

# Maximum Likelihood Analysis of Logistic Regression Models with Incomplete Covariate Data and Auxiliary Information

Nicholas J. Horton

Department of Epidemiology and Biostatistics, Boston University School of Public Health,  
715 Albany Street T3E, Boston, Massachusetts 02118, U.S.A., and  
Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, U.S.A.  
*email:* horton@bu.edu

and

Nan M. Laird

Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.

**SUMMARY.** This article presents a new method for maximum likelihood estimation of logistic regression models with incomplete covariate data where auxiliary information is available. This auxiliary information is extraneous to the regression model of interest but predictive of the covariate with missing data. Ibrahim (1990, *Journal of the American Statistical Association* **85**, 765–769) provides a general method for estimating generalized linear regression models with missing covariates using the EM algorithm that is easily implemented when there is no auxiliary data. Vach (1997, *Statistics in Medicine* **16**, 57–72) describes how the method can be extended when the outcome and auxiliary data are conditionally independent given the covariates in the model. The method allows the incorporation of auxiliary data without making the conditional independence assumption. We suggest tests of conditional independence and compare the performance of several estimators in an example concerning mental health service utilization in children. Using an artificial dataset, we compare the performance of several estimators when auxiliary data are available.

**KEY WORDS:** Conditional independence assumption; EM algorithm; Joint maximization; Logistic regression model; Missing covariates, Surrogate information; Two-stage designs.

## 1. Introduction

Much has been written about statistical methods for handling incomplete data (see Little and Rubin (1987) for a comprehensive review). Many of these approaches have focused on missing outcomes. But covariates in regression models are often missing, either by design or circumstance. Little (1992) reviewed a series of approaches to estimation of regression models with missing covariates, including complete case (CC) estimation, *ad hoc* methods, and likelihood-based methods. Robins, Rotnitzky, and Zhao (1994) suggested a class of semi-parametric estimators based on inverse probability weighted estimating equations, similar to a method proposed by Zhao and Lipsitz (1992). Ibrahim (1990) described a maximum likelihood method using the EM algorithm (Dempster, Laird, and Rubin, 1977) for generalized linear regression models with missing categorical covariates.

Many studies collect data on a large number of variables, some of which may be auxiliary (or extraneous) to the regression model of interest. These variables may be observed even when the covariates of primary interest are incomplete. This setting commonly arises by design in two-stage studies (Zhao

and Lipsitz, 1992), where, at the first stage, the response and some covariates (including some that are used for screening only) may be collected. Then, at the second stage, additional information (such as primary exposure of interest) is collected from a subset of subjects.

In the measurement error literature, the auxiliary data may be considered a surrogate for the partially observed explanatory variable. Hasabelnaby, Ware, and Fuller (1989) considered an air pollution example where surrogate measures (cigarette use) and outcome (respiratory capacity) are observed for all subjects while fine particle measurements from the home are available for a subset of the subjects. Similarly, in nutritional epidemiology, it is common to collect error-prone estimates (such as a food frequency questionnaire) from all subjects as well as a measure considered to be a gold standard (such as amount of vitamins determined from a blood sample) from a subset of subjects (Robins et al., 1994). The auxiliary variables collected on all subjects may be used to increase the efficiency of analysis if they are predictive of the covariates with missing data (Robins et al., 1994).

In this article, we consider maximum likelihood estimation of logistic regression models with missing categorical covari-

ates where categorical auxiliary information is available. The article is motivated by a study of mental health service utilization where primary interest revolves around partially observed teacher reports of child psychopathology but where auxiliary reports (parental reports of psychopathology on the same child) are fully observed. Here there is no scientific reason to consider teacher reports as a gold standard, and considerable data suggest that parent and teacher reports are not conditionally independent. Our approach will be useful in other research settings where gold standards for assessment do not exist and/or conditional independence may not hold.

## 2. A Regression Model for Auxiliary Data

For each subject, the data are  $(y_i, \mathbf{x}_i, a_i)$ , where  $y_i$  is the dichotomous outcome,  $\mathbf{x}_i$  is a vector of dimension  $p$  of discrete predictors, which may be only partially observed for a given subject, and  $a_i$  is a categorical auxiliary variable that is extraneous but always observed. Let  $y_1, \dots, y_n$  be independent binary observations. The logistic regression model of interest is given by

$$\text{logit}(E[Y_i | \mathbf{x}_i]) = \mathbf{x}_i' \boldsymbol{\beta}, \quad (1)$$

where  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of regression parameters. Maximum likelihood is based on the conditional distribution of the outcome given the predictors,  $f(Y | \mathbf{X}, \boldsymbol{\beta})$ . The contribution to the likelihood cannot be computed, however, when subjects have missing covariates.

Ibrahim (1990) suggested a maximum likelihood method for parameter estimation in generalized linear models with missing covariates and no auxiliary data, which we denote as MLNA (maximum likelihood with no auxiliary data). Ibrahim's approach estimates the covariate distribution non-parametrically using maximum likelihood via the EM algorithm. Instead of just modeling the conditional distribution of  $Y$  given the covariates  $\mathbf{X}$ , the joint distribution of  $Y$  and  $\mathbf{X}$  is modeled as

$$f(Y, \mathbf{X} | \boldsymbol{\Omega}) = f(Y | \mathbf{X}, \boldsymbol{\beta}) f(\mathbf{X} | \boldsymbol{\gamma}), \quad (2)$$

where  $\boldsymbol{\Omega} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ . If the data on  $\mathbf{X}$  are completely observed, then  $f(\mathbf{X} | \boldsymbol{\gamma})$  does not contribute to the likelihood for  $\boldsymbol{\beta}$ . With missing data in  $\mathbf{X}$ , estimating  $f(\mathbf{X})$  may be worthwhile in increasing efficiency and removing bias if missingness is related to the outcome  $Y$ .

Auxiliary information may be available for a number of reasons. Researchers often collect many covariates, though they may include only a subset in their regression models. Administrative record data (possibly from previous investigations) that can be matched to subjects in an investigation might be available. Proxy informants may be available in addition to the primary respondent. Finally, in cases where covariate exposure assessment is expensive or invasive, only a subset of subjects might be included in a validation sample, while more error-prone auxiliary data might be collected on all subjects. The selection of a validation sample often depends upon  $Y$ , components of  $\mathbf{X}$ , and  $A$ .

