Organizational Research Methods
http://orm.sagepub.com/

Using Generalized Estimating Equations for Longitudinal Data Analysis

Gary A. Ballinger

Organizational Research Methods 2004 7: 127
DOI: 10.1177/1094428104263672

The online version of this article can be found at:
http://orm.sagepub.com/content/7/2/127

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
RMD Academy of Management
The Research Methods Division of The Academy of Management

Additional services and information for *Organizational Research Methods* can be found at:

Email Alerts: http://orm.sagepub.com/cgi/alerts

Subscriptions: http://orm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://orm.sagepub.com/content/7/2/127.refs.html

>> Version of Record - Apr 1, 2004

What is This?

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Using Generalized Estimating Equations for Longitudinal Data Analysis

GARY A. BALLINGER
Purdue University

The generalized estimating equation (GEE) approach of Zeger and Liang facilitates analysis of data collected in longitudinal, nested, or repeated measures designs. GEEs use the generalized linear model to estimate more efficient and unbiased regression parameters relative to ordinary least squares regression in part because they permit specification of a working correlation matrix that accounts for the form of within-subject correlation of responses on dependent variables of many different distributions, including normal, binomial, and Poisson. The author briefly explains the theory behind GEEs and their beneficial statistical properties and limitations and compares GEEs to suboptimal approaches for analyzing longitudinal data through use of two examples. The first demonstration applies GEEs to the analysis of data from a longitudinal lab study with a counted response variable; the second demonstration applies GEEs to analysis of data with a normally distributed response variable from subjects nested within branch offices of an organization.

Keywords: *longitudinal regression; nested data analysis; generalized linear models; logistic regression; Poisson regression*

Organizational researchers who investigate topics such as absenteeism, innovation, turnover intentions, and decision making have often been forced to rely on suboptimal methods of analyzing their data because responses are generally not normally distributed. Researchers may either transform the response variable prior to conducting data analysis or use a method of aggregating their response variable so as to make the distribution of responses approximately normal. But these approaches sacrifice both precision in analysis and clarity in interpreting results (Gardner, Mulvey, & Shaw, 1995; Harrison, 2002).

A separate challenge comes in analyzing data that are correlated within subject, such as that provided in longitudinal studies and other studies in which data are clus-

---

Author's Note: The author would like to thank Bradley Alge, Jodi Goodman, Steve Green, and David Schoorman as well as two anonymous reviewers for their comments on earlier versions of this article. An earlier version of this article was presented at the 23rd annual Industrial/Organizational and Organizational Behavior Graduate Student Conference in Tampa, Florida.

Organizational Research Methods, Vol. 7 No. 2, April 2004 127-150
DOI: 10.1177/1094428104263672
© 2004 Sage Publications

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

tered within subgroups. Failure to incorporate correlation of responses can lead to incorrect estimation of regression model parameters, particularly when such correlations are large. The regression estimates (βs) are less efficient, that is, they are more widely scattered around the true population value than they would be if the within-subject correlation were incorporated in the analysis (Diggle, Heagerty, Liang, & Zeger, 2002; Fitzmaurice, 1995). To increase their confidence in regression results, researchers should use analytical methods that produce the most efficient parameter estimates that are also unbiased (cf. McCullagh & Nelder, 1989; Pindyck & Rubinfeld, 1998), that is, with an expected mean value that is the true population parameter that is being estimated. Organizational researchers faced with evaluation of dependent variables arising from longitudinal data collections instead can use a method that provides efficient and unbiased parameter estimates for analyzing data without transforming it and produces easily interpretable and communicable results that can be used to test hypotheses.

Harrison and Hulin (1989) identified generalized estimating equations (GEEs) as an analytic tool with promise for organizational research because the method accounted for correlation of responses within subject for response variables and was flexible enough for use in analyzing response variables that were not normally distributed. The GEE approach was developed by Liang and Zeger (1986) and Zeger and Liang (1986) to produce more efficient and unbiased regression estimates for use in analyzing longitudinal or repeated measures research designs with nonnormal response variables. The method has received wide use in medical and life sciences such as epidemiology, gerontology, and biology. Its application to date in organizational and psychological research has been more limited, and the purposes of this article are to describe GEEs and their statistical properties, briefly clarify the advantages and disadvantages of using GEEs over other methods for analyzing longitudinal data, and provide further information on the steps that users can take to apply GEEs to analyzing data and testing hypotheses applicable to organizational research. I also discuss some of the significant limitations of the method both as they apply to the specific context of the examples provided and as some of the general weaknesses of the approach.

The GEE algorithm has been incorporated into many major statistical software packages used by organizational researchers, including SAS, STATA, HLM, LIMDEP, and S-Plus, and the sample data sets were analyzed using both SAS and STATA. There are many similar steps that users must take to prepare their data for analysis using any software package, and in describing how to use GEEs, I will walk the reader through the various decisions that must be made to correctly specify the model and produce the appropriate results. Readers are referred to Horton and Lipsitz (1999) for a general review of software available to fit GEE regression. Hardin and Hilbe (2003) provide guidance on how to fit GEEs in STATA and SAS; Stokes, Davis, and Koch (2000) address fitting GEE regression in SAS; and Bryk, Raudenbush, and Congdon (2003) provide guidance on fitting GEEs in HLM.

Analysis of Limited-Range Dependent Variables

Organizational researchers frequently investigate choice behaviors such as absenteeism (do I show up to work today or not?) and turnover (do I leave my job or stay?). Other counted or choice behaviors investigated by researchers in organizational settings include the number of patents received by different firms (e.g., Ahuja & Katila,

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

2001; Ahuja & Lampert, 2001), market entry and exit decisions (Haveman & Nonnemaker, 2000), and the number of position upgrades granted to a division in a given year (Welbourne & Trevor, 2000). These outcomes fall into what Harrison (2002) referred to as "limited range dependent variables"; and ordinary least squares (OLS) regression analysis of them is complicated in part because the range of responses is constrained. In the case of choice behavior, we model the probability of a positive choice; the dependent variable can take on a value no larger than 1. In counted responses, the dependent variable must be greater than or equal to zero. Falsely assuming normality in such cases can lead to incorrect results (Gardner et al., 1995).

Using a linear regression approach to analyze choice behaviors is complicated as well by the fact that the regression estimates may violate other assumptions necessary for OLS. We may be faced with nonconstant variance for values of the dependent variable, particularly when modeling probabilities (McCullagh & Nelder, 1989; Pindyck & Rubinfeld, 1998) and the error terms are generally not normally distributed (Harrison, 2002). In many cases, using a power transformation on the dependent variable prior to running an OLS regression is also a suboptimal approach because it complicates interpretations of the parameter estimates. OLS may also use incorrect assumptions regarding the constancy of variance of the underlying distribution of the data that could lead to results that do not make sense given the distribution of the data (McCullagh & Nelder, 1989). For example, when modeling counted data, the dependent variable must always be positive, yet at some levels of the independent variables multiplied by the OLS regression estimates, negative values for the dependent variable could be produced (Gardner et al., 1995). Researchers have generally agreed that logit or probit models are appropriate for constructing regression models of binary choice behaviors (McCullagh & Nelder, 1989; Pindyck & Rubinfeld, 1998; Harrison, 2002), and Poisson or negative binomial regression using generalized linear models has been accepted as the best means of estimating probabilities in cases in which the dependent variable consists of counted data (Gardner et al., 1995).

