Published in final edited form as:

Lifetime Data Anal. 2006 September ; 12(3): 249–265.

Marginal Regression Models with a Time to Event Outcome and Discrete Multiple Source Predictors

Heather J. Litman¹, Nicholas J. Horton², Jane M. Murphy^{3,4}, and Nan M. Laird⁵

¹ New England Research Institutes, Watertown, MA

² Department of Mathematics and Statistics, Smith College, Northampton, MA

³ Department of Psychiatry, Harvard Medical School (Massachusetts General Hospital), Boston, MA

⁴ Department of Epidemiology, Harvard School of Public Health, Boston, MA

⁵ Department of Biostatistics, Harvard School of Public Health, Boston, MA

Abstract

Information from multiple informants is frequently used to assess psychopathology. We consider marginal regression models with multiple informants as discrete predictors and a time to event outcome. We fit these models to data from the Stirling County Study; specifically, the models predict mortality from self report of psychiatric disorders and also predict mortality from physician report of psychiatric disorders. Previously, Horton et al. found little relationship between self and physician reports of psychopathology, but that the relationship of self report of psychopathology with mortality was similar to that of physician report of psychopathology with mortality. Generalized estimating equations (GEE) have been used to fit marginal models with multiple informant covariates; here we develop a maximum likelihood (ML) approach and show how it relates to the GEE approach. In a simple setting using a saturated model, the ML approach can be constructed to provide estimates that match those found using GEE. We extend the ML technique to consider multiple informant predictors with missingness and compare the method to using inverse probability weighted (IPW) GEE. Our simulation study illustrates that IPW GEE loses little efficiency compared with ML in the presence of monotone missingness. Our example data has non-monotone missingness; in this case, ML offers a modest decrease in variance compared with IPW GEE, particularly for estimating covariates in the marginal models. In more general settings, e.g. categorical predictors and piecewise exponential models, the likelihood parameters from the ML technique do not have the same interpretation as the GEE. Thus, the GEE is recommended to fit marginal models for its flexibility, ease of interpretation and comparable efficiency to ML in the presence of missing data.

Keywords

multiple informants; censored survival data; maximum likelihood; generalized estimating equations; inverse probability weights

1 Introduction

The problem we consider is assessing the marginal effect of predictors on survival when multiple measures have been used; multiple informant reports are frequently used to measure psychopathology. Researchers have collected information using multiple informant reports

even before appropriate techniques were developed for its analysis. For instance, in the Stirling County Study (Leighton, 1959; Murphy, 1980; Murphy et al., 2000), depression and/or anxiety were assessed using self reported information; in addition, physicians reported on the same subjects for a full range of psychiatric disorders. An early analysis investigated the relationship between self report of depression and/or anxiety and mortality where they found that the standardized mortality ratio (SMR) was 1.5 (Murphy et al., 1987); another study dealt with physician report of many psychiatric disorders and found that SMR was 1.6 (Murphy et al., 1989). An analysis by Horton et al. (2001) was the first to incorporate the information from both informants and allow comparison of the marginal relationship of physician reported psychopathology and mortality to that of self reported psychopathology and mortality. While there is only modest agreement between the self reports of depression and/or anxiety and the physician reported psychiatric disorders ($\kappa = 0.18$) (Horton et al., 2001), use of marginal models illustrate that the relationship of self reported psychopathology with mortality and physician reported disorders with mortality is quite similar (Horton et al., 2001). Being able to compare the two regression results directly in the same analysis is useful in demonstrating the validity of self report.

Researchers have developed techniques to fit marginal models when treating multiple informants as covariates. A generalized estimating equations (GEE) approach (Horton et al., 1999; Pepe et al., 1999) has been developed for use with generalized linear models; a maximum likelihood (ML) technique has been implemented using a discrete response and multiple informants as discrete predictors (O'Brien, 2003), as well as for a continuous response and multiple informants as continuous predictors (Litman et al., 2005). In both situations, the ML and GEE approaches provide identical estimates. In this paper, we consider a model with a time to event outcome and multiple informants as discrete predictors. The GEE approach (Horton et al., 2001) has previously been applied in this setting but the ML technique has not; we describe how to use the latter, compare both methods and explain under what special conditions the two approaches are identical.

To demonstrate the GEE approach, Horton et al. (2001) applies the technique to the Stirling County Study (Leighton, 1959; Murphy, 1980; Murphy et al., 2000), an investigation initiated in 1952 in a community of 20,000 residents in Atlantic Canada. Horton et al. (2001) use GEE models to predict mortality in a 16 year followup period of 1079 subjects from multiple informants (self report and physician report) about psychiatric disorders. The physician reports (denoted as X_1) and self reports (denoted as X_2) have missingness while the mortality and follow-up information is complete. In particular, we compare the marginal relationship of self report of psychiatric disorders and mortality to that of physician report of psychiatric disorders and mortality using our newly developed ML technique and the GEE method.

The multiple informant covariates (physician and self reported psychopathology) in the Stirling County Study have missing values. Horton et al. (2001) found evidence that the missingness in the self reports of psychiatric disorders was not missing completely at random (MCAR); specifically, those who did not give a self report were more likely to have physician reported psychiatric disorders and were also more apt to be older and female. Thus, we investigate fitting marginal GEE and ML models for multiple informant covariates assuming missing at random (MAR) missingness. With MAR missingness where the missingness of the covariate may be dependent on the covariate itself or the response, GEE may produce inefficient estimates. We use a technique by Xie and Paik (1997) that replaces missing quantities in the estimating function with consistent estimates. In our case, the model reduces to an inverse probability weighted (IPW) GEE model for the complete observations; in particular, the weights are the inverse of the probability of missingness (Xie and Paik, 1997). The technique can be thought of as an extension of the mean score method developed by Reilly and Pepe (1995) in the GEE

context. In general, using this technique, denoted IPW GEE (Robins et al., 1994) hereafter, helps to increase efficiency.

