



Published in final edited form as:

Stat Med. 2011 September 10; 30(20): 2551–2561. doi:10.1002/sim.4280.

Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits

Ryan C. May^{*,†}, Joseph G. Ibrahim, and Haitao Chu

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

The analysis of data subject to detection limits is becoming increasingly necessary in many environmental and laboratory studies. Covariates subject to detection limits are often left censored because of a measurement device having a minimal lower limit of detection. In this paper, we propose a Monte Carlo version of the expectation–maximization algorithm to handle large number of covariates subject to detection limits in generalized linear models. We model the covariate distribution via a sequence of one-dimensional conditional distributions, and sample the covariate values using an adaptive rejection metropolis algorithm. Parameter estimation is obtained by maximization via the Monte Carlo M-step. This procedure is applied to a real dataset from the National Health and Nutrition Examination Survey, in which values of urinary heavy metals are subject to a limit of detection. Through simulation studies, we show that the proposed approach can lead to a significant reduction in variance for parameter estimates in these models, improving the power of such studies.

Keywords

EM algorithm; Gibbs sampling; logistic regression; maximum likelihood estimation; Monte Carlo EM; NHANES

1. Introduction

Data subject to detection limits are common occurrences in many environmental and laboratory studies. In such studies, outcome or covariate measures are often right or left censored because of the measuring device having a maximal upper limit of detection or minimal lower limit of detection. Although the proposed methodology in this paper can be applied to both right-censored and left-censored covariate data, the real and simulated examples presented here consider only left-censored data, as is most common in real-life studies with detection limits. To motivate these methods, we consider a study in cancer incidence conducted within the National Health and Nutrition Examination Survey (NHANES)[1]. As part of this study, levels of urinary heavy metals were recorded, along


with the presence of any form of cancer. Recorded urinary heavy metals included cadmium, uranium, tungsten, and dimethylarsonic acid. The measurement device used to examine levels of each urinary heavy metal can only be calibrated down to a specific limit of detection (i.e., only above 1.7 ug/L for dimethylarsonic acid). As a result, 24.1% of the 1350 patients had at least one covariate value that fell below the limit of detection for the measurement device. Study subjects were also surveyed as to past cancer status, the response variable for this study.

Past research on data subject to detection limits has considered models where either the response or covariates alone are subject to detection limits. The simplest and most straightforward method for dealing with such data is to remove or delete all observations falling below the limit of detection. This is known as complete-case analysis. Complete-case analysis is generally discouraged because of the loss of useful information in the data. Though complete-case analysis can give unbiased parameter estimates in regression models [2–4], the standard errors of those estimates will be inflated because of the decreased sample size. This deficiency is particularly significant for studies where a large proportion of data falls below the limit of detection. Additionally, background parameter estimates for the covariate distribution of interest will be biased [5]. Another very common approach is to use ad hoc substitution methods. These often include substituting some fraction of the limit of detection for all observations falling below the limit of detection, such as the limit of detection itself (LOD), LOD/2, LOD/3, or zero. Such methods are commonly employed because they are simple both to understand and implement. However, numerous authors have concluded that such methods are statistically inappropriate for data with censored covariates [6] or censored responses [7, 8]. Helsel [5] provided a review of several of these substitution procedures, concluding that the substitution method leads to highly biased estimates and has no theoretical basis. Singh and Nocerino [9] analyzed the substitution method on censored response values in environmental studies, concluding that highly biased estimates result even in cases with a small percent of censored values and only a single detection limit. The bias increases as more detection limits are introduced. For regression with a censored outcome, Thompson and Nelson [10] found that substitution of half the detection limit led to biased parameter estimates and artificially small standard error estimates. These results have provided strong evidence against using ad hoc substitution techniques.

In a linear regression setting, further substitution methods have been proposed for cases when a single covariate is subject to a limit of detection. Richardson and Ciampi [11] proposed substituting the conditional expected value of each censored covariate, given all observed covariates. This method relies on a specification of the underlying covariate distribution, which often is not known with certainty. When the covariate distribution is unknown, Schisterman [12] proposed substituting the average of all *observed* covariates in the model, which was shown to achieve unbiased results. Another common method is maximum likelihood (ML) estimation, which also requires knowledge of the underlying covariate distribution. These methods were compared with the previously discussed ad hoc substitution methods in Nie *et al.* [4] when only one covariate is subject to a limit of detection. It concluded that maximum likelihood performed best when the covariate

distribution is known, as ML estimation is unbiased and results in small standard errors. These results were echoed by Lynn [6], who compared substitution methods with multiple imputation and maximum likelihood estimation. Both papers noted that maximum likelihood estimation should not be attempted when the underlying covariate distribution is not known. In this case, Nie *et al.* [4] suggested using complete-case analysis.

