

Maximum Likelihood Analysis of Generalized Linear Models with Missing Covariates

Nicholas J. Horton

Nan M. Laird

Reprinted with permission, *Statistical Methods in Medical Research* 1998; **8**: 37-50

Abstract

Missing data is a common occurrence in most medical research data collection enterprises. There is an extensive literature concerning missing data, much of which has focused on missing outcomes. Covariates in regression models are often missing, particularly if information is being collected from multiple sources. The method of weights is an implementation of the EM algorithm⁸ for general maximum-likelihood analysis of regression models, including generalized linear models³² (GLMs) with incomplete covariates. In this paper, we will describe the method of weights in detail, illustrate its application with several examples, discuss its advantages and limitations, and review extensions and applications of the method.

Introduction

In this paper we will review a very general approach to estimating parameter vectors in the regression setting where some subjects may have only partially observed covariates. The problem of handling missing covariates has received much attention²⁷ because it can be very common, especially in epidemiologic surveys where information about a subject may come from multiple sources, and when the number of confounding variables is large. Provided that the missingness in the covariates is unrelated to the outcome, say Y , in the

regression model, one can always obtain unbiased estimates of regression parameters by using only subjects with complete data on all covariates in the model.¹⁸ However, this strategy may entail too much loss of information, especially if numerous variables subject to missingness are used in the analysis. It also has the undesirable feature that different models will be fit using a different subset of the data.

Other strategies sometimes considered involve use of imputed or “filled-in” values for the missing covariates. These approaches have the advantage of simplicity and including all cases, but they generally introduce bias into both the estimated regression coefficients and its standard errors.²⁷ The same can be said for methods based on indicator methods.¹⁸

Little²⁷ discussed approaches which assumed a parametric form for the joint distribution for Y and covariates X . When this joint distribution is multivariate normal, likelihood analysis is straightforward. Robins, Rotnitzky and Zhao³³ proposed a class of semiparametric estimators, based on weighted estimating equations. They specify a parametric model for the missingness law in addition to the regression model of interest. Little and Schluchter²⁹ considered a linear regression setting with missing discrete covariates. Ibrahim¹⁴ extended their approach to the generalized linear model (GLM) setting and described it as the method of weights. We will review the method of weights and its generalizations. We will assume a GLM regression model for $f(Y|X)$, assume that the covariates X are discrete, and model the covariate distribution nonparametrically using a saturated log-linear model.¹⁰ Since the class of generalized linear regression models forms the basis of our discussion, we first review the method assuming all covariates are observed.

Review of the GLM

McCullagh and Nelder³² introduced the GLM for exponential family data with the following form:

$$f_Y(y, \delta, \phi) = \exp \{ (y\delta - b(\delta)) / a(\phi) + c(y, \phi) \}, \quad (1)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, δ is the canonical parameter, and ϕ is the dispersion parameter. A sample of n independent observations $\mathbf{y} = (y_1, \dots, y_n)'$ are drawn, each with density given by (1), where in addition we assume:

$$\begin{aligned} \mu_i &= E[y_i] = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}) \\ &\Rightarrow g(E[y_i]) = \mathbf{x}_i' \boldsymbol{\beta} \end{aligned} \quad (2)$$

where $g(\cdot)$ is a given link function, \mathbf{x}_i is a $p \times 1$ vector of covariates for the i th subject, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. The link function $g(\cdot)$ determines the function

$h(\cdot)$ where for the i th subject, $\delta_i = h(\mathbf{x}'_i\boldsymbol{\beta})$ (if the link is canonical, then $\delta_i = \mathbf{x}'_i\boldsymbol{\beta}$). Writing $a(\phi) = \phi/m_i$ for some known weights m_i yields log-likelihood equations of the form:

$$l_{x,y}(\boldsymbol{\beta}, \phi|X, \mathbf{y}) = \sum_{i=1}^n l_{y|x}(\boldsymbol{\beta}, \phi|\mathbf{x}_i, y_i) = \sum_{i=1}^n \{m_i [y_i h(\mathbf{x}'_i\boldsymbol{\beta}) - b(h(\mathbf{x}'_i\boldsymbol{\beta}))] / \phi + c(y_i, \phi)\}.$$

where X is the $n \times p$ matrix of covariates for all of the observations. In general, solution of the likelihood equations requires an iterative process, but this has been widely implemented in standard computing packages. One of the attractive properties of the GLM is that it allows for linear as well as non-linear models under a single framework. It is possible to fit models where the underlying data are normal, Poisson, binomial or gamma (as well as others) by suitable choice of the functions $a(\cdot), b(\cdot), c(\cdot), g(\cdot)$ for the given sample space.

