Journal of Statistical Computation and Simulation
Vol. 77, No. 2, February 2007, 163–173

Taylor & Francis
Taylor & Francis Group

An EM algorithm for regression analysis with incomplete covariate information

ZHIWEI ZHANG† and HOWARD E. ROCKETTE*‡

†Division of Biostatistics, OSB/CDRH/FDA, 1350 Picard Drive, Rockville, MD 20850, USA
‡Department of Biostatistics, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261, USA

(Received 1 November 2004; in final form 25 May 2005)

Regression analysis is often challenged by the fact that some covariates are not completely observed. Among other approaches is a newly developed semiparametric maximum likelihood (SML) method that requires no parametric specification of the selection mechanism or the covariate distribution and that yields efficient inference, at least in some specific models. In this paper, we propose an EM algorithm for finding the SML estimate and for variance estimation. Simulation results suggest that the SML method performs reasonably well in moderate-sized samples. In contrast, the analogous parametric maximum likelihood method is subject to severe bias under model mis-specification, even in large samples.

1. Introduction

Parametric regression models such as generalized linear models are commonly used to assess the effect of a vector X of covariates on an outcome variable Y. Under such a model, the conditional distribution of Y given X is known up to a finite-dimensional regression parameter θ. Based on a random sample from (X, Y), θ can be estimated using any of the standard methods such as maximum likelihood.

Quite often, however, a portion of X is unobserved for some subjects, either by design or by happenstance. Write $X = (W, Z)$, where W is always observed and Z is possibly missing. Assume that Z is missing at random (MAR) in the sense of Rubin [1], that is, the conditional probability given (X, Y) that a subject is selected for full observation does not depend on Z. If the conditional distribution of Z given W is known can be parametrically modeled, it is straightforward to estimate θ by maximizing the likelihood for the observed data. A Monte Carlo EM algorithm has been proposed by Ibrahim *et al.* [2] for computing the maximum likelihood estimator (MLE). The resulting estimate is efficient if the covariate distribution is correctly specified, but can be biased under model mis-specification.

In practice, it is often difficult to specify a model for the covariate distribution that is nearly correct. When W is (finitely) discrete, it is possible to maximize a semiparametric likelihood

*Corresponding author. Email: herbst@pitt.edu

Journal of Statistical Computation and Simulation
ISSN 0094-9655 print/ISSN 1563-5163 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10629360600565202

where the conditional distribution of Z given W is left unspecified [3, 4]. The resulting semiparametric maximum likelihood estimator (SMLE) is efficient in the semiparametric sense, at least in some specific models. This paper is concerned with the implementation and finite sample performance of a restricted version of the SMLE, restricted in the sense that the conditional distribution of Z given $W = w$ is required to concentrate on the observed Z_i with $W_i = w$. We propose an EM algorithm for finding the SMLE and for variance estimation, and conduct simulation experiments to compare the SMLE with analogous MLEs in terms of robustness and efficiency.

The rest of the paper is organized as follows. In section 2, we formulate the problem and introduce the SMLE. In section 3, we derive an EM algorithm and discuss variance estimation. Simulation results are reported in section 4. An application is presented in section 5. The paper concludes with a discussion in section 6.

2. The (restricted) SMLE

Let X be a vector of covariates and Y be a response variable. The conditional distribution of Y given $X = x$ is specified through the conditional density $f(\cdot|x;\theta)$ with respect to some fixed measure. Here f is a known function and θ is an unknown d-dimensional regression parameter. Suppose that a portion of X is unobserved on some subjects. Write $X = (W, Z)$, where W is always observed and Z is possibly missing. Denote by $G(\cdot|w)$ the conditional distribution of Z given $W = w$. Let $R = 1$ if Z is observed, 0 otherwise. It is assumed that Z is missing at random, that is,

$$E(R|X, Y) = E(R|W, Y) =: \pi(W, Y). \tag{1}$$

The function π specifies the conditional probability of selecting a subject for complete observation, and will be refered to as the selection mechanism. Let (X_i, Y_i, R_i), $i = 1, \ldots, n$, be independent copies of (X, Y, R); however, we only observe $(R_i, W_i, R_i Z_i, Y_i)$, $i = 1, \ldots, n$.

Suppose for the moment that the conditional distribution $G(\cdot|w)$ can be parameterized, with conditional density $g(\cdot|w;\gamma)$ with respect to a fixed measure ν. Then the likelihood for (θ, γ) is given by

$$\prod_{i=1}^{n}[f(Y_i|X_i;\theta)g(Z_i|W_i;\gamma)]^{R_i}\left[\int f(Y_i|W_i, z;\theta)g(z|W_i;\gamma)\,\mathrm{d}\nu(z)\right]^{1-R_i}, \tag{2}$$

and an EM algorithm has been proposed by Ibrahim *et al.* [2] to maximize this likelihood. Note that expression (2) does not involve the selection mechanism π, by the MAR assumption (1). However, the validity of inference based on expression (2) does require correct modeling of G, which can be quite difficult in practice.