The preference for maximum likelihood approaches has also been seen in studies using logistic regression with a single covariate subject to a limit of detection. Cole *et al.* [13] compared ad hoc substitution methods with complete-case analysis and maximum likelihood estimation, concluding that maximum likelihood resulted in relatively unbiased estimates with smaller standard errors than either complete case or substitution methods, especially when the proportion of censored values was large (50% or more).

Methods have also been proposed for Cox regression models with up to two covariates subject to a lower limit of detection. D'Angelo and Weissfeld [3] presented an index-based expectation–maximization (EM) algorithm-type method for this problem. The E-step for this method involves substituting the conditional expectation of each censored covariate, whereas the M-step uses standard Cox regression. It found that the index-based approach provided improvements over complete-case analysis in terms of variance estimates, but a small bias existed in the index approach compared with the unbiased complete-case analysis. The approach was not shown to provide much improvement over the biased LOD/2 and LOD/ substitution approaches, however.

When the response variable is subject to a limit of detection, two common methods of estimation include Tobit regression [14] and multiple imputation. Generally, Tobit regression is used when interest resides primarily on the regression parameters. When interest is on estimating a 'complete' dataset, however, multiple imputation is often used to impute the missing values. Lubin *et al.* [8] developed a multiple imputation approach based on bootstrapping and compared the results with substitution methods and Tobit regression. It found that both the proposed multiple imputation approach and Tobit regression have reduced biases with respect to other ad hoc substitution methods.

All the methods previously mentioned here concern models with either a censored response and fully-observed covariates, or a fully-observed response and *at most*, two censored covariates. To the authors' knowledge, no general likelihood-based approach has been developed to account for a large number of left-censored covariates in a generalized linear model (GLM). In this paper, we investigate maximum likelihood methods for fitting models with covariates subject to a limit of detection. We show that this maximum likelihood estimation can be carried out directly via an EM algorithm called the *EM by the method of weights* [15]. For covariates subject to a limit of detection, we specify the covariate distribution via a sequence of one-dimensional conditional distributions. We discuss the missing data mechanism for censored data and explain how the notion of missingness differs from that of regular missing data problems.

In this article, we propose a method for estimating parameters in GLMs with censored covariates and an effectively ignorable missing data mechanism. We consider the case of

continuous covariates only in this paper because censored categorical covariates are unlikely to occur in real-world applications. Following Lipsitz and Ibrahim [16], the joint covariate distribution is modeled via a sequence of one-dimensional conditional distributions. Modeling the joint covariate distribution in this fashion facilitates a more straightforward specification of the distribution. The response variable is assumed to be completely observed, though our method can be easily extended to the case where the response is subject to a limit of detection. We derive the E and M steps of the EM algorithm with effectively ignorable missing covariate data. For continuous covariates, we use a Monte Carlo version of the EM algorithm to obtain the maximum likelihood estimates via the Gibbs sampler. We derive the E-step for the Monte Carlo version of EM. In addition, we show that the relevant conditional distributions needed for the E-step are log-concave, so that the Gibbs sampler is straightforward to implement when the covariates are continuous. This paper is an extension of the methods proposed for missing data in Ibrahim, Lipsitz, and Chen [17]. The proposed methods are computationally feasible and can be implemented in a straightforward fashion.

The rest of this article is organized as follows. In Section 2, we have given some general notation for GLMs. In Section 3, we discuss the proposed methods of estimation and give a detailed discussion of the various models used. In Section 4, we demonstrate the methodology with a simulation study involving a linear regression model. We also demonstrate the methodology with an example involving real data in Section 5. We conclude the article with a discussion section.

2. Notation for generalized linear models

In this paper, we will take $(x_1, y_1), \dots, (x_n, y_n)$ as a set of n independent observations, with y_i representing the response variable and x_i representing a $p \times 1$ vector of covariates. The joint distribution of (y_i, x_i) is written as a sequence of one-dimensional conditional distributions $[y_i|x_i]$ and $[x_i]$, representing the conditional distribution of y_i given x_i and the marginal distribution of x_i . The notation $p(y_i|x_i)$ is used throughout the paper to denote the conditional density of y_i given x_i .

The conditional distribution $[y_i|x_i]$ is specified by a $k \times 1$ parameter vector θ , with the conditional density being represented as $p(y_i|x_i, \theta)$. For the class of GLMs, the parameter vector θ is usually specified as $\theta = (\beta, \tau)$, with β representing the regression model coefficients and τ representing the dispersion parameter. The logistic, Poisson, and exponential models have a τ value exactly equal to one; in these cases, β and θ are equal. For nonlinear models with a normal errors, we write the parameter vector as $\theta = (\theta^*, \sigma^2)$, with θ^* representing the expectation parameters and σ^2 representing the variance of the errors.