We also consider using ML, which is consistent and efficient in the presence of MAR (or MCAR) missingness provided we model the joint distribution of the outcome and the multiple informant covariates correctly. Specifically, we use Schluchter and Jackson's (1989) ML technique to include missingness in the multiple informant covariates using the relationship of the exponential distribution with Poisson loglinear models. Other methods such as one developed by Lipsitz and Ibrahim (1996) for use with categorical covariates (potentially with ignorable missingness) and assuming a proportional hazards model could also have been used. Ibrahim (1990) extended Schluchter and Jackson's (1989) approach to the generalized linear model (GLM) setting and named it the method of weights; the technique is reviewed in Horton and Laird (1998) for models with multiple informants. Therefore, the approach we used for fitting the marginal ML models with missingness could be considered a variation of Ibrahim's method of weights technique (Ibrahim, 1990). Recently, the method of weights method has been extended for situations such as correlated censored survival data with missing covariates (Lipsitz and Ibrahim, 2000) and random effects models for survival data with missing covariates (Herring et al., 2002). Our application of the ML approach is non-standard because we focus on estimation of marginal regression models.

We begin by reviewing the GEE approach (Horton et al., 2001) for finding marginal regression models and we describe the ML technique in Section 2. Section 3 describes extensions to GEE and ML such as including covariates and addresses when the multiple informant covariates have missing values. Results of a simulation considering missingness in the multiple informant covariates comparing IPW GEE (Xie and Paik, 1997; Robins et al., 1994) and ML are given in Section 4. We apply both techniques to the Stirling County Study and present results in Section 5. We conclude with closing remarks in Section 6.

2 Methods of Obtaining Estimates

Generalized Estimating Equations Approach

In this section, we first describe the GEE approach for finding marginal regression parameter estimates with a survival outcome and multiple informants as discrete covariates; then we consider how to obtain marginal ML estimates. We define notation similar to Schluchter and Jackson's (1989) technique of loglinear analysis for censored survival data with missingness in the covariates since we implement a variation of the technique in Section 3 when considering additional covariates and incorporating partially observed subjects. Specifically, we let T_i^0 be the true failure time for subject i , T_i be the observed failure time for individual i (thus $T_i = \min(T_i^0, C_i)$ where C_i is the random censoring time for individual i) and δ_i be 1 if T_i is a failure time and 0 if individual i is right censored.

To contrast with the ML approach, we describe a simplified version of the GEE method with an exponential survival distribution and no additional covariates (see Horton et al. 2001 for a complete description). Specifically, they modelled the expected number of events as:

$$m_i = E(\delta_i \mid T_i, \mathbf{X}_i, \beta) = T_i a_i = T_i e^{\mathbf{X}_i^T \beta} \quad (1)$$

where T_i is a known offset, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ is a vector of p multiple informant covariates and a_i , the model for the rate, is specified by $e^{\mathbf{X}_i^T \beta}$ with β defined as the marginal estimates of the relationships of each informant with response.

For example, if we consider two multiple informant covariates, the models fit using a statistical package such as SAS's PROC GENMOD (SAS Institute, 2005) are

$$\ln(m_i) \mid T_i, X_{i1}, \beta = \ln(T_i) + \beta_0 + \beta_1 X_{i1} \quad (2)$$

$$\ln(m_i) \mid T_i, X_{i2}, \alpha, \beta = \ln(T_i) + (\alpha_0 + \beta_0) + (\alpha_1 + \beta_1) X_{i2} \quad (3)$$

where β_0 and β_1 are the intercept and slope respectively for the marginal relationship between the covariate value for the first informant and response and $(\alpha_0 + \beta_0)$ and $(\alpha_1 + \beta_1)$ are the intercept and slope for the marginal relationship between the covariate value for the second informant and response.

Maximum Likelihood Technique

We now describe how to obtain marginal ML estimates by defining a joint likelihood model and deriving the marginal expectations. We use the same notation for T_i^0 , T_i , C_i and δ_i as described in Section 2. For simplicity, we consider two multiple informant covariates each with two levels, $\mathbf{X}_{ijk} = (X_{ij1}, X_{ij2})^T$. The multiple informant covariates define a 2×2 contingency table with four cells where $j = 0, 1$ denotes the value for X_{ij} and $k = 0, 1$ denotes the value for X_{2k} . We define E_{ijk} to be a multinomial indicator vector whose jk^{th} component is 1 if subject i is in cell jk and 0 otherwise and W_{ijk} indicates the cells of the table subject i could possibly be in, given the observed covariates. The number of failures among people in the jk^{th} cell is defined as $d_{jk} = \sum_{i=1}^n E_{ijk} \delta_i$; exposure time among persons in the jk^{th} cell is defined as $U_{jk} = \sum_{i=1}^n E_{ijk} b_i$ where b_i is the amount of exposure time contributed by subject i and $V_{jk} = \sum_{i=1}^n E_{ijk}$ is the number of subjects in the jk^{th} cell. We define a_{jk} as the rate of failure for the jk^{th} covariate pattern (Schluchter and Jackson, 1989). We also assume that censoring is noninformative.

Because exponential and Poisson likelihood models are proportional, specifying the failure times as an exponential distribution is equivalent to modeling the number of failures as a Poisson random variable (Laird and Olivier, 1981; Aitkin and Clayton, 1980; Holford, 1976; Holford, 1980). Specifically, we assume that $d_{jk} \sim \text{Poisson}(U_{jk} a_{jk})$; this implies that $E(d_{jk} / U_{jk}, X_{1j}, X_{2k}, a_{jk}) = U_{jk} a_{jk}$. Thus, we model $\ln(E(d_{jk} / U_{jk}, \mathbf{X}_{jk}, a_{jk})) = \ln(U_{jk}) + \mathbf{v}^T \mathbf{X}_{jk}$ with each $\ln(U_{jk})$ being a known offset of the time at risk for the jk^{th} cell and $\mathbf{X}_{jk}$ being a function of X_{1j}, X_{2k} ; we consider only conditional models that are saturated.