The GLM presumes the response and all covariates are observed for each subject. The joint distribution can be partitioned such that $f(Y, X) = f(Y|X)f(X)$. Given complete data the likelihood factors and MLE's can be found by maximizing $f(Y|X)$, the conditional distribution of interest. The parameters governing the distribution of the covariates are ancillary to the parameters of interest ($\boldsymbol{\beta}$) in the GLM and can be ignored. We now consider the setting where the vector of covariates \mathbf{x}_i may not be fully observed for all subjects.

Handling missing covariates: EM via The Method of Weights

Here we assume the covariates are discrete random variables with joint distribution indexed by the parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_r)'$. For example, if there are 3 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 $f(y|\mathbf{x}, \boldsymbol{\beta}, \phi)$ be the density of the outcome, and $h(\mathbf{x}|\boldsymbol{\gamma})$ be the multinomial distribution of the covariates, with $\boldsymbol{\gamma}$ being nuisance parameters distinct from $(\boldsymbol{\beta}, \phi)$. Thus $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \boldsymbol{\gamma})$ is the set of all parameters in the model. We can write the complete data log-likelihood in two parts:

$$l_{x,y}(\boldsymbol{\theta}|\mathbf{x}, y) = \sum_{i=1}^n l_{x,y}(\boldsymbol{\theta}|\mathbf{x}_i, y_i) = \sum_{i=1}^n l_{y|x}(\boldsymbol{\beta}, \phi|\mathbf{x}_i, y_i) + \sum_{i=1}^n l_x(\boldsymbol{\gamma}|\mathbf{x}_i). \quad (3)$$

Note that if all \mathbf{x}_i are observed, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are estimated separately; $\boldsymbol{\beta}$ and ϕ are estimated using standard routines for GLM's. If we impose no restrictions on the distribution of X , then for discrete X , $\boldsymbol{\gamma}$ will be estimated by the $r+1$ observed cell counts obtained by cross-classifying the \mathbf{x}_i 's in a p -dimensional contingency table.¹⁰ If we use log-linear models to impose constraints, we can use iterative proportional fitting to maximize $l_x(\boldsymbol{\gamma}|\mathbf{x})$.

Ibrahim used the EM algorithm⁸ to estimate parameters in this likelihood where covariates were only partially observed and the missingness law for the covariates only involves the outcome and the observed covariates. The EM algorithm is a general purpose iterative algorithm for maximizing incomplete data likelihoods. The EM algorithm proceeds in 2 steps, the E-step and the M-step. At the E-step, one calculates the expectation of $l_{x,y}(\boldsymbol{\theta}|\mathbf{x}, y)$ conditioning on the current parameter vector $\boldsymbol{\theta}^{(t)}$ say, and on the observed data, in this case $(y_i, \mathbf{x}_{obs,i}), i = 1, \dots, n$ where $\mathbf{x}_{obs,i}$ represents the subset of covariates in \mathbf{x}_i which 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 this expectation by $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ we have that the E-step can be calculated using:

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^n Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) & (4) \\
&= \sum_{i=1}^n E \left[l_{x,y}(\boldsymbol{\theta}|\mathbf{x}_i, y_i) | \mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)} \right] \\
&= \sum_{i=1}^n \sum_{j=1}^{r+1} p(\mathbf{x}_i = \mathbf{x}^j | \mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)}) l_{x,y}(\boldsymbol{\theta}|\mathbf{x}^j, y_i) \\
&= \sum_{i=1}^n \sum_{j=1}^{r+1} w_{ij}^{(t)} l_{x,y}(\boldsymbol{\theta}|\mathbf{x}^j, y_i)
\end{aligned}$$

where \mathbf{x}^j is the j th possible pattern of the covariates, $l_{x,y}(\boldsymbol{\theta}|\mathbf{x}^j, y_i)$ is the complete data log-likelihood for $\boldsymbol{\beta}$ for the i th observation with \mathbf{x}_i evaluated at \mathbf{x}^j , and $w_{ij}^{(t)} = p(\mathbf{x}^j | \mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)})$ are the weights for the j th possible pattern of the covariates for the i th observation at the t th iteration. We note that $w_{i+}^{(t)} = 1$ for all t and i . Note also that most of these w_{ij} 's will be zero since $\mathbf{x}_{obs,i}$ will rule out any value of \mathbf{x}^j not compatible with $\mathbf{x}_{obs,i}$. The vectors \mathbf{x}^j and $\mathbf{x}_{obs,i}$ are compatible if the components of \mathbf{x}^j corresponding to the observed data equal $\mathbf{x}_{obs,i}$. If only the k th covariate is missing, then there are c_k non-zero terms in the inner sum. If all the covariates for the i th subject are observed, then there is just one non-zero term in the inner sum, and the weight for the observed value $\mathbf{x}^j = \mathbf{x}_{obs,i}$, is equal to 1.

This approach is similar to that used by Wood and Hinde³⁹ to fit a binomial variance component model with random effects. The unobserved random effects were assumed to have a non-parametric distribution, which leads to the same type of method of weights as suggested by Ibrahim.

