Published in final edited form as:
J Agric Biol Environ Stat. 2011 June 1; 16(2): 221–232. doi:10.1007/s13253-010-0045-3.

An EM Algorithm for Fitting a 4-Parameter Logistic Model to Binary Dose-Response Data

Gregg E. Dinse
Biostatistics Branch, National Institute of Environmental Health Sciences, Mail Drop A3-03, P.O. Box 12233, Research Triangle Park, NC 27709 USA dinse@niehs.nih.gov

Abstract

This article is motivated by the need of biological and environmental scientists to fit a popular nonlinear model to binary dose-response data. The 4-parameter logistic model, also known as the Hill model, generalizes the usual logistic regression model to allow the lower and upper response asymptotes to be greater than zero and less than one, respectively. This article develops an EM algorithm, which is naturally suited for maximum likelihood estimation under the Hill model after conceptualizing the problem as a mixture of subpopulations in which some subjects respond regardless of dose, some fail to respond regardless of dose, and some respond with a probability that depends on dose. The EM algorithm leads to a pair of functionally independent 2-parameter optimizations and is easy to program. Not only can this approach be computationally appealing compared to simultaneous optimization with respect to all four parameters, but it also facilitates estimating covariances, incorporating predictors, and imposing constraints. This article is motivated by, and the EM algorithm is illustrated with, data from a toxicology study of the dose effects of selenium on the death rates of flies. Other biological and environmental applications, as well as medical and agricultural applications, are also described briefly. Computer code for implementing the EM algorithm is available as supplemental material online.

Keywords

Binomial data; Hill model; Logistic regression; Quantal response

1. Introduction

The general problem of modeling binary data as a function of covariates is important in many research areas. This article focuses on the dose-response problem of modeling the probability of a binary response as a function of some measure of dose, which has applications in the biological and environmental sciences, as well as in many other disciplines. The data that motivated this research, and that are used to illustrate the proposed analysis, come from a toxicology study of the dose effects of selenium on the death rates of flies (Jeske *et al*, 2009). In other areas, one might have a clinical interest in the proportion of subjects experiencing pain relief after ingesting a specific dose of an analgesic drug (Finney, 1978) or an environmental interest in the dose-response relationship between dioxin-like compounds and tumor rates (Walker *et al*, 2005).

Supplemental Materials
Computer Code: A .zip archive file is available online, which contains computer code (for implementing the EM algorithm), the data from Section 4, and the output for these data.

Often, a simple 2-parameter logistic regression model provides an adequate summary of how a binary response relates to dose. This model specifies that the logit-transformed response probability is linear in the dose metric. Consequently, the parameters of interest are an intercept and a slope. Under this model, the dose-response curve has a lower asymptote of zero and an upper asymptote of one, the limits of the expected range of response probabilities.

That [0,1] range may not always be appropriate for modeling response probabilities. For example, some flies may die from causes unrelated to selenium toxicity while others may survive the study no matter how high the dose of selenium. Similarly, some patients may get pain relief from a placebo with no analgesic drug while others may get no pain relief regardless of analgesic dose; and some rodents may develop tumors from non-dioxin causes while others may remain tumor-free despite dioxin exposure. Finney (1978) gives several other biological assay examples and labels such subjects as natural responders and resistants, which we refer to as obligate responders and obligate non-responders, respectively. In these cases, the dose-response probabilities range over a subinterval of [0,1]. Thus, a natural generalization of the 2-parameter logistic model adds two more parameters so that the lower response asymptote may be greater than zero and the upper response asymptote may be less than one. The resulting 4-parameter logistic model provides increased flexibility at the cost of a higher dimensional optimization.

The notion that some subjects will or will not respond, independently of dose, while others have response probabilities that depend on dose suggests reformulating the problem as a mixture model with missing data. One observes indicators of whether or not subjects responded, but not indicators of which subjects were obligate responders and non-responders. Viewing the latter indicators as missing data, we developed an EM algorithm (Dempster *et al*, 1977) to estimate the proportions of subjects “destined” to respond and “unsusceptible” to response. Under a 2-parameter logistic model for the dose-response relationship among subjects who were neither obligate responders nor obligate non-responders, the EM algorithm provides maximum likelihood estimates (MLEs) of the intercept and slope, plus the destined and unsusceptible proportions, which together constitute the four unknowns in the full 4-parameter logistic model.