To find the marginal estimates for the relationship between each informant's covariate values and response, we use $d_{j+} \sim \text{Poisson}(\sum_k U_{jk} a_{jk})$ and $d_{+k} \sim \text{Poisson}(\sum_j U_{jk} a_{jk})$ which implies that $E(d_{j+})$ and $E(d_{+k})$ can be written in terms of $\mathbf{v}$ and U_{jk} . Then, defining the expected number of deaths for those with $X_{1j} = j$ summed over X_{2k} as $\theta_{j+} = E(d_{j+})$ for $j = 0, 1$ and similarly let $\theta_{+k} = E(d_{+k})$ for $k = 0, 1$, we write

$$\ln(\theta_{j+}) = \ln(U_{j+}) + \phi_0 + \phi_1 X_{1j} \quad (4)$$

$$\ln(\theta_{+k}) = \ln(U_{+k}) + \tau_0 + \tau_1 X_{2k} \quad (5)$$

where ϕ_0 is the baseline log mortality rate for people with $X_{1j} = 0$ and ϕ_1 describes the increase in log mortality rate for people with $X_{1j} = 1$ compared with $X_{1j} = 0$. The τ_0 and τ_1 parameters are defined similarly, except with X_{2k} . To find the marginal parameter estimates describing the relationship between each informant's covariate value and response, $\hat{\phi}_0$, $\hat{\phi}_1$, $\hat{\tau}_0$ and $\hat{\tau}_1$ are written as a function of U_{jk} and $\hat{\mathbf{v}}$ using the invariance property of maximum likelihood

estimates. We find that using ML and GEE yield identical marginal regression coefficients for this simple model.

To specify the likelihood for the complete dataset (as in Schluchter and Jackson, 1989), we consider the multiple informant covariates to be distributed as a multinomial distribution where π_{jk} is the probability associated with cell jk for $j = 0, 1$ and $k = 0, 1$ and V_{jk} is defined as in Section 2. We can estimate π_{jk} and ν separately since we have complete data; this is not the case with missingness in the multiple informant covariates (see Section 3).

3 Extensions

Including Covariates and Varying Hazard Rates over Time

We can extend the marginal model using GEE in Equations 2 and 3 to include time intervals and/or other covariates that are not multiple informants. In particular, Horton et al. (2001) fit GEE models to the Stirling County Study data with the followup period divided into $T = 4$ intervals to allow for different survival rates, e.g. fitting a piecewise exponential model; they also include covariates gender and age. The GEE model fit is

$$\ln(m_i | T_i, X_{1i}, Z_i, \beta) = \ln(T_i) + \beta_0 + \beta_1 X_{1i} + \beta_2 Z_i + \beta_3 X_{1i} Z_i \quad (6)$$

$$\ln(m_i | T_i, X_{2i}, Z_i, \alpha, \beta) = \ln(T_i) + (\alpha_0 + \beta_0) + (\alpha_1 + \beta_1) X_{2i} + (\alpha_2 + \beta_2) Z_i + (\alpha_3 + \beta_3) X_{2i} Z_i \quad (7)$$

where Z_i include the covariates not measured by multiple informants and time intervals.

We can also extend the ML technique to include time intervals and covariates; however, the resulting marginal models are somewhat limited. First, let Z be one covariate that is not a multiple informant where l denotes its level and define X_1 and X_2 as multiple informant covariates. We model $\ln(E(d_{jkl} | U_{jkl}, X_{jkl}, a_{jkl})) = \ln(U_{jkl}) + \nu^T X_{jkl}$ with each $\ln(U_{jkl})$ being a known offset of the time at risk for the jkl^{th} cell and X_{jkl} being a function of X_{1j} , X_{2k} , Z_l . The marginal models derived are equivalent to the saturated marginal model in Equations 6 and 7 with only one covariate:

$$\ln(\theta_{j+l}) = \ln(U_{j+l}) + \phi_0 + \phi_1 X_{1j} + \phi_2 Z_l + \phi_3 X_{1j} Z_l \quad (8)$$

$$\ln(\theta_{+kl}) = \ln(U_{+kl}) + \tau_0 + \tau_1 X_{2k} + \tau_2 Z_l + \tau_3 X_{2k} Z_l \quad (9)$$

As without a covariate, we find a direct correspondence between estimation using GEE and ML. Specifically, $\hat{\phi}_0 \equiv \hat{\beta}_0$, $\hat{\phi}_1 \equiv \hat{\beta}_1$, $\hat{\phi}_2 \equiv \hat{\beta}_2$, $\hat{\phi}_3 \equiv \hat{\beta}_3$, $\hat{\tau}_0 \equiv \hat{\alpha}_0 + \hat{\beta}_0$, $\hat{\tau}_1 \equiv \hat{\alpha}_1 + \hat{\beta}_1$, $\hat{\tau}_2 \equiv \hat{\alpha}_2 + \hat{\beta}_2$, $\hat{\tau}_3 \equiv \hat{\alpha}_3 + \hat{\beta}_3$. However, our likelihood approach does not permit an unsaturated marginal model without an informant by covariate interaction, as given in Equations 10 and 11:

$$\ln(\theta_{j+l}) = \ln(U_{j+l}) + \phi_0 + \phi_1 X_{1j} + \phi_2 Z_l \quad (10)$$

$$\ln(\theta_{+kl}) = \ln(U_{+kl}) + \tau_0 + \tau_1 X_{2k} + \tau_2 Z_l \quad (11)$$

To demonstrate the difficulty encountered, we begin by writing the marginal parameter estimates describing the relationship between each informant's covariate value and response, $\hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2, \hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2$, as a function of U_{jkl} and $\hat{\nu}$. For example, we find that

$$\hat{\phi}_1 = \ln \left(\frac{\hat{\theta}_{1+l} U_{0+l}}{\hat{\theta}_{0+l} U_{1+l}} \right) = \ln \left(\frac{e^{\hat{\nu}_0 + \hat{\nu}_1 + \hat{\nu}_l (U_{11l} e^{\hat{\nu}_2 + \hat{\nu}_3 + U_{10l}})}}{U_{1+l}} \right) \left/ \frac{e^{\hat{\nu}_0 + \hat{\nu}_l (U_{01l} e^{\hat{\nu}_2 + U_{00l}})}}{U_{0+l}} \right. \right). \text{ We note}$$

that the U_{jkl} terms in this expression will not cancel properly to make $\hat{\phi}_1$ from ML equivalent to $\hat{\beta}_1$ (the regression parameter from the GEE approach) since the U_{jkl} are dependent on l .