The weights can be calculated by use of Bayes rule:

$$w_{ij}^{(t)} = p(\mathbf{x}^j | \mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)}) \quad (5)$$

$$= \begin{cases} 0 & \text{if } \mathbf{x}^j \text{ is not compatible with } \mathbf{x}_{obs,i} \\ \frac{p(y_i|\mathbf{x}_{obs,i}^j, \boldsymbol{\beta}^{(t)}, \phi)p(\mathbf{x}_{obs,i}^j|\boldsymbol{\gamma}^{(t)})}{\sum_{k \in obs,i} p(y_i|\mathbf{x}_{obs,i}^k, \boldsymbol{\beta}^{(t)}, \phi)p(\mathbf{x}_{obs,i}^k|\boldsymbol{\gamma}^{(t)})} & \text{if } \mathbf{x}^j \text{ is compatible with } \mathbf{x}_{obs,i} \end{cases}$$

where $\mathbf{x}_{obs,i}^j$ denotes a p -vector whose observed components are $\mathbf{x}_{obs,i}$ 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 $\mathbf{x}_{obs,i}$.

For the M-step, we maximize equation (4) as a function of $\boldsymbol{\theta}$ by finding the solution to the complete-data log-likelihood using these weights. Since $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ can be written in the same form as (3) with the first term, $E[l_{y|x}(\boldsymbol{\beta}, \phi|\mathbf{x}, y)]$ depending only on $\boldsymbol{\beta}$ and ϕ and the second, $E[l_x(\boldsymbol{\gamma}|\mathbf{x})]$ depending only on $\boldsymbol{\gamma}$, the maximizations again can be done separately. We can use the weighted GLM to estimate $\boldsymbol{\beta}$ and ϕ and use the expected cell counts to estimate $\boldsymbol{\gamma}$. For the first iteration, the weights can be calculated using (5) with initial estimates for $\boldsymbol{\theta}$ from a complete case analysis. By repeating these steps until convergence, we obtain our parameter estimates of interest. Ibrahim referred to the implementation of the EM in this setting as the method of weights.

Standard errors for these parameter estimates can be calculated using Louis' method,³⁰ which 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 could be calculated using the relationship:

$$\begin{aligned} I(\boldsymbol{\theta}|Y_{obs}) &= E[I(\boldsymbol{\theta}|Y_{comp})|Y_{obs}] - \\ &E[S(\boldsymbol{\theta}|Y_{comp})S'(\boldsymbol{\theta}|Y_{comp})|Y_{obs}] + \\ &E[S(\boldsymbol{\theta}|Y_{comp})|Y_{obs}]E[S'(\boldsymbol{\theta}|Y_{comp})|Y_{obs}], \end{aligned}$$

where S is the score (first-derivative of the log-likelihood) and I is the information (negative of the second-derivative of the log-likelihood).

Implementation of the EM method of weights is straightforward. Consider an example with three dichotomous covariates (x_1, x_2, x_3), dichotomous y and three different missing data patterns. Figure 1 displays several observed data points and the corresponding augmented dataset, where w_{ij} are the estimated weights (i.e. $w_{31} = P(x_1 = 0|x_2 = x_3 = y = 0, \boldsymbol{\theta})$), $w_{31} + w_{32} = 1$, $w_{51} + w_{52} = 1$, and $\sum_{j=1}^8 w_{7j} = 1$. To simplify exposition, zero weights are not shown. Given these weights, the complete data log likelihood can easily be fit using existing GLM methods using the augmented dataset, and new weights can be calculated using those parameter estimates. Standard error estimates (using Louis' method) can be calculated using quantities available from weighted GLM routines.

While we have limited our discussion to generalized linear regression models, the method

Figure 1: Observed and augmented data

Original dataset					Augmented dataset				
#	y	x ₁	x ₂	x ₃	y	x ₁	x ₂	x ₃	wt
1	0	0	0	0	0	0	0	0	1
2	0	0	0	1	0	0	0	1	1
3	0	-	0	0	0	0	0	0	w ₃₁
4	0	0	1	0	0	1	0	0	w ₃₂
5	0	0	1	-	0	0	1	0	w ₅₁
6	1	0	0	0	0	0	1	1	w ₅₂
7	1	-	-	-	1	0	0	0	1
8	1	1	0	1	1	0	0	0	w ₇₁
					1	0	0	1	w ₇₂
					1	0	1	0	w ₇₃
					1	0	1	1	w ₇₄
					1	1	0	0	w ₇₅
					1	1	0	1	w ₇₆
					1	1	1	0	w ₇₇
					1	1	1	1	w ₇₈
					1	1	0	1	1

of weights can be generalized to other regression models. Schluchter and Jackson³⁴ proposed an extension of Laird and Olivier's²¹ method for log-linear analysis of survival data to a setting with missing covariates. Lipsitz and Ibrahim²⁴ proposed a straightforward extension of the method of weights to parametric survival models and have also considered estimating equation approaches to incorporate missing covariates in the Cox proportional hazards regression model.²⁵ We will now consider an example.

