Review of Software to Fit Generalized Estimating Equation Regression Models

Nicholas J. HORTON and Stuart R. LIPSITZ

Researchers are often interested in analyzing data that arise from a longitudinal or clustered design. Although there are a variety of standard likelihood-based approaches to analysis when the outcome variables are approximately multivariate normal, models for discrete-type outcomes generally require a different approach. Liang and Zeger formalized an approach to this problem using generalized estimating equations (GEEs) to extend generalized linear models (GLMs) to a regression setting with correlated observations within subjects. In this article, we briefly review GLM, the GEE methodology, introduce some examples, and compare the GEE implementations of several general purpose statistical packages (SAS, Stata, SUDAAN, and S-Plus). We focus on the user interface, accuracy, and completeness of implementations of this methodology.

KEY WORDS: Computer software for statistical analysis; Generalized estimating equations; Missing data.

1. INTRODUCTION

Generalized linear models (GLMs) (McCullagh and Nelder 1989) are a standard method used to fit regression models for univariate data that are presumed to follow an exponential family distribution. Frequently researchers are interested in analyzing data that arise from a longitudinal, repeated measures or clustered design, and there exists correlation between observations on a given subject. If the outcomes are approximately multivariate normal, then there are well established methods of analysis (Laird and Ware 1982) that have been widely implemented in general purpose statistical packages. But if the outcomes are binary or counts, general likelihood based approaches are less tractable. For clustered binary outcomes, several approaches have been suggested (e.g., Fitzmaurice and Laird 1993), but these have not been incorporated into general purpose statistical computing packages.

Generalized estimating equations (GEEs) were developed to extend the GLM to accommodate correlated data, and are widely used by researchers in a number of fields. In this article we will review GLMs and the GEE methodology, and through an example, compare the GEE implementations of several general purpose statistical packages (including SAS, Stata, SUDAAN, and S-Plus). We will begin by briefly reviewing the methodology.

2. BRIEF REVIEW OF GLM'S AND GEE'S

McCullagh and Nelder (1989) introduced the GLM for exponential family data with the form

$$f_Y(y, \theta, \phi) = \exp\left\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\right\},$$

where $a(.)$, $b(.)$, and $c(.)$ are given, θ is the canonical parameter, and ϕ is the dispersion parameter. The GLM is then given by

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

where $\mathbf{x}_i$ is a $p \times 1$ vector of covariates for the ith subject, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. One of the attractive properties of the GLM is that it allows for linear as well as non-linear models under a single framework. It is possible to fit models where the underlying data are normal, inverse Gaussian, gamma, Poisson, binomial, geometric, and negative binomial by suitable choice of the link function $g(.)$ (Hilbe 1994).

Liang and Zeger (1986) and Zeger and Liang (1986) introduced generalized estimating equations (GEEs) to account for the correlation between observations in generalized linear regression models. One aspect of their approach builds upon previous methods of variance estimation developed to protect against inappropriate assumptions about the variance (Huber 1967; White 1980, 1982). GEEs are used to characterize the marginal expectation of a set of outcomes as a function of a set of study variables. In a marginal model, the analyst is interested in modeling the marginal expectation (average response for observations sharing the same covariates) as a function of explanatory variables. Diggle, Liang, and Zeger (1994) provided a detailed review of marginal models as well as other approaches (including random effects models and transition (markov) models).

Let $Y_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, t$ be the jth outcome for the ith subject, where we assume that observations on different subjects are independent, though we allow for association between outcomes observed on the same subject. In the GEE setting, we are not assuming that Y_{ij} is a member of the exponential family, but we are assuming that the mean and variance are characterized as in the GLM. We

Nicholas J. Horton is Research Fellow, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (Email: horton@hsph.harvard.edu). Stuart R. Lipsitz is Associate Professor, Dana Farber Cancer Institute and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. We are grateful for the support provided by NIMH grant R01-MH54693 and NIH grants CA 55576 and CA 55670. We would like to thank Nan Laird, Vincent Carey, Rick Williams, James Hardin and Gordon Johnston, as well as the Section Editor for their helpful comments. We also thank Gwen Zahner for use of the Connecticut child surveys, which were conducted under contract to the Connecticut Department of Children and Youth Services.