Because it is not reasonable to make assumptions about the U_{jkl} values, using this approach to construct marginal models for ML does not accommodate unsaturated models such as the proportional hazards model where $\phi_3 = \tau_3 = 0$. Essentially, we cannot find a likelihood model that has the marginal interpretation that the relationship between an informant and covariate does not change for different values of the covariate.

We note that the covariate Z could be taken as time, which would correspond to a model for fitting a different hazard at different time points, or fitting a piecewise exponential model. Unfortunately, without a saturated model that allows the relationship between each time point and multiple informant to be left unconstrained, the likelihood method cannot be accommodated. In essence, the likelihood approach does not allow us to fit proportional hazards models, since they assume that the relationship between each time point and multiple informant is identical.

In summary, we find that this likelihood modeling approach requires saturated models if marginal and conditional parameters are to have the same interpretations. That is, the GEE approach is consistent with a likelihood model in simple saturated cases, but is not the same as the likelihood approach in general. One drawback of saturated marginal models is that the number of terms in the model becomes cumbersome as the number of covariates increases.

Incorporating Missingness in the Multiple Informant Covariates

Missingness in multiple informant covariate reports is quite common since reports must be solicited from multiple people for each subject; in the Stirling County Study data 126 observations have missing multiple informant covariate values. We now describe using IPW GEE as well as how to extend ML to include missing values. To fit IPW GEE models as in Horton et al. (2001), we define R_{hi} for $h = 1, 2$ where R_{hi} is 1 if X_{hi} is missing and zero otherwise. Then, we use the information from the study to predict $\mathbf{p}_h = \mathbf{P}(\mathbf{R}_h = \mathbf{1})$ for $h = 1, 2$ using logistic regression. That is, $\mathbf{p}_h = \mathbf{P}(\mathbf{R}_h = \mathbf{1} | \delta_h, \mathbf{T}_h, \mathbf{X}_h)$; this model need not be parsimonious as the aim is to avoid bias when predicting $\hat{\mathbf{p}}_h$. The inverses of $\hat{\mathbf{p}}_h$ for $h = 1, 2$ are used as weights in the marginal models fit using GEE. Specifically, with two multiple informants, $\hat{\mathbf{p}}_1^{-1}$ weights the equation with $\mathbf{X}_1$ as the covariate and $\hat{\mathbf{p}}_2^{-1}$ weights the equation with $\mathbf{X}_2$ as the covariate. Horton et al. assume a working independence correlation structure; thus, the model is reduced to a Poisson regression with estimated weights $\hat{\mathbf{p}}_h$. A subject contributes an observation for each informant that has complete information. The Poisson regression models with estimated weights are fit in SAS (2005), as described in Horton et al. (2001). To find estimates of the standard errors of $\hat{\alpha}$ and $\hat{\beta}$, we use the bootstrap (Efron and Tibshirani, 1993) with 1000 replications as done in Horton et al. (2001).

The preceding paragraph describes the IPW GEE technique with monotone missingness; however, the example we consider later has non-monotone missingness. In this situation, fitting missingness models correctly is difficult, particularly in the case of estimating the probability of missing observations from one covariate if the other covariate is a statistically significant predictor of the missingness in the first covariate. For this reason, we assume a variation of MAR proposed by Robins and Gill (1997) which does allow us to easily estimate appropriate weights; this approach is similar to what Horton et al. (2001) used. Specifically, in the case of two predictors, to find weights for X_2 missingness among those missing X_1 we use the entire dataset; we use only complete observations (not considering the X_1 values at all) to predict weights for X_2 missingness among those with complete X_1 observations. Although in theory,

Robins and Rotnitzky (1995) show that there is a class of efficient estimators when missingness is non-monotone, methods to utilize this result have not been fully developed.

To fit marginal models with missingness in the multiple informant covariates using ML, we adapt Schluchter and Jackson's (1989) technique for including covariates with ignorable missingness to our situation with multiple informant covariates. Essentially, we obtain estimates for ν and π_{jk} using the expectation-maximization (EM) algorithm (Dempster et al., 1977) (described in detail below); we then transform the estimates to the marginal slope and intercept estimates, $\hat{\phi}_0$, $\hat{\phi}_1$, $\hat{\tau}_0$ and $\hat{\tau}_1$. Unlike in the complete data case, the $\hat{\pi}_{jk}$ and $\hat{\nu}$ cannot be estimated separately. To find estimates of the standard errors for $\hat{\phi}_0$, $\hat{\phi}_1$, $\hat{\tau}_0$ and $\hat{\tau}_1$, we calculate the variance of the estimates from 1000 bootstrap samples (Efron and Tibshirani, 1993).

To implement the EM algorithm in the case of two multiple informant covariates, first we classify the observations according to X_1 , X_2 categories (0,1, missing). Then, depending on the missingness pattern, we calculate W_{ij} , the cell of the table subject i could possibly be in, given the observed covariates. Next, the E step of the EM algorithm is performed first by initializing the number of failures, amount of followup time and number of subjects for each X_1 , X_2 cell. Starting values for π were the number of failures divided by the followup time for each cell; when there were no failures, however, the starting value was assigned to be 0.05 to help with convergence of the algorithm. Then, we calculate $P(E_{ij} = 1/U_j, \delta_j, W_{ij}) = P_{ij}$; this is the probability of a subject being in a particular cell of the contingency table given the followup time, censoring indicator and cell of the table the subject could possibly be in given the observed covariate(s) (as described in Appendix B of Schluchter and Jackson, 1989). The fractional number of failures, exposure time and people are calculated based on the P_{ij} given above. Since we are assuming a saturated model, Schluchter and Jackson (1989) note that the M step consists of ν being the fractional number of failures divided by the total number of failures and the π as being the fractional number of failures divided by the fractional follow-up time for each X_1 ; X_2 combination. The algorithm is iterated until convergence; our stopping criteria was when the squared values between estimates was less than 10^{-8} . Using the estimates of ν and π obtained from the EM algorithm, the estimates of ϕ_0 , ϕ_1 , τ_0 , τ_1 are found.