The marginal density for x_i is taken as $p(x_i|\alpha)$, with α representing the parameters for the marginal distribution of x_i . The joint density for (y_i, x_i) can then be represented by the following sequence of conditional densities for subject i :

$$p(y_i, x_i) = p(y_i|x_i, \theta) p(x_i|\alpha) \tag{1}$$

Combining this formula for all subjects leads to the complete-data log-likelihood:

$$\ell(\theta, \alpha) = \sum_{i=1}^n \ell(x_i, y_i | \gamma) \quad (2)$$

Here, $\ell(x_i, y_i | \gamma)$ represents the log-likelihood contribution for subject i , and $\gamma = (\theta, \alpha)$. In this paper, our primary interest is in estimating θ ; here, α is considered a nuisance parameter.

Extending this notation to censored covariate data, we write $x_i = (x_{\text{cens},i}, x_{\text{obs},i})$, where $x_{\text{obs},i}$ are the fully observed covariates, and $x_{\text{cens},i}$ is a $q_i \times 1$ vector of censored covariates. For

individual censored covariate values, we use the notation c_{ij} . We allow a different censoring interval for each covariate and subject, taking (c_{lj}, c_{uj}) as the censoring interval for subject i and covariate j . We note here that the censoring intervals are considered to be fully known here. In some applications, limits of detection are not known explicitly and must be estimated. We also note that in most cases, the censoring intervals will not vary across subjects; this is included for generality. This notation is easily generalized to right or left censoring. For left-censored covariates, take $c_{lj} = 0$ (or $c_{lj} = -\infty$ if negative values are possible, i.e., when a log transformation is utilized). For right-censored covariates, take $c_{uj} = \infty$. We use the shorthand notation $(c_l < x_{\text{cens},i} < c_u)$ to denote that each element of $x_{\text{cens},i}$ takes a value within its respective censoring interval. That is,

$$c_l < x_{\text{cens},i} < c_u$$

3. Covariate data subject to a limit of detection

We now propose maximum likelihood methods for covariate data subject to a limit of detection. We will allow left, right, or interval censoring on each covariate, and for ease of exposition, will assume that $\tau = 1$. For clarity, we develop the methodology here for the class of GLMs.

Suppose y_1, \dots, y_n are independent and

$$y_i | x_i \sim \text{GLM}(\eta_i, \phi_i)$$

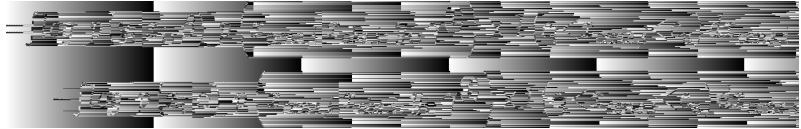
for $i = 1, \dots, n$. In general, the EM algorithm maximizes the expected value of the complete data log-likelihood of (y_i, x_i) , given the observed data, that is,




(3)

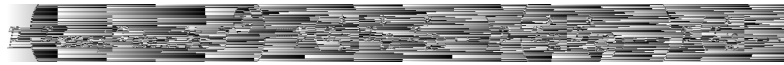
Unlike the usual missing covariate problem in which the `observed data' for subject i is $(y_i, x_{obs,i})$, in the censored covariate problem, the `observed data' are $(y_i, x_{cens,i})$ and $(c_l < x_{cens,i} < c_u)$. In the usual missing covariate problem with $x_{mis,i}$ completely missing, the `weights' in the EM by the method of weights are the conditional probabilities $p(x_{mis,i} | x_{obs,i}, y_i, \gamma)$. Now, with the additional information that $(c_l < x_{cens,i} < c_u)$ in the censored covariate problem, the weights are the conditional probabilities $p[x_{cens,i} | x_{obs,i}, (c_l < x_{cens,i} < c_u), y_i, \gamma]$.

If the censored covariates are all continuous (the most common case), then the E-step of the EM algorithm consists of an integral, which typically does not have a closed form for GLMs. We can write the E-step for the i^{th} observation as

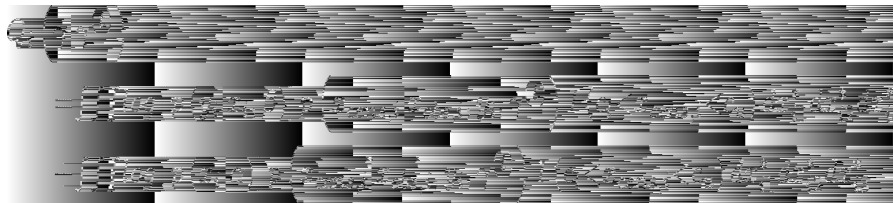


(4)

We note here that in the previous equation, $x_{cens,i}$ is a vector consisting of all covariates in observation i that fall within their respective censoring intervals. In cases where $x_{cens,i}$ contains more than a single censored covariate, Equation (4) consists of multiple integrations, one over each censored covariate, integrating over the range of the censoring interval. For example, with three censored covariates , we have the following:



and



From this, it should be clear that closed-form solutions to Equation (4), even if available (i.e., for a small number of censored covariates), are complicated, and the maximization can be very difficult. We now propose a general approach to evaluating Equation (4), regardless of the number of censored covariates.