Table 1. Common Working Correlation Models

Structure	Definition	Example	# Parameters
Independence	$R_{u,v} = 1$ if $u = v$; $= 0$ otherwise	$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$	0
Exchangeable	$R_{u,v} = 1$ if $u = v$; $= \rho$ otherwise	$\begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}$	1
Unstructured	$R_{u,v} = 1$ if $u = v$; $= \rho_{u,v}$ otherwise	$\begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,t} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,t} & \rho_{2,t} & \cdots & 1 \end{pmatrix}$	t(t-1)/2
Auto-regressive	$R_{u,v} = 1$ if $u = v$; $= \rho^{\lvert u-v \rvert}$ otherwise	$\begin{pmatrix} 1 & \rho & \cdots & \rho^{t-1} \\ \rho & 1 & \cdots & \rho^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \cdots & 1 \end{pmatrix}$	1
M-dependent	$R_{u,v} = 1$ if $u = v$; $= \rho_{\lvert u-v \rvert}$ otherwise	$\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{t-1} \\ \rho_1 & 1 & \cdots & \rho_{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{t-1} & \rho_{t-2} & \cdots & 1 \end{pmatrix}$	$0<M \leq t-1$
Fixed	$R_{u,v} = 1$ if $u = v$; $= r_{u,v}$ otherwise	$\begin{pmatrix} 1 & r_{1,2} & \cdots & r_{1,t} \\ r_{1,2} & 1 & \cdots & r_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,t} & r_{2,t} & \cdots & 1 \end{pmatrix}$	0 (User specified)

assume the marginal regression model

$$g(E[Y_{ij}]) = \mathbf{x}'_{ij}\boldsymbol{\beta}, \tag{1}$$

where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of study variables (covariates) for the ith subject at the jth outcome, $\boldsymbol{\beta}$ consists of the p regression parameters of interest and $g(.)$ is the link function. Common choices for the link function might be $g(a) = a$ for measured data (the identity link) $g(a) = \log(a)$ for count data (log link), or $g(a) = \log(a/(1-a))$ for binary data (logit link). Since likelihood methods for binary data do not commonly exist in general purpose statistical software, GEEs have been a popular approach to regression model fitting for this type of data. For binary data with the logit link, we have that

$$\log(E[Y_{ij}]/(1 - E[Y_{ij}])) = \mathbf{x}'_{ij}\boldsymbol{\beta},$$

which implies that

$$E[Y_{ij}] = \mu_{ij} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})},$$

and since the outcomes are binary, we have that

$$\text{var}(Y_{ij}) = V_{ij} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{(1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}))^2}.$$

In addition to this marginal mean model, we need to model the covariance structure of the correlated observations on a given subject. Assuming no missing data, the $t \times t$ covariance matrix of $\mathbf{Y}_i$ is modeled as

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

where $\mathbf{A}_i$ is a diagonal matrix of variance functions $v(u_{ij})$, and $\mathbf{R}(\boldsymbol{\alpha})$ is the working correlation matrix of $\mathbf{Y}_i$ indexed by a vector of parameters $\boldsymbol{\alpha}$. We will now describe specifications for $\mathbf{R}$.

2.1 Specification of Working Correlation Matrix

There are a variety of common structures that may be appropriate to use to model the working correlation matrix. Table 1 displays a number of such matrices.

Issues guiding the choice of correlation structures are beyond the scope of this article (see Diggle et al. 1994 for a readable discussion), but in general if the number of observations per cluster is small in a balanced and complete design, then an unstructured matrix is recommended. For datasets with mistimed measurements, it may be reasonable to consider a model where the correlation is a function of the time between observations (i.e., M-dependent or auto-regressive). For datasets with clustered observations (i.e., rat litters), there may be no logical ordering for observations within a cluster and an exchangeable structure may be most appropriate.

Comparisons of estimates and standard errors from several different correlation structures may indicate sensitivity

```
TYPE=FIXED(1.0 0.9 0.8 0.7
           0.9 1.0 0.9 0.8
           0.8 0.9 1.0 0.9
           0.7 0.8 0.9 1.0)
```

Figure 1. Example of Fixed Working Correlation Matrix.

to mispecification of the variance structure. For both the independence working structure and the fixed working structure, no estimation of α is performed (for the fixed structure, the user must specify a $t \times t$ matrix `mat`). We note that use of the exchangeable (also referred to as compound symmetry) working correlation matrix with measured data and identity link function is equivalent to a random effects model with a random intercept per cluster.

For the `corr=fixed` option, `mat` should be a symmetric matrix with 1's on the diagonal, as seen in Figure 1, which specifies a banded structure with a fixed correlation and linear decline as the distance between observations increases.

2.2 Empirical and Model Based Variance Estimators

Zeger and Liang (1986) referred to $\mathbf{V}_i$ as a "working" matrix because it is not required to be correctly specified for the parameter estimates and the estimated variance of the parameter estimates in model (1) to be consistent (as long as the mean model itself is correct and there is no missing data). However, Liang and Zeger (1986) showed that there can be important gains in efficiency realized by correctly specifying the working correlation matrix.

A set of estimating equations are solved (through an iterative process) to find the value of the estimator $\hat{\boldsymbol{\beta}}$. An empirical variance estimator can be used to estimate $\mathrm{var}(\hat{\boldsymbol{\beta}})$. This variance estimator is also referred to as a "sandwich" or "robust" estimator. Another variance estimate available from GEE models is the model-based (or "naive") estimate, which is consistent when both the mean model and the covariance model are correctly specified. Since in general the analyst 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 < 20, the model based variance estimator may have better properties (Prentice 1988) even if the "working variance" is wrong. This is because the robust variance estimator is asymptotically unbiased, but could be highly biased when the number of clusters is small.