Example: Mental Health Service Utilization

To illustrate these methods, we consider a study of mental health service utilization in children in urban and rural Connecticut.⁴⁰⁻⁴² 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 1, and coded 0 otherwise.

Predictors of service use included gender of the child (BOY: 1=boy, 0=otherwise), age of the child (OLD: 0=age 6 to 8, 1=age 9-11), ethnicity (BLACK: 0=non-black, 1=black and HISPANIC: 0=non-hispanic, 1=hispanic), religion (CATHOLIC: 0=non-catholic, 1=catholic) belonging to a single parent household (MOMSING: 0=father figure present, 1=no father figure present), and psychopathology of the child. One measure of psychopathology used in the study was the total problems scale of the Teacher's Report Form (TRF).¹ The raw scores were dichotomized at the cutpoint for borderline/clinical psychopathology. A score of 1 indicates borderline/clinical psychopathology, and a score of 0 indicates normal range. In this example, 43% of the teacher ratings on children were missing. Missingness of this magnitude is not uncommon: a similar rate was reported by Boyle et al in their Ontario Child Health Study.⁴ All other covariates were fully observed.

Table 1 displays the proportion of positive reports as well as the number of subjects with observed data for that variable. We fit a model using the complete case estimator, which discards all information on the 1,061 subjects who were missing teacher reports. We also fit a model via maximum likelihood (ML). Table 2 displays the parameter estimates and standard errors (using Louis' method) for these models, as well as the relative size of the complete case standard error compared to ML. Being a boy, being older, being non-black, 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 complete case estimator are larger than those

Table 1: Summary of outcome and predictors

<i>Variable</i>	<i>Mean</i>	<i># observed</i>
OUTCOME	0.185	2486
BOY	0.481	2486
OLD	0.473	2486
BLACK	0.198	2486
HISPANIC	0.068	2486
CATHOLIC	0.417	2486
MOMSING	0.207	2486
TRF	0.183	1425

Table 2: Parameter estimates (and standard errors) for models of service utilization

<i>Parameter</i>	<i>Complete case</i>	<i>Maximum likelihood</i>	<i>Relative size SE</i>
INTERCEPT	-2.336 (0.174)	-2.210 (0.132)	1.32
BOY	0.440 (0.150)	0.441 (0.111)	1.35
OLD	0.492 (0.149)	0.546 (0.111)	1.35
BLACK	-0.881 (0.239)	-0.681 (0.171)	1.39
HISPANIC	-0.332 (0.317)	-0.221 (0.231)	1.37
CATHOLIC	-0.115 (0.158)	-0.170 (0.118)	1.33
MOMSING	0.582 (0.200)	0.324 (0.148)	1.36
TRF	1.413 (0.163)	1.418 (0.162)	1.00

of the ML estimator. In this example, the complete case estimator standard errors were 32-39% larger than those of the ML estimator, with the exception of the standard error of the TRF parameter. Since no additional information is available for the covariate which is missing, there is little gain in efficiency for estimation of this parameter.

We will now consider some features and extensions of the method of weights.

Features and Extensions

The basic maximum likelihood approach is to estimate the joint distribution of Y and X without assuming any parametric form for X ; the parametric assumption for $Y|X$ is assumed known from the complete case regression setting. If the covariate sample space is large (r large), then there may be little information to estimate the nuisance parameters γ unless n is correspondingly large, even though there may be sufficient information to estimate β with X completely observed. In this case, leaving the distribution of X completely unspecified may result in little increased efficiency. For instance, if the particular pattern of observed covariates and Y for a person with missing X values is distinct from those subjects with complete data, there may be no information to estimate $P(X|X_{obs}, Y)$ and such subjects will be excluded from the analysis.

One approach to handling this problem is to use some parsimonious model for γ , such as a log-linear model¹⁰ without all higher-order interactions. Lipsitz and Ibrahim²³ suggest an alternative model for γ which reduces dimensionality. The goal of these models is to balance misspecification of the covariate distribution (with a potential increase in bias) with a parsimonious model for the covariates (with the potential for variance reduction).