To evaluate Equation (4) at the $(t + 1)^{st}$ iteration of EM, we use a Monte Carlo version of the EM algorithm [18]. To do this, we first need to generate a sample from the truncated distribution $[x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}]I(c_l < x_{cens,i} < c_u)$. This truncated distribution is log-concave in each component of $x_{cens,i}$ for most link functions. Thus, we can use the Gibbs sampler along with the adaptive rejection metropolis algorithm (ARMS) of Gilks, Best, and Tan [19] to successively sample from the truncated distribution $[x_{cens,ij}|x_{cens,ik}, k = j, x_{obs,i}, y_i, \gamma^{(t)}]I(c_l < x_{cens,i} < c_u)$, where $x_{cens,ij}$ denotes the j^{th} component of $x_{cens,i}$.

The ARMS algorithm is an extension of the adaptive rejection sampling algorithm of Gilks and Wild [20] and can sample values from complex likelihood functions, which are not required to be log-concave. ARMS works by constructing an envelope function around the desired log density. It performs rejection sampling on the envelope function, shrinking the envelope around the desired log density with each successive sample. For log densities that are not concave, the ARMS algorithm performs an additional metropolis step on each potential sampled value [21]. The shrinking envelope function provides an efficient means of sampling from a complicated log density, without having to evaluate each point of the density directly. ARMS also allows for straightforward sampling from truncated distributions, as all potential points falling outside the censoring interval are immediately rejected.

Use of the EM algorithm requires complete sampled data for each of the n observations in the dataset. For observation i , a new sample must be obtained for each of the q_i censored covariate within $x_{cens,i}$. This is done by successively sampling from the distribution of $x_{cens,ij}, j = 1, \dots, q_i$ until a new sample vector z_i is obtained for the censored vector $x_{cens,i}$. The sampled vector z_i contains q_i sampled values, one for each of censored covariates in $x_{cens,i}$. Now, suppose for the i^{th} observation, we take a sample of size $m_i, z_{i1}, \dots, z_{im_i}$ from the truncated distribution $[x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}]I(c_l < x_{cens,i} < c_u)$ via the Gibbs sampler in conjunction with the adaptive rejection algorithm. We note here that each z_{ik} is a $q_i \times 1$ vector for each $k = 1, \dots, m_i$ with q_i representing the length of $x_{cens,i}$. The E-step for the i^{th} observation at the $(t + 1)^{st}$ iteration for the GLM can be written as



(5)


We notice that this E-step is the EM by the method of weights with each $x_{cens,i}$ being filled in by a set of m_i values each contributing a weight $1/m_i$. The M-step then maximizes Equation (3), which can be expressed as




The maximization can be performed first by taking




as the $q \times 1$ gradient vector of $Q(\gamma|\gamma^{(l)})$. This can be calculated by taking

 (6)

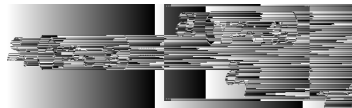
Using this procedure, the EM algorithm can then be run until convergence. In practical application, the maximization of the weighted log-likelihood (with respect to the model parameters) can often be performed by standard software.


Also here, we let  denote the $q \times q$ matrix of the second derivatives of $Q(\gamma|\gamma^{(l)})$.

Let  denote the estimate of γ at convergence. The asymptotic covariance matrix can then be calculated by the method of Louis [22]. The estimated observed information matrix of γ based on the observed data is taken as



where



The estimate of the asymptotic covariance matrix is then calculated as .

We note here that the E-step for censored data is different from the standard missing data notation. Specifically, the censored data E-step in Equation (4) omits the $\int \log[p(r_j|y_j, x_j, \phi)] \dots dx_{\text{cens},j}$ section used in missing data problems, where r_j represents an indicator for missingness. This is because the notion of ignorability is fundamentally different in detection limit problems when compared with missing data problems. In detection limit problems, it is generally assumed that the detection limits are known values. With detection limits known, the probability of censoring ('missingness' in the missing data case) clearly depends on the true value of the covariate (x_j), suggesting a non-ignorable mechanism. However, in the detection limits case, the true value of x_j explicitly determines whether or not the value is censored. The value of $p(r_j|x_j)$ is either 0 or 1, for all values of x_j . It follows

that the non-ignorable component of the E-step equation for missing data is omitted in the detection limit case.


It should be noted that having a continuous outcome variable also subject to a limit of detection only marginally complicates the situation at hand. In this case, the E-step requires an additional integration over the possible values of the censored outcome. Equation (4) then becomes

This situation is further simplified when sampling from the distribution of an outcome value is below the detection limit, however, because we are dealing with the class of GLMs. The distribution of the outcome given the covariates and parameters is assumed to come from an exponential family. Therefore, the distribution of an outcome value below the limit of detection is just a truncated form of a well-known distribution, be it normal, gamma, and others. Such sampling is straightforward.