In addition to the Zeger "robust" estimator, SUDAAN also supports the Binder (1983) estimate of variance (Taylor series approximation) and a jackknife estimator of variance can be calculated using a working independence assumption.

2.3 Missing Data Issues

Longitudinal or clustered studies often have missing data, either by design or happenstance. If a litter in a teratology study is the level of clustering, litter size may vary between litters. Patients in an observational study may miss appointments or drop out of the study. The protocol for a clinical trial may call for patients to be observed at specified intervals, but their actual observations may take place at varying times. Such unbalanced and/or incomplete data can complicate GEE analyses. If the missingness can be thought of as being missing completely at random (MCAR) in the sense of Little and Rubin (1987), then the consistency results established by Liang and Zeger (1986) hold. However, the notation and calculations for arbitrary missing data patterns are more complicated than in the balanced and complete case.

Robins, Rotnitzky, and Zhao (1995) proposed methods to allow for data that is missing at random (MAR). Their inverse probability censoring weight (IPCW) approach requires that the missingness law be modeled, and that weights corresponding to the inverse probability of missingness be included in the GEE. This will yield consistent parameter estimates, but the variance will tend to be incorrect (since the weights are being estimated but are treated as constants by default). Unfortunately, the method of Robins et al. (1995) only works well when there is dropout—that is, once a subject misses a time, that subject is not seen again. Often subjects miss a single observation, and then are seen at the next time. The probability of the missingness pattern over time is not estimable with a simple logistic regression in this case, so the Robins et al. (1995) method is more difficult to implement. Lee, Laird, and Johnston (in revision) propose a modification to the GEE approach that combines restricted maximum likelihood (REML) estimating equations for the parameters in the variance-covariance matrix. We will consider how these approaches may be carried out in existing packages.

In summary, when fitting GEEs, the analyst must consider not only the model for the mean, but the model for the variance and the underlying missingness process. We will now describe the software packages to be reviewed, and describe how to carry out an analysis in each package.

3. SOFTWARE PACKAGES TO BE REVIEWED

We review four packages that are commonly used to fit GEEs: SAS, Stata, SUDAAN, and S-Plus.

SAS—The version of SAS used for the evaluation was SAS/STAT Release 6.12 (SAS Institute 1996). GEE support has been included in PROC GENMOD. Information about SAS is available from the SAS Institute web page (http://www.sas.com).

Stata—The version of Stata used for the evaluation was 5.0 (Stata Corp 1997). GEE models can be fit in Stata using the `xtgee` command, part of the `xt` cross-sectional time-series analysis tools. Information about Stata is available from the Stata Corporation web page http://www.stata.com.

SUDAAN—The version of SUDAAN used for the evaluation was 7.5.3 (Shaw, Barnwell, and Bieler 1997). Information about SUDAAN is available from the Research Triangle Institute web page (http://www.rti.org/patents/sudaan/sudaan.html). `PROC LOGISTIC`, `PROC MULTILOG`, and `PROC REGRESS` allow fitting and evaluation of models using GEEs for binary and continuous outcomes. Support for

Table 2. Syntax to Specify a Given Correlation Structure

Structure	SAS	Stata	SUDAAN	S-Plus
Independence	`corr=inde`	`corr(ind)`	`R=INDEPENDENT`	`corstr="independent"`
Exchangeable	`corr=exch`	`corr(exc)`	`R=EXCHANGE`	`corstr="exchangeable"`
Unstructured	`corr=un`	`corr(uns)`	Not available	`corstr="unstructured"`
Auto-regressive	`corr=ar`	`corr(ar 1)`	Not available	`corstr="ar1"`
M-dependent	`corr=mdep(m)`	`corr(sta m)`	Not available	Can be done
Fixed	`corr=fixed (mat)`	`corr(fix mat)`	Not available	Can be done

count data in `PROC LOGLINK` is planned for the next release. Because SUDAAN was developed for analysis of complex survey sampling data, it is particularly well suited to the analysis of repeated measures and clustered data.

S-Plus—The version of S-Plus used for the evaluation was 3.4 (Mathsoft 1996). Information about S-Plus is available from the MathSoft web page (http://www.mathsoft.com/splus). Although GEE support is not built in to S-Plus, a package to implement GEEs (YAGS or Yet Another GEE solver) is available from Vincent Carey and easily added as a library to S-Plus. The library can be found on the web (http://www.biostat.harvard.edu/∼carey) and precompiled binaries can be found at Brian Ripley's web page (http://www.stats.ox.ac.uk/pub/SWin/). Installation of YAGS took 10 minutes and the file (which includes the source, documentation, and test data) was approximately 1/3 of a megabyte in size. Another package available for analysis of repeated measures designs within S-Plus is the Oswald system, available at http://www.maths.lancs.ac.uk/Software/Oswald. We do not further discuss Oswald, and concentrate only on YAGS.