It is therefore important to consider relaxing the parametric assumptions on G. According to Zhang and Rockette [3, 4], G can be treated nonparametrically within the maximum likelihood framework if W is discrete. Assume that W takes values in the finite set $\{w_1, \ldots, w_J\}$. Without specifying a model for G, consider the semiparametric likelihood

$$L(\theta, G) = \prod_{i=1}^{n}[f(Y_i|W_i, Z_i;\theta)G(\{Z_i\}|W_i)]^{R_i}\left[\int f(Y_i|z, W_i;\theta)G(\mathrm{d}z|W_i)\right]^{1-R_i}. \tag{3}$$

It is natural to maximize $L(\cdot, \cdot)$ over the entire parameter space, which consists of all possible values of θ and all possible conditional distributions G. This turns out to be asymptotically equivalent to a simpler maximization with the restriction that $G(\cdot|w)$ be supported

by the observed values of Z on subjects with $W = w$ ([3, Theorem 10]). Computationally, the global maximization is infinite-dimensional, whereas the restricted maximization is finite-dimensional. Therefore we focus on the restricted SMLE:

$$(\hat{\theta}, \hat{G}) = \text{argmax}_{(\theta,G):\ G(D_j|w_j)=1,\ j=1,\ldots,J} L_n(\theta, G),$$

where $D_j := \{Z_i: W_i = w_j, R_i = 1\}$, $j = 1, \ldots, J$.

It was shown in Zhang and Rockette ([4, Theorem 4.3]) that $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and variance I_e^{-1}, where I_e is the efficient information for θ with G unspecified. This result applies to such popular models as logistic, normal, and Poisson regression models. It appears difficult to estimate I_e directly using the usual plug-in method. Fortunately, a consistent estimator can be obtained by perturbing the profile log-likelihood for θ as described below. For each θ, let

$$\tilde{l}(\theta) = \log \max\{L(\theta, G): G(D_j|w_j) = 1, \quad j = 1, \ldots, J\}. \tag{4}$$

Then any quadratic form $v^T I_e v$ can be consistently estimated by

$$-2\frac{\tilde{l}(\hat{\theta} + u_n v_n) - \tilde{l}(\hat{\theta})}{nu_n^2}, \tag{5}$$

provided $v_n \xrightarrow{P} v$, $u_n \xrightarrow{P} 0$ and $(\sqrt{n}u_n)^{-1} = O_P(1)$ ([4, Theorem 5.2]).

3. The EM algorithm

We now consider how to compute $\hat{\theta}$ and its standard error. Suppose a sample of size n has been drawn. Call a subject a complete case if the corresponding value of Z is observed, or an incomplete case otherwise. Stratify the sample into J strata according to the value of $W(= w_1, \ldots, w_J)$. Denote by $z_{j1}, \ldots, z_{jK_j}$ the distinct values of Z observed in stratum j, with respective multiplicities $n_{j1}, \ldots, n_{jK_j}$, $j = 1, \ldots, J$. Let n_{j0} denote the number of incomplete cases in stratum j, so that $n_j := \sum_{k=0}^{K_j} n_{jk}$ is the size of stratum j, $j = 1, \ldots, J$. Let (j, k, l), $j = 1, \ldots, J$, $k = 1, \ldots, K_j$, $l = 1, \ldots, n_{jk}$, provide the index (in the original sample) of the lth complete case in stratum j taking the value z_{jk}. Thus we have $R_{(j,k,l)} = 1$, $W_{(j,k,l)} = w_j$, $Z_{(j,k,l)} = z_{jk}$ if $k \geq 1$. Similarly, write $(j, 0, l)$ for the index of the lth incomplete case in stratum j.

With W discrete, G is essentially a vector of distributions $(G_1, \ldots, G_J)$ with $G_j := G(\cdot|w_j)$, $j = 1, \ldots, J$. In computing the restricted SMLE, the likelihood will be maximized with the G_j supported by the observed values of Z in each stratum. Thus each G_j is identified with a probability vector $(g_{j1}, \ldots, g_{jK_j})$, where $g_{jk} := G_j(\{z_{jk}\})$, $k = 1, \ldots, K_j$, $j = 1, \ldots, J$. Under this identification, the likelihood (3) can be rewritten as

$$\prod_{j=1}^{J} \left(\left\{ \prod_{k=1}^{K_j} \left[g_{jk}^{n_{jk}} \prod_{l=1}^{n_{jk}} f(Y_{(j,k,l)}|w_j, z_{jk}; \theta) \right] \right\} \prod_{l=1}^{n_{j0}} \left[\sum_{k=1}^{K_j} g_{jk} f(Y_{(j,0,l)}|w_j, z_{jk}; \theta) \right] \right). \tag{6}$$