In this proposal, we will investigate maximum likelihood estimation with censored covariates as outlined earlier. We will study the EM algorithm for this problem and consider GLMs with covariates subject to a detection limit. Examples analyzed include both linear and logistic regression.

4. Simulation study

Here, we consider a simple linear model involving six covariates:


where . The response y_i is fully observed, as are the first three covariates $x_1; \dots, x_3$. The last three covariates $x_4; \dots, x_6$ are subject to a prespecified detection limit. Detection limits are specified according to a desired overall censoring percentage. In this case, detection limits were chosen such that 30% and 50% of observations had at least one covariate that fell below the limit of detection. The covariate distribution was specified as multivariate normal, with arbitrary prespecified parameter values and correlated observations with $0.3 \leq |\rho| \leq 0.7$ for all covariate pairs. Using this specification, datasets of size 200 were then generated; covariate values for $x_4 - x_6$ falling below the detection limit were set as missing.

Each simulation presented in this paper was performed on 1000 datasets created as described earlier, each from identical background parameter distributions and detection limits. The EM by method of weights was then applied to each dataset. Initial parameter estimates for the model and covariate distribution were taken from a complete-case analysis of the data. These were passed to the ARMS algorithm as parameters in the initial iteration of the EM algorithm. For each observation with at least one covariate falling below the limit of

detection, ARMS was used to generate $m_j = 250$ samples of complete covariate data. For observations with a single covariate falling below the limit of detection, these samples were taken from the distribution of $x_{\text{cens};i} | x_{\text{obs};i}, y_i, \gamma^{(t)}$ truncated over the censoring interval. For each observation with multiple covariates falling below the limit of detection, ARMS was used sequentially to sample from the distribution for each censored covariate until a new complete sample of covariate values was produced. The $m_j = 250$ samples from each censored observation were then combined, creating an augmented dataset of fully observed observations along with sampled values. The M-step of the EM algorithm was then performed via a weighted maximum likelihood estimation. Weights of one were used for each fully observed observation, and $1/250$ was used for each sampled observation. This weighted maximum likelihood procedure produced new estimates for β in the model, along with updated parameter estimates for the covariate distribution. The updated covariate parameter estimates were then passed back to ARMS as the estimates for the following E-step, and the procedure was run iteratively until convergence.

Convergence of this algorithm was checked by calculating the average β estimate for the previous 10 iterations. This average was compared with the β average for the 10 iterations prior. In other words, at iteration t the mean beta values from $t:(t-9)$ are compared with values from $(t-10):(t-19)$. A difference of 10^{-3} was used for convergence. After convergence was reached for all parameters, final β estimates were taken as the average of the previous 10 estimates of β in the chain.

Bootstrap standard errors were calculated for each parameter in the dataset for comparison with the standard error of the estimates obtained. For each of the 1000 datasets in a simulation, 25 bootstrapped datasets of size $n = 200$ were generated. The proposed EM algorithm was then run on each bootstrapped dataset, and final β estimates were obtained. The standard error for each population of 25 β estimates was then calculated for each parameter in the model. The mean of these standard errors were then taken as the final bootstrap standard error estimate for the model and are used for comparison with the normal β standard error from the proposed maximum likelihood approach.




Table I displays results from analysis on all 1000 datasets. Final estimates and variances for each parameter are calculated as the mean and variance of final beta estimates for all 1000 datasets. The true prespecified parameter values are given, along with variance estimates calculated using the bootstrap procedure described earlier. Results are also presented for an ad hoc substitution of  for each covariate falling below the limit of detection, along with a complete-case analysis. As expected, both the maximum likelihood approach and complete-case analysis appear to be largely unbiased, whereas the substitution approach produced very biased estimates. Maximum likelihood resulted in standard errors for the parameter estimates that were lower than those obtained with complete case analysis and similar standard errors to the substitution approach. In addition, all calculated standard error estimates for maximum likelihood are close to the asymptotic bootstrapped estimates. The reduction in standard error seen with the maximum likelihood approach was large enough to result in a change in statistical significance (here, taken at the $\alpha = 0.05$ level) for several parameters in the model when compared with the complete-case analysis. These conclusions


hold for both 30% and 50% censored observations, suggesting that the benefit seen is similar up to at least 50% censoring. The EM algorithm was also observed to converge rather quickly using the described criterion. Only 28 EM iterations were needed on average for all model parameters to converge.