To fit a GEE, the analyst must first answer several questions: what is the appropriate family of distributions for the data (i.e., binary, count, or measured)? What link function is appropriate? What is a reasonable model for the correlation between observations? What is an appropriate mean model? What variance estimator should be used?

The analyst must specify both the distribution family (which determines the approach estimator of ϕ, the scale or dispersion parameter) and the working correlation matrix. SAS, Stata, and S-Plus all support the following distribution families: Gaussian (normal), Bernoulli/binomial, Poisson, and Gamma. Stata is planning support for the inverse Gaussian and negative binomial distribution families in a forthcoming release. SUDAAN supports only the Gaussian and Bernoulli/binomial distributions, though support for Poisson regression (PROC LOGLINK) is being implemented for the next release. SUDAAN also allows a cumulative logit link (for ordered multinomial outcomes) and a generalized logit link (for nominal multinomial regression).

The packages will default to the canonical link for each distribution, but other options are available in some packages. For example, for the binomial distribution, the probit link is available in Stata, while SAS and S-Plus support the probit and complementary log-log links (though support for these links is planned for release 6.0 of Stata). SUDAAN supports only the canonical (logit) link. Some caution must be exercised when choosing distributions and links, since some combinations do not make sense. S-Plus would not allow the combination of a binomial family with an identity link, though SAS and Stata fit the model with this combination.

All packages allow the specification of the mean model in a straightforward fashion.

Table 2 displays the commands to specify a given correlation structure in the packages under review. SAS and Stata can display the estimated working correlation matrix, while SUDAAN and S-Plus will display the elements of α from which the correlation matrix can be constructed for some set of observation times or clustering values. Since clusters have no natural ordering, and because SUDAAN is designed for analysis of clustered data, it only supports the independence and exchangeable working correlation structures. Finally, all packages will display both empirical and model-based variance estimates. The default for Stata is to display the model-based estimates, while SAS and SUDAAN default to the empirical (sandwich) estimates. S-Plus displays both estimates.

We now consider an example GEE model fit using these software packages.

4. EXAMPLE: MENTAL HEALTH SERVICE UTILIZATION

To conduct the software comparison, we analyzed data from a study of mental health utilization by children. The study design has been reported elsewhere (Zahner, Pawelkiewicz, DeFrancesco, and Adnopoz 1992; Zahner, Jacobs, Freeman, and Trainor 1993), as has a substantive analysis of the service utilization data (Zahner and Daskalakis 1997). Subjects included 2,519 children, aged 6–11, who were part of two cross-sectional surveys conducted in eastern Connecticut in the late 1980s. A goal of these surveys was to study determinants of mental health service utilization.

Parents of the children completed survey questionnaires that solicited information on child characteristics. The primary outcomes were service use in three settings: general health, school, and mental health. For a given setting, service use was defined as a parental report that the child had ever seen a provider or been in a special program for a behavioral problem. If the particular service was used, the outcome (SERV) was coded 1, and coded 0 otherwise. Clearly these binary outcomes are correlated for a given child.

In this study it is of interest to relate the rate of service use in the three settings to both child and family characteristics. Covariates thought to be predictive of ser-

OBS	ID	BOY	OLD	FAMSTRS	ACADPROB	SETTING	SCHOOL	MENTAL	GENERAL	SERV
1	M0111A02	0	0	1	0	0	1	0	0	0
2	M0111A02	0	0	1	0	1	0	1	0	0
3	M0111A02	0	0	1	0	2	0	0	1	0
4	M0111A06	0	0	1	0	0	1	0	0	0
5	M0111A06	0	0	1	0	1	0	1	0	0
6	M0111A06	0	0	1	0	2	0	0	1	0
7	M0111A08	1	0	0	0	0	1	0	0	0
8	M0111A08	1	0	0	0	1	0	1	0	0
9	M0111A08	1	0	0	0	2	0	0	1	0
10	M0111A10	0	0	1	0	0	1	0	0	0
11	M0111A10	0	0	1	0	1	0	1	0	0
12	M0111A10	0	0	1	0	2	0	0	1	0

Figure 2. First 12 Observations of the Service Use Dataset.

vice use included age (OLD: 0=age 6 to 8, 1=age 9 to 11), gender (BOY: 0=female, 1=male), and academic problems (ACADPROB: 0=no academic problems, 1=repeated a grade, advised to repeat grade).

In the logistic regression analysis for repeated binary measures we adjusted for setting (using indicators for SCHOOL and MENTAL; i.e., we used general services as baseline), the above covariates, and the interaction between setting and the three covariates. Our dataset consisted of one line per setting per subject, along with a variable ID (subject identifier), and a variable SETTING which was set to equal 2 if the setting was GENERAL, 1 if MENTAL, and 0 if SCHOOL. This variable is needed to determine the proper ordering of observations when calculating the working correlation matrix. Figure 2 displays the first dozen lines of the dataset, which includes three outcomes for each of four subjects. Table 3 displays the parameter estimates and variance estimates for those parameters (both empirical and model-based) for the model with an unstructured working correlation matrix and a logit link. We fit a model using an independence and exchangeable working correlation structure, but in this example, the parameter estimates and standard error estimates were identical to the first decimal place. We also fit a model in SUDAAN using the Binder variance estimate as well as a jackknife variance estimate using an independence working correlation structure. Both yielded similar results, though we note that the jackknife estimate took 25 minutes to calculate on a Pentium PC, as opposed to approximately one minute for the other models.

Table 4 displays the estimated working correlation matrices for the independence, exchangable and unstructured working correlation structures.

The parameter estimates were identical (to the third decimal place) for all packages fitting the unstructured model. Figure 3 displays the syntax needed to specify this model for the packages under review (since SUDAAN does not support the unstructured working correlation matrix, an exchangeable model was fit).

One question of interest in this model is whether the effect of the covariates on service use is the same across service settings. If the effect is the same, another question of interest is whether the covariate is associated with the outcome. We can test the former hypothesis for the OLD covariate with the null hypothesis:

$$H_0 : \text{OLD} * \text{MENTAL} = \text{OLD} * \text{SCHOOL} = 0.$$

We can construct a Wald test statistic $T = \tilde{\boldsymbol{\beta}}'(\widehat{\text{var}}(\tilde{\boldsymbol{\beta}}))^{-1}\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is a 2×1 vector containing the parameter estimates for OLD*MENTAL and OLD*SCHOOL and $\tilde{\boldsymbol{\beta}}$ are the variances and covariances for the parameters being tested. This test statistic will have an approximate χ^2 distribution with two degrees of freedom under the null hypothesis. If we do not reject this null hypothesis, we may be interested in testing

$$H_0 : \text{OLD} = \text{OLD} * \text{MENTAL} = \text{OLD} * \text{SCHOOL} = 0.$$

Similarly, a three df test statistic may be constructed to test this hypothesis.

It was trivial to test these hypotheses in Stata. Figure 4 gives the input and output from the extremely flexible `test` command in Stata. With the `accumulate` option, multidimensional hypothesis tests can be constructed. SUDAAN also allowed for testing of arbitrary hypotheses in this fashion. SAS allows testing of contrasts using the `CONTRAST` command, though the current version will not test the GEE model (this is planned to be rectified in a future release). In addition, SAS plans to add an ESTIMATE statement to provide estimates of linear functions of the regression parameters. For other types of tests, SAS and S-Plus require

Table 3. Parameter Estimates and Estimated Standard Errors With Unstructured Working Correlation Matrix

Parameter	*Estimate*	*Empirical std err*	*Model std err*
INTERCEPT	−2.944	.149	.145
MENTAL	− .352	.193	.194
SCHOOL	.185	.174	.171
OLD	.123	.144	.144
BOY	.365	.146	.147
ACADPROB	.724	.145	.146
OLD*MENTAL	.291	.190	.190
OLD*SCHOOL	.331	.162	.163
BOY*MENTAL	−.278	.189	.193
BOY*SCHOOL	−.154	.165	.167
ACADPROB*MENTAL	.184	.191	.193
ACADPROB*SCHOOL	1.136	.167	.168

*Table 4. Estimated Working Correlation Matrices for Different Working Correlation Structures (*Independence*, Exchangeable,* **Unstructured***)*

	School	*Mental*	*General*
School	1		
Mental	0 *.196* **.165**	1	
General	0 *.196* **.198**	0 *.196* **.227**	1

SAS

```
proc genmod data=gee;
    class id;
    model serv = mental school old boy acadpro
        old*mental old*school boy*mental boy*school
        acadpro*mental acadpro*school / dist=bin;
    repeated subject=id / type=un corrw within=setting;
    make 'classlevels' noprint;
    make 'geercov' out=rcov noprint;
run;
```

Stata

```
iis numid
tis setting
xi: xtgee serv i.old*mental i.old*school i.boy*mental i.boy*school
    i.acadpro*mental i.acadpro*school, link(logit) corr(unst)
    family(binomial) robust
xtcorr
```

SUDAAN

```
proc multilog data="gee" filetype=sas semethod=zeger r=exchangeable;
recode serv = (0 1) mental = (0 1) school =(0 1) old = (0 1)
    boy = (0 1) acadpro =(0 1);
nest _one_ numid;
weight _one_;
subgroup serv mental school old boy acadpro;
levels 2 2 2 2 2 2;
reflevel mental=1 school=1 old=1 boy=1 acadpro=1;
model serv = mental school old boy acadpro old*mental old*school
    boy*mental boy*school acadpro*mental acadpro*school /genlogit;
test waldchi;
setenv labwidth=25 colwidth=8 decwidth=4 maxind=4 linesize=78 pagesize=60
    colspce=2;
print beta="beta" sebeta="sebeta" / risk=all tests=default betafmt=f8.6
    sebetafmt=f8.6;
```

S-Plus

```
mygeeun _ yags(formula = serv ~ school + mental + old + boy +
    acadpro + old*mental + old*school + boy*mental + boy*school +
    acadpro*mental + acadpro*school, cor.met = setting, id = id,
    family = binomial, corstr="unstructured")
summary(mygeeun)
```

Figure 3. Syntax to Fit GEE's for Example Model.

the analyst to perform the matrix multiplication, but they do facilitate saving the estimated covariance matrix of the parameters to automate these computations. Figure 5 displays the code needed to calculate the value of the test statistic for the OLD covariate. Note that when the user requests the covariance matrix of the parameters using the "GEERCOV" make statement, SAS actually creates a matrix with covariances in the upper triangle, and correlations in the lower triangle.

We conclude from our example that that there is a significant interaction between service setting and academic problems ($\chi^2_2 = 52.3, p < .0001$) but not for age and setting ($\chi^2_2 = 4.6, p = .10$) or gender and setting ($\chi^2_2 = 2.2, p = 0.33$). Overall, boys have a higher proportion of mental health service use than girls ($\chi^2_3 = 8.2, p = .04$) and older children tend to have used them more than younger children ($\chi^2_3 = 20.6, p = .0001$).

Table 5. Complete Data From Artificial Example

Y_1	Y_2	X	Count
0	0	0	324
0	0	1	261
0	1	0	43
0	1	1	36
1	0	0	72
1	0	1	107
1	1	0	61
1	1	1	96

We will next consider an artificial data example with missing data.

5. EXAMPLE: MISSING DATA

We constructed an artificial dataset consisting of 1000 paired binary observations from the following underlying distribution:

$$\text{logit}(E[Y_{ij}]) = \beta_0 + \beta_1 * \text{TIME} + \beta_2 * X_i \qquad (2)$$

where X_i is a dichotomous covariate; TIME $= 1$ if $j = 2$, 0 otherwise; $\beta_0 = -1$; $\beta_1 = -.50$; $\beta_2 = .50$; and $\text{corr}(Y_1, Y_2) = .40$. Table 5 displays one sample from this underlying distribution which we used as our complete case sample. We fit models with an exchangeable (saturated in this setting) correlation matrix which yielded parameter estimates (and standard errors): $\hat{\beta}_0 : -.943(.092)$, $\hat{\beta}_1 : -.500(.080)$, $\hat{\beta}_2 : .500(.118)$. We created a dataset using the following (MAR) missingness law:

$$P(Y_2 \text{ is observed}|Y_1, X) = \begin{cases} .9 & \text{if } Y_1 = 1 \text{ and } X = 1 \\ .7 & \text{if } Y_1 = 1 \text{ and } X = 0 \\ .5 & \text{if } Y_1 = 0 \text{ and } X = 1 \\ .3 & \text{if } Y_1 = 0 \text{ and } X = 0 \end{cases}.$$

The missingness mechanism is considered to be missing at random (MAR) because it does not depend on the unobserved values of Y_2. We fit models using all available cases in all packages using an independence, exchangeable and unstructured working correlation assumption. Table 6 displays the parameter estimates from one sample generated using this missingness law. Because of differences in how the unstructured and exchangeable correlation matrices are estimated in the different packages, and because the dataset was unbalanced, the results are slightly different. The estimates under the working independence assumption are much different, however, and appear highly biased. Including a correlation between Y_1 and Y_2 in the "working variance" model (either exchangeable or unstructured) appears to reduce this bias somewhat because of the high correlation (.4) between Y_1 and Y_2.

```
. test oldm=0
 ( 1)  oldm = 0.0
           chi2(  1) =    2.35
         Prob > chi2 =    0.1254
. test olds=0,accumulate
 ( 1)  oldm = 0.0
 ( 2)  olds = 0.0
           chi2(  2) =    4.55
         Prob > chi2 =    0.1029
. test old=0,accumulate
 ( 1)  oldm = 0.0
 ( 2)  olds = 0.0
 ( 3)  old = 0.0
           chi2(  3) =   20.61
         Prob > chi2 =    0.0001
```

Figure 4. Stata Code to Calculate Value of Multidimensional Wald Test Statistic.

We considered two approaches to address the bias of GEE methods with MAR data: the IPCW method of Robins, Rotnitzky, and Zhao (1995), and the normal approximation technique of Lee et al. (in revision). For the Robins approach, we fit a saturated logistic model for $\widehat{P}(Y_2 \text{ is observed}|Y_1, X)$.

We conducted a simulation by creating 500 datasets using the above missingness law, and fit all three models (available case (AC), IPCW, and REML) to each dataset. Table 7 displays the average of the parameter estimates from each of the simulations, along with the parameters from the sample generated using model (2). We note that for this missingness law, there is considerable bias in the parameter estimates using the available case technique. Use of the GEE-REML technique decreases the bias, while use of the IPCW technique tends to eliminate the bias in this example. Figure 6 displays the SAS code to fit models using the IPCW approach. We note that the standard error estimates from PROC GENMOD will be biased for the true values because by default GENMOD assumes that the weights are known, when in fact they are estimated from the data.

6. ADDITIONAL NOTES ON THE SOFTWARE PACKAGES

We now consider some additional notes on the software packages. One minor problem was found with Stata and SUDAAN, which required id variables to be numeric. It was possible to recode our id variable to a unique integer, but this was an annoyance. Stata does not support weights, which would preclude use of the IPCW method for handling MAR data.

SUDAAN's roots in complex survey sampling are both an advantage and a disadvantage. One advantage relates to descriptive tables and statistics. While not exactly GEE modeling, SUDAAN facilitates the calculation of the correct standard error for descriptive statistics. As an example in teratology, it is straightforward to calculate the proportion of fetuses (clustered by litter) that are malformed, along with a confidence interval for that proportion using a sandwich variance estimate. One disadvantage of SUDAAN is that it has limited support for correlation structures (only independence and exchangeable) though it has support for multiple nesting levels, which is not supported directly by any of the other packages. For example, consider a dataset which consisted of three repeated measurements over time

```
proc iml;
    use rcov;
    read all into rvar;
    p = ncol(rvar);
    do i=1 to p-1;  /* create covariance matrix (bottom triangle are *
        do j=i+1 to p; /* correlations, not covariances */
            rvar[ j,i] =rvar[ i,j] ;
        end;
    end;
    nvar = { 0 0,0 0} ;  /* pull off variances and covariances */
    nvar[ 1,1]  = rvar[ 7,7] ; /* that we need */
    nvar[ 1,2]  = rvar[ 7,8] ;
    nvar[ 2,1]  = rvar[ 8,7] ;
    nvar[ 2,2]  = rvar[ 8,8] ;
    beta = { 0.2905,0.3305} ; /* parameters to test */
    t = beta` * inv(nvar) * beta;
    print t;
run;
```

Figure 5. SAS Code to Calculate Value of Multidimensional Wald Test Statistic (not necessary in future versions of SAS).

Table 6. Parameter Estimates From Missing Data Example

Package	*Correlation*	*Int*	*Time*	*X*
All	Independence	−.9104	−.2508	.4424
SAS	Exchangeable	−.9574	−.5984	.5262
Stata	Exchangeable	−.9574	−.5984	.5262
SUDAAN	Exchangeable	−.9442	−.5075	.5034
S-Plus	Exchangeable	−.9574	−.5984	.5262
SAS	Unstructured	−.9574	−.5984	.5262
Stata	Unstructured	−.9571	−.5962	.5256
S-Plus	Unstructured	−.9571	−.5962	.5256

on each of two parents (mothers and fathers) in a family. Let the observed data be arranged such that $\mathbf{Y}_i = (\mathbf{Y}'_{i\mathrm{M}}, \mathbf{Y}'_{i\mathrm{F}})'$ is the vector of observations for the ith cluster, where $\mathbf{Y}_{i\mathrm{M}}$ represents the vector of three observations on the mother, and $\mathbf{Y}_{i\mathrm{F}}$ represents the vector of three observations on the father. Given enough clusters, it might be feasible to consider estimating an unstructured 6×6 correlation matrix, which can be done in SAS, S-Plus, or Stata. Use of a fixed matrix might also be reasonable to consider. While arbitrary, it might be plausible to consider that observations on the same person would be relatively highly correlated over time (.5 if one time point apart, .4 if two time points apart), that observations at the same time on different parents would be less highly correlated (.2), and that observations on different parents at different times would have a small correlation (.1). This could be implemented using a fixed working correlation matrix of the form

$$R = \begin{pmatrix} 1.0 & .5 & .4 & .2 & .1 & .1 \\ .5 & 1.0 & .5 & .1 & .2 & .1 \\ .4 & .5 & 1.0 & .1 & .1 & .2 \\ .2 & .1 & .1 & 1.0 & .5 & .4 \\ .1 & .2 & .1 & .5 & 1.0 & .5 \\ .1 & .1 & .2 & .4 & .5 & 1.0 \end{pmatrix}$$

SUDAAN allows multiple levels of clustering and avoids use of arbitrary fixed matrices or estimation of many nuisance parameters. The program allows a robust variance to be calculated at one stage of the design (e.g., parents) and an exchangeable correlation to be calculated at a lower level (time within parents) using the NEST statement.

