Kansas State University Libraries
New Prairie Press

Conference on Applied Statistics in Agriculture | 2006 - 18th Annual Conference Proceedings

DOSE-RESPONSE MODELING WITH MARGINAL INFORMATION ON A MISSING CATEGORICAL COVARIATE

John R. Stevens

David I. Schlipalius

Follow this and additional works at: http://newprairiepress.org/agstatconference

 Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

Stevens, John R. and Schlipalius, David I. (2006). "DOSE-RESPONSE MODELING WITH MARGINAL INFORMATION ON A MISSING CATEGORICAL COVARIATE," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1118

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

DOSE-RESPONSE MODELING WITH MARGINAL INFORMATION ON A MISSING CATEGORICAL COVARIATE

John R. Stevens[1] and David I. Schlipalius[2]
[1] Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322-3900 USA
[2] School of Integrative Biology, Goddard Building (8), St Lucia Campus, The University of Queensland, Brisbane, Qld 4072 Australia

Abstract

When the relationship between a dosage-type variable and a binary outcome depends on a categorical variable, a common analysis would employ a dose-response model with the categorical variable as a covariate. When the level of the categorical variable is not known for all subjects, however, the standard dose-response model alone cannot provide useful inference. We present an EM-based approach to account for the missing covariate in a dose-response model setting when additional knowledge about the marginal distribution of the covariate is available. This approach is motivated by a study of the beetle *Rhyzopertha dominica*, a pest of stored grain in Australia. Certain genotypes of this beetle have developed inheritable resistance to a widely-used insecticidal fumigant. In this study, the effects of various dosage levels of the fumigant were considered, and it was feasible to genotype only the surviving beetles.

Keywords: dose-response, EM algorithm, incomplete data

1 Introduction

The beetle *Rhyzopertha dominica* (lesser grain borer) is a pest of stored grain and is a year-round problem in moderate climates such as Australia. Australian grain is a multi-billion dollar industry with a zero-tolerance policy for insect-infested grain. Phosphine fumigant is the primary control against beetle-infested grain, but the lesser grain borer has developed inheritable resistance to phosphine. Previous work (Schlipalius et al. 2002) has identified two DNA markers (rp6.79 and rp5.11) that are closely linked to two independent resistance loci that are responsible for the majority of this inheritable resistance in a highly resistant beetle strain isolated in Australia. That is, the degree of resistance to phosphine seems to depend on the forms of these two genes in the individual beetle.

As previously reported (Schlipalius et al. 2002), susceptible and resistant beetles were crossed to produce an F1 generation which was then bred within itself to produce an F2 generation.

The F2 generation consisted of 104 individuals, 92 of which were sacrificed and used to estimate a genetic map; the other 12 were bred amongst themselves to produce the F3, F4, and F5 generations. The F5 individuals were divided into eleven groups, and each group received a different dosage of phosphine.

After 48 hours exposure to the fumigant, the surviving beetles were genotyped at both the rp6.79 and rp5.11 loci to determine what forms of the genes were present in each survivor. The rp6.79 locus was characterized as having two possible genotypes (-,+), while the rp5.11 locus had three (B, H, A). Table 1 shows the observed data. While the dose levels and exposure time used in these experiments are not the same as those used in field applications, these data can be used to characterize and better understand the varying degrees of resistance among the different genotypes.

Of immediate interest here is the relationship between mortality and dosage for each genotype. In field applications, the goal is to determine the appropriate dosage of phosphine to exterminate the beetles with high probability. In choosing the appropriate concentration of phosphine for fumigating stored grain, several considerations support the use of the lowest practical dose. Experience has shown that time is more important than concentration, so that greater exposure time at a consistently maintained lower dose is more successful in killing the beetles than is a higher dose for a shorter time. Expense is another motivation for using lower doses. In addition, the difficulty of maintaining a consistent concentration in a non-air-tight grain silo prevents a high-dose strategy from being successful. Finally, phosphine can be spontaneously combustible at high concentrations, and so safety issues also support the use of the lowest practical doses.

