Published in final edited form as:
Stat Med. 2007 February 28; 26(5): 1055–1068.

Incorporating Missingness for Estimation of Marginal Regression Models with Multiple Source Predictors

Heather J. Litman[1], **Nicholas J. Horton**[2], **Bernardo Hernández**[3], and **Nan M. Laird**[4]

1 *New England Research Institutes, 9 Galen St, Watertown, MA, USA, 02472*

2 *Department of Mathematics and Statistics, Smith College, Northampton, MA, USA, 01063*

3 *Dirección de Salud Reproductiva, Centro de Investigación en Salud Poblacional, Instituto Nacional de Salud Pública, Cuernavaca, Morelos, México 62508*

4 *Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA, 02115*

Summary

Multiple informant data refers to information obtained from different individuals or sources used to measure the same construct; for example, researchers might collect information regarding child psychopathology from the child's teacher and the child's parent. Frequently, studies with multiple informants have incomplete observations; in some cases the missingness of informants is substantial. We introduce a *Maximum Likelihood* (ML) technique to fit models with multiple informants as predictors that permits missingness in the predictors as well as the response. We provide closed form solutions when possible and analytically compare the ML technique to the existing *Generalized Estimating Equations* (GEE) approach. We demonstrate that the ML approach can be used to compare the effect of the informants on response without standardizing the data. Simulations incorporating missingness show that ML is more efficient than the existing GEE method. In the presence of MCAR missing data, we find through a simulation study that the ML approach is robust to a relatively extreme departure from the normality assumption. We implement both methods in a study investigating the association between physical activity and obesity with activity measured using multiple informants (children and their mothers).

Keywords

missingness in the response; missingness in the covariates; missing completely at random; multiple informants; maximum likelihood

1 Introduction

Multiple informant data refers to information obtained from different individuals or sources used to measure the same construct. Obtaining accurate physical activity and inactivity measurements, for example, can be difficult and time-consuming. Hernández et al. [1] developed a self-reported questionnaire to assess these measures and performed a validation study to compare responses from the questionnaire to a comparison criterion of 24-hour recalls. The validation study collected activity measurements from multiple informants: children and their mothers. In addition to using this study for validation of their scales, they also planned to design a larger study [2] to more fully investigate the relationship between physical activity and obesity. However, collecting information from mothers in a large scale study is not feasible,

Address for correspondence: Heather J. Litman, New England Research Institutes, 9 Galen St, Watertown, MA 02472, Email: hlitman@neriscience.com, Phone: 617 923 7747 x297, Fax: 617 926 8246.

so it is of interest to compare the predictive values of the mother's and child's reports from a subset of the larger study [2] in the context of study design.

Missing data is prevalent in multiple informant research; as the number of informants increases, the risk of missingness also rises. In the Hernández et al. [2] study with body mass index (BMI) as response and physical activity reported by the child and the child's mother, 26% of the cases have missingness either in the response, the multiple informant variables or both. In a study of mental health service utilization in children from Connecticut, 43% of cases had missingness [3], [4], [5], [6], [7]. Having a large amount of missingness can lead to inferential problems including bias and efficiency loss [8].

In previous work [9], we reviewed a GEE approach [10], [11] and provided a novel *Maximum Likelihood* (ML) technique for analysis of multiple informants as predictors. In addition, we showed that the GEE and ML approaches yield identical estimates for a broad range of models, but the ML technique permits more flexibility in modeling. That research did not consider missing data; however, general properties of the GEE and ML methods in the presence of missingness are well known. If nonresponse is *Missing Completely At Random* (MCAR) [12], then the GEE approach gives unbiased yet potentially inefficient estimates [11], [13], while ML leads to consistent and asymptotically normal estimates assuming *Missing At Random* (MAR) missingness [12], [14], [15] provided the likelihood model is correct. MCAR missingness does not depend on values of the variables (e.g., observed cases are a random subsample of all cases) while MAR missingness depends only on the components of the variable that are observed.

In this paper, rather than simply focus on the common situation with missingness in the response, we consider the non-standard case of fitting marginal regression models with MCAR missingness in the covariates as well as the response. We assume MCAR missingness, partly because it is a reasonable assumption for our data and also to ensure consistency for the GEE approach. We will show under what conditions the GEE and ML approaches yield the same estimates in the presence of MCAR missing data.

In general, when fitting marginal regression models with multiple informant covariates the goal is to compare the relationship of one informant with response to the other informant with response. In data without missingness, this is done by standardizing the multiple informant data and comparing the marginal regression coefficients. However, standardizing data with missingness can be undesirable, so we describe how to use ML to compare the effect of the informants on response without doing so.

Another goal of this paper is to evaluate the efficiency of using ML compared with GEE for fitting marginal regression models in the presence of MCAR missingness. It is expected that ML will offer some efficiency compared with the GEE approach. One well-known drawback to using ML is that it requires the assumption of multivariate normality; thus, we investigate how robust the ML approach is to this assumption in the presence of MCAR missingness.

We begin in Section 2 by describing how to implement the GEE approach incorporating missingness in the response and/or multiple informant covariates. In Section 3, we derive closed form solutions for the ML approach and analytically compare the two methods in the presence of missing data. We also describe how to use ML to compare the effect of each informant on response without standardizing the data. We compare the efficiency of the ML approach with the GEE technique through a simulation study in Section 4. We begin Section 5 by describing an application of the GEE and ML methods using data from Hernández’s study of physical activity/inactivity and obesity [1], [2]. We also describe a simulation that demonstrates how the ML approach is robust to the assumption of multivariate normality in the presence of MCAR missingness.