With auxiliary data, we can write the complete data likelihood as

$$f(Y, \mathbf{X}, A | \boldsymbol{\Omega}^*) = f(Y | \mathbf{X}, A, \boldsymbol{\beta}^*) f(\mathbf{X}, A | \boldsymbol{\gamma}^*), \quad (3)$$

where  $\boldsymbol{\Omega}^* = (\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$  and we use the  $*$  to indicate dependence on the distribution of  $A$ . If we are willing to assume

that  $f(Y | \mathbf{X}, A, \boldsymbol{\beta}^*) = f(Y | \mathbf{X}, \boldsymbol{\beta})$  (conditional independence of  $Y$  and  $A$  given  $X$ ), then Ibrahim's method can be extended in a straightforward way to fit model (1) by conditioning on  $A$  when estimating the distribution of the covariates. Vach (1997) suggested this estimator when the auxiliary information is used to improve the modeling of the covariate distribution. We denote this estimator as MLCI (maximum likelihood with auxiliary data assuming conditional independence). The conditional independence assumption is commonly made when  $\mathbf{X}$  is a gold standard and  $A$  is measured with error (Rosner, Willett, and Spiegelman, 1989; Reilly and Pepe, 1995; Bashir and Duffy, 1997).

But the conditional independence assumption underlying this approach is not always tenable. Diagnostic testing where true gold standards do not exist provides one example, as seen in many types of medical and epidemiological research, including occupational, environmental, and nutritional studies (Wacholder, Armstrong, and Hartage, 1993; Hui and Zhou, 1998). When such a measure (which is sometimes referred to as an alloyed gold standard) has errors that are correlated with the errors of the auxiliary measure, conditional independence will not hold (Spiegelman, Schneeweiss, and McDermott, 1997). Torrance-Rynard and Walter (1997) reviewed the bias due to such correlation on latent class models used to assess diagnostic test performance. For our example, where different informants may provide different types of information, the conditional independence assumption is often not plausible.

In our setting, it is possible to test the conditional independence assumption by assessing whether  $A$  is a significant predictor in models for  $f(Y | \mathbf{X}, A, \boldsymbol{\beta}^*)$ . Mantel et al. (Mantel et al., 1991; Singh et al., 1993) review the bias that can ensue in a survey sampling record matching setting if the assumed conditional independence model is not correct. They propose several imputation methods that incorporate auxiliary information to achieve better performance.

When conditional independence fails to hold, the factorization in (3) is not natural since the regression coefficients for  $X$  in  $E(Y | X, A)$  are not generally equal to  $\boldsymbol{\beta}$  in model (1). Another approach is suggested by the following factorization, noted by Vach (1994), of the complete data likelihood in (3):

$$f(Y, \mathbf{X}, A | \boldsymbol{\theta}) = f(A | Y, \mathbf{X}, \boldsymbol{\alpha}) f(Y | \mathbf{X}, \boldsymbol{\beta}) f(\mathbf{X} | \boldsymbol{\gamma}). \quad (4)$$

Now  $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$  is the set of parameters in the complete data log likelihood,

$$\begin{aligned} & \sum_i l_{a,y,x}(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}_i) \\ &= \sum_i \{ l_{a|y,x}(\boldsymbol{\alpha} | a_i, y_i, \mathbf{x}_i) + l_{y|x}(\boldsymbol{\beta} | y_i, \mathbf{x}_i) + l_x(\boldsymbol{\gamma} | \mathbf{x}_i) \}. \end{aligned} \quad (5)$$

Pepe (1992) described a similar factorization for missing outcomes and surrogate variables when the missingness does not depend on covariates or surrogates. This factorization is natural when we are interested in the distribution of  $f(Y | \mathbf{X}, \boldsymbol{\beta})$  and we do not impose conditional independence assumptions on  $f(Y, \mathbf{X}, A)$ .

### 3. Maximum Likelihood Estimation with Missing Covariates and Auxiliary Data

With no missing data, we can simply maximize  $\Pi f(Y_i | \mathbf{X}_i, \boldsymbol{\beta})$  to estimate  $\boldsymbol{\beta}$  because this is a recursive system. In this case,  $\boldsymbol{\beta}$  is estimated separately from  $\boldsymbol{\alpha}$  and  $\boldsymbol{\gamma}$  using standard logistic regression routines. But when  $\mathbf{X}_i$  is sometimes missing, the observed data likelihood does not factor, and all three parts of the complete data likelihood must be estimated. Throughout, we assume that the missingness law of the covariates is unrelated to the unobserved covariates, i.e., the covariates are missing at random (MAR) in the sense of Little and Rubin (1987). Vach and Blettner (1995) review the potential bias of this class of estimators when the MAR assumption is violated. Given this assumption, we have that the observed data log likelihood is

$$l_{a,y,x}^o(\boldsymbol{\theta}) = \sum_i \log \sum_{x_{\text{miss},i}} \left\{ L_{a|y,x}(\boldsymbol{\alpha} | a_i, y_i, \mathbf{x}_i) \times L_{y|x}(\boldsymbol{\beta} | y_i, \mathbf{x}_i) L_x(\boldsymbol{\gamma} | \mathbf{x}_i) \right\},$$

where  $x_{\text{miss},i}$  represents the components of  $\mathbf{x}_i$  that are missing and  $\sum_{x_{\text{miss},i}}$  is the sum over the sample space of  $x_{\text{miss},i}$ . We denote the estimator of  $\boldsymbol{\beta}$  that maximizes this likelihood by MLA (maximum likelihood with auxiliary data and no conditional independence assumption).

Following the approach of Ibrahim (1990), the covariates  $\mathbf{x} = (x_1, x_2, \dots, x_p)'$  are assumed to be discrete random variables with joint distribution indexed by the parameters  $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_r)'$ . For example, if there are three dichotomous covariates, then  $\boldsymbol{\gamma}$  is of dimension  $r = 2^3 - 1 = 7$ . More generally, let  $c_1, \dots, c_p$  represent the number of categories for each of the covariates, respectively. Then  $\boldsymbol{\gamma}$  is of dimension  $r = c_1 \times \dots \times c_p - 1$ . Let  $f(Y | \mathbf{X}, \boldsymbol{\beta})$  be the density of the outcome and let  $f(A | \mathbf{X}, Y, \boldsymbol{\alpha})$  and  $f(\mathbf{X} | \boldsymbol{\gamma})$  be the multinomial distributions of the covariates, with  $\boldsymbol{\alpha}$  and  $\boldsymbol{\gamma}$  being nuisance parameters distinct from  $\boldsymbol{\beta}$ .