Ideally in a study such as this, genotypic (or more generally, classification) information would have been recorded for all individuals. Then to determine the necessary dosage to kill a certain percentage of the beetles in each genotype, a dose-response model relating survival or death to genotype and dose level would be relatively straightforward. However, in this particular setting, genotyping all beetles was cost-prohibitive. An extraordinarily large number of beetles was used at the higher dose levels to ensure survivors at each dose level. At the same time, the large number of beetles at the high dose levels also provided large numbers of dead beetles to genotype, and resources to genotype all beetles were not available. The savings in genotyping only the 378 survivors rather than all 10,798 beetles were substantial.

Because the total number of individuals at each genotype-dose level combination is unknown, standard dose-response models can not be used for the analysis of these data. Hence it is necessary to develop a method for the analysis of these data. This method provides a way to make inference regarding the effects of different dosages on the mortality of each genotype in a dose-response model, even though the data are incomplete. One key to this method is the availability of marginal information on the missing categorical covariate, genotype, as found in the zero dose level.

2 Methods

Let $N_{.j}$ be the total number of beetles receiving dose level j ($j = 0, ..., 10$) and $n_{i,j}$ be the total number of survivors at dose level j with genotype i ($i = 1, ..., 6$). Then Table 1 shows that $N_{.j}$ and $n_{i,j}$ were observed. Let $N_{i,j}$ represent the unobserved number of beetles receiving dose level j in genotype i. Let $p_{i,j}$ be the probability of death for genotype i beetles at dose level j; then if $N_{i,j}$ had been observed, the traditional model $n_{i,j}|N_{i,j} \sim Binomial(N_{i,j}, 1-p_{i,j})$ could have been assumed. Let $\mathbf{N}_j = (N_{1,j}, ..., N_{6,j})$ be the vector of total number of beetles at dose level j. Although the $N_{i,j}$ are unobserved here, $\mathbf{N}_j$ may be [latently] considered as having come from a multinomial distribution: $\mathbf{N}_j \sim Multinomial(N_{.j}, P)$, where $P = (P_1, ..., P_6)$, and P_i is the proportion of the beetle population with genotype i.

To place this problem in a maximum likelihood context, let $\theta = (p, P)$ be the parameters (including $p_{i,j}$ and P_i) in the binomial and multinomial distributions, and let $T = (n, N)$ be the complete data. Then after some simplification of the joint likelihood $f(n, N|\theta) = f(n|N, \theta)f(N|\theta)$, the log-likelihood can be written as

$$\begin{aligned} l(\theta|\mathbf{T}) &= \sum_{j=0}^{10}[\log N_{.j}! + \sum_{i=1}^{6}\{N_{i,j}\log P_i - \log n_{i,j}! - \log(N_{i,j} - n_{i,j})! \\ &\quad + n_{i,j}\log(1 - p_{i,j}) + (N_{i,j} - n_{i,j})\log p_{i,j}\}]. \end{aligned} \tag{1}$$

In order to understand how genotype and dose level affect the probability of mortality, $p_{i,j}$ can be parameterized in terms of genotype i and dose level j using, for example, a logit link:

$$\log\frac{p_{i,j}}{1 - p_{i,j}} = \eta_{i,j} = \quad G_i + D_i d_j \tag{2}$$

Then parameter estimates can be obtained by maximizing the likelihood in equation 1, and statistical inference can be made regarding the parameters θ. The exact parameterization chosen will affect the types of inferences that can be made, and one possible parameterization is presented in section 2.3 of this paper. For our immediate needs, however, equation 1 is sufficient to consider the maximization problem.

The likelihood in equation 1 cannot be directly maximized due to the missing data (specifically, the $N_{i,j}$). The mechanism of missing information in the current data, however, suggests an implementation of the EM algorithm (Dempster et al. 1977) might be appropriate. A necessary assumption for an EM application is that of "missing at random" or MAR. In general, suppose that y is the [completely observed] response variable (here, survival/death), and that covariate x (here, genotype) is at least partially unobserved. Then covariate x is MAR if and only if the probability of observing x (conditional on y and the other observed covariates) does not depend on x or any other unobserved covariate, but may depend on y and the other observed covariates (Ibrahim 1990). In the current beetle data, genotype (x) is observed only for survivors (y="survival"), and for all beetles at zero dosage. As such, the

categorical covariate genotype is MAR, and an application of the EM algorithm to maximize equation 1 is appropriate.