Direct maximization of expression (6) with respect to (θ, G) is a constrained maximization problem of dimension $d + \sum_{j=1}^{J} K_j$, with each G_j constrained to a K_j-dimensional unit simplex. Under a suitable transformation of the G_j, this can be transformed into an unconstrained maximization problem of dimension $d - J + \sum_{j=1}^{J} K_j$. Therefore a Newton-type algorithm

is applicable, at least in principle. But note that, as the sample size n increases, the K_j increase at the same rate, unless Z truly has a finite support. Thus in a relatively large sample, Newton's method may be difficult, if not impossible, to carry out.

Expression (6) can be considered as a parametric likelihood under the *working* assumption that each G_j is concentrated on $\{z_{jk}: k = 1, \ldots, K_j\}$. As such it can be maximized by using the EM algorithm [5], which has been used in similar but different contexts [2, 6]. In the present context, the complete data comprise $\{(W_i, Z_i, Y_i): i = 1, \ldots, n\}$ and the complete-data log-likelihood is given by

$$\begin{aligned} l_c(\theta, G) &= \sum_{i=1}^{n} \big[\log f(Y_i|W_i, Z_i; \theta) + \log G(\{Z_i\}|W_i)\big] \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K_j} \Bigg\{ n_{jk} \log g_{jk} + \sum_{l=1}^{n_{jk}} \log f(Y_{(j,k,l)}|w_j, z_{jk}; \theta) \\ &\qquad + \sum_{l=1}^{n_{j0}} I(Z_{(j,0,l)} = z_{jk}) \big[\log g_{jk} + \log f(Y_{(j,0,l)}|w_j, z_{jk}; \theta)\big] \Bigg\}, \end{aligned}$$

where $I(\cdot)$ is the indicator function.

Let $(\theta^{(0)}, G^{(0)})$ be an initial guess. For example, one may take as $\theta^{(0)}$ an estimate obtained from a complete-case analysis, and set $g_{jk}^{(0)} = n_{jk}/n_j$, $j = 1, \ldots, J$, $k = 1, \ldots, K_j$. Given $(\theta^{(m)}, G^{(m)})$, $m \geq 0$, we seek to maximize

$$\begin{aligned} &E[l_c(\theta, G)|(R_i, W_i, R_i Z_i, Y_i)_{i=1}^n; \theta^{(m)}, G^{(m)}] \\ &= \sum_{i=1}^{n} \{R_i[\log f(Y_i|W_i, Z_i; \theta) + \log G(\{Z_i\}|W_i)] \\ &\quad + (1 - R_i) E[\log f(Y_i|W_i, Z_i; \theta) + \log G(\{Z_i\}|W_i)|W_i, Y_i; \theta^{(m)}, G^{(m)}]\} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K_j} \Bigg\{ n_{jk} \log g_{jk} + \sum_{l=1}^{n_{jk}} \log f(Y_{(j,k,l)}|w_j, z_{jk}; \theta) \\ &\qquad + \sum_{l=1}^{n_{j0}} h_{jkl}^{(m)} [\log g_{jk} + \log f(Y_{(j,0,l)}|w_j, z_{jk}; \theta)] \Bigg\} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K_j} \Bigg[(n_{jk} + h_{jk}^{(m)}) \log g_{jk} + \sum_{l=1}^{n_{jk}} \log f(Y_{(j,k,l)}|w_j, z_{jk}; \theta) \\ &\qquad + \sum_{l=1}^{n_{j0}} h_{jkl}^{(m)} \log f(Y_{(j,0,l)}|w_j, z_{jk}; \theta) \Bigg], \end{aligned} \tag{7}$$

where

$$h_{jkl}^{(m)} := \frac{f(Y_{(j,0,l)}|w_j, z_{jk}; \theta^{(m)}) g_{jk}^{(m)}}{\sum_{q=1}^{K_j} f(Y_{(j,0,l)}|w_j, z_{jq}; \theta^{(m)}) g_{jq}^{(m)}} \quad \text{and} \quad h_{jk\cdot}^{(m)} := \sum_{l=1}^{n_{j0}} h_{jkl}^{(m)}.$$

Note that $h_{jkl}^{(m)}$ is the conditional probability, given observed data and under current parameter estimate, that $Z = z_{jk}$ for an incomplete case in stratum j; that is,

$$h_{jkl}^{(m)} = P(Z_{(j,0,l)} = z_{jk}|W_{(j,0,l)}, Y_{(j,0,l)}; \theta^{(m)}, G^{(m)}).$$