In analyzing the selenium data, Jeske *et al* (2009) applied a probit model, which is similar to a logistic model. They assumed the upper asymptote was one, but allowed for a nonzero lower asymptote representing the proportion of deaths unrelated to selenium toxicity. They obtained an estimate of the lower asymptote from the control (dose zero) data only and treated it as a known value when estimating the intercept and slope in the probit model. Differences between probit and logistic analyses aside, this article extends the basic model of Jeske *et al* (2009) in three ways. It permits the upper asymptote to be less than one to allow a proportion of “immune” flies to survive the study regardless of the selenium dose; it simultaneously estimates the intercept, slope, and two asymptotes; and it estimates the asymptotes using data from all dose groups.

The proposed EM algorithm is easy to program and can take advantage of existing software for standard logistic regression. Specifically, we show that at each M-step, the estimates of the two asymptotes are simple proportions, and the estimates of the intercept and slope can be obtained via ordinary 2-parameter logistic regression methods. The observed information matrix and the estimated covariance matrix of the estimators are straightforward to compute using the Louis (1982) method. Finally, this EM algorithm performs a pair of 2-parameter optimizations, which may provide computational advantages over simultaneous optimization involving all four parameters. Incorporating covariates in the EM algorithm is straightforward, and perhaps more importantly, some of the required parameter constraints

are satisfied automatically. The proposed EM algorithm is illustrated with one of the selenium data sets provided by Jeske *et al* (2009).

2. Background

2.1 Observed Data

Suppose binary response data are observed from k+1 groups, say a control group and k treated groups, where all N subjects are independent and n_i subjects are randomly assigned to group I and exposed to dose d_i of the test chemical (i=0,1,…,k). Control subjects (i = 0) are unexposed, and thus $d_0 = 0$. Let Y_{ij} be a binary indicator of whether subject j (j=1,…,n_i) in group i responds ($Y_{ij} = 1$) or not ($Y_{ij} = 0$), let $\boldsymbol{Y} = \{Y_{ij}: j=1,\ldots,n_i;\ i=0,1,\ldots,k\}$ be the vector of all responses, and let $\boldsymbol{y}$ be the observed value of $\boldsymbol{Y}$.

2.2 Hill Model

Assume the probability of response for subject j in group i (j=1,…,n_i; i=0,1,…,k) is given by the Hill (1910) model, a specific form of the 4-parameter logistic model. This non-linear model often is expressed as the following monotone function of dose d_i, with parameters $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)$:

$$\Pr(Y_{ij} - 1|d_i) = \phi_1 + \frac{(\phi_2 - \phi_1) d_i^{\phi_4}}{\phi_3^{\phi_4} + d_i^{\phi_4}}, \tag{1}$$

where ϕ_1 is the baseline response probability (at dose 0), ϕ_2 is the maximum response probability (at an infinite dose), ϕ_3 is the dose producing a response probability halfway between ϕ_1 and ϕ_2, and ϕ_4 is a shape parameter. As ϕ_3 is a dose, it must be non-negative. Without loss of generality, assume the probability of response increases with dose, which implies $\phi_4 > 0$ and $0 \le \phi_1 < \phi_2 \le 1$; otherwise one can simply reverse these constraints or recode Y_{ij} as $1-Y_{ij}$. The parameter ϕ_3 is typically called the ED_{50}, or the median effective dose, and ϕ_4 is often called the Hill coefficient. The Hill model produces a sigmoidal dose-response curve, such as displayed in Figure 1.

To see that model (1) is a special case of a 4-parameter logistic model, one can rewrite it in terms of log-dose, $z_i = \ln(d_i)$, by substituting $d_i = \exp(z_i)$ into (1) and rearranging terms. If one reparameterizes by setting $\alpha = -\phi_4 \ln(\phi_3)$, $\beta = \phi_4$, $\gamma = \phi_1$, and $\delta = \phi_2 - \phi_1$, then model (1) becomes

$$\Pr(Y_{ij}=1|z_i) = \gamma + \frac{\delta}{1+\exp(-\alpha - \beta z_i)}, \tag{2}$$

a 4-parameter logistic model; see Volund (1978) for a discussion of 4-parameter logistic models for continuous responses. Set $p_i = \Pr(Y_{ij} = 1|z_i)$ and note that α and β are an intercept and slope for a response on a modified logit scale, $\ln[(p_i - \gamma)/(\gamma + \delta - p_i)]$. The bounds on ϕ imply bounds on $\Omega = (\alpha, \beta, \gamma, \delta)$: $-\infty < \alpha < \infty$, $\beta > 0$, and $0 \le \gamma < \gamma + \delta \le 1$. Note that $d_0 = 0$ implies $z_0 = -\infty$.