The EM algorithm may be summarized using the following formulation (Hastie et al. 2001):

1. *Initialization*: Select initial guess $\hat{\theta}^{(0)}$.
2. *Expectation*: At iteration k, evaluate $Q^{(k)} = Q(\theta, \hat{\theta}^{(k)}) = E[l(\theta, \mathbf{T})|\mathbf{n}, \hat{\theta}^{(k)}]$.
3. *Maximization*: Let $\hat{\theta}^{(k+1)}$ be the θ to maximize $Q^{(k)}$.
4. *Convergence*: Repeat steps (b) and (c) until convergence of $Q^{(k)}$.

2.1 Initialization

It is important to note that there are two classes of marginal information in these data. First, for all dosage levels j, $N_{\cdot j}$ is observed, providing some marginal information on the distribution of dosage. More importantly, however, at the zero dose level ($j = 0$), the $N_{i,j}$ are observed (and $N_{i,0} = n_{i,0}$), and this can be used to provide marginal information on the distribution of the missing categorical covariate genotype. Specifically, the initial estimates for $\mathbf{P}$ can be derived as the MLE's from the multinomial likelihood at the zero dose level, producing initial estimates $\hat{P}_i^{(0)} = n_{i,0}/N_{\cdot 0}$. Initial estimates for the $p_{i,j}$ can be taken to be 0.5. The initial estimates of the $p_{i,j}$ are not critical, but the presence of the zero dosage level provides the necessary marginal information on genotype to make useful initial estimates for P_i.

2.2 Expectation

Dropping constants $\log N_{\cdot j}!$ and $\log n_{i,j}!$, the expectation quantity of interest is

$$Q^{(k)} = \sum_{i,j} \tilde{N}_{i,j}^{(k)} \log P_i - L_{i,j}^{(k)} + n_{i,j} \log(1 - p_{i,j}) + (\tilde{N}_{i,j}^{(k)} - n_{i,j}) \log(1 - p_{i,j}), \qquad (3)$$

where $\tilde{N}_{i,j}^{(k)} = E[N_{i,j}|n, \hat{\theta}^{(k)}]$ and $L_{i,j}^{(k)} = E[\log(N_{i,j} - n_{i,j})!|n, \hat{\theta}^{(k)}]$. In order to evaluate $\tilde{N}_{i,j}^{(k)}$ and $L_{i,j}^{(k)}$, it is first necessary to address the distribution of $N_{i,j}$ given n and θ, or the distribution of $\mathbf{N}_j$ (the vector of total number of individuals at dose level j) given $\mathbf{n}_j$ (the vector of number of survivors at dose level j). With conditional notation temporarily suppressed, Bayes formula gives

$$\begin{aligned} h(\mathbf{N}_j|\mathbf{n}_j) &= \frac{f(\mathbf{n}_j|\mathbf{N}_j) f(\mathbf{N}_j)}{\sum_{\mathbf{N}_j} f(\mathbf{n}_j|\mathbf{N}_j) f(\mathbf{N}_j)} \\ &= \frac{f(\mathbf{n}_j|\mathbf{N}_j) f(\mathbf{N}_j)}{\sum_{N_{1j}} \sum_{N_{2j}} \cdots \sum_{N_{Ij}} f(\mathbf{n}_j|\mathbf{N}_j) f(\mathbf{N}_j)}, \end{aligned} \qquad (4)$$

where the i^{th} summation term in the denominator can be written as

$$\sum_{N_{ij}} = \sum_{N_{ij}=n_{ij}}^{N_{\cdot j}-\sum_{l<i} N_{lj}-\sum_{l>i} n_{lj}}.$$

A standard computational package such as Maple can be used to show that this simplifies to

$$h(\mathbf{N}_j|\mathbf{n}_j,\mathbf{p},\mathbf{P}) \;=\; (N_{\cdot j}-\sum_i n_{i,j})!\prod_i \left(\frac{\left(\frac{P_i p_{i,j}}{\sum_l P_l p_{l,j}}\right)^{N_{i,j}-n_{i,j}}}{(N_{i,j}-n_{i,j})!} \right). \tag{5}$$