If we impose no restrictions on the distribution of  $\mathbf{X}$ , with complete data,  $\boldsymbol{\gamma}$  will be estimated by the  $r + 1$  observed cell counts obtained by cross-classifying the  $\mathbf{X}$ 's in a  $p$ -dimensional contingency table. This approach is equivalent to fitting a saturated log-linear model (Fienberg, 1980) to the cell counts. More parsimonious representations (such as log-linear models without all higher order interactions) may be considered to reduce the number of nuisance parameters being estimated, although possibly at the expense of introducing bias into the estimation of regression coefficients. Similarly, a saturated multinomial model can be used for  $f(A | \mathbf{X}, Y, \boldsymbol{\alpha})$ , and logistic regression is used for  $\boldsymbol{\beta}$  when  $X$  is fully observed.

We use the EM algorithm (Dempster et al., 1977) to compute parameter estimates in the observed data likelihood where covariates are only partially observed. The EM algorithm is a general purpose iterative algorithm for maximizing incomplete data likelihoods and consists of two steps. At the E-step, one calculates the expectation of  $l_{a,y,x}(\boldsymbol{\theta} | a, y, \mathbf{x})$  conditioning on the current parameter vector  $\boldsymbol{\theta}^{(t)}$  and the observed data. In this case, the observed data consist of  $(y_i, \mathbf{x}_{\text{obs},i}, a_i)$ ,  $i = 1, \dots, n$ , where  $\mathbf{x}_{\text{obs},i}$  represents the subset of  $\dim \leq p$  covariates in  $\mathbf{x}_i$  that are observed on the  $i$ th subject. This is in general different for each  $i$  and will equal  $\mathbf{x}_i$  if the  $i$ th subject has no missing covariates.

Denoting  $E[l_{a,y,x}(\boldsymbol{\theta} | a, y, \mathbf{x})]$  by  $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})$ , we have that

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^n \sum_{j=1}^{r+1} w_{ij}^{(t)} l_{a,y,x}(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}^j) \\ &= \sum_{i=1}^n \sum_{j=1}^{r+1} w_{ij}^{(t)} \left\{ l_{a|y,x}(\boldsymbol{\alpha} | a_i, y_i, \mathbf{x}^j) \right. \\ &\quad \left. + l_{y|x}(\boldsymbol{\beta} | y_i, \mathbf{x}^j) + l_x(\boldsymbol{\gamma} | \mathbf{x}^j) \right\}, \end{aligned} \tag{6}$$

where  $\mathbf{x}^j$  is the  $j$ th possible pattern of the covariates,  $l_{a,y,x}(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}^j)$  is the complete data log-likelihood for  $\boldsymbol{\theta}$  for the  $i$ th observation with  $\mathbf{x}_i$  evaluated at  $\mathbf{x}^j$ , and  $w_{ij}^{(t)} = p(\mathbf{x}^j | a_i, y_i, \mathbf{x}_{\text{obs},i}, \boldsymbol{\theta}^{(t)})$  can be thought of as weights for the  $j$ th possible pattern of the covariates for the  $i$ th observation at the  $t$ th iteration since  $w_{i+}^{(t)} = 1$ . Note that most of these  $w_{ij}$ 's will be zero since  $\mathbf{x}_{\text{obs},i}$  will rule out any value of  $\mathbf{x}^j$  not compatible with  $\mathbf{x}_{\text{obs},i}$  (the vectors  $\mathbf{x}^j$  and  $\mathbf{x}_{\text{obs},i}$  are compatible if the components of  $\mathbf{x}^j$  corresponding to the observed data equal  $\mathbf{x}_{\text{obs},i}$ ). If all the covariates for the  $i$ th subject are observed, then there is just one nonzero term in the inner sum, and the weight for the observed value  $\mathbf{x}^j = \mathbf{x}_{\text{obs},i}$  is equal to one. When  $\mathbf{x}_{\text{obs},i} \neq \mathbf{x}_i$ , the weights can be calculated by use of Bayes rule,

$$w_{ij}^{(t)} = p(\mathbf{x}^j | a_i, y_i, \mathbf{x}_i, \boldsymbol{\theta}^{(t)}) = \begin{cases} 0 & \text{if } \mathbf{x}^j \text{ is not compatible with } \mathbf{x}_i \\ \frac{p(y_i | \mathbf{x}_i^j) p(a_i | \mathbf{x}_i^j, y_i) p(\mathbf{x}_i^j)}{\sum_{k \in \text{obs},i} p(y_i | \mathbf{x}_i^k) p(a_i | \mathbf{x}_i^k, y_i) p(\mathbf{x}_i^k)} & \text{if } \mathbf{x}^j \text{ is compatible with } \mathbf{x}_i, \end{cases} \tag{7}$$

where  $\mathbf{x}_{\text{obs},i}^j$  denotes a  $p$ -vector whose observed components are  $\mathbf{x}_{\text{obs},i}$  and the remaining components take on the  $j$ th pattern for the missing covariates. The range of  $k$  in the denominator is restricted so that  $\mathbf{x}^k$  is compatible with  $\mathbf{x}_{\text{obs},i}$ .

For the M-step, we maximize equation (6) as a function of  $\boldsymbol{\theta}$  by finding the solution to the complete-data log likelihood using these weights. We can use weighted logistic regression to estimate  $\boldsymbol{\beta}$  and use the expected cell counts to estimate  $\boldsymbol{\alpha}$  and  $\boldsymbol{\gamma}$ . For the first iteration, the weights can be calculated using (7) with initial estimates for  $\boldsymbol{\theta}$  from a complete case analysis. By repeating these steps until convergence, we obtain the maximum likelihood estimates (MLEs).

#### 4. Standard Errors

Standard errors for these parameter estimates can be calculated using a number of techniques, including Louis' method, the sum of squared scores method, or by bootstrapping. We will consider each of these approaches.