It follows that $h^{(m)}_{jk\cdot}$ is the expected number of z_{jk}'s among the incomplete cases in stratum j, conditional on observed data and under current parameter estimate. Because θ and G are separated in equation (7), the maximizer is readily found to be

$$\theta^{(m+1)} = \text{argmax}_\theta \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\sum_{l=1}^{n_{jk}} \log f(Y_{(j,k,l)}|w_j, z_{jk};\theta) + \sum_{l=1}^{n_{j0}} h^{(m)}_{jkl} \log f(Y_{(j,0,l)}|w_j, z_{jk};\theta)\right], \tag{8}$$

$$g^{(m+1)}_{jk} = \frac{(n_{jk} + h^{(m)}_{jk\cdot})}{n_j}, \quad k = 1, \ldots, K_j, \quad j = 1, \ldots, J. \tag{9}$$

Here $g^{(m+1)}_{jk}$ has the interpretation as the expected proportion of z_{jk}'s among all cases in stratum j, conditional on observed data and under current parameter estimates. Iterating equations (8) and (9) until convergence yields the SMLE. In many examples, the maximizer in equation (8) can be found by solving

$$\sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\sum_{l=1}^{n_{jk}} \dot{\ell}(Y_{(j,k,l)}|w_j, z_{jk};\theta) + \sum_{l=1}^{n_{j0}} h^{(m)}_{jkl}\dot{\ell}(Y_{(j,0,l)}|w_j, z_{jk};\theta)\right] = 0$$

for θ, where $\dot{\ell}(y|x;\theta) := \partial \log f(y|x;\theta)/\partial\theta$. The above equation can be solved analytically for the normal linear model. In general, a Newton-type algorithm can be used. This application of Newton's method differs from the one mentioned earlier in that the dimension of the current problem is d, regardless of n or the K_j.

A slightly modified version of this EM algorithm can be used to evaluate the profile likelihood for θ. For each θ, let $\hat{G}(\theta)$ be any maximizer in equation (4), so that $\tilde{l}(\theta) = \log L(\theta, \hat{G}(\theta))$. For a given θ, $\hat{G}(\theta)$ can be found by iterating until convergence a simpler version of equation (9) with $\theta^{(m)}$ in the definition of $h^{(m)}_{jkl}$ replaced by θ. In light of the discussion in Section 2, a consistent estimate of the efficient information I_e is now available. Consider first the diagonal elements $I_e(s,s)$, $s = 1, \ldots, d$. Let e_s be a d-vector with 1 as the sth element and 0 everywhere else. Set $v_n \equiv v = e_s$ and $u_n = an^{-1/2}$ for some constant $a > 0$. Then a consistent estimate of $I_e(s,s)$ is obtained from expression (5) as

$$2a^{-2}\left[\tilde{l}(\hat{\theta}) - \tilde{l}(\hat{\theta} + an^{-1/2}e_s)\right].$$

This can be interpreted as a numerical second-order partial derivative. Naturally the desired derivative can be approximated from the opposite direction as well. In other words, e_s can be replaced by its negative to yield

$$2a^{-2}\left[\tilde{l}(\hat{\theta}) - \tilde{l}(\hat{\theta} - an^{-1/2}e_s)\right].$$

Common wisdom then suggests taking the average of the two and estimating $I_e(s,s)$ by

$$a^{-2}\left[2\tilde{l}(\hat{\theta}) - \tilde{l}(\hat{\theta} + an^{-1/2}e_s) - \tilde{l}(\hat{\theta} - an^{-1/2}e_s)\right].$$

For an off-diagonal element $I_e(s,t)$, $s \neq t$, let $e_{st} = e_s + e_t$. Then a consistent estimate of $e_{st}^{\mathrm{T}} I_e e_{st} = I_e(s,s) + I_e(t,t) + 2I_e(s,t)$ is given by

$$a^{-2}\left[2\tilde{l}(\hat{\theta}) - \tilde{l}(\hat{\theta} + an^{-1/2}e_{st}) - \tilde{l}(\hat{\theta} - an^{-1/2}e_{st})\right].$$

It follows that $I_e(s, t)$ can be consistently estimated by

$$\frac{1}{2a^2}\Big[\tilde{l}(\hat{\theta}+an^{-1/2}e_s)+\tilde{l}(\hat{\theta}-an^{-1/2}e_s)+\tilde{l}(\hat{\theta}+an^{-1/2}e_t)+\tilde{l}(\hat{\theta}-an^{-1/2}e_t) -2\tilde{l}(\hat{\theta})-\tilde{l}(\hat{\theta}+an^{-1/2}e_{st})-\tilde{l}(\hat{\theta}-an^{-1/2}e_{st})\Big].$$

Inverting the estimate of I_e gives a consistent estimate of the asymptotic variance of $\hat{\theta}$.