Another feature of the the maximum likelihood implementation is the requirement that covariates be discrete. In principle it is possible to estimate the distribution of X completely nonparametrically, allowing continuous covariates, but if the values of the continuous variables are truly distinct there may be insufficient information to estimate the required conditional distributions, depending upon where the missingness occurs. A general approach to this problem, besides categorizing the continuous covariates, is to assume a parametric model for the continuous X 's. If the covariates are continuous and subject to missingness, the form of (4) is no longer a sum but an integral:

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int p(\mathbf{x}|\mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)})l(\boldsymbol{\theta}|\mathbf{x}, y_i)d\mathbf{x}, \quad (6)$$

where the integral is taken over the subspace of \mathcal{X} (the complete data sample space) where \mathbf{x} is known to lie, having observed $(y_i, \mathbf{x}_{obs,i})$. Ibrahim and Weisberg¹⁶ considered

a solution to this problem by approximating the E-step using Gaussian quadrature. In effect, the covariate is discretized using an approximating orthogonal polynomial, and the simple form of (4) is restored. Ibrahim, Chen and Lipsitz¹⁵ describe a method to fit parametric regression models for $f(Y|X)$ with continuous missing covariates using a Monte Carlo EM algorithm. If certain parametric distributions are assumed for the covariates (a much stronger assumption than the nonparametric form assumed earlier), then a Gibbs sampler can be used to take a series of samples from this distribution. These g_i samples can be used in the likelihood (4) with weights equal to $1/g_i$. This technique is even more computationally expensive, since within each step of the EM algorithm many samples must be taken from the distribution and the number of observations in the complete data GLM can be quite large if g_i is large. For example, in their simulation study Lipsitz and Ibrahim²³ performed 2000 Monte Carlo iterations within each iteration of the EM algorithm.

If both categorical and continuous covariates are missing, an assumption of joint multivariate normality may not be appropriate. There are few models for the joint distribution of mixtures of categorical and continuous variables, though some have been proposed (consider Little and Schluchter,²⁹ Cox and Wermuth⁷ or Fitzmaurice and Laird¹¹). Lipsitz and Ibrahim²³ considered factoring the distribution of the covariates into a part involving the categorical covariates and one involving the continuous covariates. Without loss of generality, order the covariates such that the d categorical covariates are given by the vector $(x_{i1}, \dots, x_{id})'$ and the $p-d$ continuous covariates are given by the vector $(x_{i,d+1}, \dots, x_{ip})'$. It may be reasonable to partition the nuisance covariate distribution:

$$p(\mathbf{x}_i|\boldsymbol{\gamma}) = p(x_{i,d+1}, \dots, x_{ip}|x_1, \dots, x_{id}, \boldsymbol{\gamma}_1)p(x_{i1}, \dots, x_{id}|\boldsymbol{\gamma}_2), \quad (7)$$

where the first term on the right hand side of the equation is modeled assuming conditional joint multivariate normality and the second term is modeled using a log-linear model.^{7,11}

Underlying the validity of the maximum likelihood analysis is the assumption that the missingness law for the covariates depends only on the outcome and the observed covariates (missing at random (MAR) in the sense of Little and Rubin²⁸). This assumption is inherently untestable using the observed data and external information is needed to consider whether the assumption is true. If the missingness is due to design, then the assumption may be warranted. As an example of this situation, Leong et al²² applied the method of weights in a genetic study where for cost reasons, a complete genetic evaluation of all patients was not carried out. The investigators were interested in determining the association between two disease cell types and mutations of the RAS family of oncogenes in a sample of patients with multiple myeloma. Five locations (codons) for mutations were examined. Patients were considered to have a mutation if any of the codons were found to have a mutation. But only 20% of the patients had all five locations tested. The probability that a location was tested generally depended on whether a mutation had

been found previously.

In other settings, the appropriateness of this missingness assumption may be less straightforward to determine. Vach and others³⁶⁻³⁸ have considered the sensitivity of the method of weights as well as other approaches to violations of this assumption. Ibrahim, Lipsitz and Chen¹⁷ and Lipsitz, Ibrahim, Chen and Peterson²⁶ generalized this approach to situations where the missingness law may relate to the unobserved covariates and outcome. Little and Rubin describe this as non-ignorable non-response (NINR). In this situation, the joint distribution (Y, X, R) needs to be specified and estimated, where R is a $p \times 1$ random vector of indicator variables which equal 1 if the p th element of X is observed and 0 otherwise. In addition to the regression model of interest and a model for the covariates, a model must also be specified for the missing data mechanism. For a single missing covariate, Lipsitz et al²⁶ suggested a logistic regression model for this component of the likelihood. For multiple missing covariates, Ibrahim et al¹⁷ suggested a series of conditional logistic regression models. While there are no formal methods to aid in model selection for this missingness model, comparisons of the parameter estimates for β under different missingness models may be made. Ibrahim et al¹⁷ note that characterization of the estimability of models for the joint distribution of X , Y , and R is not easily determined. Baker and Laird² discuss this problem when outcomes are nonignorably missing.

Thus far we have focused on missing covariate values. We will now review applications and extensions of the method of weights which arise with other types of incompleteness in the covariates, including measurement error, regression on latent variables, and use of auxiliary information.