From equation 5, it can be seen that

$$(\mathbf{N}_j-\mathbf{n}_j)|\mathbf{n}_j,\theta \;\sim\; Multinomial((N_{\cdot j}-\sum_i n_{i,j}),\lambda_j), \tag{6}$$

where λ_j is a length six vector with

$$\lambda_{i,j} \;=\; \frac{P_i p_{i,j}}{\sum_l P_l p_{l,j}}. \tag{7}$$

Using this multinomial distribution,

$$\tilde{N}_{i,j}^{(k)} = E[N_{i,j}|n,\hat{\theta}^{(k)}] = \frac{\hat{P}_i^{(k)}\hat{p}_{i,j}^{(k)}}{\sum_l \hat{P}_l^{(k)}\hat{p}_{l,j}^{(k)}}(N_{\cdot j}-\sum_l n_{l,j})+n_{i,j}. \tag{8}$$

Note that the θ maximizing $Q^{(k)}$ will not depend on $L_{i,j}^{(k)} = E[\log(N_{i,j}-n_{i,j})!|n,\hat{\theta}^{(k)}]$, and so evaluation of $L_{i,j}^{(k)}$ is not necessary for the maximization step. In fact, the evaluation of $L_{i,j}^{(k)}$ is only necessary because it will affect the rate of convergence of $Q^{(k)}$. Unlike $\tilde{N}_{i,j}^{(k)}$, $L_{i,j}^{(k)}$ has no closed form, and direct evaluation from the multinomial distribution is computationally prohibitive. Instead, an approximation strategy is necessary, first making use of Binet's formula

$$\log(N-n)! \;\approx\; (N-n+0.5)\log(N-n+1)-(N-n+1)+0.5\log 2\pi. \tag{9}$$

A graphical check reveals that as a function of $N-n$, Binet's formula is a locally linear function for $N-n \geq 20$, and at most locally quadratic for $N-n < 20$. As such, the expected value of Binet's formula may be approximated using a second-order Taylor series taken about $\tilde{N}_{i,j}^{(k)} - n_{i,j}$. From the variance of the multinomial distribution of $\mathbf{N}_{i,j}-n_{i,j}$ it can be shown that

$$E[N_{i,j}^2|\mathbf{n}_j,\theta] \;=\; (N_{\cdot j}-\sum_l n_{l,j})\hat{\lambda}_{i,j}^{(k)}(1-\hat{\lambda}_{i,j}^{(k)})+(\tilde{N}_{i,j}^{(k)})^2, \tag{10}$$

where $\hat{\lambda}_{i,j}^{(k)}(1-\hat{\lambda}_{i,j}^{(k)})$ can be rewritten as

$$\hat{\lambda}_{i,j}^{(k)}(1-\hat{\lambda}_{i,j}^{(k)}) = \frac{\hat{P}_i^{(k)}\hat{p}_{i,j}^{(k)}\sum_{l\neq i}\hat{P}_l^{(k)}\hat{p}_{l,j}^{(k)}}{(\sum_l \hat{P}_l^{(k)}\hat{p}_{l,j}^{(k)})^2}. \quad (11)$$