2 Generalized Estimating Equations Approach Incorporating Missingness

Pepe et al. [11] and Horton et al. [10] describe a nonstandard GEE approach for fitting multiple informants as covariates assuming no missing data; we review the method and extend it to include missingness in the response and the multiple informants. Let Y be the outcome and $X_1, \ldots, X_K$ be the K multiple informant predictors. The GEE approach models the univariate associations between Y and X_k, defined as $E(Y|X_k)$ for $k = 1, ..., K$. The model with no covariates and distinct parameters for each informant assuming an identity link is $E(Y|X_k) = \alpha_k + \beta_k X_k$ for $k = 1, \ldots, K$ where α_k is the intercept and β_k is the slope in the k^{th} regression. Defining

$$\mathcal{Y}_i = \begin{pmatrix} Y_i \\ Y_i \\ \vdots \\ Y_i \end{pmatrix}_{(K\times 1)} \quad X_i = \begin{pmatrix} 1 & X_{i1} & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & X_{i2} & \ldots & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & 1 & X_{iK} \end{pmatrix}_{(K\times 2K)} \quad \beta = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \vdots \\ \alpha_K \\ \beta_K \end{pmatrix}_{(2K\times 1)}$$

the GEE equations assuming an identity link, constant variance and a working independence correlation matrix are

$$\sum_{i=1}^{n} X_i^T(\mathcal{Y}_i - X_i\beta) = 0. \qquad (1)$$

The GEE approach generalizes easily when X_{ik} is missing by removing the corresponding rows from $\mathcal{Y}_i$and X_i. However, missing $Y'_i s$ mean the entire $\mathcal{Y}_i$ is omitted from the estimating equations. Specifically, α_k and β;$_k$ are estimated by *Ordinary Least Squares* (OLS) using only cases with both observed Y_i and X_{ik} values. For example, with two multiple informant covariates ($K = 2$), consider six simple monotone and MCAR missingness patterns: data missing X_1 only, X_2 only, both Y and X_1, both Y and X_2, Y only and both X_1 and X_2. When X_1 only is missing, the estimate of β_1 is based on cases with observed Y_i and X_{i1} (*Complete Cases* (CC)) whereas the estimate of β_2 is based on cases with observed Y_i and X_{i2} (*Available Cases* (AC)) . A similar pattern occurs when X_2 only is missing. In all other situations considered, GEE estimates of β_1 and β_2 are based on complete cases (see Table 1).

As previously done we assume working independence [11] and use the model-based estimate of variance:

$$\widehat{var}(\hat{\beta}) = (\sum_{i=1}^{n} X_i^T X_i)^{-1}(\sum_{i=1}^{n} X_i^T \hat{\Sigma} X_i)(\sum_{i=1}^{n} X_i^T X_i)^{-1} \qquad (2)$$

where

$$\hat{\Sigma} = \frac{\sum_{i=1}^{n} (\mathcal{Y}_i - X_i\hat{\beta})(\mathcal{Y}_i - X_i\hat{\beta})^T}{n}.$$

As described in Pepe et al. [11], assuming working independence treats data from each subject as independent data clusters and uses independence as the working correlation matrix. This must be done for the model to be valid [16]; however, we have shown that the use of the independence working correlation matrix is optimal for certain models when assuming normality where the GEE and ML approaches yield identical estimates and standard errors [9]. The variance in Equation 2 is model-based since it assumes the same $\hat{\Sigma}$ for each individual and $\hat{\Sigma}$ does not depend on the design matrix. To accommodate missingness, estimates of *var* ($\hat{\beta}$) are based only on available cases or complete cases depending on the missingness pattern, as when obtaining $\hat{\beta}$. We can fit the GEE model and model-based estimates of standard error with missingness using a standard statistical package such as R [17].

Often data from multiple informants are standardized to the same scale and a model with equal slope coefficients (e.g., setting $\beta_1 = \beta_2 = \ldots = \beta_K = \beta_C$) may be desirable; this can lead to more precise parameter estimates [9]. To fit this constrained model in the presence of missing data, $\bar{Y}_i$ and X_i change depending on the observed data. With $K = 2$ informants, we solve for β assuming $\beta_1 = \beta_2 = \beta_C$ and find $\hat{\alpha}_1 = \bar{Y}_1 - \hat{\beta}_C \bar{X}_1$, $\hat{\alpha}_2 = \bar{Y}_2 - \hat{\beta}_C \bar{X}_2$ and

$$\hat{\beta}_C = \frac{\sum_{i=1}^{n} I_{i1}(X_{i1} - \bar{X}_1)(Y_i - \bar{Y}_1) + \sum_{i=1}^{n} I_{i2}(X_{i2} - \bar{X}_2)(Y_i - \bar{Y}_2)}{\sum_{i=1}^{n} I_{i1}(X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^{n} I_{i2}(X_{i2} - \bar{X}_2)^2} \qquad (3)$$

where

$$\bar{Y}_k = \frac{\sum_{i=1}^{n} I_{ik} Y_i}{\sum_{i=1}^{n} I_{ik}}, \quad \bar{X}_k = \frac{\sum_{i=1}^{n} I_{ik} X_{ik}}{\sum_{i=1}^{n} I_{ik}}$$

and I_{ik} is an indicator function for those observations with observed Y_i and X_{ik} for $k = 1, 2$. In the Hernández et al. [2] study, the physical activity measurements from the mother and child have different variances; hence, it is desirable to first standardize the X_{ik} values and then find estimates under the constrained model. However, in the presence of non-monotone missing data, standardizing a priori on the basis of observed responses can be unattractive since standardization of different multiple informant covariates may deal with different subsets of the data. We provide an alternative technique for dealing with this issue in Section 3.