Maximum Likelihood for Related Incomplete Covariate Problems

There is an extensive literature concerning statistical methods which incorporate measurement error of covariates into regression models.^{3,6} For discrete predictors, measurement error can be thought of as misclassification error. We will review how methods proposed to address misclassification of exposures in the epidemiological literature, including latent class analysis,^{12,19} can be implemented in a straightforward fashion using the method of weights. Kosinski and Flanders²⁰ considered estimation of odds ratios for disease given true exposure where subjects were given two (imperfect) tests to detect the presence or absence of an exposure. For simplicity of exposition, we assume that the true exposure is a dichotomous variable, and the two tests each generate a binary response. Extension to more than two unobserved (latent) classes, or more than two tests is straightforward.

Let T_{i1} and T_{i2} represent the results of the tests for the i th subject, and let E_i represent the true (unobserved) latent class for that individual. Kosinski and Flander considered a dichotomous outcome (disease) corresponding to an epidemiologic setting (such as a case-control study). Primary interest revolves around the odds ratio of disease given true exposure:

$$OR = \frac{P(Y = 1, E = 1|X)P(Y = 0, E = 0|X)}{P(Y = 1, E = 0|X)P(Y = 0, E = 1|X)}, \quad (8)$$

where X denotes other covariates, which for simplicity of exposition are presumed to be fully observed. The above odds ratio corresponds to a parameter in the logistic regression model $f(Y|E, X)$. As in the missing data setting, it is necessary to estimate $f(E, T_1, T_2|X)$ and calculate $P(E|T_1, T_2, X, Y)$ to determine the weights. Because E is never observed, a number of assumptions are needed to identify this model. Frequently it is assumed that there is non-differential misclassification (i.e. the sensitivities and specificities do not depend on the levels of the outcome²⁰). In addition, conditional independence of the two tests is often assumed, given the true exposure. Finally, it is assumed that given the true exposure, the outcome is conditionally independent of the measurement error. These assumptions allow the factorization of the joint likelihood:

$$\begin{aligned} P(Y, E, T_1, T_2|X) &= P(Y|E, T_1, T_2, X)P(E, T_1, T_2|X) \\ &= P(Y|E, X)P(E, T_1, T_2|X) \end{aligned} \quad (9)$$

This can be fit by maximum likelihood using the method of weights by conditioning on the observed mismeasured covariates T_1 and T_2 when calculating the weights.

The assumptions underlying this misclassification latent class model—like any modeling assumptions—must be carefully considered to ensure valid results. Brenner⁵ found situations where positive error correlation between the dual measurements led to biased latent class estimates.

Another setting allows for the inclusion of auxiliary (extraneous) covariates to improve the efficiency of the generalized linear regression model, as suggested by Horton and Laird.¹³ 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) might be available which can be matched to subjects in an investigation. 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 subjected to the high-quality assessment, while more error-prone auxiliary data might be collected on all subjects.

Let V be a discrete auxiliary variable. A commonly used approach in this setting is to

assume a conditional independence model. similar to (9):

$$f(Y, V|X) = f(Y|X)f(V|X).$$

Multiplying both sides by $f(X)$ yields the full likelihood and maximum likelihood estimation via EM leads to the method of weights. Mantel et al^{31,35} review the bias that can ensue if the conditional independence model is not correct. Horton and Laird considered a more general factorization of the joint likelihood which does not rely on conditional independence:

$$f(Y, V, X) = f(V|Y, X)f(Y|X)f(X) \tag{10}$$

where primary interest is in the regression $f(Y|X)$. Now $\theta = (\alpha, \beta, \gamma, \phi)$ is the set of all parameters in the complete data log-likelihood, which for the i th individual is given by:

$$l_{v,y,x}(\theta|y_i, v_i, x_i) = l_{v|y,x}(\alpha|v_i, y_i, x_i) + l_{y|x}(\beta, \phi|y_i, x_i) + l_x(\gamma|x_i). \tag{11}$$

To estimate the parameters of interest in this model we must simultaneously maximize the likelihoods of the other two terms. As before, when X is categorical, $f(X)$ can be specified nonparametrically. We must specify a model for $f(V|Y, X)$. If Y is discrete, we can consider fitting a saturated model. In other cases, a model will need to be specified. With no missing data, we can just maximize $\prod f(y_i|\mathbf{x}_i)$ to estimate β and ϕ because this is a recursive system. But since components of \mathbf{x}_i are sometimes missing, we must simultaneously maximize the likelihoods of the other terms. Horton and Laird proposed use of the EM algorithm and the method of weights to fit this model by estimating $f(\mathbf{x}_i|v_i, y_i)$ for all possible covariate patterns (the E-step) and then fitting the model for the three parts of the likelihood separately (the M-step) with saturated models (if possible) for the two nuisance parts of the likelihood. Bootstrap⁹ samples from the observed contingency table were used to generate standard error estimates for the parameters. Several approaches were compared, including the complete case estimator, the maximum likelihood approach (method of weights) ignoring the auxiliary information, the conditional independence maximum likelihood model along with the joint maximization approach with auxiliary data.