4 Simulation

In this section, we compare the variance of the marginal estimates from ML and IPW GEE by performing a small simulation study in situations where the appropriate underlying model is fit by both. Specifically, we consider two multiple informants, X_1 and X_2 , each with two levels as well as a dichotomous age covariate, Z . To generate simulated data, we draw X_1 and X_2 for $Z = 0$ (age < 50) and then $Z = 1$ (age 50+) from multinomial distributions with probabilities calculated from the Stirling County Study example. Specifically, for $Z = 0$, the probability of people with $X_1 = X_2 = 1$ is 0.0244; for $X_1 = 0$; $X_2 = 1$ is 0.0518; for $X_1 = 1$, $X_2 = 0$ is 0.0308; and for $X_1 = X_2 = 0$ is 0.4481. The calculated probabilities for $Z = 1$ with $X_1 = X_2 = 1$ is 0.0140; for $X_1 = 0$; $X_2 = 1$ is 0.0538; for $X_1 = 1$; $X_2 = 0$ is 0.0478 and for $X_1 = X_2 = 0$ is 0.3293. To find the survival times, we draw from exponential distributions with rates also calculated from the example. Specifically, the rate for each of the 8 cells a_{jkl} where $j, k, l = 0, 1$ are $a_{110} = 0.0163$; $a_{010} = 0.0062$; $a_{100} = 0.0102$; $a_{000} = 0.0018$; $a_{111} = 0.0862$; $a_{011} = 0.0391$; $a_{101} = 0.0529$; $a_{001} = 0.0381$; these were calculated from the EM algorithm with 1079 observations and are the number of failures divided by the follow-up time. Furthermore, the values for ν calculated using a_{jkl} are: $\nu_0 = -6.3200$; $\nu_1 = 1.7346$; $\nu_2 = 1.2368$; $\nu_3 = -0.7680$; $\nu_4 = 3.0524$; $\nu_5 = -1.4064$; $\nu_6 = -1.2109$; $\nu_7 = 1.2303$. Using U_{jkl} and ν to calculate the marginal regression parameters from Equations 8 and 9, we find $\phi_0 = -6.1744$; $\phi_1 = 1.7562$; $\phi_2 = 2.9105$; $\phi_3 = -1.3318$; $\tau_0 = -5.9549$; $\tau_1 = 1.3730$; $\tau_2 = 2.7315$; $\tau_3 = -1.2348$. Censoring is determined by

the survival times. Specifically, the survival status is considered censored if the survival time generated is 16 or more, since the range of survival times in the example is from 0 to 16.

We generate data where X_2 has MAR missingness and X_1 is complete. We consider simulations with monotone missingness to make it straightforward to compute the weights in IPW GEE; however, ML can handle more general missingness patterns easily. Specifically, we consider four different missingness scenarios: missingness in X_2 depending on X_1 (10% and 20% missingness in X_2) and missingness in X_2 depending on Z (10% and 20% missingness in X_2) (Horton and Lipsitz, 1999). With ignorable missingness in X_2 we assume the probability of missingness for ages younger than 50 is 0.05 (0.10), while the probability of missingness for ages 50 and older is 0.15 (0.30) when considering 10% and 20% missingness, respectively; therefore, we assume that the probability of missing X_2 increases with age. Because of the unequal distribution of X_1 values in the Stirling County data, we assume the probability of missingness is 0.09 (0.18) for $X_1 = 0$ (no psychiatric diagnosis), while the probability of missingness for $X_1 = 1$ (psychiatric diagnosis) is 0.19 (0.38) for 10% and 20% missingness, respectively; therefore, we assume that those with a psychiatric diagnosis have a higher proportion of missing X_2 values. With nonignorable missingness, we assume that the probability of X_2 missingness is 0 when $X_2 = 0$ and is 10% when $X_2 = 1$ and that the amount of missingness is either 10% or 20%.

After generating the data, we fit GEE (SAS Institute, 2005) and ML (R Development Core Team, 2003) models based on the complete cases. We also fit IPW GEE using PROC LOGISTIC in SAS to create weights and then PROC GENMOD with the SCWGT weight option to fit the IPW GEE models (SAS Institute, 2005) and ML models by a user-written EM algorithm (R Development Core Team, 2003) based on all cases. We generate 500 random samples for each scenario; standard errors of the parameter estimates are calculated using the sample variance from the 500 data sets.

When considering only the complete cases, the estimates of the marginal regression parameters are identical when found using GEE and ML; the estimates of their standard errors are also identical (not reported). When comparing estimates found from IPW GEE compared with ML, the marginal estimates of $\phi_0, \phi_1, \phi_2, \phi_3$ describing the relationship between X_1 and survival are identical, but the estimates of $\tau_0, \tau_1, \tau_2, \tau_3$ describing the relationship between X_2 and survival differ since we have considered a situation with missingness only in X_2 . Thus, in Table 1, we present the latter only. The means of the ML and GEE estimates of the parameters, $\tau_0, \tau_1, \tau_2, \tau_3$, did not differ significantly based on paired t tests. This implies that bias will be similar for the two approaches. Therefore, in Table 1, we only present ratios of the standard errors with the standard error from ML in the numerator and the standard error from IPW GEE in the denominator. Assuming an exponential model with ignorable missingness, we find that ML offers a small amount of efficiency for estimating $\hat{\tau}_0$ and $\hat{\tau}_2$, particularly when the missingness in X_2 depends on X_1 . According to Table 1, ML and IPW GEE also perform similarly in the nonignorable missingness situation. Therefore, in general the efficiency for estimating the variance of the parameter estimates using ML and IPW GEE is similar.