3 Maximum Likelihood Approach Incorporating Missingness

To use the ML approach we assume a joint multivariate distribution for the outcome and multiple informant predictors. We assume for simplicity only two predictors here. The model can be easily extended to accommodate more predictors. For each of n observations with complete data, let $Q_i^T = (Y_i, X_{1i}, X_{2i})^T$ and thus

$$Q_i \sim MVN\left(\begin{pmatrix} \mu_Y \\ \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{X_1,Y} & \sigma_{X_2,Y} \\ \sigma_{X_1,Y} & \sigma_{X_1}^2 & \sigma_{X_1,X_2} \\ \sigma_{X_2,Y} & \sigma_{X_1,X_2} & \sigma_{X_2}^2 \end{pmatrix}\right).$$

From this distribution, we find estimates for $\theta = (\mu_Y, \mu_{X_1}, \mu_{X_2}, \sigma_Y^2, \sigma_{X_1,Y}, \sigma_{X_2,Y}, \sigma_{X_1,X_2}, \sigma_{X_1}^2, \sigma_{X_2}^2)^T$. Here we consider the regression parameter estimates, $\hat{\beta}$, found from solving Equation 1 when $K = 2$. Using conditional mean formulas for the multivariate normal distribution, we find $E(Y \mid X_i) = \mu_Y + \sigma_{X_i,Y}(X_i - \mu_{X_i}) \big/ \sigma_{X_i}^2$ where $i = 1, 2$. We define $\alpha_i = \mu_Y - \beta_i \mu_{Xi}$ and

$\beta_i = \sigma_{X_i,Y} \Big/ \sigma^2_{X_i}$ where $i = 1, 2$. We also define V_{11}, V_{22} and V_{12} in terms of θ by utilizing conditional variance formulas for the multivariate normal distribution, e.g., $V_{11} = var(Y|X_1)$, $V_{22} = var(Y|X_2)$ and $V_{12} = cov(Y|X_1, Y|X_2)$. To make the full rank transformation, we include two parameters, μ_Y and σ^2_Y, from θ into $\tau = (\alpha_1, \beta_1, \alpha_2, \beta_2, V_{11}, V_{22}, V_{12}, \mu_Y, \sigma^2_Y)^T$. Estimates of θ have closed form solutions with complete data in this case; furthermore, estimates of α_1, β_1, α_2, β_2 are identical to those obtained using GEE. This implies that although the derivation of ML estimates assumes multivariate normality, because the ML and GEE estimates are the same, both estimates have the same robustness properties [9].

To find solutions when any element of Q can have missingness, we use the expectation-maximization (EM) algorithm [18], consisting of an expectation (E) and a maximization (M) step. The ML approach is performed as if there are no missing observations in the M step and the E step finds the conditional expectation of the missing data given the observed data and the current parameter estimates. These steps are iterated until convergence. In particular, we use R [17] for implementation; we stratify by whether or not an observation is complete and then use the EM algorithm (see Appendix). Details are also provided in the Appendix for a case with monotone missingness where closed form solutions can be found. While other techniques (e.g. Newton-Raphson, Fisher scoring) have quadratic convergence rates [19], the EM algorithm has a linear convergence rate since it does not depend on second derivatives for its calculation. Consequently, the algorithm does not automatically provide the asymptotic standard errors of the ML estimates, so we use the bootstrap [20] to obtain estimates of $var(\hat{\beta})$.

For data with simple monotone missingness patterns considered in Section 2 (data missing X_1 only, X_2 only, both Y and X_1, both Y and X_2, Y only and both X_1 and X_2), some closed form solutions exist. Using the factorization theorem for finding ML estimates with monotone missingness patterns, in some cases we find explicit solutions for $\hat{\beta}_1$ and $\hat{\beta}_2$ (results in Table 1). Furthermore, when solutions involve only complete or available cases for ML, the estimates are the same as GEE, but in general, they differ.

We now describe using the factorization theorem in the specific monotone situations to derive ML estimates. First we consider when X_1 alone is missing. According to the factorization theorem, we write $f(Y, X_1, X_2) = f(X_1|Y, X_2)f(Y, X_2)$ where parameters from $f(X_1|Y, X_2)$ are estimated from complete cases and parameters from $f(Y, X_2)$ are estimated from all observed cases (available cases). We show how to use the EM algorithm to obtain the estimates in the Appendix. The estimate of β_1 calculated from estimates of $\sigma_{X_1,Y}$ and $\sigma^2_{X_1}$ is based on data from a combination of expressions involving all observed data, but $\hat{\beta}_2$ is calculated from estimates of $\sigma_{X_2,Y}$ and $\sigma^2_{X_2}$ and is based on available cases for Y and X_2. Therefore, $\hat{\beta}_2$ from ML is equivalent to that obtained using GEE, but $\hat{\beta}_1$ is not. The argument with missing X_2 is derived similarly.

In the case of missingness in both Y and X_1, we write $f(Y, X_1, X_2) = f(Y, X_1|X_2)f(X_2)$, hence parameters from $f(Y, X_1|X_2)$ are estimated from complete cases and parameters from $f(X_2)$ come from all observed cases. In deriving $f(Y|X_1)$, we find that X_2 does give information about the joint distribution of Y and X_1, so the estimate of β_1 is based on all observed cases. However, the estimate of β_2 is the same using complete cases as when implementing ML. As expected, the converse is true with data where both Y and X_2 are missing.