5. National Health and Nutrition Examination Survey data

Here, we consider data from the National Health and Nutrition Examination Survey (NHANES)[1] concerning the effect of urinary heavy metal levels on cancer status. The survey years considered here are 2005–2006. The outcome variable in this study is cancer status, a binary variable recorded via questionnaire to the question ‘Have you ever been told by a doctor or other health professional that you had cancer or malignancy of any kind?’. Urinary heavy metals were recorded via physical examination. The measurement device for each urinary heavy metal in the study can only be calibrated down to a specific limit of detection (LOD), leading to many left-censored observations. The degree of censoring varied by each covariate. The urinary heavy metals analyzed in this study include dimethylarsonic acid (13.7% below LOD), cadmium (5.3% below LOD), tungsten (10.7% below LOD), and uranium (9.6% below LOD). In total, 24.1% of the 1350 patients in the study had at least one urinary heavy metal value that fell below a limit of detection. A logistic regression model was chosen for analysis to predict the binary outcome measure of cancer status. Other covariates included in the model are gender, race (dichotomized to White/non-White), physical activity (dichotomized survey response for any physical activity during an average day), and current nicotine use (yes/no). A log transformation was performed on each of the urinary heavy metals variables prior to modeling, and a multivariate normal prior distribution was assumed for these continuous covariates. An independent Bernoulli prior was assumed for the binary covariates gender, race, and smoking status.

Initial parameter estimates for the model were taken from a complete-case analysis. Every observation with a urinary heavy metal covariate value falling below the LOD was then sampled $m_i = 250$ times using the ARMS algorithm. For observations with multiple covariate values below the LOD, each missing covariate value was consecutively sampled until a complete sampled observation was obtained. In such cases, 250 complete sampled observations were recorded. A weighted logistic regression model was then fit to the data, and ML estimates and standard errors were obtained. Parameter estimates for the prior distributions were updated, and the procedure was run iteratively until convergence of the logistic model parameter estimates. The convergence criterion used here was identical to the procedure detailed in Section 4. Upon convergence, final β estimates and standard errors were taken as the average estimates of the previous 10 iterations.

Table II summarizes the results of this study again comparing the maximum likelihood approach with both a complete-case analysis and ad hoc substitution of . The substitution of  is particularly relevant in this case, as urinary heavy metals falling below the limit of detection are actually reported by the NHANES researchers as  in the available public data releases. As can be seen, the maximum likelihood approach

results in significantly smaller standard errors for the parameter estimates when compared with complete-case analysis and very similar to those obtained via substitution with . In this simulation, 20 EM iterations were needed for convergence of all model parameters. It should be noted here that the maximum likelihood standard errors reported in Table II are based on only one simulation and are calculated via a straightforward fitting of the weighted logistic regression model at convergence. Standard errors for the simulation study reported in Table I were calculated as the standard error of the population of 1000 final β estimates, one for each simulated dataset. These estimation procedures are not equivalent, and it is important to note this difference.

It should also be noted that the fitted model used here does not include age as a covariate in the prediction of cancer status. A logistic model including the age covariate was also fit to this data, and age was found to be highly significant. The current model (without an age covariate) has been included here to more clearly display the potential benefits of the proposed methodology.

6. Discussion

In this paper, we have proposed a method of maximum likelihood estimation in GLMs with an unlimited number of covariates subject to a limit of detection. We have proposed models for the joint covariate distribution, which is based on a sequence of one-dimensional conditional distributions. The methodology presented here can be easily extended to cases where both the response and the covariates are subject to a limit of detection. The maximum likelihood approach presented here is much simpler computationally than a direct computation by way of the observed data likelihood, especially for cases with multiple covariates subject to an LOD. When only a single covariate (or just the response) is subject to an LOD, closed-form solutions can often be used.

For the simulation considered in Section 4, the parameter estimates for β using the maximum likelihood approach and complete-case appear similar and largely unbiased. This similarity was also seen in the real-life NHANES example, though with a slightly larger degree of variability. This variability can be attributed to the NHANES analysis only being performed on a single dataset, whereas the simulation results are averaged over 1000 datasets. The substitution estimates appeared quite similar to the maximum likelihood estimates in the NHANES study, but differed somewhat in the simulation study, where the substitution estimates appeared biased. This is likely due to the possibility that the chosen substitution values in the NHANES study were very close to the expected covariate means below the limit of detection. A different choice of substitution points (such as LOD or zero) would likely result in estimates that were not as close to maximum likelihood.

Both the simulation study and the NHANES example gave variance estimates for β that were significantly improved over the complete-case analysis. This improvement can clearly lead to higher statistical power in studies that include data subject to detection limits.

A consistent drawback to maximum likelihood estimation in GLMs with data subject to detection limits is that a new algorithm needs to be created for each individual analysis that

is performed. For sampling within ARMS, the log-likelihood function for the model of interest needs to be explicitly specified. In cases where the covariates are considered to follow a multivariate normal distribution, for example, the log-likelihood function is consistent and straightforward. However, more complicated covariate distributions will require a less standard computation of the log-likelihood, which can take significant additional time and can lead to error.