5 Application to the Stirling County Study

Of the 1079 Stirling County subjects, 953 had complete X_1 (physician reported psychopathology) and X_2 (self reported psychopathology) information (Table 2) (Horton et al., 2001). Additional description of the data is reported in Horton et al. (2001); instead we focus on comparing the IPW GEE and ML model results. We start by considering only multiple informant covariates in the models. Specifically, we fit the model in Equations 2 and 3 for GEE and Equations 4 and 5 for ML. We present estimates of the marginal models predicting survival from physician reported psychiatric disorders and also predicting survival from self reported

psychiatric disorders $\hat{\varphi}_0$ gives the baseline log mortality rate for people whose physicians did not report as having a disorder $\hat{\varphi}_1$ gives the increase in log mortality rate for people whose physicians did report as having a disorder. Similarly $\hat{\tau}_0$ and $\hat{\tau}_1$ are the intercept and slope for the marginal relationship between self reported psychopathology and survival. In Table 3, we present estimates using GEE and ML considering only complete cases ($n = 953$). A complete case (CC) analysis can lead to biased estimates and is generally not the most efficient technique for data with missingness; here 126 cases are simply not considered in the analysis. Since we showed in Section 2 that GEE and ML yield the same estimates when considering a saturated marginal model, we find that the estimates are identical for the GEE and ML approaches. Variance of the estimates from GEE is obtained using the empirical variance formula (Horton et al., 2001) and the variances of the estimates from ML are calculated from the bootstrap; we find similar estimates for the slope variances.

In Table 3, we also consider all 1079 cases and fit models using IPW GEE and ML; standard errors of the estimates using IPW GEE and ML are both calculated using the bootstrap. To implement IPW GEE in the Stirling County Study, Horton et al. (2001) use age, survival status and gender to predict the probability of observing physician reports, X_1 . They use main effects of age, gender, survival status, X_1 and interactions of gender and age as well as X_1 and age to predict the probability of observing X_2 reports. We use these weights for our model, even though we consider a model without covariates since we want to avoid introducing bias (Horton et al., 2001).

One difficulty that arises in this dataset is that the missingness is not monotone. Thus, fitting the missingness models correctly is difficult, particularly in the case of estimating the probability of missing X_2 since X_1 is a statistically significant predictor of the missingness in X_2 . We use an adhoc technique for finding estimates in Table 3 (similar to what Horton et al. (2001) used). Specifically, to find weights for X_2 missingness among those missing X_1 we use the entire dataset; we use only complete observations (not considering the X_1 values at all) to predict weights for X_2 missingness among those with complete X_1 observations. Alternatively, Robins and Rotnitzky (1995) show that there is a class of efficient estimators when missingness is non-monotone; however, the methods to utilize this result have not been fully developed.

In Table 3, we find similar marginal estimates using the ML and IPW GEE approaches; the standard error of the estimates obtained from ML and IPW GEE is also similar. The ratios of ML compared to IPW GEE include 95% for $\widehat{se}(\hat{\varphi}_1)$ and 96% for $\widehat{se}(\hat{\tau}_1)$. Furthermore, $\hat{\varphi}_1$ and $\hat{\tau}_1$ are similar (based on a test of $\alpha_1 = 0$ in the GEE setting), indicating that the marginal relationship between physician report of psychopathology and mortality is similar to that between self report of psychopathology and mortality. This is similar to the results reported in Horton et al. (2001); they also found that the two informants agree in terms of mortality risk.

Next we consider addition of the age covariate in the models; instead of considering three levels of age (as in Horton et al. (2001)), for simplicity, we define it in two groups ($< 50, 50+$). Specifically, we fit the model in Equations 6 and 7 for GEE assuming one covariate and the model in Equations 8 and 9 for ML. Again we fit four models (GEE and ML for complete cases, IPW GEE and ML for all cases). Results are in Tables 4 and 5 where $\hat{\varphi}_0$ is the baseline log mortality rate for people aged younger than 50 and without physician reported psychiatric disorders; $\hat{\varphi}_1$ is the increase in log mortality rate for those with physician-reported psychiatric disorders holding age constant; $\hat{\varphi}_2$ is the increase in log mortality rate for those aged 50 years or older holding physician diagnosis constant and $\hat{\varphi}_3$ is the interaction between physician reported psychopathology and age. $\hat{\tau}_0$, $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}_3$ are the corresponding estimates involving

self reported psychopathology. Similar to the model without age, $\hat{\phi}_1$ and $\hat{\tau}_1$ are similar, indicating that self reports of psychopathology have a similar effect on mortality as physician reports. The estimates of the age coefficients ($\hat{\phi}_2$, $\hat{\tau}_2$) are also similar and are both statistically significant predictors of survival; that is, older patients have lower survival rates. The interactions between diagnosis and age ($\hat{\phi}_3$, $\hat{\tau}_3$) are significant as well, indicating that the association of psychiatric morbidity and mortality is strongest for those who died before the age of 50. When we compare the standard error of the estimates derived using ML compared with those from IPW GEE in the presence of the age covariate, the ratio of standard errors for ML to IPW GEE is 95% for $\hat{\phi}_1$, 86% for $\hat{\phi}_2$, 94% for $\hat{\phi}_3$, 96% for $\hat{\tau}_1$, 88% for $\hat{\tau}_2$ and 93% for $\hat{\tau}_3$. The results are somewhat similar to the simulation; in particular, the decrease in estimated standard error for $\hat{\phi}_2$ and $\hat{\tau}_2$ is more pronounced than for $\hat{\phi}_1$, $\hat{\phi}_3$, $\hat{\tau}_1$ and $\hat{\tau}_3$. Therefore, in this example, ML offers a modest decrease in estimated variance for the estimated regression coefficients which do not involve multiple informants.

This section has focused on comparison of the GEE and ML approaches for fitting marginal models comparing the relationship of physician reported psychopathology and mortality with self report of psychopathology and mortality with inclusion of an age covariate. Because the marginal regression parameter estimates remain largely unchanged in the ML compared with the GEE approach, our analysis yields results similar to that reported in Horton et al. (2001). Specifically, we find that although physician reported psychiatric disorders and self reported psychiatric disorders are not highly correlated, the relationship of each individually with mortality is similar. With respect to the interaction between diagnosis of psychiatric disorders and age, we find that for younger people, the diagnosis of psychiatric disorders has a particularly high effect on mortality.