Correlation of Responses

When faced with data that consist of repeated measures that may be correlated within a subject over repeated measures or within a cluster of observations in a particular group, researchers must account for the correlation within responses when estimating regression parameters. Otherwise, they can make incorrect inferences about the regression coefficients (because of incorrect estimation of the variances) and inefficient or biased estimates of the regression coefficients (Diggle et al., 2002) that could lead to incorrect conclusions regarding their research questions. Fitzmaurice (1995) demonstrated that when faced with an independent variable that varies within a cluster (referred to as a time-dependent covariate in longitudinal studies), "the efficiency of . . . estimators declines with increasing correlation, and the decline is most notable when the correlation is greater than .4" (p. 313). Efficiency losses were large as correlation increased, as the asymptotic relative efficiency of parameter estimates assuming independence fell to approximately 40% for within-cluster correlations of .5 or more. The errors are particularly large for cases in which the correlation within subject is highly positive or highly negative.

Repeated measures ANOVA approaches to the problem are inadequate because they do not use a model of the covariance among repeated observations to increase the

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

efficiency of the parameter estimates, they generally require balanced and complete data sets, they are restricted to the analysis of normally distributed response variables, and they do not allow for the analysis of covariates that change over time (Diggle et al., 2002). Given that users of statistical estimation approaches should be highly concerned with efficiency in the estimators, researchers conducting longitudinal research have relied on several tools to approach these problems, including time series regression (Pindyck & Rubinfeld, 1998) and linear regression with adjustments for nonindependence. OLS regression models have been adapted for analysis of correlated responses when the dependent variable is normally distributed. But in conducting regression analysis of panel-correlated binary or counted dependent variables, one needs to use the quasi-likelihood method based on generalized linear models (McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972; Wedderburn, 1974) known as GEEs.

GEEs

GEEs were developed by Liang and Zeger (1986) and Zeger and Liang (1986) as a means of testing hypotheses regarding the influence of factors on binary and other exponentially (e.g., Poisson, Gamma, negative binomial) distributed response variables collected within subjects across time. They are an extension of generalized linear models, which facilitate regression analyses on dependent variables that are not normally distributed (McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972). The approach that I focus on mostly in this article is the one in which GEE develops a population average or marginal model. Marginal models give an average response for observations sharing the same covariates as a function of the covariates (Zeger, Liang, & Albert, 1988). In other words, for every one-unit increase in a covariate across the population, GEE tells the user how much the average response would change (Zorn, 2001). GEEs estimate regression coefficients and standard errors with sampling distributions that are asymptotically normal (Liang & Zeger, 1986), can be applied to test main effects and interactions, and can be used to evaluate categorical or continuous independent variables. GEE estimates are the same as those produced by OLS regression when the dependent variable is normally distributed and no correlation within response is assumed. Test statistics have been developed that allow users to test hypotheses regarding parameter estimates in a method analogous to those used in testing coefficients from normal-errors regression methods (Rotnitzky & Jewell, 1990), including linear regression and repeated-measures ANOVA.

GEEs start with maximum-likelihood estimation of our regression parameters (β) and the variance calculated using a link function, which is a transformation function that allows the dependent variable to be expressed as a vector of parameter estimates ($y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots$) in the form of an additive model. The GEEs also use a variance function that is a transformation matrix with a value calculated from the observed mean that is used in calculating the variances of the parameters that permit nonconstant variances for values of the mean because they can depend on a specified function of the mean (McCullagh & Nelder, 1989). The outcome produces both a matrix of the βs and a matrix with the inverse of the variance. If we assume the data are correlated, the variances are multiplied against a working matrix of correlation coefficients that corrects for correlation within subjects or panels. This matrix can be either specified by the researcher or estimated by the GEE model in a form that matches the

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

expected correlation structure within the subject or cluster. If we assume the observations are independent, this variance matrix is then transformed into a column of error terms through matrix multiplication with an identity matrix, which has all 1s on the diagonal and therefore produces no change in the parameter estimates.

The output from these equations is then used in starting the procedure all over again in an iteratively reweighted least squares procedure that involves minimizing the extent of change in the parameter estimates from a perfectly fitted regression model (Gardner et al., 1995; Hardin & Hilbe, 2003; Liang & Zeger, 1986; McCullagh & Nelder, 1989). As the size of these changes compared to the prior iteration approaches zero, the parameter estimates (βs and standard errors) stabilize. Specification of the correct form of the correlation of responses increases the efficiency of these estimates (Fitzmaurice, 1995; Hardin & Hilbe, 2003), which is of concern particularly when the correlation within responses is high (Diggle et al., 2002; Zorn, 2001). However, the model is robust to errors in the specification of correlation structure because estimates of the regression parameters remain consistent; therefore, the efficiency gains from exact specification of the structure are usually slight (Liang & Zeger, 1986). Fitting a GEE model requires the user to specify (a) the link function to be used, (b) the distribution of the dependent variable, and (c) the correlation structure of the dependent variable. Details on how to make these three decisions that have to be made will be discussed in turn below.

What Is the Best Link Function?

To model the expected value of the marginal response for the population $\mu_i = E(y_i)$ as a linear combination of the covariates, the user must specify a link transformation function that will allow the dependent variable to be expressed as a vector of parameter estimates (β) in the form of an additive model (McCullagh & Nelder, 1989). Harrison (2002) noted that the link function is what "makes [generalized linear modeling] techniques part of a larger family of log-linear models; nonlinear and distinct from multiple linear regression in the link function but linear and familiar in terms of the string of regression parameters" (p. 454). An example of a link function would be the logit link for binary response variables. In this case, the covariates would be transformed by the log of the odds ratio (the ratio of a response of "1" in the data to a response of "0"). Users are not necessarily restricted to a single link function for the distribution of the data specified in the next step.

The choices available for the link function are shown in the appendix. The basic link function is the identity link function, which involves no transformation of μ before construction of the matrix of βs, and this is used for normally distributed data. The distribution of a dependent variable generally limits the user's choices with reference to the link function used. The logit link is the standard linking function for binary dependent variables. This link allows for the regression equation to map the interval from 0 to 1 and is expressed as $g(x) = \log[\mu/(1 - \mu)]$. In cases in which counted data are being modeled with Poisson regression, the most appropriate link function involves modeling the logarithm of the mean. Regression coefficients represent the expected change in the log of the mean of the dependent variable for each change in a covariate (Gardner et al., 1995; McCullagh & Nelder, 1989). Other link functions available to users include the probit link for conducting cumulative predictive analysis of binary or ordered dependent variables and the cumulative logit, which is useful for analysis of

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

ordered multinomial data. The regression coefficients that result from GEE models for logit, probit, and log links need to be exponentiated before they are meaningful.

What Is the Distribution of My Response Variable?

A second step involves specifying the distribution of the outcome variable so that the variance might be calculated as a function of the mean response calculated above (Hardin & Hilbe, 2003). GEEs permit specification of distributions from the exponential family of distributions, which includes normal, inverse normal, binomial, Poisson, negative binomial, and Gamma distributions. As in generalized linear models, the variance needs to be expressed as a function of the mean; this is then incorporated in the calculation of the covariance matrix by multiplying the components against an $N \times N$ matrix (W) with a value W_i on the diagonal that is determined by the variance function. For Poisson distributions, that figure is μ; for binary data, it is $\mu\ (1 - \mu)$; and for normally distributed data, it is 1 (Gardner et al., 1995; McCullagh & Nelder, 1989). As Gardner and colleagues (1995) showed, misspecification of the link function or the variance function can have important consequences, for example, specifying a normal distribution when the data are counted can lead to incorrect statistical conclusions. The appendix displays appropriate link and variance functions for specific distributions of the outcome variable.