Louis' method (Louis, 1982) partitions the complete-data information into two parts: the information associated with the observed data and the information associated with the missing data. Louis showed that a consistent estimate of the second-derivative matrix can be calculated using complete-

data quantities. Following the development of Lipsitz, Ibrahim, and Fitzmaurice (1999), we have that the information matrix of the observed data is given by

$$I(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^{r+1} -w_{ij} \frac{\partial^2 l(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}^j)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

$$- \sum_{i=1}^n \left( \sum_{j=1}^{r+1} w_{ij} S_{ij}(a_i, y_i, \mathbf{x}^j, \boldsymbol{\theta}) S_{ij}(a_i, y_i, \mathbf{x}^j, \boldsymbol{\theta})' \right)$$

$$+ \sum_{i=1}^n \left( \sum_{j=1}^{r+1} w_{ij} S_{ij}(a_i, y_i, \mathbf{x}^j, \boldsymbol{\theta}) \right)$$

$$\times \left( \sum_{j=1}^{r+1} w_{ij} S_{ij}(a_i, y_i, \mathbf{x}^j, \boldsymbol{\theta}) \right)',$$

where  $S_{ij}(a_i, y_i, \mathbf{x}^j, \boldsymbol{\theta}) = \partial l(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}^j) / \partial \boldsymbol{\theta}$  is the complete data score vector for the  $j$ th covariate pattern for the  $i$ th observation.  $I(\boldsymbol{\theta})$  can be estimated by

$$I(\hat{\boldsymbol{\theta}}) = -\ddot{Q}(\hat{\boldsymbol{\theta}})$$

$$- \sum_{i=1}^n \left( \sum_{j=1}^{r+1} \hat{w}_{ij} S_{ij}(a_i, y_i, \mathbf{x}^j, \hat{\boldsymbol{\theta}}) S_{ij}(a_i, y_i, \mathbf{x}^j, \hat{\boldsymbol{\theta}})' \right)$$

$$+ \sum_{i=1}^n (\dot{Q}_i(\hat{\boldsymbol{\theta}})) (\dot{Q}_i(\hat{\boldsymbol{\theta}}))',$$

where  $\hat{\boldsymbol{\theta}}$  is the MLE at convergence,

$$\dot{Q}_i(\hat{\boldsymbol{\theta}}) = \left( \sum_{j=1}^{r+1} \hat{w}_{ij} \frac{\partial l(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}^j)}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right),$$

and

$$\ddot{Q}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \sum_{j=1}^{r+1} \hat{w}_{ij}^{(t)} \frac{\partial^2 l(\boldsymbol{\theta} | a_i, y_i, \mathbf{x}^j)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

The sum of squared scores method (Meilijson, 1989; Ruud, 1991; Fitzmaurice, Laird, and Lipsitz, 1994) exploits the relationship between the observed and the sample empirical covariance matrix of the individual scores. Thus, a consistent estimator of the asymptotic information matrix for  $\boldsymbol{\theta}$  in a correctly specified model is

$$I(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n (\dot{Q}_i(\hat{\boldsymbol{\theta}})) (\dot{Q}_i(\hat{\boldsymbol{\theta}}))'.$$

This approach has the advantage of only involving the first derivatives of the log likelihood.

Finally, in settings where  $r$  (the dimension of  $\boldsymbol{\gamma}$ ) is large, it may be impractical to estimate the entire covariance matrix of the parameters using these methods. In this situation, it may be computationally convenient to assume that  $I(\boldsymbol{\theta})$  is block diagonal in  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\gamma}$  or to utilize a bootstrap (Efron and Tibshirani, 1993) estimate of the covariance matrix by resampling from the underlying contingency table.

To illustrate these methods, we consider a study of mental health service utilization in children in urban and rural

Connecticut (Zahner et al., 1992; Zahner et al., 1993; Zahner and Daskalakis, 1997). One of the outcomes of interest in this study was mental health service utilization in school-based settings. Service use was defined as a parental report that the child had ever seen a provider or been in a special program at school for a behavioral problem. If the particular service was used, the outcome was coded one and coded zero otherwise.

Predictors of service use included gender of the child (BOY: 1 = boy, 0 = otherwise), age of the child (OLD: 0 = age 6–8, 1 = age 9–11), ethnicity (BLACK: 0 = nonblack, 1 = black; HISPANIC: 0 = nonhispanic, 1 = hispanic), belonging to a single-parent household (MOMSING: 0 = father figure present, 1 = no father figure present), and psychopathology of the child. The measure of psychopathology used in the study was the total problems scale of the Child Behavioral Checklist (CBCL; Achenbach, 1991a) and the Teacher’s Report Form (TRF; Achenbach, 1991b), which were completed by parents and teachers, respectively. The raw scores were dichotomized at the cutpoint for borderline/clinical psychopathology. A score of one indicates borderline/clinical psychopathology, and a score of zero indicates normal range.

In the survey we considered, only a few of the parent reports were missing, while 43% of teacher ratings on children were unobserved. Missingness of this magnitude is not uncommon; a similar rate was reported by Boyle et al. (1993) in their Ontario Child Health Study. We include in our analysis the 2486 children with complete data on parent reports. Table 1 displays the proportion of positive reports as well as the number of subjects with observed data for that variable.

In our notation, we let  $A$  be the parent report (CBCL), while  $\mathbf{X}$  is composed of the teacher report and other covariates (including age, sex, and family demographic characteristics). We fit the logistic regression model  $f(Y | \mathbf{X})$  utilizing the information in the observed  $A$ ’s (parents) as auxiliary information. This model is of interest in predictions of school service use using only school-based information.

Fitzmaurice, Laird, and Zahner (1996) considered the question of whether the missingness in this dataset is related to the unobserved teacher’s rating and found no evidence for this hypothesis. Thus, the missing-at-random assumption appears to be reasonable. We fit a main effects linear logistic regression model for missingness of the teacher report. There is strong evidence that missingness is related to the outcome ( $p = 0.004$ ) and some mild evidence that missingness is related to the auxiliary variable (CBCL,  $p = 0.099$ ). Missingness relating to the outcome will yield biased estimates for the intercept in the complete case (CC) estimator, while missing-

**Table 1**  
Summary of outcome and predictors

Variable	Mean	Number observed
OUTCOME	0.185	2486
BOY	0.481	2486
OLD	0.473	2486
BLACK	0.198	2486
HISPANIC	0.068	2486
MOMSING	0.207	2486
CBCL	0.186	2486
TRF	0.183	1425

**Table 2**  
Correlations between variables

	OUTCOME	BOY	OLD	BLACK	HISPANIC	MOMSING	CBCL
BOY	0.100						
OLD	0.097	-0.028					
BLACK	-0.034	-0.026	-0.026				
HISPANIC	-0.005	0.007	-0.046	-0.134			
MOMSING	0.030	-0.015	-0.028	0.377	0.171		
CBCL	0.240	0.018	0.028	0.100	0.090	0.111	
TRF	0.245	0.088	-0.001	0.132	0.006	0.091	0.274

ness related to the auxiliary data will yield biased estimates for the MLNA estimator since it does not condition on this variable.