6 Conclusion

We have described using GEE and ML for fitting marginal models with multiple informants. Our example includes only one covariate (age), although GEE can be easily modified to incorporate additional covariates and ML can be extended to include covariates as long as we fit ML models leading to saturated marginal models. Another model to consider is where the marginal slope coefficients are equal, e.g., $\phi_1 = \tau_1$. This constrained model is easily fit using GEE by assuming $\alpha_1 = 0$ in Equations 2 and 3; however, in the ML setting, implementation is not straightforward. For example, with two informants, to constrain the marginal slope coefficients, we need $E(d_{1+}) = E(d_{+1})$. However, the obvious choice of parameters for the conditional model, $v_1 = v_2$ and $v_3 = 0$, with $U_{10} = U_{01}$ leads us to a model with both constrained slopes and intercepts. Therefore, fitting constrained slope ML models is a subject for additional research.

Using our likelihood model, it is difficult to find models in which estimates from ML and GEE match; this is in contrast to other settings. For example, when Y is a continuous response and the multiple informant covariates are also continuous or when Y is discrete and the multiple informant covariates are also discrete, the ML and GEE approaches yield the same estimates in complete case situations (Litman et al., 2005; O'Brien, 2003). In addition, we find that in this situation with a time to event response and discrete multiple informant covariates likelihood parameters with the same marginal interpretation as in GEE can only be found in special situations. However, in the continuous and discrete cases described above, marginal models using ML can be developed that match the GEE approach. We speculate that one reason the difficulty arises is that we assume with ML that the exponential distribution is proportional to the Poisson distribution which may not be the case in practice.

We compare efficiency of the estimates from the IPW GEE method to those from the ML technique in the presence of monotone missingness in our simulation and find that estimates from the IPW GEE method do not lose much efficiency compared with ML for multiple informant parameters. Therefore, we infer that the IPW GEE method is close to fully efficient in the presence of monotone missingness. We also compare the two methods with the Stirling County data that has non-monotone missingness. In this case, we find that estimates from the IPW GEE method are also similar to those from ML in terms of efficiency, although the efficiency losses due to IPW GEE compared with ML are larger than in the simulation. In the example, we used an adhoc method that is not fully efficient for calculating the weights for the IPW GEE with non-monotone missingness. While Robins and Rotnitzky (1995) have shown that there are optimal estimates using IPW GEE with non-monotone missingness, methods to find them have not yet been developed and generating them would be beyond the scope of this manuscript.

In general, we find that IPW GEE does not lose much efficiency compared with ML (9% was the most lost) in the situation with monotone missingness under the exponential model. However, we note that in our example the slight differences in efficiency occur in the intercept and age coefficients, not the main effect of psychopathology or the interaction between psychopathology and age. This is consistent with results of a study investigating the relationship between teacher's report of psychopathology with mental health service utilization to the relationship between parent's report of psychopathology and utilization where the teacher reports had missing information but the parent reports were complete (Horton and Laird, 1998; Horton and Laird, 2001). In particular, they found efficiency gains for estimating all coefficients except those for which the information was partly observed. In our example, self reported psychiatric disorders, X_2 , are assumed to have missingness and these coefficients reveal no efficiency gain. It is thought that since no additional information is available for the missing covariate, little gain in efficiency for estimation of the parameter is found (Horton and Laird, 1998).

While our example is in the psychiatric realm, time to event data with multiple informant covariates may arise in other settings. For instance, clinical trials frequently have survival outcomes and assessment of quality of life could be performed by multiple informants. In this setting, marginal regression models for each informant's measure of quality of life could be used to predict survival. Also, most studies are subject to missing data; hence, consideration of efficient methods for dealing with the missingness is important.

Acknowledgements

We are grateful for the support provided by the National Institute of Mental Health (NIMH) grant number MH54693 and National Institute of Health (NIH) grant number T32-MH017119. Support for the Stirling County Study was provided by NIH grant number R01-MH39576.

References

- Aitkin M, Clayton D. "The fitting of exponential, weibull and extreme value distributions to complex censored survival data using GLIM". *Journal of the Royal Statistical Society* 1980;29:156–163.
- Dempster AP, Laird NM, Rubin DB. "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)". *Journal of the Royal Statistical Society Series B* 1977;39:1–38.
- Efron, B.; Tibshirani, RJ. Chapman and Hall; New York, NY: 1993. An introduction to the bootstrap.
- Herring AH, Ibrahim JG, Lipsitz SR. "Frailty models with missing covariates". *Biometrics* 2002;58:98–109. [PubMed: 11890332]
- Holford TR. "Life tables with concomitant information". *Biometrics* 1976;32:587–597. [PubMed: 963172]