Although the specification of the distribution is important, users do not need to be precise in the specification of the variance functions for the parameter estimates to have a sampling distribution that is approximately normal (Liang & Zeger, 1986). This is helpful because it is difficult to know the exact covariance structure (Horton & Lipsitz, 1999); the variances derived from the data may be lower or higher than those assumed in the model (the data would be underdispersed or overdispersed) and therefore the data may not exactly fit a distribution assumed—the variance estimate should account for this (Gardner et al., 1995). Because the variance estimator that is used in generalized linear models assumes independence of observations, in developing the GEE model, Zeger and Liang (1986) extended use of a method of estimating the variance that incorporates the correlation of the observations and produces variance estimates (but not regression coefficients) that are consistent in cases in which the specification of the variance function is not exactly correct (Diggle et al., 2002; Hardin & Hilbe, 2003; Wedderburn, 1974; Zeger & Liang, 1986).

In fitting a GEE (or any generalized linear model), the user should make every reasonable effort to correctly specify the distribution of the response variable so that the variance can be efficiently calculated as a function of the mean and the regression coefficients can be properly interpreted (McCullagh & Nelder, 1989). Specifying a Poisson distribution with a binary distribution (and vice versa) is a major error that can lead to mistakes regarding inferences about regression parameters. In Example 1 below, I show how a change in the distribution specification for the GEE models leads to a researcher's reaching different conclusions.

Generally, the user will have some prior knowledge of the distribution of the response variable. As a rule, if the responses are binary data, users should specify the binomial distribution. In cases in which the responses are counted, users should first select a Poisson distribution and then examine the extent of dispersion in the outcome predictor. When the variances derived from the data are higher or lower than those assumed in the model, the data may be over- or underdispersed. When analyzing

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

counted data, a negative binomial distribution should be specified in cases in which the dispersion is high (Gardner et al., 1995). In cases in which users are faced with a dependent variable that is an ordered multinomial response, multinomial distributions can be specified.

What Is the Likely Form of Correlation Within My Response Variable?

A third step involves specification of the form of correlation of responses within subjects or nested within group in the sample. It is this working correlation matrix that allows GEEs to estimate models that account for the correlation of the responses (Liang & Zeger, 1986). Users of GEEs have several options to select from in specifying the form of the correlation matrix. This specification will differ based on the nature of the data collected. "The goal of selecting a working correlation structure is to estimate β more efficiently" (Pan, 2001, p. 122), and incorrect specification of the correlation structure can affect the efficiency of the parameter estimates (Fitzmaurice, 1995). Although GEE models are generally robust to misspecification of the correlation structure (Liang & Zeger, 1986), in cases in which the specified structure does not incorporate all of the information on the correlation of measurements within the cluster, we can expect that inefficient estimators will result. Below, I review four common options for the specification of the correlation structure of the data.

For data that are correlated within cluster over time, an autoregressive correlation structure is specified to set the within-subject correlations as an exponential function of this lag period, which is determined by the user. Users may specify that the within-subject observations are equally correlated, which is referred to as an "exchangeable" correlation structure. Where there is no logical ordering for observations within a cluster (such as when data are clustered within subject or within an organizational unit but not necessarily collected over time), an exchangeable correlation matrix should be used (Horton & Lipsitz, 1999). This method is more appropriate in situations in which data are clustered within a particular subject but are not time-series data. An example of this could be where responses are correlated within an industry group on cross-sectional data. Users may also permit the free estimation on the within-subject correlation from the data. Such an unstructured working correlation matrix estimates all possible correlations between within-subject responses and includes them in the estimation of the variances (Fitzmaurice, Laird, & Rotnitzky, 1993). Finally, users may assume that the responses within subject are independent of each other; this approach sacrifices one of the two benefits of using GEE in that it does not account for within-subject correlation but is still useful in model fitting (as a base model) and is currently the only correlation matrix permitted for the analysis of ordered multinomial responses in SAS.

Fitting GEE Regressions

Attention to issues of model selection and fitting issues for GEEs has lagged behind the attention paid to extending distributional assumptions and other refinements in the variance estimation processes (Pan, 2001). The process of selecting model terms and the appropriate correlation structure for GEE models is complicated by the correlations within subject. Because the observations are not independent of each other, the

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

residuals are not independent, and therefore, common likelihood-based methods and other measures of model fit from ordinary linear regression need to be adjusted. Zheng (2000) introduced a simple extension of R^2 statistics for GEE models of continuous, binary, and counted responses that is referred to as "marginal R-square." The test measures improvement in fit between the estimated model and the intercept-only (null) model. The formula is very straightforward and requires the user to obtain predicted values from the model after it is estimated and compare these against the actual values and against the squared deviations of the observations from mean values for the response variable.

$$R^2m = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(Y_{it} - \hat{Y}_{it})^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}(Y_{it} = \overline{Y}_{it})^2}.$$

In this equation, $\overline{Y}$ is the marginal mean across all time periods $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}Y_{it}$. The marginal R^2 is interpreted as the amount of variance in the response variable that is explained by the fitted model (Hardin & Hilbe, 2003). It has roughly the same properties as R^2 measures do for regression models, with the exception that it can take a value of less than zero when the model that is estimated does a worse job of predicting than the intercept-only model and it reduces to an R^2 measure when there is one measurement per panel ($T = 1$). The GEE covariance matrix is not explicitly included in the calculation of this statistic. Zheng (2000) argued that "goodness of fit [for marginal models] is concerned with the agreement between the response and the prediction. The covariance matrix is only relevant to the point that it affects the fitted value through the parameter estimates" (p. 1270). As shown in Example 1 below, this statistic is useful in making decisions regarding the terms to add to a GEE model using a stepwise regression method.

In general, decisions about correlation structure should be guided first by theory; as noted earlier, there are specific correlation structures that are appropriate for time-dependent correlation structures (e.g., autoregressive) and some that are not (e.g., exchangeable). For cases in which users may be undecided between two structures, Pan (2001) proposed a test that extends Akaike's information criterion to allow comparison of covariance matrices under GEE models to the covariance matrix generated from a model that assumes no correlation within cluster. Because these compare the variance and magnitude of the squared deviations for an independence model to models that assume different sorts of correlation (exchangeable, unstructured, autoregressive, etc.) within subjects, Pan's quasilikelihood under the independence model information criterion (QIC) measure is helpful in selecting the appropriate correlation structure. The correlation structure with the QIC score that is lowest (closest to zero) is judged to be the best, although the QIC scores may not be meaningfully different. In that case, users should select the model that makes the most sense theoretically. Neither the QIC nor the marginal R^2 measures are currently automatically computed by the major statistical software packages such as SAS or STATA.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Decisions about testing whether coefficients are equal to zero are most commonly made using a Wald chi-square statistic first proposed by Rotnitzky and Jewell (1990). The generalized Wald test statistic can be calculated as $T_w = K(\hat{\gamma} - \gamma_0)'(\hat{V}_R)^{-1}(\hat{\gamma} - \gamma_0)$, where the variance matrix ($\hat{V}_R$) is a variance estimate that incorporates the correlation structure within subjects. This test is distributed as a chi-square test statistic with degrees of freedom equal to the number of parameters being tested. It can be used to test the significance of individual parameters or several parameters. The generalized Wald test statistic for single-variable tests in GEE can generally be calculated by squaring the z score for the parameter to be tested. In cases in which the researcher faces singularity in the robust variance matrix and it cannot be inverted, when there are fewer covariates than observations per group, a "working" Wald test is available that is calculated using the inverse of the initial, model-based variance matrix (Hardin & Hilbe, 2003). Researchers should take care in using either of these Wald test statistics when the value of the regression parameter is large (Harrison, 2002), as the standard errors used in calculating the test score may be too large. A case in which this might become an issue would be if an unadjusted income covariate were included in a regression along with a series of 5-point scales.