When Y only is missing, we find $f(Y, X_1, X_2) = f(Y|X_1, X_2)f(X_1, X_2)$ where parameters from $f(Y|X_1, X_2)$ are estimated using complete cases and parameters in $f(X_1, X_2)$ are estimated using all observations. Therefore, estimates of β_1 and β_2 are found from a combination of all observed data. This implies that even when Y is missing, observations with observed X_1 and X_2 values contribute to the likelihood; this is contrary to previous work on estimating parameters in ordinary regression models with missing Y values [21]. Specifically, they found that when interested in inferences about a response with missing data given complete covariates, the missingness in Y did not affect the regression estimates obtained; that is, the same regression estimates were found when analyzing the data with and without the missing Y values.

When both X_1 and X_2 are missing, using the same technique as above, $f(Y, X_1, X_2) = f(X_1, X_2|Y)f(Y)$ where parameters from $f(X_1, X_2|Y)$ are estimated using complete cases and parameters from $f(Y)$ are from all observations. Therefore, estimates of both slopes will be different when calculated using GEE and ML.

While in many examples comparing the slopes is desirable, in the Hernández et al. [2] study, the variances of the two multiple informants are not the same, so comparing the predictive values of the multiple informant reports is more appropriate. For example, we are interested if the relationship of vigorous exercise reported by the child and BMI is similar to the relationship of vigorous exercise reported by the mother without assuming that the variance of vigorous exercise reported by each informant is the same. If the relationship is indeed similar, this implies that using the child's report of vigorous exercise is adequate to assess the relationship in future studies. To compare the two marginal relationships, we test $\rho_{X_1,Y} = \rho_{X_2,Y}$ where $\rho_{X_1,Y} = \frac{\sigma_{X_1,Y}}{\sigma_Y \sigma_{X_1}}$ and $\rho_{X_2,Y} = \frac{\sigma_{X_2,Y}}{\sigma_Y \sigma_{X_2}}$; we use the bootstrap [20] to obtain an appropriate standard error for the difference in correlations. This ability to easily compare the correlation between each informant report and response is an advantage of using ML with the bootstrap variance estimate.

4 Simulations: Efficiency of ML Compared with GEE

MCAR missingness means that the outcomes (response and multiple informant covariates) are a random subsample of all outcomes [12]. For simplicity, we do not include additional covariates with missingness, although the techniques will extend to do so. We consider MCAR missingness only in the simulations since GEE may be biased when missingness is not MCAR and in our example the MCAR assumption appears reasonable since missing data arose due to logistical reasons rather than refusals. Thus, we use the simulations to compare the efficiency of the ML and GEE approaches.

In the case of two multiple informant predictors, monotone missingness often occurs when one predictor has missingness and the other is complete; for example, a mental health service utilization study had many teacher ratings of children missing but all parents responded about their children [3], [4], [5], [6], [7]. In general, multiple informant data missingness is not necessarily monotone and missing responses are also possible. For example, in a study predicting utilization of health services obtained from multiple sources [22], both self report and administrative source responses had missingness. The Hernández et al. [2] example has missingness in the response in addition to the multiple informant covariates.

In our simulations, we compare the GEE and ML models under similar situations as described above. We assume a multivariate normal distribution for Q_i^T and generate 500 observations assuming $\sigma_Y^2 = \sigma_{X_1}^2 = \sigma_{X_2}^2 = 1$ for different values of $\sigma_{X_1,Y}$, $\sigma_{X_2,Y}$ and σ_{X_1, X_2}. Specifically, we consider situations with zero covariance between the two informants, $\sigma_{X_1, X_2} = 0.3334$ (as calculated from the Hernández et al. dataset [2] without missingness) and a moderately large value, $\sigma_{X_1, X_2} = 0.6$. For comparison, we also consider $\sigma_{X_1,Y}$ and $\sigma_{X_2,Y}$ for the same values (see Table 2). For each set of values, we perform 10,000 simulations with four scenarios: 1.) data with no missingness, 2.) data where X_1 is missing 50% of its observations at random (MCAR and monotone missingness in one multiple informant), 3.) data where X_1 is missing 25% of its observations at random and X_2 is missing 25% of its observations at random (non-monotone MCAR missingness in the multiple informants) and 4.) data where Y is missing 25% of its observations at random (MCAR and monotone missingness in the response). This last case is designed to illustrate loss of efficiency when respondents with Y missing are discarded.

With the monotone missingness in the multiple informants situation, we calculate $\hat{\beta}_1$ and *var* $(\hat{\beta}_1)$ only since we illustrated in Section 3 that the ML estimate of β_2 is equivalent to that obtained from GEE. We also calculate the ratio of *var* $(\hat{\beta}_1)$ from ML to *var* $(\hat{\beta}_1)$ from GEE and compare the two methods where 100% means that ML and GEE are equivalent in terms of efficiency (Table 2). For the non-monotone missingness in the multiple informants situation and with missingness in Y, both efficiency ratios for *var* $(\hat{\beta}_1)$ and *var* $(\hat{\beta}_2)$ are affected similarly by the missingness so we present only *var* $(\hat{\beta}_1)$ (also in Table 2). We repeat the simulations to compare ML and GEE assuming the constrained model with equal variances under similar scenarios (not presented).