Tables 3 and 4 summarize the settings, factorizations of the likelihoods and assumptions underlying these methods for the various approaches reviewed in this paper. These factorizations are attractive because they all have the regression model of interest, $f(Y|X)$ or $f(Y|E, X)$, as part of the model.

Table 3: Complete-data likelihood factorizations for different methods and settings

<i>Method/Setting</i>	<i>Likelihood to Maximize</i>	<i>Factorization</i>
complete case	$f(Y X)$	$f(Y X)$
ML under MAR	$f(Y, X)$	$f(Y X)f(X)$
ML under NINR	$f(Y, X, R)$	$f(Y X)f(R Y, X)f(X)$
ML with auxiliary data under MAR/CI	$f(Y, V, X)$	$f(Y X)f(V X)f(X)$
ML with auxiliary data under MAR	$f(Y, V, X)$	$f(Y X)f(V Y, X)f(X)$
ML latent variable CI	$f(Y, E, T, X)$	$f(Y E, X)f(T, E X)f(X)$

Table 4: Assumptions underlying different methods

<i>Method/Setting</i>	<i>Underlying Assumptions</i>
complete case	$f(R Y, X) = f(R X)$
ML under MAR	$f(R Y, X) = f(R Y, X_{obs})$
ML under NINR	$f(R Y, X)$ correctly specified
ML with auxiliary data under MAR/CI	$f(Y V, X) = f(Y X), f(R Y, V, X) = f(R Y, V, X_{obs})$
ML with auxiliary data under MAR	$f(R Y, V, X) = f(R Y, V, X_{obs})$
ML latent class CI	$f(Y E, T, X) = f(Y E, X)$

Discussion

The EM method of weights provides maximum likelihood estimates of the parameters governing the joint distribution of the outcome and the covariates. If the missingness law involves only the unobserved covariates and the outcome, the method yields unbiased estimates of the parameters of interest when the likelihood for $l(\boldsymbol{\beta}, \phi|X, Y)$ is correct. If the covariates were fully observed, the distribution of the covariates would be ancillary to the regression model of interest and only the regression model needs to be correct for consistent estimates of $\boldsymbol{\beta}$. But when covariates are missing, estimation of this nuisance distribution takes into account the partial information. The method of weights was originally proposed for logistic regression, but has been generalized to the generalized linear model, parametric and semi-parametric censored survival models as well as longitudinal binary regression models.

When the missing covariates are categorical or discrete, it is straightforward to fit a saturated non-parametric model for the distribution of the covariates. If there are many covariates, it may be necessary to specify a more parsimonious model to avoid a proliferation of nuisance parameters. If the nuisance parameter distribution is not correctly modeled (which is only guaranteed if a saturated model is specified), then the method of weights may introduce bias in the regression model of interest. Further work is needed to consider how estimation of the parameters of interest are affected by missmodeling of the nuisance distribution.

For models with discrete covariates, it is straightforward to compute maximum likelihood estimates of the joint distribution using existing statistical regression packages which allow for weights in the regression analysis, under the assumption that the missingness depends only upon observed covariates and the outcome. Work is underway to incorporate this method more generally in statistical packages (Lipsitz SL, personal communication 1998).

Extensions to the method which allow for continuous covariates and mixed continuous and categorical covariates add additional computational complexity to model-fitting, but are computationally feasible if the number of missing covariates is small. Other statistical methods (such as models for misclassified exposure and latent class models) can be considered as missing covariate problems where the true covariate is never observed for any subject; here maximum likelihood via the method of weights can be straightforwardly applied. The method also permits one to incorporate auxiliary information V into the model, where it is assumed that V gives information about X and also possibly $f(Y|X, V)$.

Acknowledgements

We are grateful for the support provided by NIMH grants R01-MH54693 and T32-MH17119. We would like to thank Joseph Ibrahim for useful discussions about the method of weights, the reviewers for useful comments and suggestions, and Gwendolyn Zahner for providing access to the data from the example. The Connecticut child surveys were conducted by Dr. Zahner under contract to the Connecticut Department of Children and Youth Services.