2.3 Likelihood of Observed Data

Conditional on the dose values and ignoring combinatoric factors, the likelihood of the observed response data is proportional to

$$\prod_{i=0}^{k}\{(p_i)^{y_{i+}}(1-p_i)^{n_i-y_{i+}}\}, \tag{3}$$

where $y_{i+}=\sum_{j=1}^{n_i} y_{ij}$. Note that $z_0 = -\infty$ and $\beta > 0$ imply that $p_0 = \gamma$. Thus, the log-likelihood of the parameter vector $\Omega = (\alpha, \beta, \gamma, \delta)$, apart from additive constants, is $L_Y(\Omega; \boldsymbol{y}) = L_Y$, where

$$L_y=y_{0+}\ln(\gamma)+(n_0-y_{0+})\ln(1-\gamma)+\sum_{i=1}^{k}\left\{y_{i+}\ln\left(\gamma+\frac{\delta}{1+\exp(-\alpha-\beta z_i)}\right)+(n_i-y_{i+})\ln\left(1-\gamma-\frac{\delta}{1+\exp(-\alpha-\beta z_i)}\right)\right\}. \tag{4}$$

2.4 Maximum Likelihood Analysis

The maximum likelihood estimates (MLEs) are usually calculated by iteratively optimizing L_Y. For example, one might use a Newton-Raphson method, which requires both first and second derivatives of L_Y; a quasi-Newton method, which only requires first derivatives of L_Y; or a downhill simplex method, which does not require any derivatives. These approaches typically work well unless the data are too sparse and lead to ill-conditioned matrices or the starting values are too far from the MLEs. With any of these methods, however, constraints to honor the bounds on Ω must be imposed explicitly or circumvented through reparameterization.

3. Missing-Data Reformulation

3.1 Complete Data

The original problem can be reformulated into one amenable to EM iterations by incorporating latent variables. Suppose one observes whether or not each subject responded, but not whether a responder was destined to respond, nor whether a non-responder was unsusceptible to response. Thus, each subject is regarded as belonging to one of four mutually exclusive categories, but exact category membership is unknown. Regardless of dose, subjects in Category 1 are destined to respond, whereas subjects in Category 4 are unsusceptible and will not respond. All other subjects are susceptible to response but not destined to respond; they may respond (Category 2) or not respond (Category 3), and the probability of response can depend on dose.

Define a collection of latent indicators (X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij}), where X_{hij} is 1 if subject j from group i belongs to Category h (j=1,…,n_i; i=0,1,…,k; h=1,2,3,4) and is 0 otherwise. The observed indicators (Y_{ij}) and their additive complements ($1-Y_{ij}$) can be partitioned into sums of unobserved indicators: $Y_{ij} = X_{1ij} + X_{2ij}$ and $1-Y_{ij} = X_{3ij} + X_{4ij}$. One observes whether a subject responded (Y_{ij} =1) or not ($1-Y_{ij}$ = 1), but not whether a responder was destined to respond (X_{1ij} = 1) or not (X_{2ij} = 1), nor whether a non-responder was susceptible (X_{3ij} = 1) or not (X_{4ij}= 1).

3.2 Relationship to Hill Model

Let γ be the proportion of subjects in the population who are destined to respond (Category 1), let δ be the proportion who are susceptible but not destined to respond (Categories 2 and 3), and let $1 - \gamma - \delta$ be the proportion who are unsusceptible to response (Category 4). Among subjects who are susceptible but not destined to respond, let $\theta(z_i)$ and $1 - \theta(z_i)$ denote the dose-dependent proportions who respond (Category 2) and do not respond (Category 3), respectively.

For the j^{th} subject in the i^{th} group, the expected values of X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij} are

$$\gamma, \quad \delta\theta_i, \quad \delta(1-\theta_i), \quad 1-\gamma-\delta, \tag{5}$$

respectively, where $\theta_i = \theta(z_i)$. Note that X_{1ij}, X_{2ij}, X_{3ij}, and X_{4ij} sum to 1, as do their expected values. Also, the fact that X_{1ij} and X_{2ij} are binary and mutually exclusive implies that

$$Pr(Y_{ij}=1|z_i)=Pr(X_{1ij}+X_{2ij}=1|z_i)=Pr(X_{1ij}=1|z_i)+Pr(X_{2ij}=1|z_i)=\gamma+\delta\theta_i, \tag{6}$$

which reduces to the Hill model in (2) under the logistic model: $\theta_i = [1 + \exp(-\alpha - \beta z_i)]^{-1}$.