In Table 2, simulations with 50% missingness in X_1 show that the correlation between informant one and response, $\sigma_{X_1,Y}$, is the largest determinant of the efficiency of ML compared with GEE whereas $\sigma_{X_2,Y}$ has the smallest effect. Specifically, the largest efficiency gains (22 – 32%) for ML compared with GEE are found when $\sigma_{X_1,Y} = 0.6$. ML is also more efficient than GEE with moderate effects ($\sigma_{X_1,Y} = 0.3334$) (efficiency gains of 9 – 24%). For a nonzero $\sigma_{X_1,Y}$ value, gains are largest when $\sigma_{X_2,Y} = 0$ and $\sigma_{X_1,X_2} = 0.6$; in this case, X_2 and Y provide independent information about X_1.Under the null hypothesis of no relationship between the first informant and response ($\sigma_{X_1,Y} = 0$), with $\sigma_{X_1,X_2} = 0$ or $\sigma_{X_2,X_2} = 0.3334$, small efficiency gains (0 – 8%) are realized, but when the relationship between the two informants is large ($\sigma_{X_1,X_2} = 0.6$), the efficiency gain is as high as 27%.

Regarding bias, we find no evidence that GEE or ML produce biased estimates in these situations. Thus, we present only the efficiency ratios not $\hat{\beta}$ or *var* $(\hat{\beta})$. The ratio of variances for data missing 50% of its X_1 observations to data with no missingness for all scenarios is about two for GEE; this follows since approximately 50% of the data is not being used. As described earlier, the loss is not as great when using ML.

Efficiency gains for the non-monotone missingness in the multiple informants pattern are less than for the previous pattern (Table 2). For *var* $(\hat{\beta}_1)$, gains are largest (10 – 13%) when $\sigma_{X_1,Y} = 0.6$; it appears to make little difference what the $\sigma_{X_2,Y}$ and σ_{X_1,X_2} values are. Similarly, for *var* $(\hat{\beta}_2)$, the biggest gains occur when $\sigma_{X_2,Y} = 0.6$ with a range of 11 – 13% (not presented). Thus, the non-monotonicity of the missingness pattern does not affect the efficiency gains, but the amount of missingness in the X_1 and X_2 values does. Specifically, in the monotone scenario first considered with 50% missingness in X_1, more efficiency is found than in the non-monotone case with only 25% missingness in X_1 (and 25% missingness in X_2). When 25% are missing Y (Table 2), the largest efficiency gain for estimation of β_1 using ML (14%) is found when $\sigma_{X_1,Y} = 0$, $\sigma_{X_2,Y} = 0.6$, $\sigma_{X_1,X_2} = 0.6$ or when $\sigma_{X_1,Y} = 0.6$, $\sigma_{X_2,Y} = 0.6$, $\sigma_{X_1,X_2} = 0.6$ A maximum efficiency gain of the same magnitude occurs for estimation of β_2 (not presented).

Under the constrained model with monotone missingness in the multiple informants, the efficiency gains of ML compared with GEE are small (0 – 7%) since the constrained estimate averages over both X_1 and X_2. Instead, if we consider the non-monotone missingness pattern where both X_1 and X_2 are missing 25% of their observations, efficiency gains for ML compared with GEE are 7 – 15% with the largest gains found when $\sigma_{X_1,Y} = \sigma_{X_2,Y} = 0.6$. When Y is missing 25% of its observations, as in the unconstrained case, the largest efficiency gain (18%) is found when $\sigma_{X_1,X_2} = 0$, $\sigma_{X_1,Y} = 0.6$, $\sigma_{X_2,Y} = 0.6$. While generally the efficiency gains of ML compared with GEE under the constrained model are equal or smaller than the unconstrained model, the efficiency loss when comparing models with missing values and no missing values is less than in the unconstrained case.

5 Illustration

We begin this section by describing results of applying the GEE and ML methods and end by investigating the robustness of ML to the assumption of multivariate normality in the presence of MCAR missingness. In 1996, a study investigating the association between physical activity/ inactivity and obesity was performed in two towns of Mexico City [1], [2]. We illustrate ML to evaluate the relationship between BMI (Y) and vigorous exercise as reported by the child (X_1) and the child's mother (X_2) and compare to the GEE approach. One question we consider is the consequence of relying on the child's information for larger surveys. Partial information is available for 29 of the 111 observations (3 are missing X_1 only, 13 are missing X_2 only, 10 are missing Y only and 3 are missing Y and X_1). We find that X_1 and X_2 are not predictive of missingness in Y; similar conclusions are drawn for the missingness in X_1 and X_2. Because missing data arose due to logistical reasons rather than refusals, the MCAR assumption is reasonable. To calculate standard errors for both approaches, we use 2000 bootstrap [20] samples.

We begin by presenting the estimated means and covariance matrix using GEE and when using ML (Table 3) ; both the means and variances are similar from both approaches. However, while the ML and GEE estimates in this analysis of 111 cases (Table 4) are similar to the previous complete case analysis of 82 cases (data not shown), the GEE model-based standard errors are 6–18% smaller whereas the ML standard errors are 15–20% smaller. This indicates that missingness has not substantially biased our estimates, but by incorporating information from the missing data, we gain efficiency when using ML or GEE. In addition, we can directly compare the ML estimates of standard error to the model-based GEE estimates for our analysis of the 111 cases. We find estimated efficiency gains of ML compared with GEE of 4% for estimating $var\,(\hat{\beta}_1)$ and 20% for estimating $var\,(\hat{\beta}_2)$. Generally consistent with our simulations (Section 4), our example demonstrates that compared with using GEE, ML is more efficient.

Our primary goal is to compare the relationship of child's report of vigorous exercise and BMI with mother's report and BMI; however, the variances of the two reports are different according to an F test ($p < 0.0001$). Thus, rather than constraining the two slopes to be equal, we test whether the two correlations are equal ($\rho_{X_1,Y} = \rho_{X_2,Y}$). We find $\hat{\rho}_{X_1,Y} = -0.096$, $\hat{\rho}_{X_2,Y} = -0.161$ and the difference between the two correlations, $\hat{\rho}_{X_1,Y} - \hat{\rho}_{X_2,Y} = 0.065$ with standard error 0.131, is not statistically significant. Thus, using either mother's or child's report of vigorous exercise gives similar predictive power and would be appropriate to use in a subsequent study.

When using the ML approach, we must be concerned about deviations from the assumption of normality. In this example, the distributions of vigorous exercise reported by the child and the child's mother are skewed to the right (not presented) with many children receiving no exercise. Residual plots (not presented) show some evidence of increasing heterogeneity in the residuals as the predicted Y values increase. Therefore, it is probable that the multivariate distribution of BMI, vigorous activity reported by the child and the child's mother is likely not normally distributed. We therefore describe a simulation done to address whether a deviation from normality impacts results from the ML model.

Although ML assumes normality, with complete data the ML estimates are not sensitive to this assumption since they are identical to the GEE estimates. To investigate this in the presence of MCAR missingness, we simulated data from the observed underlying distribution of 82 complete cases in the Hernández et al. dataset [2]. Specifically, for each of the 2000 simulated datasets, we drew with replacement 111 cases from this complete case distribution. For each of these simulated datasets, we created missing data in 29 cases randomly according to the same general pattern of missingness in the original dataset of 111 observations. We calculated estimates of β using ML from the simulated data and compared these to the original complete case dataset to test if there was bias (Table 5); indeed, we find no evidence of bias. In summary, we generated data which were not multivariate normal, used an MCAR missingness mechanism to create missing data, assumed ML to estimate parameters and found no bias in the ML estimates. Hence, we have confirmed the robustness of ML to the assumption of multivariate normality in the presence of MCAR missingness in our data.

6 Conclusion

Both ML and GEE marginal regression models can be used to combine information from multiple informants into one analysis; here we describe how to extend them to deal with missing data. In the Hernández et al. dataset [2], researchers collected activity measurements from children and their mothers. We found that the relationship between BMI and physical activity as reported by the child is similar to the relationship of BMI and physical activity as reported by the child's mother in this small study. Thus, a subsequent larger study (where it is not feasible to obtain mother's reports) would provide reasonable estimates of the relationship between physical activity and BMI. In addition, our simulations illustrated that the ML approach can be more efficient than the GEE technique.

Although we have found that the ML approach for estimation of marginal regression models with multiple source predictors incorporating missingness is more efficient than using GEE, ML does, however, require additional assumptions. For instance, ML assumes multivariate normality whereas GEE only assumes that the mean model is correctly specified. Without missingness, the ML and GEE approaches are identical in many cases indicating that ML is not at all sensitive to the normality assumption in the absence of missing data. The vigorous activity multiple informant measurements in the Hernández et al. [2] dataset were skewed to the right, but the residuals from the model were approximately normal. Analyzing the data without missingness reveals that ML is still equivalent to GEE, thus confirming the robustness of ML to deviations from normality. We also performed a simulation study based on the Hernández et al. data [2] confirming the robustness of the ML estimates to the assumption of normality with MCAR missing data. Hence, there is evidence that ML does not require the additional assumption of normality, has little potential for bias due to normality assumptions and does provide more efficient estimates than GEE for estimation of marginal regression models with multiple source predictors incorporating MCAR missingness.

We have considered MCAR missingness in the simulations since GEE can give biased estimates under more general missingness patterns and this was a reasonable assumption for the Hernández et al. dataset [2]. However, the ML technique is also valid assuming MAR missingness. While the GEE method presented is generally not appropriate for MAR missingness, an extension of the method (inverse probability weighted GEE [23], [24]) is. The weighted GEE technique has been applied in the presence of missing $Y\,'s$ [24] or missing $X's$ [23], not both (as in our example); construction of appropriate weights has not yet been investigated. The weights are designed to correct the bias of the estimates and their effect on efficiency has yet to be investigated.

In our example, the variances of the two multiple informants are not equal; to alleviate this problem, data can be standardized. However, with missingness, standardizing data a priori is not attractive and we prefer to leave the data unstandardized. When the scale of measurement of the informants is an issue and standardizing is not feasible, using ML allows us to compute correlations to compare the effect of the informants on response without standardizing the data. Therefore, the ML technique has the advantage of providing a technique to compare the relationships of informants on response without having to standardize the data. Comparing the relationship between one informant with response to another informant with response could alleviate the need to collect information from both informants in future studies.

We have considered a simple case with one response and two multiple informant covariates. The ML method extends straightforwardly to include more than two sets of multiple informants or to measure one construct with more than two multiple informants. For example, the Hernández et al. study [2] also included measures such as video viewing, moderate exercise and videogame playing. We can also envision a study collecting information from additional informants, e.g., children, the child's mother and the child's teacher. In addition, the ML model extends readily to include covariates that are not measured by multiple informants as in Litman et al [9].

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.

References

1. Hernández B, Gortmaker SL, Laird NM, Colditz GA, Parra-Cabrera S, Peterson KE. Validity and reproducibility of a physical activity and inactivity questionnaire for Mexico City's schoolchildren. Salud Publica de Mexico 2000;42:315–323. [PubMed: 11026073]
2. Hernández B, Gortmaker SL, Colditz GA, Peterson KE, Laird NM, Parra-Cabrera S. Association of obesity with physical activity, television programs and other forms of video viewing among children in Mexico City. International Journal of Obesity 1999;23:845–854. [PubMed: 10490786]
3. Achenbach, TM. University of Vermont: Department of Psychiatry; 1991. Manual for the teacher's report form and 1991 profile.
4. 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]
5. Zahner GEP, Daskalakis C. Factors associated with mental health, general health and school-based service use for psychopathology. American Journal of Public Health 1997;87:1440–1448. [PubMed: 9314794]
6. Zahner GEP, Jacobs JH, Freeman DH, Trainor K. Rural-urban child psychopathology in a north-eastern US state: 1986–1989. Journal of the American Academy of Child Adolescent Psychiatry 1993;32:378–387.
7. Zahner GEP, Pawelkiewicz W, DeFrancesco JJ, Adnopoz J. Children's mental health service needs and utilization patterns in an urban community. Journal of the American Academy of Child Adolescent Psychiatry 1992;31:951–960.
8. 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]
9. Litman HJ, Horton NJ, Hernández B, Laird NM. Estimation of Marginal Regression Models with Multiple Source Predictors. Handbook of Statistics 2005. Submitted
10. 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.

11. 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]

12. Little, RJA.; Rubin, DB. 2nd Edition. John Wiley; New Jersey: 2002. Statistical Analysis with Missing Data.

13. Fitzmaurice GM, Laird NM, Zahner GEP, Daskalakis C. Bivariate logistic regression analysis of child psychopathology ratings using multiple informants. American Journal of Epidemiology 1995;142:1194–1203. [PubMed: 7485066]

14. Goldwasser MA, Fitzmaurice GM. Multivariate linear regression of childhood psychopathology using multiple informant data. International Journal of Methods in Psychiatric Research 2001;20:1–11.

15. Little RJA. Regression with missing Xs: A review. Journal of the American Statistical Association 1992;87:1227–1237.

16. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics 1994;23(4): 939–951.

17. R Development Core Team. R Foundation for Statistical Computing; Vienna, Austria: 2004. R: A language and environment for statistical computing.

18. 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.

19. Laird, NM. 8. Institute of Mathematical Statistics and the American Statistical Association; Ohio: 2004. Analysis of longitudinal and cluster-correlated data.

20. Efron, B.; Tibshirani, RJ. Chapman and Hall; New York, NY: 1993. An introduction to the bootstrap.

21. Baker SG, Laird NM. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. Journal of the American Statistical Association 1988;83:62–69.

22. Horton NJ, Saitz R, Laird NM, Samet JH. A method for modeling utilization data from multiple sources: Application in a study of linkage to primary care. Health Services and Outcomes Research Methodology 2002;3:211–223.

23. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 1994;89:846–866.

24. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995;90:106–121.

Appendix

We begin by giving details on how the EM algorithm is implemented for a case with two multiple informant predictors where Y and X_2 are complete but X_1 has MCAR missingness [12]. As in Section 3, we let $Q_i^T = (Y_i, \ X_{1i}, \ X_{2i})^T$ with Q_i having a trivariate normal distribution. We note that Q_i is composed of an observed portion, $Q_{obs;i}$ and a missing portion, $Q_{mis;i}$. We aim to maximize the log likelihood based on the observed data; the complete data belong to an exponential family with the following sufficient statistics:

$$\sum_{i=1}^{n} Y_i, \ \sum_{i=1}^{n} X_{1i}, \ \sum_{i=1}^{n} X_{2i}, \ \sum_{i=1}^{n} Y_i^2, \ \sum_{i=1}^{n} X_{1i}^2, \ \sum_{i=1}^{n} X_{2i}^2, \ \sum_{i=1}^{n} Y_i X_{1i}, \ \sum_{i=1}^{n} Y_i X_{2i}, \ \sum_{i=1}^{n} X_{1i} X_{2i}$$

In general, the estimates of θ are found by equating the observed sufficient statistics with the expected values, defined as the M step. We let the current estimates at the t^{th} iteration be $\theta^{(t)}$. The E step finds the conditional expectation of the missing data given the observed data and the current parameter estimates. For example in the case with X_{1i}, the E step is:

$$E(\sum_{i=1}^{n} X_{1i} \mid Q_{obs,i}, \theta^{(t)}) = \sum_{i=1}^{n} X_{1i}^{(t)}$$

$$\text{where } X_{1i}^{(t)} = X_{1i} \text{ if } X_{1i} \text{ is observed,}$$

$$X_{1i}^{(t)} = E(X_{1i} \mid Q_{obs,i}, \theta^{(t)}) \text{ if } X_{1i} \text{ is missing}$$