References

- [1] Achenbach TM. *Manual for the Teacher's Report Form and 1991 Profile*. University of Vermont: Department of Psychiatry, 1991.
- [2] Baker SG and Laird NM. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83(401):62–69, 1988.
- [3] Bashir SA and Duffy SW. The correction of risk estimates for measurement error. *Annals of Epidemiology*, 7(2):154–164, 1997.
- [4] Boyle MH, Offord DR, Racine YA, Fleming JE, Szatmari P, and Links PS. 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(4):535–544, 1993.
- [5] Brenner H. Use and limitations of dual measurements in correcting for nondifferential exposure misclassification. *Epidemiology*, 3:216–222, 1992.
- [6] Carroll RJ, Ruppert D, and Stefanski LA. *Measurement Error in Nonlinear Models*. Chapman & Hall, 1995.
- [7] Cox DR and Wermuth N. Response models for mixed binary and quantitative variables. *Biometrika*, 79:441–461, 1992.
- [8] Dempster AP, Laird NM, and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22, 1977.
- [9] Efron B and Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

- [10] Fienberg SE. *The Analysis of Cross-classified categorical data*. Massachusetts Institute of Technology, 1980.
- [11] Fitzmaurice GM and Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90:845–852, 1995.
- [12] Formann AK and Kohlmann T. Latent class analysis in medical research. *Statistical Methods in Medical Research*, 5:179–211, 1996.
- [13] Horton NJ and Laird NM. Using auxiliary information to estimate generalized linear regression models with missing covariate data. submitted.
- [14] Ibrahim JG. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.
- [15] Ibrahim JG, Chen M-H, and Lipsitz SR. Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, 55(2):591–596, 1999.
- [16] Ibrahim JG and Weisberg S. Incomplete data in generalized linear models with continuous covariates. *The Australian Journal of Statistics*, 34:461–470, 1992.
- [17] Ibrahim JG, Lipsitz SR, and Chen M-H. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B*, 61:173–190, 1999.
- [18] Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91(433):222–230, 1996.
- [19] Kaldor J and Clayton D. Latent class analysis in chronic disease epidemiology. *Statistics in Medicine*, 4:327–335, 1985.
- [20] Kosinski AS and Flanders WD. Regression model for estimation of odds ratios with misclassified exposure. Technical Report 96-5, Department of Biostatistics, Rollins School of Public Health, Emory University, 1996.
- [21] Laird NM and Olivier D. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76:231–240, 1981.
- [22] Leong T, Lipsitz SR, and Ibrahim JG. Using missing data methods in genetic studies with missing mutation status. *Statistics in Medicine*, 18:473–485, 1999.
- [23] Lipsitz SR and Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83(4):916–922, 1996.

- [24] Lipsitz SR and Ibrahim JG. Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2:5–14, 1996.
- [25] Lipsitz SR and Ibrahim JG. Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, 54(3):1002–1013, 1998.
- [26] Lipsitz SR, Ibrahim JG, Chen M-H, and Peterson H. Non-ignorable missing covariates in generalized linear models. *Statistics in Medicine*, 1999. To appear.
- [27] Little RJA. Regression with missing X’s: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.
- [28] Little RJA and Rubin DB. *Statistical Analysis With Missing Data*. John Wiley & Sons, 1987.
- [29] Little RJA and Schluchter MD. Maximum likelihood estimation with mixed continuous and categorical data with missing values. *Biometrika*, 72:497–512, 1985.
- [30] Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44(2):226–233, 1982.
- [31] Mantel H, Singh A, Kinack M, and Rowe G. Statistical matching: Use of auxiliary information to avoid the conditional independence assumption. In *Proceedings of the Bureau of the Census Annual Research Conference*, pages 688–711, 1991.
- [32] McCullagh P and Nelder JA. *Generalized Linear Models*. Chapman & Hall, 1989.
- [33] Robins JM, Rotnitzky A, and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.
- [34] Schluchter MD and Jackson KL. Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, 84:42–52, 1989.
- [35] Singh AC, Mantel HJ, Kinack MD, and Rowe G. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19:59–79, 1993.
- [36] Vach W. Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Statistics in Medicine*, 16:57–72, 1997.
- [37] Vach W and Blettner M. Logistic regression with incompletely observed categorical covariates – Investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, 14:1315–1329, 1995.

- [38] Vach W and Illi S. Biased estimation of adjusted odds ratios from incomplete covariate data due to violation of the missing at random assumption. *Biometrical Journal*, 39(1):13–28, 1997.
- [39] Wood A and Hinde J. Binomial variance component models with a non-parametric assumption concerning random effects. In *Longitudinal Data Analysis: Surrey Conference on Sociological Theory and Method 4*, pages 110–128. Avebury, Aldershot, 1987.
- [40] Zahner GEP and Daskalakis C. Factors associated with mental health, general health and school-based service use for psychopathology. *American Journal of Public Health*, 87(9):1440–1448, 1997.
- [41] Zahner GEP, Jacobs JH, Freeman DH, and Trainor K. Rural-urban child psychopathology in a northeastern U.S. state: 1986-1989. *Journal of the American Academy of Child Adolescent Psychiatry*, 32:378–387, 1993.
- [42] Zahner GEP, Pawelkiewicz W, DeFrancesco JJ, and Adnopoiz J. 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, 1992.