3.3 Likelihood of Complete Data

The likelihood of the complete data is proportional to a product of terms such as those in (5). Note that $z_0 = -\infty$ and $\beta > 0$ imply $\theta_0 = 0$ and $X_{20j} = 0$ for j=1,…,n_0. Apart from additive constants, the log-likelihood of the complete data is $L_X(\Omega; \boldsymbol{x}) = L_X$, where $\boldsymbol{X} = \{\boldsymbol{X}_{ij}: j=1, \ldots, n_i; i=0,1,\ldots,k\}$, $\boldsymbol{X}_{ij} = (X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij})$, $\boldsymbol{x}$ is a particular realization of $\boldsymbol{X}$, and

$$L_X=\sum_{j=1}^{n_0}\left\{x_{10j}\ln(\gamma)+x_{30j}\ln(\delta)+x_{40j}\ln(1-\gamma-\delta)\right\}+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{x_{1ij}\ln(\gamma)+x_{2ij}\ln(\delta\theta_i)+x_{3ij}\ln[\delta(1-\theta_i)]+x_{4ij}\ln(1-\gamma-\delta)\right\}. \tag{7}$$

Modeling θ_i by $[1 + \exp(-\alpha - \beta z_i)]^{-1}$ and collecting terms yields $L_X = L_{X1} + L_{X2}$, where

$$L_{X1}=-\sum_{i=1}^{k}\left\{x_{3i+}\alpha+x_{3i+}z_i\beta+(x_{2i+}+x_{3i+})\ln(1+e^{-\alpha-\beta z_i})\right\}, \tag{8}$$

$$L_{X2}=x_{1++}\ln(\gamma)+(x_{2++}+x_{3++})\ln(\delta)+x_{4++}\ln(1-\gamma-\delta), \tag{9}$$

and the "+" subscript indicates summation over the corresponding index. Note that L_{X1} and L_{X2} are functionally independent, with the former involving only α and β, and the latter involving only γ and δ, which simplifies the maximization of L_X and the calculation of the information matrix. Also, L_{X2} does not involve z_i, consistent with the asymptotes being dose-independent.

3.4 EM Algorithm

The MLE of Ω can be obtained via an EM algorithm (Dempster *et al*, 1977). After choosing a starting value for Ω, the EM algorithm iterates between expectation (E) and maximization (M) steps until convergence. At each iteration, the E-step calculates the expectations of the sufficient statistics for the complete data, conditional on the observed data and the current parameter estimates, and the M-step calculates the value of Ω that maximizes the log-likelihood of the current complete data. Each EM iteration increases the likelihood of the observed data.

At the E-step, conditional on the observed response y_{ij} and the current parameter estimate $\hat{\Omega} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$, the expected values $E(X_{1ij}|y_{ij}, \hat{\Omega})$ and $E(X_{4ij}|y_{ij}, \hat{\Omega})$ are estimated by

$$\widehat{x}_{1ij=y_{ij}}\left(\frac{\widehat{\gamma}}{\widehat{\gamma}+\widehat{\delta}\widehat{\theta}_i}\right) \quad \text{and} \quad \widehat{x}_{4ij}=(1-y_{ij})\left(\frac{1-\widehat{\gamma}-\widehat{\delta}}{1-\widehat{\gamma}-\widehat{\delta}\widehat{\theta}_i}\right), \tag{10}$$

respectively, where $\hat{\theta}_0 = 0$ and $\hat{\theta}_i = [1+ \exp(-\hat{\alpha} - \hat{\beta}z_i)]^{-1}$ for $i > 0$. By subtraction, estimates of the expected values of X_{2ij} and X_{3ij} are $\hat{x}_{2ij} = y_{ij} - \hat{x}_{1ij}$ and $= \hat{x}_{3ij} = 1 - y_{ij} - \hat{x}_{4ij}$, respectively.

At the M-step, conditional on $\hat{x}_{2i+}$ and $\hat{x}_{3i+}$, the estimates of α and β that maximize L_{X1} are the MLEs for a 2-parameter logistic regression problem with log-likelihood (8). Furthermore, substituting $\hat{x}_{1++}$, $\hat{x}_{2++}$, $\hat{x}_{3++}$, and $\hat{x}_{4++}$ into (9), the estimates of γ and δ that maximize the trinomial log-likelihood L_{X2} are

$$\widehat{\gamma}=\widehat{x}_{1++}/N \quad \text{and} \quad \widehat{\delta}=(\widehat{x}_{2++}+\widehat{x}_{3++})/N, \tag{11}$$

where $N = n_+ = \hat{x}_{+++}$ is the total number of subjects. Although only two of the four complete-data MLEs ($\hat{\gamma}$, $\hat{\delta}$) are available in closed form, the iterative procedure for obtaining the other pair ($\hat{\alpha}$, $\hat{\beta}$) is simpler than maximizing the entire 4-parameter observed-data log-likelihood L_Y.

Continue iterating until successive differences are suitably small for both the observed-data log-likelihood L_Y and the estimate $\hat{\Omega}$, and then declare the latter to be the MLE of Ω.

Computer code for implementing the EM algorithm is available online.

