$ based on the standardized residuals, the current estimate of $\boldsymbol{\beta}$, and the specified structure of $\mathbf{R}$.

4. Compute the $T \times T$ estimated covariance matrix:

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

5. Update $\hat{\boldsymbol{\beta}}$:

   $$\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + \left[\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right]$$

6. Repeat steps 3–5 until convergence.

You could obtain observation-specific weighted GEE estimates for $\boldsymbol{\beta}$ with two steps: First, you fit a logistic regression to estimate the weights. Second, you estimate $\boldsymbol{\beta}$ by specifying the estimated weights in the WEIGHT statement in the GENMOD procedure. This is sometimes done in practice. However, there are two issues with this approach:

1. This strategy is only appropriate for an independent working correlation matrix structure. The WEIGHT statement in the GENMOD procedure doesn't incorporate the weights properly for other correlation structures.

2. The weights are treated as fixed and known by the GENMOD procedure. Consequently, the standard errors of the regression parameters from the two-step approach are conservative, which leads to wider confidence intervals and conservative inference (Fitzmaurice, Laird, and Ware 2011).

Both of these issues are solved with the GEE procedure in SAS/STAT 13.2, which produces the appropriate standard errors; PROC GEE also handles a variety of working correlation structures.

Subject-Specific Weighted GEE Method

In contrast to the observation-specific weighted method, the subject-specific weighted method assigns a single weight to each subject. In other words, all the observations from a subject receive the same weight. Specifically, the subject-specific weighted method obtains the regression parameter estimates by solving the equations

$$\mathbf{S}_{sw}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} w_i (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$

where the weight w_i for subject i is the inverse probability of a subject i dropping out at the observed time (Fitzmaurice, Molenberghs, and Lipsitz 1995; Preisser, Lohman, and Rathouz 2002). Note that w_i is a scalar, as opposed to the weight matrix $\mathbf{W}_i$ in the observation-specific weighted GEE method.

The subject-specific weighted estimating equations are also unbiased after the observations have been appropriately weighted, producing consistent estimates for the regression parameters $\boldsymbol{\beta}$.

The weight w_i is usually unknown in practice and needs to be estimated. Suppose m_i is a dropout indicator for subject i, where $m_i = \sum_{j=1}^{T} r_{ij} + 1$. Assume that the first visit y_{i1} is always observed with $r_{i1} = 1$. Thus, the values of m_i are $2, \cdots, T+1$. Note that $m_i = T+1$ indicates that subject i completes all the T visits.

The weight w_i is defined as follows: If subject i drops out before completing the last visit (that is, $m_i \le T$), then $w_i = P(r_{im_i} = 0, r_{im_i-1} = 1 | X_i, Y_i)^{-1}$. Otherwise, the subject completes all the T visits (that is, $m_i = T+1$), and $w_i = P(r_{iT} = 1 | X_i, Y_i)^{-1}$.

As with observation-specific weights, the dropout process implies that subject-specific weights can be estimated as a cumulative product of conditional probabilities:

- $\hat{w}_i = P(r_{im_i} = 0, r_{im_i-1} = 1 | X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{im_i-1} \times (1 - \lambda_{im_i}(\hat{\boldsymbol{\alpha}}))]^{-1}$, if $m_i \le T$
- $\hat{w}_i = P(r_{iT} = 1 | X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{iT}(\hat{\boldsymbol{\alpha}})]^{-1}$, if $m_i = T+1$

Thus, the subject-specific weights $\hat{w}_i$ can be obtained after λ_{ij} is estimated by using logistic regression as described in the previous section. The algorithm that fits marginal models for the subject-specific weighted GEE method is similar to the algorithm for the observation-specific weighted GEE method and is not repeated here.

Example: Weighted GEE for Longitudinal Data with Missing Values

This example shows how you can use the GEE procedure to analyze longitudinal data that contain missing values. The data set is taken from a longitudinal study of women who continued to use contraception during four consecutive months (Fitzmaurice, Laird, and Ware 2011). In this study 1,151 women were randomly assigned to one of two treatments: 100 mg or 150 mg of depot-medroxyprogesterone acetate (DPMA). The response variable indicates their amenorrhea status in each of the four months. The question of interest is whether the treatment has an effect on the rate of the amenorrhea over time, and the example follows the analysis done by Fitzmaurice, Laird, and Ware (2011).

The following statements create the data set Amenorrhea:

```
data Amenorrhea;
   input id dose time y@@;
   datalines;
1      0        1      0
1      0        2      .
1      0        3      .
1      0        4      .

   ... more lines ...

1150     1        4      1
1151     1        1      1
1151     1        2      1
1151     1        3      1
1151     1        4      1
;
```

Prior to the analysis, two additional variables (**prevy** and **ctime**) are created. The value of **prevy** is set to the patient's previous amenorrhea symptoms, and its value is set to missing for the first visit of the patient. Because **time** is used as a continuous variable in the marginal model of interest and you need a similar classification variable in the missingness model, the variable, **ctime** is created. The following statements add these two variables to the data set:

```
data Amenorrhea;
   set Amenorrhea;
   by id;
   prevy=lag(y);
   if first.id then prevy=.;
   time=time-1;
   ctime=time;
run;
```

The variables in the data set Amenorrhea include:

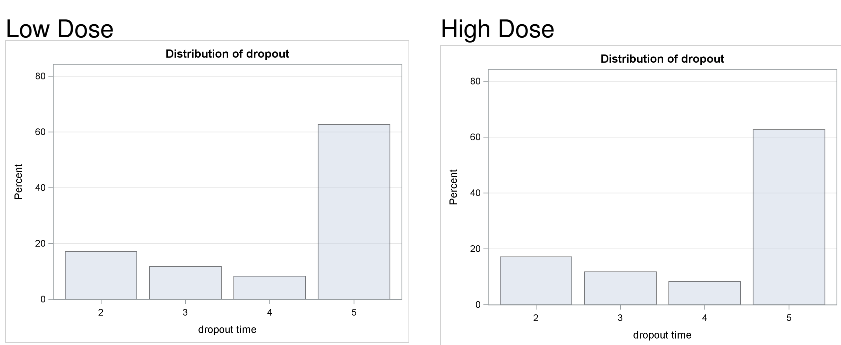
- **id**: patient's ID
- **y**: indicator of amenorrhea symptoms (1 for amenorrhea; 0 otherwise)
- **prevy**: patient's previous amenorrhea symptoms
- **time**: month (0, 1, 2, or 3)
- **ctime**: a copy of **time** that is used as a class variable in the missingness model (0, 1, 2, or 3)
- **dose**: administered dose (0 for treatment with 100 mg injection; 1 for treatment with 150 mg injection)

Suppose y_{ij} denotes the amenorrhea status of the ith woman i at the jth visit, $j = 1, \ldots, 4$, and suppose $\mu_{ij} = \mathrm{P}(y_{ij} = 1)$ denotes the average rate of this status. To determine whether the treatment has an effect on the rate of amenorrhea over time, consider the following marginal model:

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_1 \mathrm{time}_{ij} + \beta_2 \mathrm{time}_{ij}^2 + \beta_3 \mathrm{dose}_i + \beta_4 \mathrm{dose}_i \times \mathrm{time}_{ij} + \beta_5 \mathrm{dose}_i \times \mathrm{time}_{ij}^2$$

Of the 1,151 women in this study, 576 are from the low-dose group, and 575 are from the high-dose group. Figure 1 shows the distribution of dropout rates in the two groups. The maximum number of visits is 4, and a dropout indicator m_i of 5 means that all visits were complete. For the low-dose group, 62.67% of the women completed the trial; for the high-dose group, 61.39% of the women completed the trial. Both groups have substantial dropouts.

Figure 1 Dropout Distributions by Dose



To obtain the weights for the weighted GEE estimates, consider the following logistic regression model for missingness:

$$\begin{aligned}\mathrm{logit}\, p(r_{ij} = 1 | r_{ij-1} = 1, \mathrm{dose}_i, \mathrm{time}_{ij}, y_{ij-1}) =& \alpha_0 + \alpha_1 I(\mathrm{time}_{ij} = 2) + \alpha_2 I(\mathrm{time}_{ij} = 3) \\ &+ \alpha_3 \mathrm{dose}_i + \alpha_4 y_{ij-1} + \alpha_5 \mathrm{dose}_i \times y_{ij-1}\end{aligned}$$

The following statements request the observation-specific weighted GEE method to analyze the data with the previously defined response and missingness models:

```
proc gee data=Amenorrhea descending;
   class id ctime;
   missmodel ctime prevy dose dose*prevy / type=obslevel; /* missingness model */
   model y=time dose time*time dose*time dose*time*time / dist=bin; /* marginal model */
   repeated subject=id/within=ctime corr=cs;
run;
```

You need to specify both the MODEL and MISSMODEL statements to invoke the weighted GEE method.

The MODEL statement specifies the regression model for the mean with the binomial distribution variance function. The DESCENDING option in the PROC GEE statement specifies that the model be based on the probability that y = 1. If this option had not been specified, PROC GEE would base the model on the probability that y = 0.

The REPEATED statement requests the GEE approach for correlated data analysis. The SUBJECT=ID option specifies that observations from the same subject are identified by ID. The TYPE=CS option specifies the compound symmetric working correlation structure.

The MISSMODEL statement specifies the logistic regression model for missingness. No response is needed to specify a missingness model because the response is completely determined by whether the response variable y in the MODEL statement has a missing value. Without the MISSMODEL statement, PROC GEE would implement the standard GEE approach provided by PROC GENMOD. The TYPE=OBSLEVEL option requests observation-specific weights.

Figure 2 shows the parameter estimates for the missingness model. The estimate of α_4 is –0.4514 with a p-value of 0.0053, which suggests that the probability that a participant will drop out is related to her previous amenorrhea status. These results suggest that the assumption of MAR is more appropriate than that of MCAR.

Figure 2 Parameter Estimates for the Missingness Model

Maximum Likelihood Parameter Estimates for Missingness Model

Parameter		Estimate	Standard Error	Wald 95% Confidence Limits		Z	Pr > |Z|
Intercept		2.3967	0.1438	2.1149	2.6785	16.67	<.0001
ctime	1	-0.7286	0.1439	-1.0106	-0.4466	-5.06	<.0001
ctime	2	-0.5919	0.1469	-0.8798	-0.3040	-4.03	<.0001
ctime	3	0.0000	0.0000	0.0000	0.0000	.	.
prevy		-0.4514	0.1619	-0.7687	-0.1341	-2.79	0.0053
dose		0.0680	0.1313	-0.1893	0.3253	0.52	0.6046
prevy*dose		-0.2381	0.2196	-0.6685	0.1923	-1.08	0.2782

Figure 3 displays the parameter estimates for the marginal model.

Figure 3 Parameter Estimates for Marginal Model Using Weighted GEE

Analysis of GEE Parameter Estimates
Empirical Standard Error Estimates

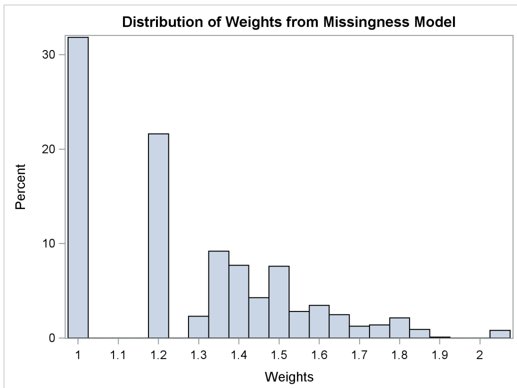
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.4965	0.1072	-1.7067	-1.2863	-13.95	<.0001
time	0.5379	0.1334	0.2764	0.7994	4.03	<.0001
dose	0.1061	0.1491	-0.1861	0.3983	0.71	0.4767
time*time	-0.0037	0.0405	-0.0831	0.0757	-0.09	0.9275
dose*time	0.4092	0.1903	0.0362	0.7823	2.15	0.0315
dose*time*time	-0.1264	0.0577	-0.2395	-0.0134	-2.19	0.0284

The estimates of β_4 and β_5 are 0.4092 and –0.1264, which indicate that the change of amenorrhea rate over time depends on the dose. For the low-dose group, the rates at the four time intervals, as estimated from the fitted marginal model, are 0.1830, 0.2764, 0.3928, and 0.5210. For the high-dose group, the rates are 0.1997, 0.3609, 0.4963, and 0.5701. In other words, the rate increases over time for both treatments, and the rates of increase are slightly different.

You can request subject-specific weights by specifying the TYPE=SUBLEVEL option. The results (not shown here) are similar to the results from the observation-specific weighted method. Both methods provide unbiased estimates if the missingness model is specified correctly.

Because large weights can impact the parameter estimates, you should check the distribution of the estimated weights. You can trim large weights by specifying the MAXWEIGHT= option in the MISSMODEL statement. Figure 4 shows that the weights in this example range between 1 and 2.1, so no trimming is needed.

Figure 4 Distribution of Estimated Weights



Syntax for the GEE Procedure

The syntax for PROC GEE closely follows that of PROC GENMOD. The main statements available in the GEE procedure are as follows:

> **PROC GEE** < *options* > **;**
> **CLASS** *variable* < **(***options***)** > . . . < *variable* < **(***options***)** > > < / *options* > **;**
> **MISSMODEL** < *effects* > < / *options* > **;**
> **MODEL** *response* **=** < *effects* > < / *options* > **;**
> **REPEATED SUBJECT=***subject-effect* < / *options* > **;**

The PROC GEE statement invokes the procedure. The PROC GEE and MODEL statements are required. The MODEL statement specifies the variables to be used in the regression. The CLASS statement specifies which explanatory variables to treat as categorical variables.

To request the usual GEE approach for longitudinal data analysis, you use the REPEATED statement to specify the working correlation structure of repeated responses. When there are missing observations caused by dropouts, you can request the weighted GEE approach by using the MISSMODEL statement. This statement specifies a logistic regression model for missingness to obtain the weights.

For the full syntax, see the chapter “The GEE Procedure” in *SAS/STAT User’s Guide*.

Summary

The GEE procedure extends the generalized estimating equation functionality of the GENMOD procedure to include weighted methods that provide valid inference when missing data occur under the missing at random (MAR) assumption. The generalized estimating equation approach in the GENMOD procedure is commonly used to estimate population-averaged effects in a marginal model for longitudinal data. When missing data that are caused by dropouts are MAR, this approach can lead to biased parameter estimates. Using the weighted GEE methods in the GEE procedure can produce unbiased estimates.

PROC GEE is experimental in SAS/STAT 13.2. This release does not include all of the capabilities provided by the REPEATED statement in the GENMOD procedure, and additional features will be included in future releases.

REFERENCES

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). “Regression Models for Longitudinal Binary Responses with Informative Drop-Outs.” *Journal of the Royal Statistical Society, Series B* 57:691–704.

Liang, K.-Y., and Zeger, S. L. (1986). “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73:13–22.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: National Academies Press.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). “Performance of Weighted Estimating Equations for Longitudinal Binary Data with Drop-Outs Missing at Random.” *Statistics in Medicine* 21:3035–3054.

Robins, J. M., and Rotnitzky, A. (1995). “Semiparametric Efficiency in Multivariate Regression Models with Missing Data.” *Journal of the American Statistical Association* 90:122–129.

Rubin, D. B. (1976). “Inference and Missing Data.” *Biometrika* 63:581–592.

Acknowledgments

The authors thank Anne Baxter and Ed Huddleston for editorial assistance.

Contact Information

Your comments and questions are valued and encouraged. You can contact the authors at the following addresses:

Guixian Lin	Robert N. Rodriguez
SAS Institute Inc.	SAS Institute Inc.
SAS Campus Drive	SAS Campus Drive
Cary, NC 27513	Cary, NC 27513
Guixian.Lin@sas.com	Bob.Rodriguez@sas.com