For both the simulation study and real-data analysis presented here, $m_i = 250$ samples were taken for each observation with covariates below a limit of detection. Based on the authors experience and other extensive simulations performed with this type of data, we feel that a sample size of at least $m_i = 100$ is necessary for accurate inference.

The computing time required to achieve EM convergence here clearly depends on the number of covariates in a model, the degree of censoring that is observed, and the number of samples that are taken for each censored observation. The simulation presented in Section 4 tended to converge quickly, with only an average of 28 iterations performed per dataset. This simulation of 1000 datasets took about 16 h to complete on a Lenovo laptop (Lenovo Group Ltd, Morrisville, NC, USA) with a dual-core pentium processor, making this approach very computationally feasible.

Although the analyses presented here discuss applications to GLMs, much interest exists in studies of longitudinal and survival data, where covariates are subject to a limit of detection. Further research will look at applying the methods presented here to such data.

Acknowledgments

Dr Ibrahim's research for this paper was partially supported by the NIH grant nos. GM 70335 and CA 74015.

References

1. Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey Data. 2005–2006
2. Rigobon R, Stoker TM. Estimation with censored regressors: basic issues. *International Economic Review*. 2007; 48(4):1441–1467.
3. D'Angelo G, Weissfeld L. An index approach for the Cox model with left censored covariates. *Statistics in Medicine*. 2008; 27:4502–4514. [PubMed: 18407573]
4. Nie L, Chu H, Liu C, Cole SR, Vexler A, Schisterman EF. Linear regression with an independent variable subject to a detection limit. *Epidemiology*. 2010; 21(4):S17–S24. [PubMed: 21422965]
5. Helsel, DR. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. New York: 2004.
6. Lynn H. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*. 2001; 20:33–45. [PubMed: 11135346]
7. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*. 2006; 65:2434–2439. [PubMed: 16737727]
8. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*. 2004; 112(17):1691–1696. [PubMed: 15579415]
9. Singh A, Nocerino J. *Chemometrics and Intelligent Laboratory Systems*. 2002; 60
10. Thompson M, Nelson KP. Linear regression with Type I interval- and left-censored response data. *Environmental and Ecological Statistics*. 2003; 10:221–230.

11. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*. 2003; 157(4):355–363. [PubMed: 12578806]
12. Schisterman E, Vexler A, Whitcomb B, Liu A. The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology*. 2006; 163(4):374–383. [PubMed: 16394206]
13. Cole SR, Chu H, Nie L, Schisterman EF. Estimating the odds ratio when exposure has a limit of detection. *International Journal of Epidemiology*. 2009; 38:1674–1680. [PubMed: 19667054]
14. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*. 1958; 26:24–36.
15. Ibrahim JG. Incomplete data in generalized linear models. *Journal of the American Statistical Association*. 1990; 85:765–769.
16. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. *Biometrika*. 1996; 72:916–922.
17. Ibrahim JG, Lipsitz SR, Chen M. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society, Series B*. 1999; 61:173–190.
18. Wei GC, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*. 1990; 85:699–704.
19. Gilks WR, Best NG, Tan KKC. Adaptive rejection metropolis sampling within Gibbs sampling. *Applied Statistics*. 1995; 44:455–472.
20. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*. 1992; 41:337–348.
21. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*. 1953; 21:1087–1092.
22. Louis T. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1982; 44:226–233.

Table 1

Parameter estimates and standard errors, comparing expectation-maximization algorithm approach with complete-case analysis and substitution of LOD/1000. One thousand datasets were used for each analysis, with 250 samples taken for each observation below the limit of detection.

Parameter	True	30% Below limit of detection												
		Maximum likelihood estimation (ML)				Complete case (CC)				Substitution LOD/ $\sqrt{2}$				Significant? ML/CC
		Estimate	SE	Boot SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	
β_0	1.00	1.0105	0.3161	0.3275	0.0014	1.0029	0.3771	0.0078	0.9242	0.1996	<.0001	Yes/Yes		
β_1	-0.75	-0.7517	0.0782	0.0794	<.0001	-0.7504	0.0910	<.0001	-0.6875	0.0842	<.0001	Yes/Yes		
β_2	0.26	0.2576	0.1089	0.1090	0.0180	0.2604	0.1241	0.0359	0.0684	0.1118	0.5405	Yes/Yes		
β_3	-0.17	-0.1696	0.0881	0.0853	0.0541	-0.1677	0.1003	0.0944	-0.3128	0.0830	0.0002	No/No		
β_4	3.00	3.0042	0.1191	0.1191	<.0001	3.0001	0.1416	<.0001	2.9359	0.1175	<.0001	Yes/Yes		
β_5	0.20	0.2001	0.0841	0.0821	0.0173	0.1990	0.1024	0.0520	0.1661	0.1020	0.1036	Yes/No		
β_6	-0.60	-0.6000	0.0853	0.0831	<.0001	-0.5988	0.1040	<.0001	-0.8589	0.0637	<.0001	Yes/Yes		
σ_y^2	0.50	0.4795	0.0519	0.0499	<.0001	0.4958	0.0598	<.0001	0.4255	0.0512	<.0001	Yes/Yes		