Note that if

$$g(x) = (x+0.5)\log(x+1) - (x+1) + 0.5\log 2\pi, \quad (12)$$

then $g''(x)$ can be written as

$$g''(x) = \frac{x+1.5}{(x+1)^2}. \quad (13)$$

Then the second-order Taylor series approximation to Binet's formula yields

$$\begin{aligned} L_{i,j}^{(k)} &= E[\log(N_{i,j}-n_{i,j})!|n,\hat{\theta}^{(k)}] \\ &\approx E[g(\tilde{N}_{i,j}^{(k)}-n_{i,j}) + g'(\tilde{N}_{i,j}^{(k)}-n_{i,j})((N_{i,j}-n_{i,j})-(\tilde{N}_{i,j}^{(k)}-n_{i,j})) \\ &\quad +0.5g''(\tilde{N}_{i,j}^{(k)}-n_{i,j})((N_{i,j}-n_{i,j})-(\tilde{N}_{i,j}^{(k)}-n_{i,j}))^2|n,\hat{\theta}^{(k)}] \\ &= g(\tilde{N}_{i,j}^{(k)}-n_{i,j}) + 0 + 0.5g''(\tilde{N}_{i,j}^{(k)}-n_{i,j})(E[N_{i,j}^2|n,\hat{\theta}^{(k)}] - (\tilde{N}_{i,j}^{(k)})^2) \\ &= (\tilde{N}_{i,j}^{(k)}-n_{i,j}+0.5)\log(\tilde{N}_{i,j}^{(k)}-n_{i,j}+1) - (\tilde{N}_{i,j}^{(k)}-n_{i,j}+1) + 0.5\log 2\pi \\ &\quad + \frac{\tilde{N}_{i,j}^{(k)}-n_{i,j}+1.5}{2(\tilde{N}_{i,j}^{(k)}-n_{i,j}+1)^2} \times \frac{(N_{.j}-\sum_l n_{l,j})\hat{P}_i^{(k)}\hat{p}_{i,j}^{(k)}\sum_{l\neq i}\hat{P}_l^{(k)}\hat{p}_{l,j}^{(k)}}{(\sum_l \hat{P}_l^{(k)}\hat{p}_{l,j}^{(k)})^2} \\ &= \tilde{L}_{i,j}^{(k)}. \end{aligned} \quad (14)$$

Thus the expectation step yields

$$\tilde{Q}^{(k)} = \sum_{i,j}\left(\tilde{N}_{i,j}^{(k)}\log P_i - \tilde{L}_{i,j}^{(k)} + n_{i,j}\log(1-p_{i,j}) + (\tilde{N}_{i,j}^{(k)}-n_{i,j})\log(1-p_{i,j})\right), \quad (15)$$

where $\tilde{N}_{i,j}^{(k)}$ and $\tilde{L}_{i,j}^{(k)}$ are calculated from equations 8 and 14, respectively.

2.3 Maximization

The portion of $\tilde{Q}^{(k)}$ involving P is

$$\tilde{Q}_P^{(k)} = \sum_{i,j}\tilde{N}_{i,j}^{(k)}\log P_i. \quad (16)$$

Using the method of Lagrange multipliers to maximize $Q^{(k)}$ with respect to P, subject to the constraint that $\sum_i P_i = 1$, yields

$$\hat{P}_i^{(k+1)} = \frac{\sum_j \tilde{N}_{i,j}^{(k)}}{\sum_{l,j}\tilde{N}_{l,j}^{(k)}}. \quad (17)$$

The portion of $\tilde{Q}^{(k)}$ involving p is

$$\tilde{Q}_p^{(k)} = \sum_{i,j} \left(n_{i,j} \log(1 - p_{i,j}) + (\tilde{N}_{i,j}^{(k)} - n_{i,j}) \log(1 - p_{i,j}) \right). \quad (18)$$

Note that this is just the binomial log likelihood, which can be maximized with respect to p by first specifying a parameterization of p in terms of some vector ϑ of genotype and dosage effects. There are of course may possible parameterizations, and Table 2 gives one. The parameter estimates $\hat{\vartheta}$ may be obtained using an approach such as Newton-Raphson iterations to maximize the log likelihood, producing both $\hat{p}_{i,j}^{(k)}$ and $\hat{\vartheta}^{(k)}$.

2.4 Convergence

This implementation of the EM algorithm is made in R, and iterations run until $\tilde{Q}^{(k)} - \tilde{Q}^{(k-1)} < \epsilon$. Here, the convergence criterion is chosen to be $\epsilon = 1e - 12$. Smaller values of ϵ do not give substantially different results. The implementation converges in 1639 iterations, taking 53 seconds.