Thus, missing values of X_{1i} are replaced by the conditional mean of X_{1i} given the set of values $Q_{obs,i}$ that are observed for that case. From the M step, closed form solutions for $\hat{\theta}$ can be found. In particular, for the parameters that have no missing data from the joint distribution of Y and X_2 such as the mean of Y, the M step is:

$$\mu_Y^{(t+1)} = n^{-1} \sum_{i=1}^{n} Y_i^{(t)}$$

This leads to the solution that

$$\hat{\mu}_Y = n^{-1} \sum_{i=1}^{n} Y_i$$

For the parameters from the conditional distribution of X_1 given Y that do have missing data, the same technique is used; however, the solutions are more complicated. For example, the M step involving $\sum_{i=1}^{n} X_{1i}$ is:

$$\mu_{X_1}^{(t+1)} = n^{-1} \sum_{i=1}^{n} (b_0 + b_1 Y_i + b_2 X_{2i})$$

which leads to the solution that

$$\hat{\mu}_{X_1} = \bar{X}_1 + \hat{b}_1(\hat{\mu}_Y - \bar{Y}) + \hat{b}_2(\hat{\mu}_{X_2} - \bar{X}_2)$$

where $\bar{Y}, \bar{X}_1, \bar{X}_2$ are sample means based on only the observed values, $\hat{\mu}_Y, \hat{\mu}_{X_2}$ are sample means based on all observations and $\hat{b}_i, i = 1, 2$ are functions of the sample variance estimates involving only observed values. Once all the parameters are found directly through the M step, a transformation is made to find estimates of $\hat{\theta}$. Another transformation is made to obtain $\hat{\tau}$ since interest lies in the marginal regression parameter estimates, e.g. $\hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2$.

In our situation where any of the Y, X_1, X_2 can have missingness, the sufficient statistics and the M step remain the same as in the case just described, but the E step is more complicated because all of the sufficient statistics involve missing data. Thus, none of the θ parameters have simple solutions from the M step; all involve combinations of expressions involving all observed data. However, the general technique to obtain estimates using the EM algorithm remains the same. That is, the expectations of the sufficient statistics are found as sums of the quantities over all observations and the M step calculates the moment based estimates from the filled-in sufficient statistics [12]. We also note that although we have demonstrated using the EM algorithm in a situation with one response and two multiple informant covariates, it can be similarly used in situations with more than two covariates, but finding closed form estimates becomes increasingly complicated.

Table 1

Data Used by GEE and ML for Parameter Estimates by Missingness Pattern

Variable(s) Missing	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_1^{(2)}$
X_1 only	*†	‡ available cases of Y and X_2
X_2 only	‡ available cases of Y and X_1	§†
Both Y and X_1	*†	‡ complete cases of Y and X_2
Both Y and X_2	‡ complete cases of Y and X_1	§†
Y only	*†	§†
Both X_1 and X_2	*†	§†

[*] GEE is based on complete cases of Y and X_1

[§] GEE is based on complete cases of Y and X_2

[†] ML is a combination of expressions involving all observed data

[‡] ML estimate is the same as the GEE estimate

Table 2

Variance Ratios ($var\,(\hat{\beta}_1)$) for Unconstrained Models Comparing ML to GEE

		50% missing X_1			25% missing X_1 and 25% missing X_2			25% missing Y		
$\sigma_{X_1,X_2} = 0$										
			$\sigma_{X_1,Y}$			$\sigma_{X_1,Y}$			$\sigma_{X_1,Y}$	
		0	0.3334	0.6	0	0.3334	0.6	0	0.3334	0.6
	0	101%	91%	77%	100%	100%	100%	100%	100%	100%
$\sigma_{X_2,Y}$	0.3334	101%	91%	77%	100%	95%	96%	97%	97%	96%
	0.6	101%	88%	74%	100%	96%	88%	90%	90%	85%
$\sigma_{X_1,X_2} = 0.3334$										
			$\sigma_{X_1,Y}$			$\sigma_{X_1,Y}$			$\sigma_{X_1,Y}$	
		0	0.3334	0.6	0	0.3334	0.6	0	0.3334	0.6
	0	96%	88%	77%	100%	100%	100%	100%	100%	99%
$\sigma_{X_2,Y}$	0.3334	95%	89%	78%	100%	95%	95%	97%	98%	99%
	0.6	92%	90%	76%	100%	95%	88%	89%	92%	93%
$\sigma_{X_1,X_2} = 0.6$										
			$\sigma_{X_1,Y}$			$\sigma_{X_1,Y}$			$\sigma_{X_1,Y}$	
		0	0.3334	0.6	0	0.3334	0.6	0	0.3334	0.6
	0	83%	76%	68%	100%	100%	100%	100%	99%	92%
$\sigma_{X_2,Y}$	0.3334	83%	79%	73%	100%	98%	95%	96%	99%	100%
	0.6	73%	81%	76%	100%	95%	87%	86%	93%	97%

Table 3

Estimated means and variance-covariance matrix for vigorous exercise

Variable	Estimated Mean using ACs	Estimated Mean from ML
BMI (Y)	21.294	21.287
Vigorous exercise reported by child (X_1)	0.972	0.974
Vigorous exercise reported by mother (X_2)	0.791	0.787

∑ **using ACs**	$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
Y	10.975	−0.282	−0.391
X_1	−0.282	0.939	0.187
X_2	−0.391	0.187	0.439

∑ **from ML**	$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
Y	10.989	−0.317	−0.358
X_1	−0.317	0.937	0.170
X_2	−0.358	0.170	0.436

Table 4

Regression coefficients and bootstrapped standard error from unconstrained models predicting Y from X_1 and Y from X_2 for the GEE and ML methods

Method	$\hat{\beta}_1$	$\widehat{se}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\widehat{se}(\hat{\beta}_2)$
GEE	−0.333	0.331	−0.889	0.529
ML	−0.338	0.324	−0.820	0.474

Table 5

Regression coefficients and 95% confidence intervals (CI)

	$\hat{\beta}_1$	95% CI for β_1	$\hat{\beta}_2$	95% CI for β_2
Original dataset	−0.417		−0.908	
Simulated dataset	−0.417	(−0.433,−0.401)	−0.888	(−0.911,−0.865)