Parameter	True	50% Below limit of detection												
		Maximum likelihood estimation (ML)				Complete case (CC)				Substitution LOD/ $\sqrt{2}$				Significant? ML/CC
		Estimate	SE	Boot SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	
β_0	1.00	1.0215	0.3360	0.3440	0.0024	1.0097	0.4788	0.0350	0.3729	0.1068	0.0005	Yes/Yes		
β_1	-0.75	-0.7531	0.0848	0.0842	<.0001	-0.7473	0.1123	<.0001	-0.5715	0.1005	<.0001	Yes/Yes		
β_2	0.30	0.2928	0.1153	0.1144	0.0111	0.2982	0.1502	0.0471	0.0721	0.1505	0.6318	Yes/Yes		
β_3	-0.19	-0.1966	0.0933	0.0896	0.0350	-0.1937	0.1165	0.0964	-0.3495	0.1051	0.0009	Yes/No		
β_4	3.00	3.0100	0.1283	0.1270	<.0001	3.0053	0.1712	<.0001	2.7380	0.1381	<.0001	Yes/Yes		
β_5	0.20	0.1999	0.0925	0.0888	0.0307	0.1929	0.1343	0.1511	0.2046	0.1410	0.1469	Yes/No		
β_6	-0.60	-0.6062	0.0921	0.0893	<.0001	-0.6001	0.1297	<.0001	-0.9548	0.0889	<.0001	Yes/Yes		
σ_y^2	0.50	0.4786	0.0551	0.0538	<.0001	0.4994	0.0720	<.0001	0.3352	0.0512	<.0001	Yes/Yes		

Note: SE, standard error.

Table II

Logistic regression model summary for National Health and Nutrition Examination Survey data, comparing maximum likelihood approach with complete case analysis and ad hoc substitution of LOD.

Parameter	Method	Estimate	SE	P-value	Significant
Intercept	Complete case	-0.6047	0.7954	0.4471	No
	Substitution	-0.6562	0.7009	0.3492	No
	ML	-0.6377	0.7049	0.3656	No
Gender	Complete case	-0.0035	0.2305	0.9879	No
	Substitution	-0.0934	0.2041	0.6472	No
	ML	-0.0939	0.2041	0.6453	No
Race	Complete case	-1.6468	0.2789	<.0001	Yes
	Substitution	-1.6022	0.2516	<.0001	Yes
	ML	-1.6029	0.2516	<.0001	Yes
Physical activity	Complete case	-0.2641	0.2379	0.2671	No
	Substitution	-0.1984	0.2121	0.3494	No
	ML	-0.1986	0.2121	0.3491	No
Nicotine	Complete case	-1.1471	0.3221	0.0004	Yes
	Substitution	-1.1103	0.2798	0.0001	Yes
	ML	-1.1086	0.2797	0.0001	Yes
Dimethylarsonic acid	Complete case	-0.2309	0.1856	0.2133	No
	Substitution	-0.1969	0.1515	0.1936	No
	ML	-0.1975	0.1540	0.1996	No
13.7% below LOD	Complete case	0.6172	0.1465	<.0001	Yes
	Substitution	0.7236	0.1225	<.0001	Yes
	ML	0.7245	0.1229	<.0001	Yes
Cadmium	Complete case	-0.2689	0.1493	0.0717	No
	Substitution	-0.1958	0.1234	0.1126	No
	ML	-0.2000	0.1240	0.1068	No
5.3% below LOD	Complete case	0.0769	0.1396	0.5817	No
	Substitution	0.0297	0.1249	0.8120	No
	ML	0.0348	0.1257	0.7821	No
Tungsten	Complete case	-0.2689	0.1493	0.0717	No
	Substitution	-0.1958	0.1234	0.1126	No
	ML	-0.2000	0.1240	0.1068	No
10.7% below LOD	Complete case	0.0769	0.1396	0.5817	No
	Substitution	0.0297	0.1249	0.8120	No
	ML	0.0348	0.1257	0.7821	No
Uranium	Complete case	0.0769	0.1396	0.5817	No
	Substitution	0.0297	0.1249	0.8120	No
	ML	0.0348	0.1257	0.7821	No
9.6% below LOD	Complete case	0.0769	0.1396	0.5817	No
	Substitution	0.0297	0.1249	0.8120	No
	ML	0.0348	0.1257	0.7821	No

Note: SE, standard error; ML, maximum likelihood; LOD, limit of detection.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript