Behavior Research Methods, Instruments, & Computers
2000, 32 (2), 269-273

Regression analysis of correlated binary outcomes

CHING-FAN SHEU
DePaul University, Chicago, Illinois

The purpose of this paper is to describe and illustrate a regression approach to the analysis of correlated binary outcomes (Liang & Zeger, 1986). Ignoring the correlations between repeated observations can lead to invalid inferences. This approach extends logistic regression to account for repeated observations in each of a series of individuals. In this paper, I present a nontechnical introduction to the generalized estimating equations (GEE) approach. A fictitious example is used to demonstrate that GEE regression correctly adjusts for the correlations between repeated binary observations. The approach is illustrated with an analysis of safer sex practices among high-risk teenagers.

Repeated measures and longitudinal designs are important research methods in the study of behavior and development (Nesselroade & Baltes, 1979). For the analysis of repeated (continuous) measurements, the linear models (linear regression and analysis of variance) have been extended to deal with longitudinal data assuming normality (Ware, 1985).

The outcome variables in behavioral research are often binary or discrete (Likert scale). Consider, for example, that two groups of subjects are observed repeatedly over time when the behavioral response is dichotomous (the use of condoms, say), and time-dependent covariates (such as drug use, gang membership) are also collected. The researcher may wish to know, for instance, whether the propensity of condom use differs between gender groups and whether the difference changes with drug use or being a gang member. When the research question is to express the change in average response over all individuals in the population with the same covariate values, then regression models are suitable to address it. Because multiple observations on the same individual tend to be correlated, extensions of the regression model to account for repeated measurements require advances in both estimation procedures and in software implementation.

Although the generalized estimating equations (GEE) approach (Liang & Zeger, 1986) to fit regression models to repeated binary data has become readily accessible in many computer packages (Burton, Gurrin, & Sly, 1998; Horton & Lipsitz, 1999), it has not yet been widely adopted in psychological research. A goal of this paper is to facilitate the application of this useful methodology in longitudinal studies of behavior.

The author thanks Gary Harper and Lisa Carver for the use of safer sex data, Gwowen Shieh, George Michel, and Sue O'Curry for comments on an earlier version of this paper, and Patrick Onghena and two anonymous referees for comments and suggestions, which resulted in a clearer exposition. Correspondence should be addressed to C.-F. Sheu, Department of Psychology, DePaul University, 2219 North Kenmore Ave., Chicago, IL 60614-3522 (e-mail:csheu@depaul.edu).

I begin by reviewing why the correlations between repeated observations must be taken into proper account to avoid making misleading conclusions. The second section describes the GEE approach in nontechnical terms. In the third section, I show, by a fictitious example, that GEE yields correct regression estimates and standard error estimates for correlated binary data. The fourth section illustrates the method by an analysis of a longitudinal study of safer sex practices among at-risk teenagers.

Repeated Measures and Correlated Data

Repeated measures designs are familiar research tools to psychologists. The repeated observations over experimental conditions on each of many participants bring efficiency and increased power to treatment comparisons. The advantage of this experimental design arises from allowing each participant to serve as his or her own control. The gain, however, comes with an analytic cost associated with the need to account for the dependence between observations from the same individuals. The following example is a simplified version given by Dunlop (1994) to illustrate the incorrect inferences that can result from ignoring the correlations between repeated observations.

Consider a simple experiment in which we have 1 male subject and 1 female subject, and each subject is measured twice, once before and once after a treatment. This is a two-factor (one within and one between) repeated measures design. Denote the responses by $Y_{m,b}$, $Y_{m,a}$, $Y_{f,b}$, and $Y_{f,a}$. The subscripts are the first letters of the factor labels. We assume that the measurements between subjects are independent; however, within the same subject, the correlation between two measurements equals ρ. The gender effect can be estimated by $\frac{1}{2}[(Y_{m,b} + Y_{m,a}) - (Y_{f,b} + Y_{f,a})]$. The variance of the gender effect estimate is $\sigma^2(1 + \rho)$, where σ^2 is the error variance of the measurement. On the other hand, the treatment effect can be estimated by $\frac{1}{2}[(Y_{m,a} - Y_{m,b}) + (Y_{f,a} - Y_{f,b})]$. The variance of the treatment effect estimate is $\sigma^2(1 - \rho)$. Typically, multiple measurements within the same subject are positively correlated ($\rho > 0$). Thus, the variance of gender effect is underestimated if within-subjects responses are

Table 1
Artificial Data of 8 Subjects in a Two-Time-Period Longitudinal Study

1	0	0
1	1	0
2	0	0
2	1	0
3	0	0
3	1	1
4	0	0
4	1	1
5	0	0
5	1	1
6	0	0
6	1	1
7	0	1
7	1	1
8	0	1
8	1	1

Note—The variables are, columnwise from left to right, subject identification number, time (0 = first period, 1 = second period), and binary response.

assumed to be independent ($\rho = 0$). With a smaller standard error estimate, a significance test will reject the null hypothesis of no gender effect too often (larger Type I error). On the other hand, if the positive correlation is ignored, the variance of treatment effect is overestimated. Testing the null hypothesis of no treatment effect will result in a larger Type II error.

Since analysis of variance models can be expressed as regression models, the above argument extends to ordinary linear regression for many observations. In short, ignoring correlations between repeated observations may lead to invalid inferences about the regression coefficients.

Generalized Estimating Equations

The GEE approach to modeling longitudinal data relates the population means of a set of responses as a function of the explanatory variables. The approach focuses on regression for discrete and continuous outcomes and treats the association across time of the repeated responses for a subject as a nuisance. The regression coefficients are estimated without completely specifying the joint distribution of the multivariate responses; but the parameters of the within-subjects correlations are explicitly accounted for in the estimating procedure. A detailed review of this approach to analysis of longitudinal data is provided by Diggle, Liang, and Zeger (1994). Here, we give a brief explanation of GEE in nontechnical terms.

Estimating equations refer to a set of equations the solutions of which give estimates of parameters. For example, the least squares estimates for linear regression are obtained from solving the normal equations. In generalized linear models (for binary and count data), the parameter estimates are obtained by solving likelihood equations (McCullagh & Nelder, 1989). To account for correlated measurements in longitudinal designs, Liang and Zeger (1986) incorporated a correlation matrix of the outcomes on the same individual into the estimating functions of generalized linear models and showed that the solution to the GEE gives a consistent estimate of the regression parameters that is multivariate normal for large samples. The GEE parameter estimates are computed using the method of iteratively reweighted least squares. Two different types of standard errors of the regression parameters are available. The first is a model-based estimator that assumes a correctly specified correlation matrix. The second is an empirically based estimator that uses a robust variance estimator to allow for the possibility that the choice of the "working" correlation matrix may be incorrect. Zeger and Liang (1986) showed that the confidence intervals for the regression parameter estimates will be correct for large samples even if the correlation structure is misspecified. Because the GEE approach is not a maximum likelihood method, it is not possible to derive goodness-of-fit measures, such as deviance, to compare different working correlation matrices to determine which one is most suitable.

Horton and Lipsitz (1999) provided a list of common working correlation structures and reviewed the GEE implementations of general-purpose statistical packages, such as SAS, Stata, and S-Plus.

Correlated Binary Regression

We analyze a fictitious data set to illustrate the effect of ignoring correlation on the variance of the regression slope estimate and to verify that the GEE empirical-based variance estimator is not sensitive to the choice of the correlated structure. Table 1 lists binary responses at two time points for a group of 8 subjects. The data set consists of one line per time per subject, along with a subject identification number.

Following a logistic regression framework, we let the responses of a subject $Y_{i,j}$ be a two-dimensional binary vector and let the mean response vector $E(Y_{i,j}) = \pi_{i,j}$, where $\pi_{i,j}$ is the probability of a "success" response for subject i at the jth time period. The correlation between two responses of the same subject $Y_{i,1}$ and $Y_{i,2}$ equals ρ

Listing 1
Specifications of Logistic Regression and GEE Regression Models in SAS

```
PROC LOGISTIC;
  MODEL response = time;
  TITLE2 'Logistic Regression';

PROC GENMOD;
  CLASS subject time;
  MODEL response = time / DIST=BINOMIAL LINK=LOGIT;
  REPEATED SUBJECT=subject / TYPE=IND CORRW MODELSE;
  TITLE2 'Independent correlation';

PROC GENMOD;
  CLASS subject time;
  MODEL response = time / DIST=BINOMIAL LINK=LOGIT;
  REPEATED SUBJECT=subject / TYPE=EXCH CORRW MODELSE;
  TITLE2 'Exchangeable correlation';

RUN;
```

Table 2
Comparison of Slope Estimates and Standard Error (*SE*) Estimates

	Model-Based		Empirical-Based	
Model	Estimate	*SE*	Estimate	*SE*
Logistic regression	−2.1927	1.1547		
GEE regression				
Independent	−2.1927	1.1547	−2.1972	0.9428
Exchangeable	−2.1927	0.9428	−2.1972	0.9428

(ρ = .3333 for the data). Responses across subjects are assumed to be independent. With this set-up, a regression model for the mean response vector of any subject is

$$\ln\left(\frac{\pi_{i,j}}{1-\pi_{i,j}}\right) = \beta_0 + \beta_1 x_{i,j},$$

where $x_{i,j}$ is coded 0 for the first response of subject i and is coded 1 for the second response. Our interest in this example is to compare the slope parameter (time effect) estimates and standard error estimates using three different models.

First, we estimate the regression coefficients of a simple logistic regression in which the repeated observations within the same individual are assumed to be independent (i.e., ignoring the correlation). We then obtain the slope estimates and standard error estimates based on the GEE regression approach: (1) assuming an independent correlation structure and (2) assuming an exchangeable (compound symmetry) correlation structure where all the off-diagonal elements in the correlation matrix have the same value. This means that the correlation between distinct observations on the same individual is the same regardless of when in time the observations were taken. In repeated measures analysis of variance, the correlation matrix is often assumed to have this structure (see Max & Onghena, 1999, for a discussion on why this assumption might not be appropriate).

The GEE analysis based on the independent correlation structure proceeds as if the observations were independent (ρ = 0), except that using the empirical-based estimator will ensure a consistent variance estimation regardless of the actual degree of dependence. Using the model-based variance estimator produces estimates identical to those obtained from fitting the logistic regression to the data (which ignores the dependence of measurements from the same individual).

We use SAS PROC LOGISTIC to fit a logistic regression model and PROC GENMOD to fit GEE regression models to the fictitious data set. The model specifications in SAS statements are shown in Listing 1. A syntax synopsis to implement the GEE procedure can be obtained from ftp.sas.com/techsup/download/stat/gee.txt.

Table 2 presents the slope estimate (time effect) and standard error estimate for each of the three models. We summarize as follows: (1) When a model-based variance estimator is used, the GEE regression with independent correlation yields results identical to those of the logistic regression. (2) When an empirical-based variance estimator is used, the two GEE regression models produce the same results regardless of which correlation structure is specified. (3) The estimated standard error of the GEE regression (slope) coefficient is related to the logistic standard error estimate by $.9428 = 1.1547\sqrt{1 - .3333}$, where .3333 is the value of estimated working correlation with the exchangeable structure. This verifies the formula for the variance estimate of the treatment (time) effect presented in the first section.

Application: Condom Use Among At-Risk Teenagers

As an illustration, we analyzed a binary longitudinal data set collected by Harper and Carver (1999) in a study of safer sex practices among high-risk teenagers. This longitudinal study targeted high-risk youth from six suburban neighborhoods who were chronically truant or had dropped out of school, were displaced from their homes (i.e., run away or kicked out), and were using substances and/or were involved with gangs. Those youth who agreed to participate in the study completed a baseline interview prior to participating in a safer sex education workshop and then were followed for a year, with four subsequent follow-up surveys administered every 3 months. Each binary series indicates whether a teenager had abstained from sex or had always used a condom in intercourse in the past 3 months (1) or not (0) for 15 consecutive months. Two hundred twenty-seven teenagers participated in the study. A complete survey from a teenager has five repeated binary outcomes (safe), on the basis of self-ratings. The covariates of interest are the following (variable names are in brackets): gender (gender: 0 = female, 1 = male), and whether or not the teenager had been arrested (arrest), had run away from home (run-

Table 3
Data of the First 4 Subjects in the Safer Sex Practices Study

1	1	1	1	1	1	1	1	1	0	0
2	1	1	1	2	0	1	1	1	1	0
3	1	1	0	3	0	1	0	0	0	0
4	1	1	0	4	0	1	1	1	0	0
5	1	1	0	5	0	1	0	1	1	0
6	2	1	1	1	0	1	1	1	0	0
7	2	1	0	2	0	0	0	0	0	0
8	2	1	1	3	0	0	1	1	0	0
9	2	1	1	4	0	0	0	1	0	0
10	3	1	1	1	0	0	0	1	0	0
11	3	1	1	2	0	0	0	1	0	0
12	3	1	0	3	1	0	0	1	0	0
13	3	1	1	4	0	0	0	1	0	0
14	3	1	1	5	0	0	0	1	0	0
15	4	1	0	1	0	0	0	0	0	0
16	4	1	0	2	0	0	0	1	0	0
17	4	1	0	3	0	0	0	0	0	0
18	4	1	0	4	0	0	0	0	0	0
19	4	1	0	5	0	1	0	1	0	0

Note—The variables are, columnwise from left to right, observation number, subject identification number, gender, safe, time, arrest, runaway, gang, marijua, ampheta, lsd.

Listing 2
Specifications of Regression Models for Repeated Binary Data Using Generalized Estimating Equations Approach

```
PROC GENMOD DATA=condom;
 CLASS gender time subject;
 MODEL safe = gender arrest runaway gang marijua ampheta lsd
              / DIST=BINOMIAL LINK=LOGIT;
 REPEATED SUBJECT=subject / TYPE=IND CORRW WITHIN=time;
 TITLE2 'Independent correlation';

PROC GENMOD DATA=condom;
 CLASS gender time subject;
 MODEL safe = gender arrest runaway gang marijua ampheta lsd
              / DIST=BINOMIAL LINK=LOGIT;
 REPEATED SUBJECT=subject / TYPE=EXCH CORRW WITHIN=time;
 TITLE2 'Exchangeable correlation';

PROC GENMOD DATA=condom;
 CLASS gender time subject;
 MODEL safe = gender arrest runaway gang marijua ampheta lsd
              / DIST=BINOMIAL LINK=LOGIT;
 REPEATED SUBJECT=subject / TYPE=AR(1) CORRW WITHIN=time;
 TITLE2 'AR(1) correlation';

RUN;
```

away), had been in a gang (gang), had used marijuana (marija), had used amphetamines (ampheta), and had used LSD (lsd) (0, no; or 1, yes) in the past 3 months. Except for gender, the covariates are all within subject and time dependent. Table 3 displays data of the first 4 subjects in the study. Each row represents the vector of response and covariate information of a subject at a particular time period. Notice that the second subject did not have a record for the fifth time period. However, the interpretation and computation of the GEE regression estimates are not affected by the number of repeated observations, which may vary among subjects. Here, the missing data are assumed to be missing completely at random so that the results established by Liang and Zeger (1986) are still applicable. The modification of the GEE approach to allow for arbitrary missing data pattern is a complicated topic beyond the scope of this article (Robins, Rotnitzky, & Zhao, 1995).

The primary goals of the longitudinal study were to describe the risk behaviors and life experiences of high-risk youth in a suburban community and to determine various factors that impact their participation in sexual risk behavior. Here, we examine whether these covariates are adequate explanatory variables of safer sex practices. We fit a GEE (logistic) regression model for repeated binary (safe) responses, including all the covariates. To compare regression estimates under different working correlation structures, we use independent, exchangeable, first-order autoregressive, and unstructured working correlation structures. An unstructured correlation matrix places no constraint on the correlation between observations. For the autoregressive correlation matrix, the correlation among observations becomes smaller as the number of time lags increases.

Listing 2 displays the model specifications in SAS syntax segments. The code for the choice of an unstructured correlation is not shown. It can be obtained by using any code segment with the option TYPE=UN in the REPEATED statement. Table 4 shows the estimated exchangeable, autoregressive, and unstructured correlation matrices. It appears that the correlation estimates are moderate and nonnegligible. The unstructured and exchangeable correlation estimates are not very different.

Table 4
Estimated Correlation Matrices for Different Working Correlation Structures

$$\text{Exchangeable} = \begin{pmatrix} 1.0000 & .4223 & .4223 & .4223 & .4223 \\ .4223 & 1.0000 & .4223 & .4223 & .4223 \\ .4223 & .4223 & 1.0000 & .4223 & .4223 \\ .4223 & .4223 & .4223 & 1.0000 & .4223 \\ .4223 & .4223 & .4223 & .4223 & 1.0000 \end{pmatrix}$$

$$\text{Autoregressive} = \begin{pmatrix} 1.0000 & .4897 & .2398 & .1174 & .0575 \\ .4897 & 1.0000 & .4897 & .2389 & .1174 \\ .2398 & .4897 & 1.0000 & .4897 & .2389 \\ .1174 & .2398 & .4897 & 1.0000 & .4897 \\ .0575 & .1174 & .2389 & .4897 & 1.0000 \end{pmatrix}$$

$$\text{Unstructured} = \begin{pmatrix} 1.0000 & .4461 & .3816 & .3724 & .3468 \\ .4461 & 1.0000 & .6030 & .4505 & .5245 \\ .3816 & .6030 & 1.0000 & .5825 & .4185 \\ .3724 & .4505 & .5825 & 1.0000 & .5669 \\ .3468 & .5245 & .4185 & .5669 & 1.0000 \end{pmatrix}$$

Table 5
Parameter Estimates and Standard Error (*SE*) Estimates With Different Working Correlation Structures

	Independent		Exchangeable		Autoregressive	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	.3629	.2167	.5394	.1894	.5321	.1954
Sex	.3131	.2338	.2918	.2286	.2802	.2272
Arrest	−.3723	.2194	−.1168	.1835	−.1630	.1760
Run away	.4299	.1961	.1597	.1623	.2093	.1547
Gang	.0602	.2910	.0250	.2063	.0320	.2286
Marijuana	.0208	.2297	−.2415	.1701	−.2016	.1793
Amphetamine	−.7452	.2084	−.5740	.1793	−.6438	.1699
LSD	−.3687	.2014	−.4553	.1604	−.3833	.1663

Table 5 displays the parameter estimates and standard error estimates for GEE regression coefficients under the independent, exchangeable, and autoregression correlation matrices. We do not report the estimates using the unstructured correlation because they are very similar to those obtained using the exchangeable working correlation structure. Note that the regression estimates and the standard error estimates under the first-order autoregressive and exchangeable correlation assumption are quite comparable. Given the other covariates in the model, the regression estimate for amphetamine use is significantly different from zero ($p < .001$) under both exchangeable and autoregressive assumptions. With LSD use, the regression estimate is significantly different from zero at the .05 level. Both estimates are of negative value, indicating that the use of these two types of drugs is strongly associated with a reduced probability of engaging in safer sex practices. In particular, the odds are about 1.8 to 1 for amphetamine-free teenagers to be practicing safer sex, relative to their amphetamine-using counterparts. The other covariates do not appear to be significant predictors of safer sex practices among at-risk teenagers. However, a slightly different conclusion is drawn if the independent working correlation matrix is used. Under this model, the impact of LSD use on safer sex practice appears only marginally significant ($p \approx .067$), whereas that of running away from home becomes significant ($p \approx .028$). Amphetamine use remains the most important explanatory variable, and the other covariates are still not statistically significant. Currently, there are no established goodness-of-fit statistics to compare different working correlation matrices to determine which is most suitable. It seems wise not to downplay the importance of LSD use as a risk factor when interpreting the results of this study.

Conclusion

Repeated measures designs play an important role in psychological research. Psychologists are familiar with analysis of variance, both univariate and multivariate, for normal data collected from such experimental designs. When the research objective is to relate an outcome to other variables, regression methods are often the statistical models of choice. With nonnormal data, generalized linear models are employed. The extension of generalized linear models to account for correlated binary (and other discrete) responses means that the data collected in longitudinal behavioral studies can now be analyzed readily and in a valid manner. It is hoped that this paper has provided sufficient information for potential users of the GEE to tackle the computer packages for the purpose of their own research.

REFERENCES

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine*, **17**, 1261-1291.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford: Oxford University Press, Clarendon Press.

Dunlop, D. D. (1994). Regression for longitudinal data: A bridge from least square regression. *American Statistician*, **48**, 299-303.

Harper, G. W., & Carver, L. J. (1999). "Out-of-the-mainstream" youth as partners in collaborative research: Exploring the benefits and challenges. *Health Education & Behavior*, **26**, 250-265.

Horton, N. J., & Lipsitz, S. (1999). Review of software to fit generalized estimating equation regression models. *American Statistician*, **53**, 160-169.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Max, L., & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, & Hearing Research*, **42**, 261-270.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman & Hall.

Nesselroade, J. R., & Baltes, P. B. (Eds.) (1979). *Longitudinal research in the study of behavior and development*. New York: Academic Press.

Robins, J. M., Rotnitzky, A., & 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.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician*, **39**, 95-101.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-130.

(Manuscript received October 29, 1999;
revision accepted for publication February 25, 2000.)