3.5 Variance Estimation

Let $\boldsymbol{G}_X(\Omega; \boldsymbol{X})$ and $\boldsymbol{H}_X(\Omega; \boldsymbol{X})$ be the gradient (first derivative) vector and negative Hessian (second derivative) matrix, respectively, of $L_X(\Omega; \boldsymbol{X})$ with respect to Ω, and define $\boldsymbol{G}_Y$ and $\boldsymbol{H}_Y$ similarly. Louis (1982) showed that the observed information matrix for $\boldsymbol{Y}$ at the MLE $\hat{\Omega}$, say $\boldsymbol{H}Y(\hat{\Omega}; \boldsymbol{y})$, is

$$\mathbf{I}_Y(\widehat{\boldsymbol{\Omega}})=E_{X|Y}[\mathbf{H}_X(\boldsymbol{\Omega};\mathbf{X})|\boldsymbol{Y}=\mathbf{y}]\Big|_{\Omega=\hat{\Omega}} - E_{X|Y}[\mathbf{G}_X(\boldsymbol{\Omega};\mathbf{X})\mathbf{G}_X^T(\boldsymbol{\Omega};\mathbf{X})|\boldsymbol{Y}=y]\Big|_{\Omega=\hat{\Omega}}. \tag{12}$$

Simplification of the observed information matrix in (12) is possible because $\boldsymbol{X}$ is a multinomial. The variance-covariance matrix for $\hat{\Omega}$, say Σ, can be estimated by $\widehat{\boldsymbol{\Sigma}}=\left[\mathbf{I}_Y(\widehat{\boldsymbol{\Omega}})\right]^{-1}$. This method of estimating Σ involves only L_X and is generally simpler than working with L_Y directly.

4. Application to Selenium Data

Jeske *et al* (2009) presented data from a toxicology study of the dose effects of four types of selenium on the death rates of flies. We focused on selenocysteine, which they labeled as type 4 selenium, and fitted the Hill model via the EM algorithm. The data are given in Table 1. Of the n_i flies receiving dose d_i of selenocysteine (i = 0,1,2,3,4), let Y_{ij} indicate whether fly j died during the study (j = 1,…, n_i). Specify the probability of dying during the study by the Hill model in (2), where γ and 1 − γ − δ are the respective proportions of flies destined to die from causes other than selenocysteine toxicity and to survive the study despite selenocysteine toxicity. The remaining proportion δ die during the study with dose-dependent probability $\theta_i = [1 + \exp(-\alpha - \beta z_i)]^{-1}$.

We selected EM starting values for α and β by fitting a 2-parameter logistic model to the observed data, assuming no predestined or unsusceptible subpopulations. However, one cannot set γ=0 and δ=1 as starting values because the EM algorithm will not move from these boundary values. Instead, we defined $\tilde{p}_i = (y_{i+}+½)/(n_i + 1)$ to guarantee estimated response rates in (0,1), and then we initially set the lower asymptote (γ) to the smallest $\tilde{p}_i$ and the upper asymptote (γ+δ) to the largest $\tilde{p}_i$, with the initial value of δ being the difference. This procedure produced starting values Ω = (−5.814, 1.289, 0.023, 0.568). The resulting MLEs of (α, β, γ, δ) are given in Table 2, along with estimates of their standard errors based on the Louis (1982) method. The MLEs of (ϕ_1, ϕ_2, ϕ_3, ϕ_4), which are simple transformations of (α, β, γ, δ), are also given in Table 2, along with estimates of their standard errors based on applying the delta method (Rao, 1973) to $\hat{\Sigma}$.

The Hill model fits these data well, as seen from the empirical (symbols) and fitted (solid curve) death rates in Figure 1; the dashed curves show pointwise 95% confidence bands obtained by applying the delta method. The usual observed-minus-expected goodness-of-fit statistic is 1.59 (Table 1), which suggests no significant lack of fit (P = 0.21, based on the chi-squared distribution with one degree of freedom). The MLEs of the lower ($\hat{\phi}_1 = 0.033$) and upper ($\hat{\phi}_2 = 0.673$) asymptotes are more than 1.96 standard errors above zero and below one (Table 2), respectively, suggesting that the full 4-parameter Hill model fits better than a reduced model.

Though Jeske *et al* (2009) fitted a 3-parameter probit model rather than a 4-parameter logistic model, they obtained similar results for the median effective dose. In their second table, they reported an MLE of 4.42 with a standard error of 0.19 for $\ln(ED_{50})$. After taking natural logs, the MLE and standard error from the EM algorithm are 4.02 and 0.17, respectively.

As a check, a quasi-Newton method gave the same estimate of Ω as the EM algorithm. Also, as a further check, we verified that the first derivative of the observed-data log-likelihood with respect to each parameter was zero when evaluated at the MLE: $\boldsymbol{G}_Y(\hat{\boldsymbol{\Omega}}; \boldsymbol{y}) = \mathbf{0}$.