An alternative measure is a likelihood-based measure that is calculated assuming independence. The test statistic, called a "naïve likelihood ratio test" is calculated as 2 times the difference between the likelihood score for the unconstrained (intercept-only) model and the likelihood for the constrained model with the covariates: $T^*_{LR} = 2[\lambda(\hat{\beta}_t) - \lambda(\hat{\beta}_{tC})]$ (Hardin & Hilbe, 2003, p. 170). For population-averaged GEE models, the statistic follows a chi-square distribution with an adjusted degrees of freedom that for a single covariate is calculated as the ratio of the variance in the GEE model being tested to the variance of the variable in the independence model.

Residuals from GEE regression models should be checked for the presence of outlier values that may seriously affect the results (Diggle et al., 2002). Measures that test for the influence of a panel or case in the regression equation are extensions of those used in generalized linear models and are similar to those used in OLS regression (Cook & Weisberg, 1982; Preisser & Qaqish, 1996). DFBETA measures the change in the fitted coefficient vector when a case is removed and is a measure of influence that can be used to analyze outliers and determine whether there are issues in the data that need further investigation. A valuable visual test of the GEE model that has been estimated is to request residual versus fitted plots for each individual panel. In visually testing the residuals, a researcher should look for patterns that suggest a random distribution of residuals; they should not be clustered around certain values (Hardin & Hilbe, 2003). For example, if a researcher saw that there were a large number of residuals with small negative values and a small number of high positive values, then different distribution and correlation structures should be examined. Another example would be the case in which there are changes in the pattern of the residuals across the time periods; this could indicate that they depend on the panel identifier and/or on the time identifier, and a different correlation structure should then be specified (Hardin & Hilbe, 2003). Software programs that fit GEEs (including SAS and STATA) provide users with the functionality to display residuals and DFBETA diagnostic statistics for observations in the data set.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Table 1
Sample Data for the First Three Groups in Simulation Example 1

Team	*Trial*	*Object*	*New Choice*	*Group Size*	*Group Cohesion*	*Visits*
1	1	1	0	3	4.5	3
1	2	1	0	3	5	1
1	3	1	0	3	5	0
1	4	1	0	3	4	0
1	5	4	1	3	3	0
2	1	2	0	5	4.67	7
2	2	4	1	5	5	3
2	3	4	0	5	5	0
2	4	4	0	5	4.67	0
2	5	4	0	5	4.67	5
3	1	1	0	5	4.33	9
3	2	1	0	5	5	6
3	3	1	0	5	4.5	4
3	4	1	0	5	4.67	2
3	5	1	0	5	3.67	1

Using GEEs to Analyze Data: Two Examples

As a demonstration, I will first demonstrate the application of GEEs in analyzing the results of a longitudinal experiment with counted responses that are correlated within subjects over time. The second analysis will involve normally distributed response terms that are correlated not over time but within branch offices of an organization. These two examples will provide ample opportunity to demonstrate the steps users should go through to determine when GEE models are needed and how to go about fitting the most appropriate model based on the hypotheses to be tested and the data that have been collected.

Example 1: Longitudinal Data With Counted Responses

An analysis of data from a laboratory experiment offers the opportunity to show how use of GEEs facilitates the analysis of correlated longitudinal data that are collected with limited-range dependent variables. In this case, I will use data collected from a recently conducted laboratory study that involved groups assembling Lego objects over five consecutive sessions. The 52 groups were given 1 minute to view four objects and select an object to assemble. During the assembly task, the group was allowed to send only one person out of the room at a time to view the object. These data were collected along with data on the time required to assemble the object, the size of the group, and whether the group had selected a new object this week. Data were also collected from group members on their level of satisfaction with other group members. (See Table 1 for a data sample from three teams.) The hypotheses of interest involved whether there were main effects for the object on the number of trips out of the room, whether time (and therefore group learning) affected the number of trips across trials, and whether group satisfaction and group size affected the number of

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

trips. Three features of this data set require us to select a method different from OLS for our analysis.

1. The responses are not normally distributed because they consist of a count of the number of trips out of the room. (See Figure 1 for a histogram of the distribution counts of trips out of the room per trial.)
2. There are five responses within subject, and they are not independent (they are correlated with each other).
3. There are time-dependent covariates. Group satisfaction changed over the course of the study, and the groups were allowed to select a new object at each trial.

The model needs to take into account the fact that the number of trips for a group at any trial is likely to be highly associated with prior observations for that group (see Item 2 above), and therefore it is not an independent observation. As we noted earlier, failure to account for nonindependence of observations can result in biased estimates for the regression parameters and variances, especially when the responses are highly correlated within subject. Each team can be expected to learn at an individual rate, and thus each response across the five trials should be highly correlated over time.

For purposes of comparison of regression results under different distributional assumptions, five regressions were run. The first three assumed no correlation within subject and are designed to compare an OLS regression approach when assuming a normal distribution of data against a logistic regression with a binomial dependent variable created using a median split of the number of trips and a regression approach that assumes a Poisson distribution. Then I compared the Poisson regression against two models that assumed an autoregressive correlation structure and an unstructured correlation structure within subject (see Table 2, columns 1-5).

The panel regressions that assume normal and binary distributions of the data are clearly unsatisfactory. In the normal OLS regression model, we see that at certain levels of the covariates, multiplying all of the observed values of the covariates by the βs produces a negative value (the minimum fitted value that results is –1.06, see column 1), which is not appropriate for counted data. As Gardner and colleagues (1995) noted,

> Regression . . . models data in the sense that it maps vectors of predictors into a space of expected values such that the ensemble of μ_is resembles the observations. A minimal criterion for resemblance is that the range of the μ_is should correspond to the range of *y*. (p. 394)

The OLS regression model assuming normality would lead us to substantively incorrect conclusions regarding the data and therefore must not be used. The binary model with a dichotomized dependent variable failed to converge after 100 iterations, and in viewing the incomplete results, it is clear that we would reach a different conclusion about the significant impact of selecting a new object to build on the group's performance compared to the Poisson and the normal regression models (see Table 2, column 2). It is clear that specification of the proper (Poisson) distribution will increase the usefulness of the regression estimates. Below, I outline the steps taken in fitting the Poisson GEE regressions.