The S-Plus GEE implementation offers the most general coding for working correlation matrices, which allows for arbitrarily complex or special-case models. The disadvantage of this approach is that some additional work is required by the analyst to take advantage of this versatility.

Table 7. Results From Missing Data Simulation

Parameter	*Actual sample*	*AC*	*IPCW*	*REML*
β_0	−.943	−.980	−.941	−.980
β_1	−.500	−.596	−.502	−.540
β_2	.500	.569	.498	.570

```
libname d '.';
data miss; set d.artif; /* create missing data */
    unif = ranuni(0);
    if (unif<.1 and t1=1 and x=1) then t2=.;
    if (unif<.3 and t1=1 and x=0) then t2=.;
    if (unif<.5 and t1=0 and x=1) then t2=.;
    if (unif<.7 and t1=0 and x=0) then t2=.;
run;
data miss; set miss; /* create indicators of missingness plus interaction */
    r=0;
    if t2=. then r=1;
    t1x = t1*x;
    keep r t1 t2 x t1x;
run;
proc logistic data=miss descending; /* calculate probability of missingness */
    model r = t1 x t1x; /* saturated model */
    output out=pred pred=pred; /* save model-based probabilities */
run;
data weight; set pred; /* add weights and transpose */
    weight=1/(1-pred); id=_N_; time=0; y=t1; output;
    weight=1/(1-pred); id=_N_; time=1; y=t2; output;
run;
data weight; set weight; /* only keep complete cases */
    if r=0;
run;
proc genmod data=weight;
    class id;
    scwgt weight; /* weight observations in GEE */
    model y = time x /dist=bin;
    repeated subject=id / type=un corrw;
    make 'classlevels' noprint;
    make 'GEEEmpPest' out=wt; /* save parameter estimates */
run;
```

Figure 6. Syntax to Fit IPCW GEEs With MAR Data in SAS.

All vendors provided quick (within four to five hours in all cases), cordial, and accurate responses to our emailed requests for technical support. SUDAAN had the most extensive documentation regarding GEE's, though the offerings for both SAS and Stata were complete and well organized. The documentation for YAGS was more fragmentary and required more basic knowledge on the part of the analyst. All packages provided examples to augment their reference documentation.

7. DISCUSSION

In this article we have reviewed the methodology underlying GEEs, fit models for two examples using four software packages, and considered extensions to handle missing at random data. In general, the packages were easy to use, the implementations were similar, and the results calculated were similar. Analysis of a dataset with missing observations yielded somewhat differing results between packages, and may warrant further investigation. Certain aspects of the interface (i.e., calculating multidimensional hypothesis tests, use of weights, etc.) were particularly well-implemented in certain packages. On the whole, GEEs are well-supported by all of these software packages, and are straightforward to use.

[Received January 1999. Revised February 1999.]

REFERENCES

Binder, D. A. (1983), "On the Variances of Asympototically Normal Estimators From Complex Surveys," *International Statistical Review*, 51, 279–292.

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

Fitzmaurice, G. M., and Laird, N. M. (1993), "A Likelihood-Based Method for Analysing Longitudinal Binary Responses," *Biometrika*, 80, 141–151.

Hilbe, J. M. (1994), "Generalized Linear Models," *The American Statistician*, 48, 255–265.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 221–233.

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Lee, H., Laird, N. M., and Johnston, G. (in revision), "Combining GEE and REML for Estimation of Generalized Linear Models With Incomplete Multivariate Data."

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. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

MathSoft (1996), *Splus Version 3.4, Supplement*, Seattle, WA: Data Analysis Products Division.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Binary Observation," *Biometrics*, 44, 1033–1048.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

SAS Institute (1996), *SAS/STAT Software: Changes and Enhancements for Release 6.12*, Cary, NC: SAS Institute, Inc.

Shah, B. V., Barnwell, B. G., Bieler, G. S. (1997), *SUDAAN User's Manual, Release 7.5*, Research Triangle Park, NC: Research Triangle Institute.

StataCorp (1997), *Stata Statistical Software: Release 5.0*, College Station, TX: Stata Corporation.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

——— (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Zahner, G. E. P., and Daskalakis, C. (1997), "Factors Associated With Mental Health, General Health and School-Based Service Use for Psychopathology," *American Journal of Public Health*, 87, 1440–1448.

Zahner, G. E. P., Jacobs, J. H., Freeman, D. H., and Trainor, K. (1993), "Rural-Urban Child Psychopathology in a Northeastern U.S. State: 1986–1989," *Journal of the American Academy of Child Adolescent Psychiatry*, 32, 378–387.

Zahner, G. E. P., Pawelkiewicz, W., DeFrancesco, J. J., and Adnopoz, J. (1992), "Children's Mental Health Service Needs and Utilization Patterns in an Urban Community," *Journal of the American Academy of Child Adolescent Psychiatry*, 31, 951–960.

Zeger, S. L., and Liang, K.-Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 42, 121–130.