Optimization procedures can be sensitive to initial values, so we tried several sets. First, we set the lower asymptote (ϕ_1) to 0.023 and the upper asymptote (ϕ_2) to 0.591, which were the starting values used earlier, and then investigated a grid of starting values for the ED_{50} (ϕ_3) and shape (ϕ_4) parameters. The MLEs of ϕ_3 and ϕ_4 were roughly 56 and 3, so we examined starting values of 40, 60, 80, and 100 for ϕ_3 and values of 1, 2, 3, and 4 for ϕ_4. All 16 combinations of these starting values gave the same final estimates as before, as did several other sets of starting values, suggesting that the EM algorithm is not overly sensitive to the choice of initial values.

5. Discussion

We developed an EM algorithm for fitting a Hill model, or more generally a 4-parameter logistic model, to binary (quantal) dose-response data. The EM algorithm is simple to program and leads to a pair of 2-parameter optimizations at each iteration, one of which has a closed-form solution. Thus, in this non-linear setting, the EM approach may provide computational advantages over conventional iterative approaches that optimize with respect to all 4 parameters simultaneously, at least for some data sets, though a rigorous investigation was not performed. Also, certain constraints that other methods impose explicitly are satisfied automatically in the EM algorithm.

Expanding on this last point, estimates of the lower (ϕ_1) and upper (ϕ_2) asymptotes must satisfy $0 \le \phi_1 < \phi_2 \le 1$ if the dose-response curve is increasing, as assumed in the development. The EM algorithm produces estimates of ϕ_1 and ϕ_2 (or γ and δ) that are simple

proportions, which always lie in the unit interval. In contrast, conventional methods must either explicitly restrict the asymptotes to fall in [0,1] or else circumvent constraints via reparameterization. For example, ϕ_1 and ϕ_2 can be forced to lie in (0,1) by applying a logistic transform to each. Typically, both EM and conventional methods will satisfy $\phi_1 < \phi_2$ and $\phi_4 > 0$ if the observed response rates mostly increase with dose, or $\phi_1 > \phi_2$ and $\phi_4 < 0$ if the rates mostly decrease with dose.

This article focused on the Hill model, a special case of the 4-parameter logistic model in which the dose metric is the natural logarithm of dose. The same methods can be applied with other dose metrics, though, such as $z_i = d_i$ or $z_i = i$. Also, the notation was developed to allow for a control group having a dose of zero ($d_0 = 0$), but the same methods can be adapted easily to handle studies without a control group by simply ignoring the terms with a subscript of $i = 0$. Furthermore, although the usual dose-response study involves multiple observations per dose group, the proposed approach can still be applied with only $n_i = 1$ observation per dose group. Finally, the EM algorithm can be modified trivially to fit a reduced 3-parameter logistic model, such as under the constraint $\phi_2 = 1$ (i.e., $\delta = 1 - \gamma$) used by Jeske *et al* (2009). However, the 2-parameter model, which constrains $\phi_1 = \gamma = 0$ and $\phi_2 = \delta = 1$, does not require any EM iterations.

Note that the proposed conceptualization, involving a mixture model with missing data, need not correspond precisely to reality; it is simply a convenient construction for calculating the MLEs under a 4-parameter logistic (or Hill) model. We hypothesize three mutually exclusive groups of subjects: those who always respond, those who never respond, and those who respond with a dose-dependent probability specified by a logistic curve with asymptotes of 0 and 1. This formulation may not mimic reality, but the MLEs it produces are identical to those obtained by other methods under a 4-parameter logistic model with asymptotes that need not equal 0 and 1.

Several extensions of the proposed method are possible. One is the incorporation of additional covariates. The M-step of the EM algorithm maximizes a 2-parameter logistic regression likelihood with an intercept and a slope for a single covariate equal to ln(dose). The incorporation of more covariates is straightforward when modeling response rates of susceptible subjects who are not destined to respond; that is, when maximizing L_{X1}. Also, since the two pieces of the complete-data log-likelihood are functionally independent, polytomous regression methods can be used to separately maximize L_{X2} after modeling the destined and unsusceptible proportions as functions of covariates unrelated to dose. This would allow formal assessment of explanatory variable effects on the destined and unsusceptible proportions, as well as on the response rate of susceptible subjects who are not destined to respond, say through a likelihood ratio test of whether certain regression coefficients are zero.