Table 2 displays the Pearson product-moment correlation matrix for these variables. We note that the correlation between the parent and teacher reports of psychopathology is the third largest correlation in the table, though it is relatively small in magnitude ( $\rho = 0.274$ ).

We can test the conditional independence assumption in a number of ways. One straightforward (albeit inefficient) approach is to fit a model for  $f(A | Y, \mathbf{X})$  using complete case ordinary logistic regression and testing whether  $Y$  is a significant predictor of the auxiliary variable after controlling for  $\mathbf{X}$ . Table 3 displays the parameter estimates for a main effects model with the auxiliary variable as the outcome. There is strong evidence ( $p = 0.0001$ ) that the conditional independence assumption is not met in this example. We can also test for conditional independence by calculating twice the difference in log likelihoods for the maximum likelihood conditional independence model and the maximum likelihood joint maximization model ( $\chi^2_{40} = 103.7$ ,  $p < 0.0001$ ).

For comparison, we display four sets of regression parameters. One is ordinary least squares (OLS) based on complete cases (CC), which discards all information on the 1061 subjects missing teacher reports. A second uses the maximum likelihood approach of Ibrahim ignoring the auxiliary information (MLNA). The third approach uses the maximum likelihood conditional independence approach (MLCI), which assumes  $f(Y | \mathbf{X}, A) = f(Y | \mathbf{X})$ . Finally, we fit a model using our maximum likelihood approach incorporating the auxiliary information (MLA). An S-plus macro (MathSoft, 1996) was used to fit this model. Table 4 displays the parameter estimates and standard errors (using 2000 bootstrap samples from the observed data) for these models.

Being a boy, being older, being nonblack, being in a single-parent household, and being above the cutscore for total psychopathology problems were all significantly associated with increased levels of school-based mental health service utilization. We note that the parameter estimates remain similar but that the standard errors of the MLNA, MLCI, and MLA estimators are smaller than those of the CC estimator, with the exception of the covariate that is partially observed (TRF). The parameter estimate for the teacher report (TRF) changes by a standard error between the MLCI and MLA models. Because we do not believe the assumption of conditional independence, we suggest that the MLA or MLNA results are most appropriate to report.

## 5. Artificial Data Example

We constructed a probability distribution using a dichotomous outcome  $Y$ , a single dichotomous auxiliary variable  $A$ , a single dichotomous covariate  $X$  (coded 1 and  $-1$ ) that is sometimes missing, and a dichotomous missingness indicator  $R$  ( $=1$  if  $X$  is missing). Table 5 displays the  $2 \times 2 \times 2$  contingency table with a supplemental  $2 \times 2$  table, where we denote by  $n_i$ ,  $i = 1, \dots, 12$ , the observed count of subjects with the  $i$ th pattern of outcome, auxiliary data, and covariate.

We considered the following factorization and parameterization of this joint distribution:

$$P(Y = y, A = a, X = x, R = r) = f(R | Y, A, X, \tau) f(A | Y, X, \alpha) f(Y | X, \beta) f(X | \gamma), \quad (8)$$

where

$$f(R | Y, A, X, \tau) = \frac{\exp(\tau_0 + \tau_1 y + \tau_2 a + \tau_3 x)^r}{1 + \exp(\tau_0 + \tau_1 y + \tau_2 a + \tau_3 x)},$$

$$f(A | Y, X, \alpha) = \frac{\exp(\alpha_0 + \alpha_1 y + \alpha_2 x + \alpha_3 xy)^a}{1 + \exp(\alpha_0 + \alpha_1 y + \alpha_2 x + \alpha_3 xy)},$$

$f(Y | X, \beta)$  is given by (1), and  $f(X | \gamma) = \gamma^{I(x=1)} \times (1 - \gamma)^{1 - I(x=1)}$ . In two-stage studies,  $\tau_1$  and  $\tau_2$  will often be nonzero by design. When  $\tau_1 = \tau_2 = 0$ , the CC estimator will be consistent. The MLNA method will be consistent for  $\beta_1$  when  $\tau_2 = \tau_3 = 0$ . The MLCI method will be consistent when  $\tau_3 = \alpha_1 = \alpha_3 = 0$ , while the MLA method will be consistent when  $\tau_3 = 0$ . To simplify our exposition, we set  $\tau_1 = \tau_2 = \tau_3 = \alpha_3 = 0$  in these simulations, fixed the intercept of our auxiliary variable distribution ( $\alpha_0 = 0$ ), and set

**Table 3**  
Test of conditional independence assumption

Variable	Parameter estimate	Standard error	Wald chi-square	Pr > chi-square
INTERCEPT	-2.265	0.150	228.55	0.0001
OUTCOME	1.311	0.169	60.46	0.0001
BOY	-0.072	0.147	0.24	0.6239
OLD	0.003	0.147	0.0003	0.9851
BLACK	0.466	0.193	5.85	0.0156
HISPANIC	0.959	0.264	13.22	0.0003
MOMSING	0.335	0.186	3.23	0.0724
TRF	1.177	0.163	52.11	0.0001

**Table 4**  
Parameter estimates (and bootstrap standard errors) for models of service utilization<sup>a</sup>

Parameter	CC	MLNA	MLCI	MLA
INTERCEPT	-2.391 (0.146)	-2.291 (0.120)	-2.323 (0.120)	-2.258 (0.114)
BOY	0.442 (0.150)	0.444 (0.114)	0.441 (0.114)	0.449 (0.116)
OLD	0.488 (0.148)	0.543 (0.112)	0.549 (0.115)	0.540 (0.115)
BLACK	-0.831 (0.250)	-0.603 (0.174)	-0.627 (0.179)	-0.563 (0.179)
HISPANIC	-0.335 (0.357)	-0.215 (0.245)	-0.225 (0.246)	-0.194 (0.246)
MOMSING	0.578 (0.223)	0.316 (0.157)	0.324 (0.162)	0.311 (0.162)
TRF	1.416 (0.164)	1.413 (0.169)	1.523 (0.166)	1.298 (0.165)

<sup>a</sup> CC, complete case; MLNA, maximum likelihood with no auxiliary data; MLCI, ML with CI assumption; MLA, ML with no CI assumption.

$\gamma = 0.5$ . Then our joint distribution could be written

$$P(Y = y, A = a, X = x, R = r) = \frac{\exp(\tau_0)^r \exp(\alpha_1 y + \alpha_2 x)^a \exp(\beta_0 + \beta_1 x)^y}{1 + \exp(\tau_0)} \frac{1}{1 + \exp(\alpha_1 y + \alpha_2 x)} \frac{1}{1 + \exp(\beta_0 + \beta_1 x)} \frac{1}{2}$$

For a given set of parameter values, it is straightforward to calculate the probability of a subject falling into a given cell in Table 5. These probabilities were used as the observed cell counts for our example. We maximized the observed-data log likelihood using Newton’s method. We calculated the mean squared error (MSE: variance + bias<sup>2</sup>) of our estimators under a variety of parameter values using  $(\hat{\beta} - \beta)$  as an approximation to the bias and the inverse of the negative of the second derivative of the observed-data log likelihood evaluated at the true parameters to calculate the variance.

We began by considering the MSE of our estimators when the conditional independence assumption was true ( $\alpha_1 = 0$ ) with a moderate association between  $A$  and  $X$  ( $\alpha_2 = 2$ ) and our parameters of interest fixed ( $\beta_0 = \beta_1 = 1$ ). We varied  $\tau$  (which controls the proportion missing) and calculated the MSE for five estimators: the unobserved truth with no missing data (NO MISSINGNESS), the complete case estimator (CC), the maximum likelihood conditional independence estimator (MLCI), the maximum likelihood auxiliary data estimator (MLA), and the maximum likelihood estimator ignoring the auxiliary data (MLNA). All estimators are consistent in this setting; thus, the relative MSE approximates the asymptotic relative efficiency of these estimators.

Ordinarily, one is not interested in bias and variance for the intercept. We report these since the results should be similar to those for a second covariate, orthogonal to  $X$ , that has no missing data.

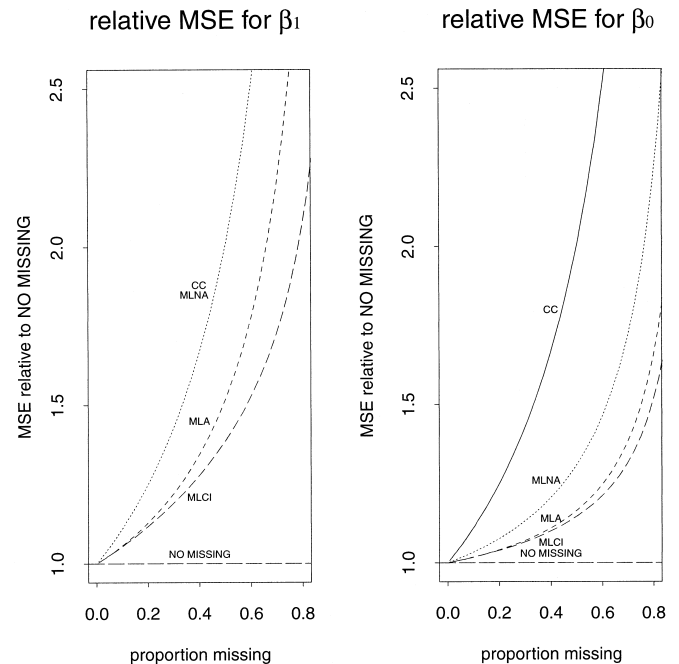
Figure 1 displays the relative MSE (compared with no missing data) for these estimators for a range of missing data pro-

portions. Since the conditional independence assumption is true, the MLCI estimator has smaller MSE than the MLA estimator. We also note that there is more recovery of information for the intercept ( $\beta_0$ ) than for the parameter of the missing covariate ( $\beta_1$ ) for all missing data estimators. The MLNA estimator is unable to recover any information relative to the CC for  $\beta_1$  since it ignores the auxiliary information, although it recovers considerable information about  $\beta_0$  relative to the CC estimator.

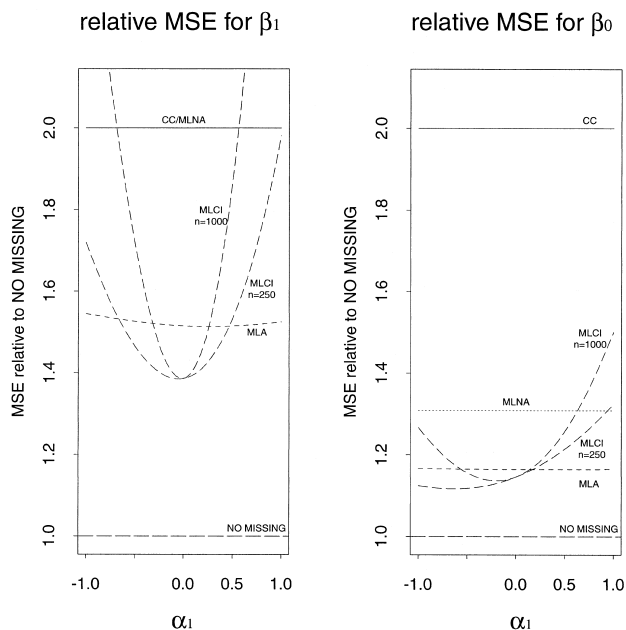
We next consider a model where the conditional independence model was not true, which will yield bias for the MLCI method. Because of this bias, the relative MSE for MLCI will be a function of sample size. We considered two sample sizes for the total number of subjects in the contingency table:  $n = 250$  and  $n = 1000$ . We fixed the missingness proportion at 50% ( $\tau = 0$ ) and left the other parameters unchanged, with the exception of  $\alpha_1$ . Figure 2 displays relative MSE when  $\alpha_1$

**Table 5**  
Contingency table for artificial example

	A	Y = 0			Y = 1		
		X			X		
		-1	1	?	-1	1	?
0	$n_1$	$n_2$	$n_9$	0	$n_5$	$n_6$	$n_{11}$
1	$n_3$	$n_4$	$n_{10}$	1	$n_7$	$n_8$	$n_{12}$



**Figure 1.** MSE (relative to NO MISSINGNESS) for different proportions of missing data.



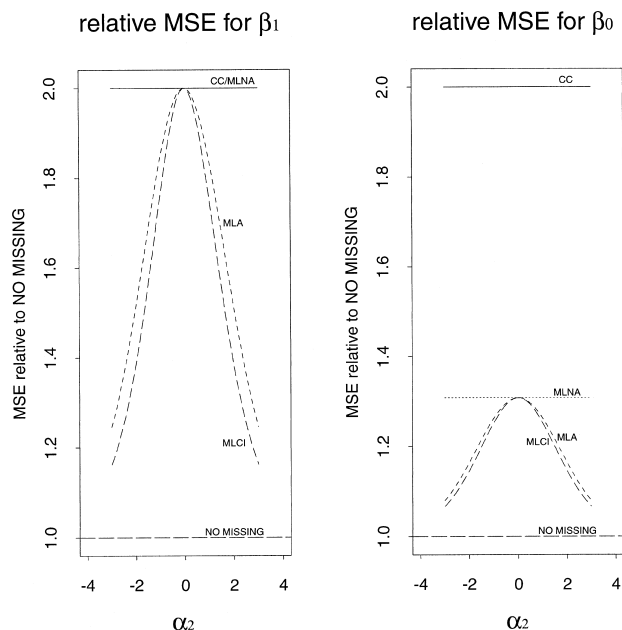
**Figure 2.** MSE (relative to NO MISSINGNESS) when conditional independence is not true.

is varied. We see that, when conditional independence is true, the MLCI estimator has the smallest relative MSE, but when the assumption is false, the bias quickly offsets the gains in efficiency from not estimating the full distribution of the auxiliary variable. The bias of MLCI is more pronounced with larger sample sizes due to the smaller variance with more data.

We next varied the association between the auxiliary variable and the missing covariate ( $\alpha_2$ ) while fixing  $\alpha_1 = 0$  (i.e., conditional independence is true). For this and the remaining simulations, all estimators are consistent, so the relative MSE approximates the asymptotic relative efficiency. Figure 3 displays relative MSE over different values of the  $\alpha_2$  parameter. When  $\alpha_2 = 0$ , there is no information being recovered about the missing covariate and the relative MSE for  $\beta_1$  is the same as for the CC estimator. The relative MSE for  $\beta_0$  for the MLCI and MLA estimators is the same as that for the MLNA estimator when  $\alpha_2 = 0$ . The recovery of information is always good for  $\beta_0$  but is not impressive for  $\beta_1$  unless  $\alpha_2$  is very large.

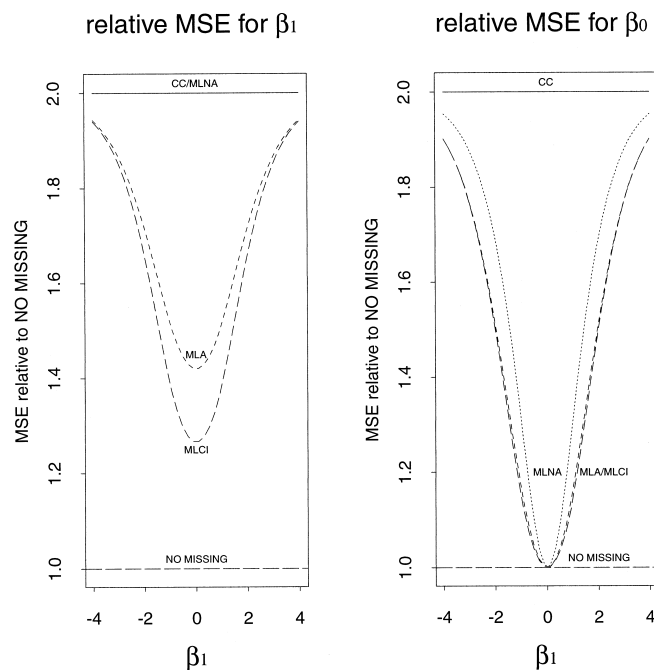
Finally, we varied the magnitude of  $\beta_1$  while fixing  $\alpha_2 = 2$ . Figure 4 displays relative MSE over different values of the  $\beta_1$  parameter. The efficiency of the maximum likelihood estimators depends on the strength of the association between the missing covariate and the outcome, particularly for the  $\beta_0$  parameter. All methods except CC recover 100% of the information about  $\beta_0$  if  $\beta_1 = 0$ . When the association between the outcome and the missing covariate is stronger, efficiency is degraded.

These results show that, in settings where there is a moderate association between the auxiliary variable and the partially observed covariate, the MLCI and MLA estimators have greater efficiency than methods that ignore the auxiliary variable. When the conditional independence is true, the MLCI



**Figure 3.** MSE (relative to NO MISSINGNESS) for different associations between the auxiliary variable and the partially observed covariate.

estimator is preferable, but this assumption is not always tenable. In those settings, the MLA estimator remains consistent and efficient.



**Figure 4.** MSE (relative to NO MISSINGNESS) for different associations between the outcome and the partially observed covariate.

## 6. Discussion

We have considered generalizations of methods for maximum likelihood analysis of logistic regression models with missing covariates that take into account auxiliary information that is not part of the regression model of interest. Our extensions allow use of auxiliary information, which may further improve the efficiency of these methods without making strong assumptions of conditional independence. Our method will be most useful in situations where there is a moderate association between the auxiliary variable and the partially observed covariate and it is not tenable to assume that the conditional independence assumption holds (or our proposed tests of conditional independence reject the null hypothesis). The method is also attractive when it is known that missingness depends on a factor extraneous to the regression model of interest.

Our model can be thought of as semiparametric since no assumptions are made about the nuisance distributions  $f(X)$  and  $f(A | X, Y)$  in calculating the likelihood. It is straightforward, in principle, to extend these methods to a generalized linear model (McCullagh and Nelder, 1989) setting. In the case where  $Y$  is continuous, a model for  $f(A | Y, \mathbf{X})$  would need to be specified or one could consider discretizing  $Y$  for this part of the model. Extensions to other regression models (i.e., for failure time data or longitudinal responses) should be feasible. Such extensions in a setting without auxiliary data have been considered by Lipsitz, Ibrahim, and others (Lipsitz and Ibrahim, 1996b; Lipsitz and Ibrahim, 1998; Lipsitz et al., 1999).

We note that, if the law of the missingness involves unobserved quantities, maximum likelihood estimators that do not model the missingness law will be biased. In our notation, this occurs when the missingness depends on the unobserved  $\mathbf{X}$ 's. The complete case estimator, while potentially inefficient, will remain consistent for the parameters of interest if the missingness only depends on the  $\mathbf{X}$ 's. Also, if the model for the covariates is not correct (which is only guaranteed if a saturated model is fit), then the ML approach may introduce bias.

Because our joint maximization requires separate maximization of each part of the likelihood and because distributions must be estimated for the nuisance terms  $f(X)$  and  $f(A | Y, X)$ , there are limits to the size of the model that is feasible to analyze given finite samples. If one models the nuisance distributions, it is possible to relax the assumptions that the covariates and auxiliary data are discrete. Methods for this case that do not incorporate auxiliary data have been developed to accommodate continuous covariates (Ibrahim and Weisberg, 1992; Ibrahim, Chen, and Lipsitz, 1999) or to provide more parsimonious models for these nuisance parts of the joint likelihood (Lipsitz and Ibrahim, 1996a). These approaches may be useful in the auxiliary data setting, at the risk of potentially introducing additional bias into the overall model because of misspecification of the distributions of  $f(X)$  and  $f(A | Y, X)$ .

## ACKNOWLEDGEMENTS

We are grateful for the support provided by NIMH grants T32-MH17119 and R01-MH54693-02. We would also like to thank Stuart Lipsitz, Garrett Fitzmaurice, the associate editor, and a referee for their helpful comments and Gwendolyn Zahner for use of the dataset.

## RÉSUMÉ

Cet article présente une nouvelle méthode d'estimation des paramètres de modèles logistiques par maximisation de la vraisemblance quand des covariables, pour lesquelles une information auxiliaire est disponible, sont incomplètement observées. Cette information auxiliaire est prédictive de la variable incomplète sans rien apporter au modèle d'intérêt. Ibrahim (*Journal of the American Statistical Association*, 1990) a proposé une méthode générale qui utilise l'algorithme EM pour estimer des modèles linéaires généralisés avec covariables incomplètes. Sa mise en œuvre est facile en l'absence de données auxiliaires. Vach (*Statistics in Medicine*, 1997) a montré comment l'étendre au cas où la réponse est indépendante des données auxiliaires conditionnellement aux covariables. Notre méthode permet d'incorporer l'information auxiliaire sans exiger l'indépendance conditionnelle. Nous proposons des tests d'indépendance conditionnelle et nous comparons plusieurs estimateurs sur un exemple concernant l'utilisation par des enfants d'un service de santé mentale. Nous utilisons des données artificielles pour comparer les performances de plusieurs estimateurs quand l'information auxiliaire est disponible.

## REFERENCES

- Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, Vermont: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington, Vermont: University of Vermont, Department of Psychiatry.
- Bashir, S. A. and Duffy, S. W. (1997). The correction of risk estimates for measurement error. *Annals of Epidemiology* **7**, 154–164.
- Boyle, M. H., Offord, D. R., Racine, Y. A., Fleming, J. E., Szatmari, P., and Links, P. S. (1993). Predicting substance use in early adolescence based on parent and teacher assessments of childhood psychiatric disorder: Results from the Ontario Child Health Study follow-up. *Journal of Child Psychology and Psychiatry* **34**, 535–544.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–22.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge, Massachusetts: Massachusetts Institute of Technology.
- Fitzmaurice, G. M., Laird, N. M., and Lipsitz, S. R. (1994). Analysing incomplete longitudinal binary responses: A likelihood-based approach. *Biometrics* **50**, 601–612.
- Fitzmaurice, G. M., Laird, N. M., and Zahner, G. E. P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association* **91**, 99–108.
- Hasabelnaby, N. A., Ware, J. H., and Fuller, W. A. (1989). Indoor air pollution and pulmonary performance: Investigating errors in exposure assessment. *Statistics in Medicine* **8**, 1109–1126.

- Hui, S. L. and Zhou, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**, 354–370.
- Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**, 765–769.
- Ibrahim, J. G. and Weisberg, S. (1992). Incomplete data in generalized linear models with continuous covariates. *The Australian Journal of Statistics* **34**, 461–470.
- Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591–596.
- Lipsitz, S. R. and Ibrahim, J. G. (1996a). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83**, 916–922.
- Lipsitz, S. R. and Ibrahim, J. G. (1996b). Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis* **2**, 5–14.
- Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* **54**, 1002–1013.
- Lipsitz, S. R., Ibrahim, J. G., and Fitzmaurice, G. M. (1999). Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates. *Biometrics* **55**, 214–223.
- Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* **87**, 1227–1237.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226–233.
- Mantel, H., Singh, A., Kinack, M., and Rowe, G. (1991). Statistical matching: Use of auxiliary information to avoid the conditional independence assumption. *Proceedings of the Bureau of the Census Annual Research Conference*, 688–711.
- MathSoft. (1996). *Splus*, Version 3.4 Supplement. Seattle, Washington: Mathsoft, Data Analysis Products Division.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.
- Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B* **51**, 127–138.
- Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365.
- Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.
- Rosner, B., Willett, W., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8**, 1051–1069.
- Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics* **49**, 305–341.
- Singh, A. C., Mantel, H. J., Kinack, M. D., and Rowe, G. (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology* **19**, 59–79.
- Spiegelman, D., Schneeweiss, S., and McDermott, A. (1997). Measurement error correction for logistic regression models with an “alloyed gold standard.” *American Journal of Epidemiology* **145**, 184–196.
- Torrance-Rynard, V. L. and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* **16**, 2157–2175.
- Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. Berlin: Springer-Verlag.
- Vach, W. (1997). Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Statistics in Medicine* **16**, 57–72.
- Vach, W. and Blettner, M. (1995). Logistic regression with incompletely observed categorical covariates—Investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine* **14**, 1315–1329.
- Wachholder, S., Armstrong, B., and Hartage, P. (1993). Validation studies using an alloyed gold standard. *American Journal of Epidemiology* **137**, 1251–1258.
- Zahner, G. E. P. and Daskalakis, C. (1997). Factors associated with mental health, general health and school-based service use for psychopathology. *American Journal of Public Health* **87**, 1440–1448.
- Zahner, G. E. P., Pawelkiewicz, W., DeFrancesco, J. J., and Adnopoz, J. (1992). Children's mental health service needs and utilization patterns in an urban community. *Journal of the American Academy of Child Adolescent Psychiatry* **31**, 951–960.
- Zahner, G. E. P., Jacobs, J. H., Freeman, D. H., and Trainor, K. (1993). Rural-urban child psychopathology in a north-eastern U.S. state: 1986–1989. *Journal of the American Academy of Child Adolescent Psychiatry* **32**, 378–387.
- Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine* **11**, 769–782.

Received November 1999. Revised July 2000.

Accepted August 2000.