- Holford TR. "The analysis of rates and survivorship using loglinear models". *Biometrics* 1980;36:299–306. [PubMed: 7407317]
- Horton NJ, Laird NM. "Maximum likelihood analysis of generalized linear models with missing covariates". *Statistical Methods in Medical Research* 1999;8:37–50. [PubMed: 10347859]
- Horton NJ, Laird NM, Zahner GEP. "Use of multiple informant data as a predictor in psychiatric epidemiology". *International Journal of Methods in Psychiatric Research* 1999;8:6–18.
- Horton NJ, Lipsitz SR. "Review of software to fit generalized estimating equation regression models". *The American Statistician* 1999;53:160–169.
- Horton NJ, Laird NM. "Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information". *Biometrics* 2001;57:34–42. [PubMed: 11252616]
- Horton NJ, Laird NM, Murphy JM, Monson RR, Sobol AM, Leighton AH. "Multiple informants: Mortality associated with psychiatric disorders in the Stirling County Study". *American Journal of Epidemiology* 2001;154:649–656. [PubMed: 11581099]
- Ibrahim JG. "Incomplete data in generalized linear models". *Journal of the American Statistical Association* 1990;85:765–769.
- Laird NM, Olivier D. "Covariance analysis of censored survival data using log-linear analysis techniques". *Journal of the American Statistical Association* 1981;76:231–240.
- Leighton, AH. Basic Books, Inc; New York: 1959. *My name is Legion: The Stirling County Study of psychiatric disorder and sociocultural environment.*
- Lipsitz SR, Ibrahim JG. "Using the EM algorithm for survival data with incomplete categorical covariates". *Lifetime Data Analysis* 1996;2:5–14. [PubMed: 9384645]
- Lipsitz SR, Ibrahim JG. "Estimation with correlated censored survival data with missing covariates". *Biostatistics* 2000;1:315–327. [PubMed: 12933512]
- Litman HJ, Horton NJ, Hernández B, Laird NM. "Estimation of marginal regression models with multiple source predictors". *Handbook of Statistics*. 2005Submitted
- Murphy JM. "Continuities in community-based psychiatric epidemiology". *Archives of General Psychiatry* 1980;37:1215–1223. [PubMed: 7436683]
- Murphy JM, Monson RR, Olivier DC, Sobol AM, Leighton AH. "Affective disorders and mortality: A general population study". *Archives of General Psychiatry* 1987;44:473–480. [PubMed: 3555383]
- Murphy JM, Monson RR, Olivier DC, Sobol AM, Pratt LA, Leighton AH. "Mortality risk and psychiatric disorders: Results of a general physician survey". *Social Psychiatry and Psychiatric Epidemiology* 1989;24:134–142. [PubMed: 2500711]
- Murphy JM, Laird NM, Monson RR, Sobol AM, Leighton AH. "A 40-year perspective on the prevalence of depression: The Stirling County Study". *Archives of General Psychiatry* 2000;57:209–215. [PubMed: 10711905]
- O'Brien LM, Fitzmaurice GM, Horton NJ. "Maximum likelihood estimation of marginal pairwise associations with multiple source predictors". Currently under review after revision
- Pepe MS, Whitaker RC, Seidel K. "Estimating and comparing univariate associations with application to the prediction of adult obesity". *Statistics in Medicine* 1999;18:163–173. [PubMed: 10028137]
- R Development Core Team. Vienna; Austria: 2003. *R Foundation for Statistical Computing, R: A language and environment for statistical computing, Version 1.8.1.*
- Reilly M, Pepe M. "A mean score method for missing and auxiliary covariates data in regression models". *Biometrika* 1995;82:299–314.
- Robins JM, Gill RD. "Non-response models for the analysis of non-monotone ignorable missing data". *Statistics in Medicine* 1997;16:39–56. [PubMed: 9004382]
- Robins JM, Rotnitzky A, Zhao LP. "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American Statistical Association* 1994;89:836–866.
- Robins JM, Rotnitzky A. "Semiparametric efficiency in multivariate regression models with missing data". *Journal of the American Statistical Association* 1995;90:122–129.
- SAS Institute. Cary; NC, USA: 2005. *SAS/STAT Software Version 9.1.*
- Schluchter MD, Jackson KL. "Log-linear analysis of censored survival data with partially observed covariates". *Journal of the American Statistical Association* 1989;84:42–52.

Xie F, Paik MC. "Generalized estimating equation models for binary outcomes with missing covariates". *Biometrics* 1997;53:1458–1466. [PubMed: 9423260]

Table 1

Simulation Results: Ratios of Standard Errors from ML Compared to IPW GEE

Distribution	Missingness Type	% Missing X_2	Depending on	Self Reported Psychopathology (PP)			
				Intercept τ_0^*	PP τ_1^*	Age τ_2^*	Interaction τ_3^*
Exponential	Ignorable	10	X_1	99%	99%	98%	99%
Exponential	Ignorable	20	X_1	92%	99%	91%	98%
Exponential	Ignorable	10	Z	98%	100%	98%	100%
Exponential	Ignorable	20	Z	96%	100%	95%	100%
Exponential	Nonignorable	10	X_2	99%	97%	99%	98%
Exponential	Nonignorable	20	X_2	99%	100%	99%	100%

Table 2

Distribution of Physician and Self Reports of Psychopathology from the Stirling County Study

Physician report	Missing report	Self report		Total
		No disorder	Report of disorder	
Missing report	0	44	6	50
No disorder	58	748	81	887
Report of disorder	18	90	34	142
Total	76	882	121	1079

Table 3
GEE and ML Model Slope Estimates and their Standard Errors (se)

Model	Physician Reported Psychopathology		Self Reported Psychopathology	
	ϕ_1	$se(\phi_1)$	τ_1	$se(\tau_1)$
GEE (complete cases)	0.495	0.166	0.473	0.175
ML (complete cases)	0.495	0.171	0.473	0.178
IPW GEE (all cases)	0.596	0.159	0.504	0.170
ML (all cases)	0.597	0.151	0.525	0.163

Table 4
GEE and ML Model Estimates Including the Age Covariate

Model	Physician Reported Psychopathology			Self Reported Psychopathology		
	Psychopathology ϕ_1	Age ϕ_2	Interaction ϕ_3	Psychopathology τ_1	Age τ_2	Interaction τ_3
GEE (complete cases)	1.797	3.111	-1.508	1.172	2.838	-1.041
ML (complete cases)	1.797	3.111	-1.508	1.172	2.838	-1.041
IPW GEE (all cases)	1.706	2.953	-1.274	1.384	2.793	-1.257
ML (all cases)	1.738	2.892	-1.308	1.368	2.706	-1.220

Table 5
GEE and ML Model Standard Errors Including the Age Covariate

Model	Physician Reported Psychopathology			Self Reported Psychopathology		
	Psychopathology $se(\phi_1)$	Age $se(\phi_2)$	Interaction $se(\phi_3)$	Psychopathology $se(\tau_1)$	Age $se(\tau_2)$	Interaction $se(\tau_3)$
GEE (complete cases)	0.407	0.285	0.445	0.468	0.245	0.503
ML (complete cases)	0.420	0.276	0.460	0.485	0.245	0.504
IPW GEE (all cases)	0.396	0.279	0.433	0.443	0.259	0.476
ML (all cases)	0.375	0.240	0.408	0.426	0.228	0.442