Another extension is incorporation of survival adjustments in studies where estimation of dose-response relationships might be biased by differential mortality. For example, Walker *et al* (2005) fitted a 3-parameter logistic model to binary tumor incidence data from a carcinogenicity study and incorporated a poly-3 survival adjustment (Bailer and Portier, 1988) to account for the reduced tumor risk of animals dying before the end of the study. The EM algorithm can be easily adapted to incorporate this same survival adjustment. As an alternative survival adjustment, one could use time as a covariate explaining response. This approach would generalize the nonlethal tumor analysis of Dinse and Lagakos (1983), which applied standard logistic regression methods. By including both dose and time metrics as covariates, one could allow non-boundary values for the asymptotes and also could provide an alternative to the poly-3 correction to adjust for survival effects on the incidence rates of

nonlethal tumors. This extension represents ongoing research and will be the subject of a future article.

In summary, the EM algorithm provides a natural solution to the problem of modeling binary responses when some subjects are obligate responders or obligate non-responders. This approach leads to a straightforward way to estimate the covariance matrix of the MLEs and to incorporate explanatory variables. Furthermore, as seen in other contexts (e.g., forcing positive variance component estimates), the EM algorithm automatically satisfies certain constraints that are more complicated to implement with other methods. Though the example and some of the terminology focused on dose-response analysis of toxicology data, the proposed EM algorithm has general applications for various binary outcomes observed in a broad range of research areas. For instance, consider a clinical trial evaluating a new therapy where the probability of disease remission generally increases with dose, but some patients improve even if not treated, while others regress no matter how high the dose. Or, consider an agricultural study of an herbicide, where the death rates of targeted plants generally increase with dose, but some plants may die from causes unrelated to the herbicide, while others may appear resistant within the range of doses applied. The proposed EM algorithm should handle these situations and many others.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES-102685). I am very grateful to Shyamal Peddada, David Umbach, Clarice Weinberg, the editors, and the referees for their valuable suggestions.

References

Bailer AJ, Portier CJ. Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics. 1988; 44:417–431. [PubMed: 3390507]

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39:1–38.

Dinse GE, Lagakos SW. Regression analysis of tumour prevalence data. Journal of the Royal Statistical Society, Series C. 1983; 32:236–248. [Corrigenda, Vol. 33, 79–80, 1984.].

Finney, DJ. Statistical Method in Biological Assay. London: Oxford University Press; 1978.

Hill AV. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. Journal of Physiology. 1910; 40 Suppl.:iv–vii.

Jeske DR, Xu HK, Blessinger T, Jensen P, Trumble J. Testing for the equality of EC50 values in the presence of unequal slopes with application to toxicity of selenium types. Journal of Agricultural, Biological, and Environmental Statistics. 2009; 14:469–483.

Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B. 1982; 44:226–233.

Rao, CR. Linear Statistical Inference and Its Applications. New York: John Wiley; 1973.

Volund A. Application of the four-parameter logistic model to bioassay: comparison with slope ratio and parallel line models. Biometrics. 1978; 34:357–365. [PubMed: 719119]

Walker NJ, Crockett PW, Nyska A, Brix AE, Jokinen MP, Sells DM, Hailey JR, Easterling M, Haseman JK, Yin M, Wyde ME, Bucher JR, Portier CJ. Dose-additive carcinogenicity of a defined mixture of “dioxin-like compounds.”. Environmental Health Perspectives. 2005; 113:43–48. [PubMed: 15626646]

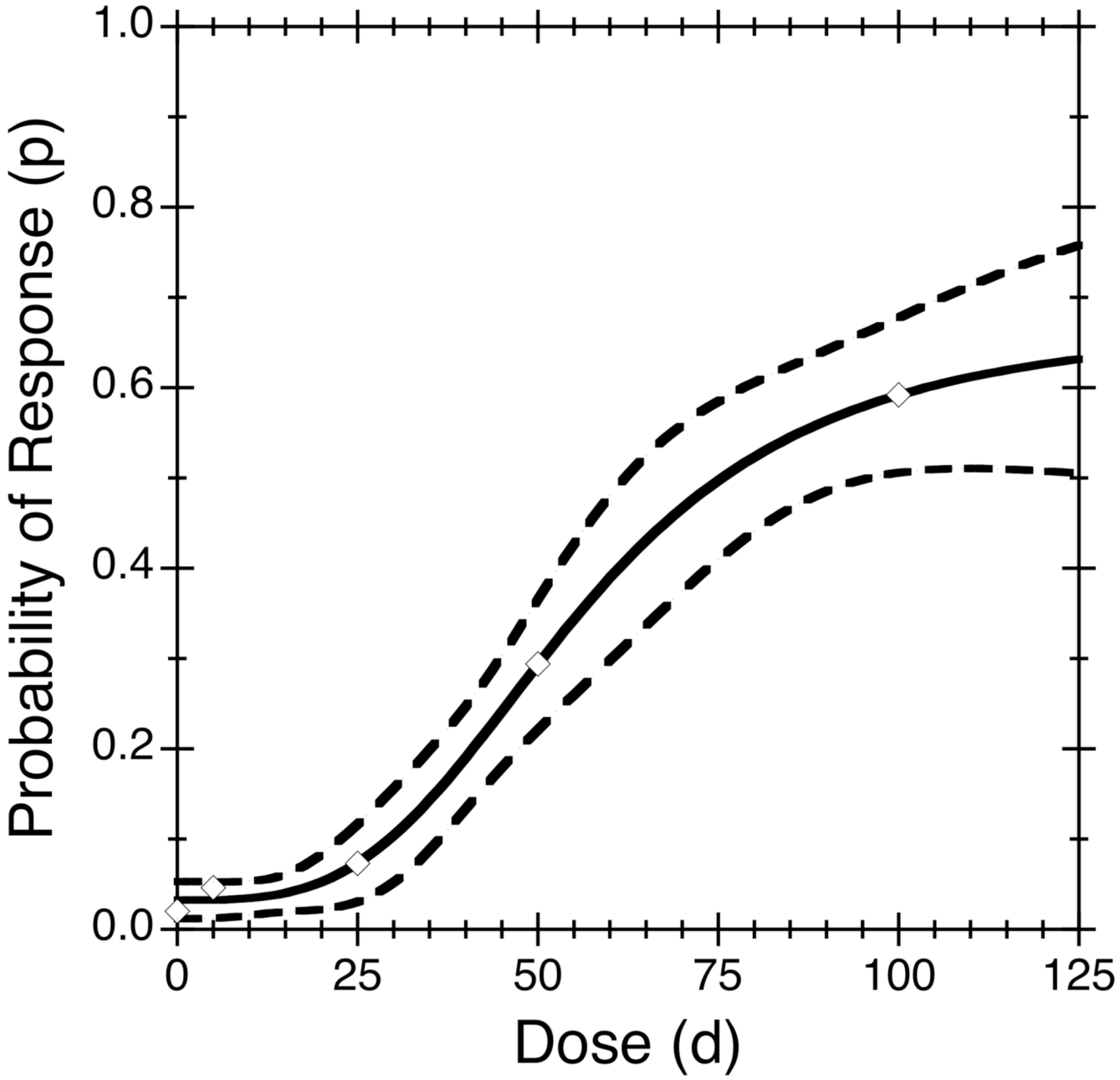

Figure 1.
Probability of response as a function of dose for the selenocysteine data. The empirical response rate for each dose group is shown by a diamond, the dose-response curve fitted under a 4-parameter Hill model is shown by a solid curve, and the pointwise 95% confidence bands are shown by dashed curves.

Table 1

Selenocysteine data from Jeske *et al* (2009). For group i, d_i is the dose, y_{i+} is the observed number of deaths, n_i is the number of subjects, y_{i+}/n_i is the empirical death rate, $\hat{P}_i$ is the predicted death rate under the Hill model, E_i is the expected number of deaths (rounded) under the Hill model, and the last column is a measure of model goodness-of-fit.

i	d_i	y_{i+}	n_i	y_{i+}/n_i	$\hat{P}_i$	E_i	$\frac{(y_{i+}-E_i)^2}{E_i}$
0	0	3	152	0.020	0.033	5.0	0.78
1	5	7	152	0.046	0.033	5.0	0.81
2	25	11	150	0.073	0.074	11.1	0.00
3	50	45	153	0.294	0.294	45.0	0.00
4	100	74	125	0.592	0.592	74.0	0.00
		140	732			140.1	1.59

Table 2

Maximum likelihood estimates (MLEs) and estimated standard errors (S.E.s) under the Hill model, expressed in terms of either dose or the natural logarithm of dose, for the data of Jeske *et al* (2009) on the effect of selenocysteine on the death rates of flies.

Hill model (1), in terms of dose d_i			**Hill model (2), in terms of $z_i = \ln(d_i)$**		
Parameter	**MLE**	**S.E.**	**Parameter**	**MLE**	**S.E.**
ϕ_1 (min)	0.033	0.010	γ (min)	0.033	0.010
ϕ_2 (max)	0.673	0.115	δ (max–min)	0.641	0.118
ϕ_3 (ED_{50})	55.962	9.392	α (intercept)	−13.368	3.923
ϕ_4 (shape)	3.322	1.084	β (slope)	3.322	1.084

The parameterizations are related as follows: $\phi_1 = \gamma$, $\phi_2 = \gamma + \delta$, $\phi_3 = \exp(-\alpha/\beta)$, $\phi_4 = \beta$.