Before using GEEs in any software package, the researcher needs to do the following (Stokes, 1999):

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

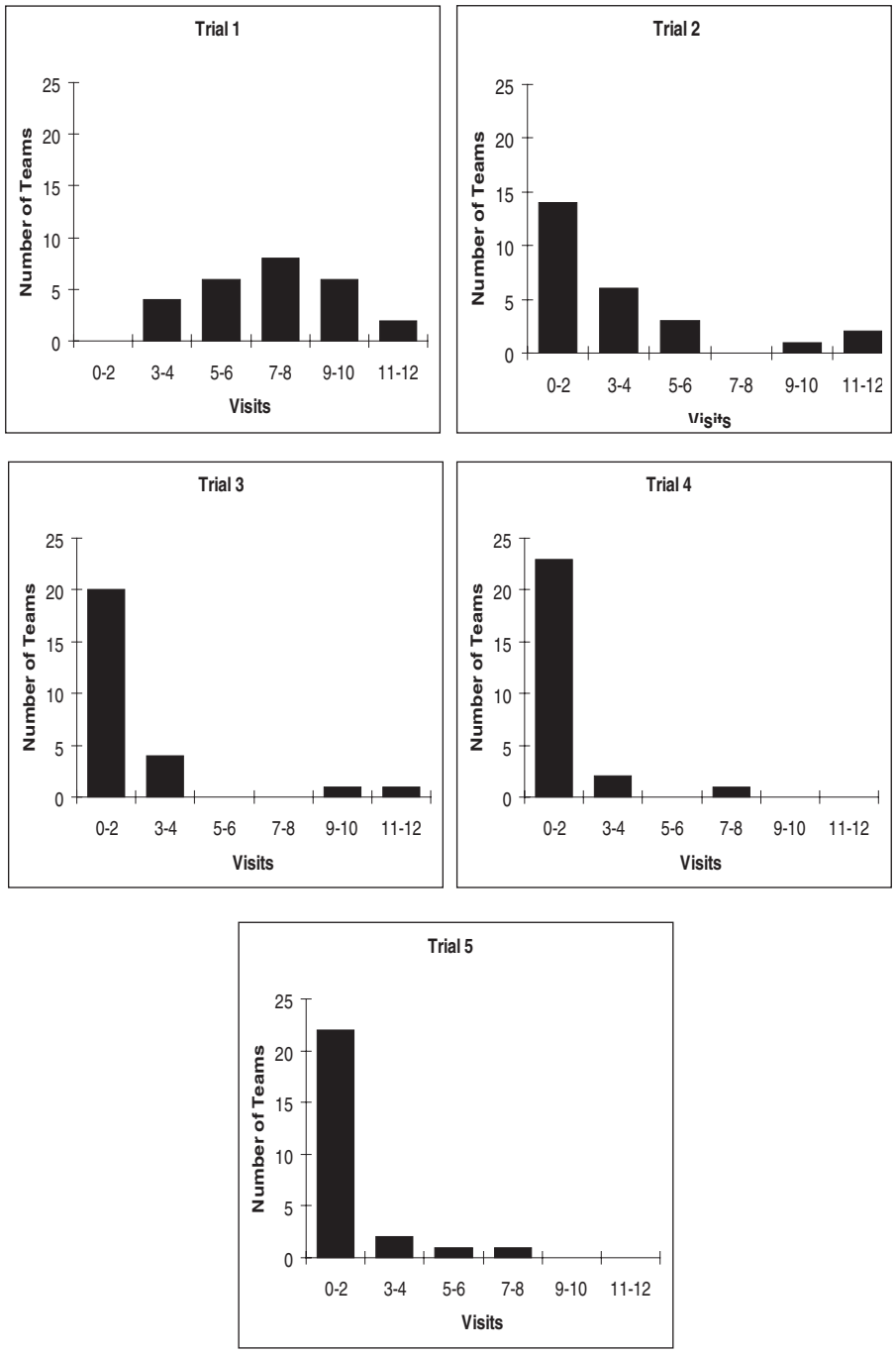

Figure 1: Histograms for Dependent Variable in Example 1 by Trial

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Table 2
Comparisons of Generalized Estimating Equation (GEE) Regressions for Example 1 ($N = 52$ Groups)

	Method									
	Ordinary Least Squares, Normal Distribution, Independent Correlation		*Logistic, Binomial Distribution, Independent Correlation*		*GEE, Poisson Distribution, Independent Correlation*		*GEE, Poisson Distribution, Unstructured Correlation*		*GEE, Poisson Distribution, One-Period Autoregressive Correlation*	
	Unstandardized Coefficient	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE
Object	–2.01	.29***	–.32	.14*	–.29	.10**	–.61	.22**	–.29	.10**
New choice	–1.49	.26***	–.10	.08	1.29	.17***	.97	.17***	1.17	.17***
Trial	3.52	.36***	.36	.16*	–.20	.14	–.35	.15*	–.20	.13
Size	.26	.28	.04	.09	.05	.09	.18	.07*	.04	.09
Group satisfaction	–.54	.23*	–.00	.11	–.02	.09	–.09	.10	–.06	.08
Trial × Object	.31	.10**	.05	.04	–.04	.06	–.00	.08	–.05	.06
Constant	9.41	2.19	–.32	.99	1.42	.74	1.97	.75	1.72	.72
R^2_{Marg}					.69		.69		.62	
Wald χ^2 (6 *df*)	788.07		26.32		408.61		522.37		388.54	
Dispersion	4.25		1.22		1.84					
Pan statistic					0.00		18.29		24.85	
Mean fitted	2.82				2.82					
Minimum fitted	–1.06				0.22					

Note. Binomial model failed to converge after 100 iterations.
*$p < .05$. **$p < .01$. ***$p < .001$.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

1. specify the model parameter(s) of interest,
2. specify any interaction terms of interest,
3. specify the variables that indicate the clustering of the dependent variable responses in the data (e.g., by case and trial, or by hospital or work unit),
4. specify the link function that will "linearize" the regression equation,
5. specify the distribution of the dependent variable,
6. specify the structure of the correlation of within-subject responses (the "working" correlation matrix), and
7. identify and request the appropriate test statistics to be generated.

Our outcome of interest is the number of times the group sent someone out of the room to look at the finished object (trips); these data are counted and therefore are not normally distributed. Our covariates (Step 1) are selected based on the hypotheses of the study; these variables include trial, group size, and group satisfaction. We also include two control variables: one for the difficulty of the object (1 through 4) and one for whether the group selected a new object during the specific trial (a 0/1) variable. An interaction term (Step 2) for Trial × Object was included to control for whether groups constructing more complex objects experienced more difficulty in learning throughout the experiment. The clustering variable (Step 3) is the group; each group will have five (possibly) correlated responses, one for each trial. In statistical software packages such as SAS or STATA, users need to specify both the cluster identifier (in this case, the group number variable, ranging from 1 to 52) and the time variable (in this case, the trial variable, ranging from 1 to 5). GEE output from software packages provides information on the form and number of the clusters used in the analysis to help the user check whether this has been properly specified.

We will make tentative assumptions for the link, distribution, and correlation structures and use model fit and dispersion statistics to determine whether the mean and variance are equivalent (a necessary assumption for Poisson regression) as well as our theory to help us determine which model fits best. The link function (Step 4) for Poisson regression is generally a log link (McCullagh & Nelder, 1989); $g(\mu) = \log(\mu)$, and the variance function (Step 5) is specified as the mean of the data: $v(\mu) = \mu$ for Poisson distributions. In setting the specification of the correlation of the data (Step 6), we will test an independence model and compare it against two other models that are logically based on the nature of the within-group correlation of the responses. Because the data were collected in a longitudinal design, we expected the responses within the groups to be correlated with each other over time. It makes sense that effective teams may require less views of the object in Trial 1, and these teams will also require fewer views in Trial 2, Trial 3, and so on.

For the Poisson GEE models shown in columns 3 through 5 of Table 2, the interpretation of the signs of the raw coefficients of the main effects is straightforward. Positive values mean that increases in the covariate result in increased number of trips; the bigger the group, for example, the greater the number of trips to view the model. A negative sign for the coefficient of trial implies that the number of trips required decreased over time. Because the log link function was specified, interpretation of the value of the parameter estimates requires that they be exponentiated by taking the log of the β coefficient estimates. These values are then interpreted as incidence rate ratios similar to Poisson regression (Gardner et al., 1995) and are shown in Table 3. The coefficients in a model using a logit link are different: They are odds ratios and represent the odds that a value is 1 as opposed to 0 for increasing values of the covariate. In gen-

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Table 3
Exponentiated Coefficients for Poisson Regressions for Example 1

	Method		
	GEE, Poisson Distribution, Independent Correlation	*GEE, Poisson Distribution, Unstructured Correlation*	*GEE, Poisson Distribution, One-Period Autoregressive Correlation*
Object	0.75**	0.541**	0.75**
New choice	0.82	0.706*	0.82
Trial	3.62***	2.65***	3.24***
Size	1.04	1.20*	1.04
Group	0.98	0.92	0.94
Trial × Object	0.96	1.00	0.95

*$p < .05$. **$p < .01$. ***$p < .001$.

eral, exponentiated coefficients are interpreted in GEE models in the same way they are in nonlinear models such as logistic, Poisson, and probit analysis.

In comparing the results between the Poisson regression and the normal and binary variables seem in columns 1 and 2 of Table 2, we see that we would reach substantively different conclusions by using a regression approach that specified the wrong distribution. Note that the sign changes from positive for the variable trial (3.52) from the normal to negative in the three Poisson models seen in columns 3 through 5. The same thing happens for the variable new choice, which was a time-dependent variable indicating that the group had selected a different object in a particular trial over the previous trial. In the normal model, we would conclude that a new choice led to fewer trips ($\beta = -1.49$, $SE = 0.26$, $p < .001$); in the Poisson models (columns 3-5), the sign is reversed (and significant at $p < .001$), which implies (correctly) that groups facing a novel task would require more trips outside the room to learn about the new object to complete the task. These findings exemplify major points made by Liang and Zeger (1986) and others regarding the critical nature of the task of specifying the distribution in correctly modeling limited-range dependent variables (Diggle et al., 2002; Harrison, 2002; McCullagh & Nelder, 1989).

The next step is to compare the GEE Poisson regression models and select the appropriate model structure. Three models were tested against each other: the independence, unstructured, and autoregressive patterns. The first thing to notice is that we reach different statistical conclusions about the relationship between group size and the number of trips out of the room. Under the independence and autoregressive models, the variable is nonsignificant ($p = .60$ and $p = .62$, respectively) compared to the unstructured model, where it is significant at the $p = .01$ level. The size of the raw coefficient is more than 4 times greater (.18 vs. .04) in the unstructured model. The difference in the coefficients is created primarily by the differences in the estimated correlation structures used by the model. In Table 4, I compare the correlation structures that were used in the models. The unstructured model estimates within-panel correlations that are generally higher over time and that do not decline over time when compared against the autoregressive model, which requires a declining level of correlation within subject over time.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Table 4
Comparison of Within-Group Correlation Estimates for Example 1

	Trial 1	*Trial 2*	*Trial 3*	*Trial 4*	*Trial 5*
Unstructured correlation					
Trial 1	1				
Trial 2	.17	1			
Trial 3	.24	.06	1		
Trial 4	.20	.35	.28	1	
Trial 5	.38	.66	.13	.22	1
One-period autoregressive correlation					
Trial 1	1				
Trial 2	.15	1			
Trial 3	.02	.15	1		
Trial 4	.00	.02	.15	1	
Trial 5	.00	.00	.02	.15	1

We conclude that the unstructured model is superior for two reasons. First, it is the least restrictive in terms of modeling the true correlation structure within subject. The values of the correlations are not trivial in this case, sometimes reaching as high as .67 (Trial 2-Trial 5). The correlations do not decrease over time, as assumed by an autoregressive correlation structure. Ignoring the magnitude of these correlations means the autoregressive model creates estimates that are less efficient than the unstructured model because they do not use all the information about the parameters (Fitzmaurice, 1995). Second, the nature of the data involves groups' repeating the same act over subsequent events, and there is no reason to expect that the correlation of the responses between trials would decrease over time as it would in an autoregressive model. No treatment was introduced during the course of the trials, and so the responses within the group should remain consistent across several trials. Finally, comparison of Pan's (2001) QIC statistic for the unstructured model shows that it is closer to zero relative to the scores for the independence and autoregressive models, implying that the model is a better fit.

Example 2: Normally Distributed Responses Correlated Within Branch Offices

In organizational research, clustered data arise outside of longitudinal study designs. In a cross-sectional study of an organization with a number of branch offices, a researcher may believe that clustering of responses within each branch office may create nonindependence in the responses and therefore affect the ability to test hypotheses in the same way that correlation of responses over time affects longitudinal designs. GEE regression models are appropriate for estimating relationships in such cases; indeed, there are few differences in constructing the model in panel-clustered data cases compared to cases of longitudinal clustering. This example is helpful in that it also shows how the benefits of using GEE over OLS approaches are smaller when correlations within the panel are low as opposed to the higher correlations seen in Example 1.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Hierarchical linear modeling (HLM) is an alternative approach to analyzing these data. The primary difference between HLM and GEE approaches lies in HLM's assumption that the random effects are normally distributed and on the researchers' assumptions regarding the structure of the variances and covariances estimated. When the number of clusters is large, GEEs do not require such assumptions to be made to produce estimates with good statistical properties. HLM software (release 5.05) provides users with a helpful comparison table of the estimates from GEEs and HLM approaches to nested data so that users can test and compare the extent to which such violations lead to changes in conclusions.

As an illustration of the application of GEE models, I analyzed data on employee attitudes that were collected from 50 hospitals of a medical firm. A simple regression test was constructed to test a hypothesis relating the quality of self-reported satisfaction with supervision, growth, pay, and security to a measure of values commitment relative to the parent corporation. But because there are different supervisory referents at each location, there may be clustering of responses within hospitals that will affect the estimates, which will affect the efficiency of our parameter estimates. In this case, I will run a normal regression model and a regression model with robust standard errors estimated and compare it against two GEE models: one that accounts for correlation within hospital and one that does not.

The first step is to identify the terms in the model; we are fitting a very basic model with a 5-point scale (commitment to the parent company) as the dependent variable and 5-point scale responses on the facet satisfaction measures. There are no interactions proposed in any of the hypotheses, and so we will not estimate interaction terms in the model (Step 2). Our responses may be correlated within hospital, so our clusters (Step 3) are specified to be on that variable. This will result in 50 clusters ranging in size from 1 to 50 employees, with 411 total observations.

We assume that the data from our 5-point scale are normally distributed, which follows traditional assumptions regarding the distribution of responses in industrial/organizational psychology and organizational research. The link function (Step 4) for normally distributed dependent variables is specified as the "identity" link (which involves transformation of regression coefficients by multiplying them by a factor of 1), and our variance function (Step 5) is set as that for a normally distributed response variable. Our correlation within hospital is not time dependent and therefore we will test (Step 6) an exchangeable working correlation matrix against an independence correlation matrix and an unstructured correlation matrix for the model that fits best. No special test statistics (Step 7) will be needed to test our hypotheses, as the βs and standard errors produced by the models are adequate for testing this simple hypothesis.

Results from the normal regression model are compared against the GEE model in Table 5. As expected, the inclusion of robust standard errors as opposed to standard errors that do not incorporate the correlation within hospital (also referred to as "naïve" standard errors) makes both the normal regression model and the two GEE models more conservative; the standard errors of the parameter estimates are all higher than the naïve model (Model 1). The differences between the estimates from the GEE model (Model 3) that assumes independence and the GEE model that assumes an exchangeable correlation (Model 4) are quite small, but they are meaningful. The reason for this is that the correlation within clusters is estimated to be quite small: In this case, it is estimated by the GEE model as .0244. The differences in analytic power that come from the incorporation of the correlation structure in the GEE regression are

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Table 5
Comparison of Regression Results for Example 2

	Parameter							
	OLS, Naïve SE		*OLS, Robust* SE		*GEE, Independence*		*GEE, Exchangeable*	
	Unstandardized Coefficient	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE
Supervision	.042	.051	.042	.055	.042	.054	.040	.055
Pay	.235	.040***	.235	.044***	.235	.039***	.228	.039***
Growth	.133	.063*	.133	.073	.133	.068*	.138	.067*
Security	.146	.048**	.146	.054**	.146	.059*	.152	.058*
Constant	1.212	.201	1.212	.239	1.212	.281	1.212	.282
R^2 (R^2_{Marg})	.2454		.2454		.2454		.2451	
$F(4, 406)$	33.01		30.51					
Wald χ^2 (4 *df*)					75.14		72.82	

Note. OLS = ordinary least squares. $N = 50$ hospitals; 411 employees.
*$p < .05$. **$p < .01$. ***$p < .001$.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

small in this case. Recall that when correlation levels within the cluster are low, the difference in parameter estimates between GEE regression and models (such as OLS) that assume independence will be small (Diggle et al., 2002). Although this might lead to our reaching a different conclusion with regard to satisfaction with growth opportunities (GROWTH) and commitment to the parent in that the z score under the independence model is 1.96 ($p = .05$) and under the exchangeable model is 2.04 ($p = .04$), in general, GEEs do not result in much change in results when correlation within clusters is low. In testing the difference in the marginal R^2 between the independence and exchangeable models, we find that the independence model provides a higher score. But in principle, one should reject an OLS or independence model in favor of a GEE regression whenever there is reason to suspect correlation within subject because for higher values of this correlation, the efficiency losses in the parameter estimates is much greater, and researchers as a rule should favor methods that are expected to produce the most efficient and unbiased parameter estimates (Diggle et al., 2002; Fitzmaurice, 1995).

Cautions Regarding GEE

Users should be cautioned that the estimate of the variance produced under GEE models could be highly biased when the number of subjects within which observations are nested is small (Prentice, 1988). Horton and Lipsitz (1999) suggested that the GEE variance estimate be used only when there are more than 20 such clusters. In data sets in which there are a low number of these clusters, the standard errors that are constructed ignoring correlation within subject (naïve/model-based standard errors) may have better statistical properties in that they will have sampling distributions closer to normal than the empirical variance estimates that incorporate the correlation within subject.

Despite the advances of Zheng (2000) and Pan (2001), goodness-of-fit statistics for GEEs that would function as the equivalent to measures such as the magnitude of the squared differences of observed versus predicted values or dispersion measures are not widely accepted for most classes of dependent variables beyond binary data or for different correlation structures (Barnhart & Williamson, 1998; Horton et al., 1999; Sheu, 2000; Stokes, 1999). The statistics may be calculated only for certain distributions when making certain assumptions about within-subject correlation (e.g., independence or exchangeable equal correlation) that do not permit the user to fully benefit from GEE modeling. Because the response variables in GEEs are generally not independent, the residuals from models fitted to these responses are not independent and thus are not appropriate for use in the development of these statistics (Barnhart & Williamson, 1998; Zorn, 2001). The goodness-of-fit measures of Zheng (2000) (marginal R^2 and the concordance correlation) and Pan (2001) presented here have the benefit of simplicity and ease of interpretation, but they have not been used extensively in biostatistics and health research literature, in which GEEs originated and are most widely used. Another measure that shows some promise is chi-square distributed goodness-of-fit statistics for binary response variables developed by Barnhart and Williamson (1998). Given the uncertain status of this area of research on GEE applications, users of current statistical programs should be cautioned that although the gen-

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

eralized linear model algorithm that is used when running GEEs may produce a deviance or chi-square statistic for a GEE model, such a statistic is interpretable only under certain conditions. For example, the Wald chi-square statistic presented by STATA in the output for GEEs is a test of whether all of the variables in the estimate are different from each other and different from zero and is not a goodness-of-fit measure. It is generally not interpretable when one wants to model different correlations within subjects across different time periods, such as in an autoregressive correlation structure in which, for example, the $t_1 - t_2$ correlation is accounted for differently from the $t_1 - t_3$ correlation. In addition, as noted earlier, the Wald tests for individual parameters may be sensitive to large differences in the scale of the different independent variables (Harrison, 2002).

The scale parameter specified also has an important impact on the GEE parameter estimates, and users must be cautioned that the two software programs (SAS and STATA) used to estimate the models shown in this article have different default settings for the scale parameter that result in the programs' producing different results from the same data set unless this is changed. SAS users can use the "V6CORR" command to obtain similar results from the two programs.

There are other issues for which users of GEEs should be alert. Errors in the specification of the relationship of responses within subjects in the form of the working correlation matrix can lead to a loss of efficiency in models and lead to different assessments of standard errors. However, because the method uses the initial parameter estimates and residuals to reset the covariance matrix, it is robust to misspecification of the initial relationship of the within-subject correlations (Zeger & Liang, 1986). As noted earlier, researchers should pay close attention to specifying the distribution of the dependent variable and the link function that will be used to linearize the regression equation. Errors in the calculation of parameter estimates can be made if these are incorrectly specified (Gardner et al., 1995; McCullagh & Nelder, 1989). If the researcher is unclear as to the form of the distribution of the dependent variable, it is a good practice to use statistical tests to remove this doubt. For example, prior to running a GEE model to analyze their data, Welbourne and Trevor (2000) first ran a regression-based test to determine that the distribution of their dependent variable fit a negative binomial as opposed to a Poisson distribution. GEEs can handle missing data in longitudinal studies under the assumption that such data are missing completely at random, but when the probability of missing data may depend on previous values of the dependent variable, the parameter estimates may be compromised (Zorn, 2001). Another issue is that in cases in which there is high correlation within clusters, GEE models estimated using an unstructured correlation matrix might take a long time to converge. Software packages such as SAS or STATA allow the user to increase the maximum number of iterations that the model will go through to generate the best set of parameter estimates.

Users should also be alert for the form of the missing data in their analysis, especially the relationships that may be compromised by subject attrition. GEE assumes that the data are missing completely at random, and the model results may not be interpretable if the attrition in the data set is associated with one of the covariates or the dependent variable. Researchers using longitudinal data sets are referred to Goodman and Blum (1996) for a framework for assessing the effects of attrition on data relationships in longitudinal studies.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Conclusion

GEEs offer researchers and managers the opportunity to use longitudinal designs with organizational topics that do not lend themselves to normally distributed responses. Management researchers are frequently faced with dependent variables that do not follow normal distributions, and in the past, researchers applied suboptimal methods such as transformations to convert their data prior to analysis. When faced with the challenge of collecting data such as event count gs and "yes" or "no" response variables, greater analytic precision can be gained by instead applying GEEs to the data. The method is already being applied in social sciences such as political science (Zorn, 2001) and criminology (Conaway & Lohr, 1994), and it is extensively in use in the life sciences such as epidemiology and gerontology research. There have been several recent examples of the use of the method in longitudinal organizational research in field settings, including Welbourne and Trevor (2000) and in repeated measure laboratory experiments, such as those conducted by Lepine, Colquitt, and Erez (2000).

The use of GEE regression models in management research has been limited since their introduction in 1986. Writing in 1989, Harrison and Hulin favorably highlighted the potential of GEEs in their review of the applicability of event history models for studying absenteeism. They noted that GEE models use all the data available for each subject, account for correlations between binary outcomes across time within the same individual, and allow for specification of both time-varying and individual difference variables. They pointed out that the models have a "strong potential" for application to attendance data (Harrison & Hulin, 1989, p. 315). Despite reiteration of the point 4 years later (Martocchio & Harrison, 1993), GEE regression is still not generally used in research on absenteeism.

As with any method that is still being perfected, researchers are advised to pay close attention to the emerging body of literature on analysis of GEE data. One area that researchers should pay close attention to is the development of goodness-of-fit tests for GEE models. Despite the introduction of several new methods and recent advances over the past several years (Barnhart & Williamson, 1998; Horton et al., 1999; Zheng, 2000), there is still no universally accepted test for goodness of fit for GEE models in use that extends beyond binary dependent variables. Another area in which GEE research and theory is still developing is in the stability of the models in handling missing data.

Research that uses longitudinal designs in these areas will increase the strength of findings of relationships and perhaps uncover new relationships that have been missed because of suboptimal treatment of response variable data. GEE approaches to regression analysis provide researchers with a means of reaching easily interpretable conclusions regarding limited-range dependent variables. Users of GEEs can also be more confident in their statistical conclusions regarding data that arise from longitudinal and nested research designs, particularly when the dependent variable is highly correlated within subject because the method produces parameter estimates that are more efficient and unbiased than is OLS regression. The increased use of this method in organizational research can facilitate expanded use of longitudinal research in fields such as absenteeism, strategic management, innovation, strategy, and organizational theory, where counted data are frequently used as dependent variables. Researchers in each of these areas will benefit from the versatility of this emerging approach to data analysis.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

APPENDIX
Distribution Choices and Link Functions Available in Generalized Estimating Equations (GEEs)

The link function that is selected will vary depending on the distribution of the underlying dependent variable. Certain dependent variables permit multiple link functions depending on how the user wishes to interpret the coefficients (as cumulative probabilities, for example). This appendix provides some brief guidance on the different link functions available in GEE models.

Normal Distribution

Identity link: This fits the same model as the general linear model
Power link: Any power transformation (e.g., square root, square of variable)
Reciprocal link: Links using reciprocal of dependent variable ($1/\mu$)

Binomial Distribution (1/0 Data)

Logit link: Fits logistic regression models
Probit link: Fits cumulative probability functions
Power link: Any power transformation (e.g., square root, square of variable)
Reciprocal link: Links using reciprocal of dependent variable ($1/\mu$)

Poisson Distribution (Counted Data)

Log link
Power link: Any power transformation (e.g., square root, square of variable)
Reciprocal link: Links using reciprocal of dependent variable ($1/\mu$)

Negative Binomial Distribution

Power link: Any power transformation (e.g., square root, square of variable)

Gamma Distribution

Power link: Any power transformation (e.g., square root, square of variable)
Reciprocal link: Links using reciprocal of dependent variable ($1/\mu$)

Multinomial Distribution*

Cumulative logit link

*At the time of this writing, analysis of multinomial distributed dependent variables was permitted in SAS using only the cumulative logit link and the independence assumption regarding correlation of responses.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

References

Ahuja, G., & Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study. *Strategic Management Journal*, *22*, 197-220.

Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, *22*, 521-543.

Barnhart, H. X., & Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics*, *54*, 720-729.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (2003). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. St. Paul, MN: Assessment Systems Corporation.

Cook, D., & Weisberg, J. (1982). *Residuals and influence in regression.* London: Chapman and Hall.

Conaway, M. R., & Lohr, S. L. (1994). A longitudinal analysis of factors associated with reporting violent crimes to the police. *Journal of Quantitative Criminology*, *10*, 23-39.

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed). Oxford, UK: Oxford University Press.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, *51*, 309-317.

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, *8*, 284-309.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson and negative binomial models. *Psychological Bulletin*, *118*, 392-404.

Goodman, J. S., & Blum, T. C. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. *Journal of Management*, *22*, 627-652.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. Boca Raton, FL: Chapman and Hall/CRC Press.

Harrison, D. A. (2002) Structure and timing in limited range dependent variables: Regression models for predicting if and when. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 446-497). San Francisco: Jossey-Bass.

Harrison, D. A., & Hulin, C. L. (1989). Investigations of absenteeism: Using event-history models to study the absence-taking process. *Journal of Applied Psychology*, *74*, 300-316.

Haveman, H. A., & Nonnemaker, L. (2000). Competition in multiple markets: The impact on growth and market entry. *Administrative Science Quarterly*, *45*, 232-267.

Horton, N. J., Bebchuk, J. D., Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P., & Fitzmaurice, G. M. (1999). Goodness-of-fit for GEE: An example with mental health service utilization. *Statistics in Medicine*, *18*, 213-222.

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

Lepine, J. A., Colquitt, J. A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive ability, conscientiousness, and openness to experience. *Personnel Psychology*, *53*, 563-593.

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

Martocchio, J. J., & Harrison, D. A. (1993). To be there or not to be there? Questions, theories and methods in absenteeism research. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 11, pp. 259-328). Greenwich, CT: JAI.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370-384.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*, 120-125.

Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (4th ed.). Boston: Irwin, McGraw-Hill.

Preisser, J. S., & Qaqish, B. F. (1996). Deletion diagnostics for generalized estimating equations. *Biometrika*, *83*, 551-562.

Prentice, R. L.(1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics, 44*, 1033-1048.

Rotnitzky, A., & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, *77*, 485-497.

Sheu, C.-F. (2000). Regression analysis of correlated binary outcomes. *Behavior Research Methods, Instruments, & Computers*, *32*, 269-273.

Stokes, M. E. (1999). Recent advances in categorical data analysis. *24th Annual Meeting of the SAS Users Group International*. Cary, NC: SAS Institute.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 439-447.

Welbourne, T. M., & Trevor, C. O. (2000). The roles of departmental and position power in job evaluation. *Academy of Management Journal*, *43*, 761-771.

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

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*, 1049-1060.

Zheng, B. (2000). Summarizing the goodness of fit on generalized linear models for longitudinal data. *Statistics in Medicine*, *19*, 1265-1275.

Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, *45*, 470-490.

Gary A. Ballinger is a doctoral student at Purdue University. His research interests include leadership succession, the role of technology in the relationship between supervisors and workers, and issues surrounding longitudinal research methods.

Downloaded from orm.sagepub.com at UNIV OF MICHIGAN on February 10, 2014