4. Simulation studies

Simulation experiments are conducted under a normal linear model and a Poisson regression model. In each model, W is assumed empty and Z one-dimensional. Then G is just the marginal distribution of Z and the selection mechanism π is a function of y only. Under the linear model, data are generated according to the following mechanism:

$$Z \sim \text{Beta}\ (\alpha, 1), \tag{10}$$

$$Y|Z = z \sim \text{Normal}\ (\beta_0 + \beta_1 z, \sigma^2), \tag{11}$$

$$\text{logit}[\pi(y)] = y + \gamma, \tag{12}$$

where $\alpha \in \{0.5, 1, 2\}$, $\beta_0 = 0$, $\beta_1 \in \{0, 5\}$, $\sigma^2 = 1$, and γ is chosen such that $E(R) = 0.5$. A sample consists of $n = 100$ or 200 independent copies of (Z, Y, R). For each sample size, 1000 replicates (samples) are generated under each of the six scenarios (combinations of parameter values).

Given a sample, $\theta = (\beta_0, \beta_1, \sigma^2)$ is estimated using the following five methods. FD (full data) is the usual least-squares procedure applied to $\{(Z_i, Y_i): i = 1, \ldots, n\}$, as if they were all observed. This is not a competitor method for missing covariates. Rather, it serves as an indicator for the total amount of information about θ contained in the data generated. CC (complete case) is the least-squares procedure applied to $\{(Z_i, Y_i): R_i = 1\}$, as if they were the original sample. Under an outcome-dependent selection mechanism, this approach is invalid. It is included in this study to illustrate the potential bias and loss of efficiency and also to provide initial parameter values for the iterative procedures. ML0 is the standard maximum likelihood procedure under the parametric model defined by expressions (10) and (11). The relative (in)efficiency of ML0 to FD indicates the amount of information lost due to missing values of Z, with G known up to a finite-dimensional parameter. On the other hand, the Fisher information for θ in this model (or any other correct parametric model) is larger in the sense of non-negative definiteness than the efficient Fisher information for θ in the semiparametric model where G is unspecified. Therefore ML0 is expected to be more efficient than a semiparametric method. Of interest to us is the amount of efficiency gain that comes with a detailed knowledge of the covariate distribution. In practice, it is often difficult to specify a parametric model that is nearly correct. In the present setting, a data analyst without sufficient information about G might simply specify a normal model:

$$Z \sim \text{Normal}\ (\nu, \tau^2), \tag{13}$$

which may be called common practice. Denote by ML1 the maximum likelihood procedure under (11) and (13). We would like to quantify the bias of ML1 due to model mis-specification

and hence the robustness achieved by sparing a parametric specification of G. Lastly, SML is the semiparametric maximum likelihood method defined earlier.

ML0 and ML1 are both implemented using an EM algorithm similar to the Monte Carlo EM algorithm of Ibrahim *et al.* [2]. Preliminary simulation results suggest that, even for ML0 and ML1, the EM algorithm is more stable than a quasi-Newton algorithm where variable scaling can be a serious problem. In the implementation of ML0, the conditional expectation in the E-step has no closed form and is evaluated via numerical integration. In the implementation of ML1, $(\theta^{(m+1)}, \nu^{(m+1)}, \tau^{(m+1)})$ can be found in closed form.

Each method gives for each regression parameter a point estimate, a standard error (standard deviation estimate), and a Wald confidence interval. The only exception here is that, under the least-squares approach, inference about σ^2 is based on a χ^2 distribution and does not involve variance estimation. Empirical bias and standard deviation (SD) of a point estimate are calculated using knowledge of the true parameter value and standard formulas applied to the different replicates. Standard errors (SEs) are averaged across replicates and compared with the empirical standard deviation. Empirical coverage probabilities (CPs) are calculated for (intended) 95% confidence intervals.

Tables 1 and 2 summarize numerical results obtained under different scenarios (described earlier) at $n = 100$, 200. In all scenarios studied here, CC is associated with a large bias.

Table 1. Linear regression with $n = 100$.

Scenario			Bias (×1000)			SD (×1000)			SE (×1000)			CP (×100)		
β_1	α	Method	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
0	0.5	FD	−4	5	−2	149	339	134	150	337		95	94	96
		CC	416	8	−170	197	457	166	196	443		43	94	89
		ML0	24	−52	−31	203	522	141	194	495	141	93	92	93
		ML1	8	7	−38	207	552	133	199	520	141	92	91	93
		SML	7	8	−37	209	555	133	200	521	141	93	91	93
	1	FD	1	5	−3	202	347	144	201	347		94	95	95
		CC	408	7	−172	265	470	173	261	451		64	94	86
		ML0	18	−5	−34	294	547	142	278	518	141	93	93	91
		ML1	9	11	−39	297	560	143	284	531	141	92	92	91
		SML	10	10	−38	298	562	143	285	532	140	92	92	91
	2	FD	−4	8	−4	302	428	140	300	425		94	94	96
		CC	410	7	−171	406	572	168	392	555		80	94	88
		ML0	18	−7	−26	461	665	137	447	652	143	93	93	94
		ML1	12	3	−39	475	695	140	446	651	141	92	93	92
		SML	11	4	−39	473	694	140	445	651	141	92	93	92
5	0.5	FD	1	1	−7	141	340	142	149	338		96	94	95
		CC	644	−683	−132	257	473	172	252	445		29	65	92
		ML0	−4	14	−30	169	370	165	175	380	170	96	95	92
		ML1	178	−353	−100	200	375	182	215	385	187	86	86	87
		SML	−8	13	−24	176	382	172	177	374	170	95	94	93
	1	FD	7	−19	4	203	358	149	201	348		95	95	94
		CC	795	−752	−134	383	551	189	371	536		42	68	89
		ML0	−12	7	−28	260	402	193	260	418	192	95	96	91
		ML1	−70	52	−4	349	484	239	314	464	219	93	94	91
		SML	−47	54	−9	332	482	217	272	431	191	89	92	90
	2	FD	−14	15	5	304	428	149	303	429		95	95	94
		CC	950	−793	−144	584	736	182	576	730		61	79	89
		ML0	−31	44	−20	414	562	219	422	565	212	96	95	90
		ML1	−438	502	89	565	710	259	482	652	242	85	87	95
		SML	−177	197	33	568	713	240	438	598	210	86	89	91

Table 2. Linear regression with $n = 200$.

Scenario			Bias (×1000)			SD (×1000)			SE (×1000)			CP (×100)		
β_1	α	Method	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
0	0.5	FD	2	−3	1	108	234	101	106	239		95	95	94
		CC	419	−13	−168	136	309	123	137	309		13	95	76
		ML0	3	−5	−14	140	355	98	137	350	100	94	94	93
		ML1	11	−15	−17	142	371	101	141	368	100	94	94	93
		SML	11	−15	−17	143	372	101	142	368	100	94	94	93
	1	FD	−2	−3	2	141	249	101	142	246		96	94	95
		CC	416	−5	−167	190	331	116	184	319		38	94	79
		ML0	22	−25	−21	211	396	99	197	367	99	92	92	93
		ML1	6	−7	−16	212	401	99	202	379	100	93	93	94
		SML	6	−7	−16	212	401	99	202	379	100	93	93	94
	2	FD	−5	9	3	222	316	101	213	301		94	93	95
		CC	416	−1	−170	277	388	122	276	390		66	95	76
		ML0	20	−20	−25	328	483	100	310	454	99	93	92	93
		ML1	7	−3	−14	326	474	100	320	467	100	94	94	94
		SML	7	−3	−14	325	474	100	319	467	100	94	94	94
5	0.5	FD	−3	6	−6	106	235	102	106	237		94	95	94
		CC	631	−657	−127	170	299	123	178	312		7	42	87
		ML0	−4	5	−10	119	258	126	129	284	124	97	97	94
		ML1	171	−339	−81	138	253	132	153	272	134	81	77	86
		SML	−8	22	−20	123	258	121	125	261	120	95	95	93
	1	FD	−4	6	0	146	249	102	142	246		95	94	95
		CC	777	−718	−130	269	396	123	262	379		17	52	84
		ML0	−12	13	−12	185	289	139	193	311	139	96	97	93
		ML1	−81	72	14	239	341	158	223	329	157	93	95	94
		SML	−32	44	−7	215	328	142	197	308	138	94	94	93
	2	FD	−1	−3	2	212	296	101	213	302		95	95	94
		CC	948	−789	−146	398	503	127	399	506		34	65	81
		ML0	−2	5	−17	263	360	154	322	433	151	97	98	93
		ML1	−403	463	92	383	483	178	331	448	170	79	82	95
		SML	−82	88	16	384	487	160	312	418	149	89	92	93

In the presence of a strong regression relationship ($\beta_1 = 5$), it also tends to have a large standard deviation. In contrast, all three methods (ML0, ML1, SML) that explicitly adjust for missing data generally perform better, at least in terms of bias. We now turn to the comparison of ML0, ML1, and SML, with ML0 being an ideal that cannot be achieved (without a good knowledge of G). It appears that, under weak regression ($\beta_1 = 0$), the three methods are nearly equivalent in terms of the few criteria considered here. In that case, it does not seem to matter how to deal with the covariate distribution – parametrically or nonparametrically, correctly or incorrectly – as long as we do deal with it. In the case of strong regression ($\beta_1 = 5$), however, the ML1 estimates can be seen to carry a significant bias. In fact, strong regression also has the effect of setting a higher sample size requirement for the asymptotic properties of SML to take effect. Indeed, for each fixed n, one can make ML1 and SML perform arbitrarily poorly by choosing large values of β_1. Note, for example, the biases of ML1 and SML in the scenario where $\beta_1 = 5$ and $\alpha = 2$, at a sample size of $n = 100$. On the other hand, in each fixed scenario, the bias of SML eventually vanishes with increasing n, whereas that of ML1 does not. In the same scenario as noted above, but at $n = 200$, ML1 remains severely biased, whereas SML becomes much less so. The (in)efficiency of SML relative to ML0 quantifies the statistical buying power of an accurate knowledge of G in the presence of

missing values of Z. β_1 is again an important factor in this assessment: the larger it is, the less efficient SML is relative to ML0. In most cases, the SML standard errors estimate the true standard deviations reasonably well and the associated confidence intervals enjoy good coverage probabilities.

The simulation experiments for Poisson regression are conducted in a similar fashion and yield similar results. Data are generated according to equations (10) and (12) and, of course, a Poisson regression model:

$$Y|Z = z \sim \text{Poisson } (\exp(\beta_0 + \beta_1 z)), \tag{14}$$

where $\beta_0 = 0$ and $\beta_1 \in \{0, 2\}$. Again, 1000 replicates are generated in each scenario at each sample size. Here FD and CC refer to the standard maximum likelihood procedure applied to $\{(Z_i, Y_i)\colon i = 1, \ldots, n\}$ and $\{(Z_i, Y_i)\colon R_i = 1\}$, respectively. ML0 is the maximum likelihood method under expressions (10) and (14). ML1 is the maximum likelihood method under expressions (13) and (14). Both ML0 and ML1 are computed using an EM algorithm with numerical integration in the E-step. Numerical results are reported in tables 3 and 4. All the qualitative remarks in the preceding paragraph remain valid here.

Table 3. Poisson regression with $n = 100$.

Scenario			Bias (×1000)		SD (×1000)		SE (×1000)		CP (×100)	
β_1	α	Method	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
0	0.5	FD	−2	−32	154	340	152	343	94	95
		CC	331	−27	175	399	184	417	53	97
		ML0	7	−57	186	464	178	439	93	94
		ML1	3	−36	192	485	179	447	94	93
		SML	2	−33	190	481	182	455	93	93
	1	FD	−12	3	210	376	202	351	95	95
		CC	328	2	233	407	245	425	70	97
		ML0	−2	−14	265	487	250	456	94	94
		ML1	−9	3	271	499	251	459	95	95
		SML	−9	3	270	496	255	466	94	94
	2	FD	−14	10	309	438	304	430	95	95
		CC	328	2	337	475	370	523	85	98
		ML0	7	−18	396	570	393	567	95	95
		ML1	−9	6	400	578	387	560	95	95
		SML	−7	4	398	575	396	573	95	95
2	0.5	FD	−2	−4	126	208	128	208	95	96
		CC	492	−512	147	225	170	254	16	46
		ML0	−3	−9	135	226	141	232	96	97
		ML1	−6	−15	144	230	161	259	97	98
		SML	−3	−7	136	225	138	225	95	96
	1	FD	0	−3	148	215	151	215	95	95
		CC	607	−594	184	251	218	288	17	45
		ML0	−8	3	168	240	175	247	96	96
		ML1	−75	88	193	266	221	310	96	95
		SML	−10	8	173	245	174	245	95	95
	2	FD	−4	1	200	254	196	249	95	95
		CC	767	−715	262	321	306	369	25	48
		ML0	−7	−2	240	307	237	301	95	95
		ML1	−195	223	308	381	318	407	69	69
		SML	−42	42	279	348	245	311	92	93

Table 4. Poisson regression with $n = 200$.

Scenario			Bias (×1000)		SD (×1000)		SE (×1000)		CP (×100)	
β_1	α	Method	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
0	0.5	FD	−3	−4	107	230	107	239	95	96
		CC	338	−11	114	255	127	287	23	98
		ML0	5	−25	124	300	124	307	94	96
		ML1	2	−13	127	313	126	312	95	96
		SML	1	−12	127	311	127	317	95	95
	1	FD	−6	−2	147	254	142	247	95	94
		CC	334	−5	167	284	172	298	48	97
		ML0	4	−20	188	341	176	322	93	94
		ML1	−2	−7	190	346	178	326	94	95
		SML	−2	−7	190	344	179	328	93	94
	2	FD	−9	8	214	297	214	302	95	96
		CC	329	10	229	321	257	364	76	98
		ML0	−3	1	277	396	276	399	95	95
		ML1	−13	15	279	400	272	394	95	95
		SML	−11	14	278	398	277	402	95	94
2	0.5	FD	−11	13	90	142	90	147	96	96
		CC	480	−489	101	151	119	179	2	17
		ML0	−12	13	96	153	102	168	97	97
		ML1	−14	4	102	156	126	203	99	99
		SML	−12	15	97	154	97	158	95	96
	1	FD	−2	3	107	151	107	151	95	95
		CC	608	−594	130	179	153	202	2	14
		ML0	−3	2	121	170	126	178	96	96
		ML1	−68	83	139	189	187	263	95	95
		SML	−5	6	126	175	122	172	95	95
	2	FD	−1	−2	141	178	138	175	95	94
		CC	769	−719	181	221	213	257	5	18
		ML0	−5	−1	167	209	171	217	97	96
		ML1	−181	211	210	257	162	206	33	32
		SML	−17	15	183	227	171	217	95	94

5. Application

The proposed method is applied to data from the Breast Cancer Prevention Trial at the National Surgical Adjuvant and Bowel Project. The aim of this trial is to evaluate the effect of Tamoxifen for preventing breast cancer. As each patient enters the trial, a variety of measurements are made, collectively known as baseline information. This includes body weight and alanine aminotransferase (ALT) level, an index of liver functioning.

At one point, investigators are interested in relating body weight to ALT level, for which a simple linear regression model is deemed plausible. Specifically, let Z denote ALT level in units per liter, and Y body weight in pounds. It is postulated that

$$Y|Z = z \sim N(\beta_0 + \beta_1 z, \sigma^2),$$

and the scientific focus is on β_1. Available for this analysis are records of the first 1000 patients in the trial. Unfortunately, although Y is recorded for every subject, Z is measured only for some 50% of the subjects. This missingness is attributed to financial and other nonbiological factors. In statistical terms, there are reasons to believe that Z is missing completely at random. Hence the CC analysis is valid. Nevertheless, it is certainly desirable to take into account the information in the incomplete cases and obtain a more accurate estimate. One way to do this

is to specify a normal model for Z and maximize the parametric likelihood; this is denoted by ML1 in Section 4.2. Another possibility is the proposed SML method.

All three methods are used to analyze this data. The point estimates of β_1 are 1.01 (CC), 0.97 (ML1), and 0.96 (SML), with respective standard errors 0.24, 0.18, and 0.22. There is a visible difference in the point estimate with and without using the incomplete cases. As expected, SML has a smaller standard error than CC. The standard error for ML1 is even smaller. Without validating the normal model, inference based on the small standard error of ML1 could be overoptimistic.

6. Discussion

An EM algorithm proposed in this paper implements an SML method for parametric regression problems where some covariates are missing at random. This method requires no parametric specification of the selection mechanism or the covariate distribution, and the EM algorithm yields reasonable numerical results in moderate-sized samples. Fortran programs are available from the first author.

Simulation experiments are carried out to compare the proposed method (SML) with the CC analysis and another maximum likelihood method (ML) based on a parametric model for the covariate distribution. It is clear that the CC analysis tends to be more biased than both ML and SML, even if the covariate distribution is mis-specified in ML. This is further evidence that missing data should be dealt with explicitly rather than ignored. SML is less efficient than ML when the covariate distribution is correctly specified. However, if the covariate distribution is mis-specified, then ML can be severely biased. We therefore recommend the SML method when the covariate distribution is difficult to model. The strength of the regression relationship has an effect on the performance of both ML and SML. Under weak regression, both methods perform well, whether the covariate model in ML is correct or not. Under strong regression, ML is sensitive to mis-specification of the covariate distribution and SML requires a large sample. Thus if a strong regression relationship is expected, an ML user should be very careful about the covariate model and an SML user may wish to collect a large sample. The meaning of "large" here is admittedly vague, and in practice we recommend further simulation studies targeted at the application at hand.

The proposed method applies when a portion (Z) of X is missing as a whole and the observed portion (W) is finitely discrete. If W is not discrete or, more generally, if multiple patterns of missing covariates can occur, then it seems difficult to treat the covariate distribution completely nonparametrically within the ML framework. The method of Ibrahim *et al.* [2] is available in the more general setting and is subject to the usual mis-specification bias. We are currently exploring a different approach based on nonparametric regression ideas.

References

[1] Rubin, D.B., 1976, Inference and missing data. *Biometrika*, **63**, 581–592.
[2] Ibrahim, J.G., Chen, M.H. and Lipsitz, S.R., 1999, Monte Carlod EM for missing covariates in parametric regression models. *Biometrics*, **55**, 591–596.
[3] Zhang, Z. and Rockette, H.E., 2005, On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, **134**, 206–223.
[4] Zhang, Z. and Rockette, H.E., 2004, Semiparametric maximum likelihood for missing covariates in parametric regression. *Annals of the Institute of Statistical Mathematics*, in press.
[5] 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**, 1–38.
[6] Lipsitz, S.R. and Ibrahim, J.G., 1996, A conditional model for incomplete covariates in parametric regression models. *Biometrika*, **83**, 916–922.