2.5 Covariance Matrix for Parameter Estimates

Each iteration of the EM algorithm will produce updated estimates of the parameter vector $\phi = (P, \vartheta)$, with $\hat{\phi}$ as the estimate at the last EM iteration where convergence is achieved. In order to make statistical inference about the dose-response model, it is necessary to estimate $V = Var[\hat{\phi}]$, the covariance matrix of $\hat{\phi}$. The missingness mechanism in the data is not accounted for by the estimation approaches such as Newton-Raphson in the maximization step. Various approaches have been proposed for estimating the variance-covariance matrix in EM applications, such as EM by method of weights (Ibrahim 1990), Supplemented EM (SEM) (Meng and Rubin 1991), and direct calculation of the information matrix (Oakes 1999). Of these, the direct calculation was found to be the most straightforward for these data, and the most generalizable to alternative parameterizations ϑ.

In the direct calculation approach, let the expected log-likelihood Q be reparameterized in terms of $\phi = (P, \vartheta)$, so that $\tilde{Q}^{(k)} = \tilde{Q}(\phi, \hat{\phi}^{(k)})$. Then of interest is the variance-covariance matrix $V = Var[\hat{\phi}] = I_m^{-1}$, where

$$I_m = -\frac{\partial^2 l(\hat{\phi}|T)}{\partial \hat{\phi}^2}. \quad (19)$$

Due to the missing data, the differentiation required in equation 19 is not practical. Instead, the "direct calculation" approach (Oakes 1999) can be used, with I_m expressed in terms of the expected log-likelihood quantity:

$$\frac{\partial^2 l(\hat{\phi}|T)}{\partial \hat{\phi}^2} = \left[\frac{\partial^2 \tilde{Q}(\phi, \hat{\phi})}{\partial \phi^2} + \frac{\partial^2 \tilde{Q}(\phi, \hat{\phi})}{\partial \phi \partial \hat{\phi}} \right]_{\phi = \hat{\phi}}. \quad (20)$$

On the right-hand side of equation 20, the first term is often referred to as the "expected complete data information," and the second term is the "missing information." The differentiation and evaluation required for these two terms is relatively straightforward using a symbolic computation package such as Maple.

With these data, the resulting I_m matrix is not positive definite, which is a necessary condition for the covariance matrix V to be positive definite (i.e., symmetric and all eigenvalues positive). However, this non-positive-definiteness could easily be attributed to small numerical imprecisions, as the sole non-positive eigenvalue is on the order of $-1e-13$, and there are slight differences between off-diagonal elements. The smallest absolute diagonal element of I_m was approximately 0.07, and the largest relative difference between off-diagonal elements was on the order of $1e-04$ (0.0001745709 vs. 0.00017454). The approach employed here to "force" symmetry was to average the off-diagonal elements, to adjust for differences in rounding. To "force" positive definiteness, an approach similar to the Levenberg-Marquardt adjustment in section 4.5.3.3 of Thisted 1988 was adopted, adding a small positive constant $(1e-08)$ to the diagonal entries of I_m. With the I_m thus calculated, the variance-covariance matrix $V = I_m^{-1}$ can be obtained, and then statistical inferences can be made, with the missingness of the data appropriately taken into account.

3 Results

As an example of the type of statistical inferences that can be made once the EM parameter estimates and corresponding covariance matrix have been obtained, consider the standard errors of the predicted probabilities of mortality. For each of the six genotypes, there is some vector a such that the predicted probability of death at some dosage d can be expressed as

$$\hat{p} = \left(1 + \exp(-a'\hat{\phi})\right)^{-1}, \tag{21}$$

where a involves d, and ϕ is the vector of EM parameter estimates. Then by the delta method,

$$Var[\hat{p}] = \hat{p}^4 \exp(-2a'\hat{\phi})a'Va, \tag{22}$$

where $V = Var[\hat{\phi}]$. Then the dose-response curves ($\hat{p}$ as a function of dosage d) for each genotype can be visualized, with $\pm$ 2 standard errors for approximate pointwise 95% confidence bounds. These can be seen in Figure 1.

From Figure 1 it can be seen that for genotypes -/B, -/H, +/H, and +/A, there are observed dosage levels across the regions of dosage where the dose-response curve moves from 0 to 1. For genotype -/A, the dose-response curve appears to "jump" between observed dosage levels, and for genotype +/B the dose response curve appears to "jump" at a single observed dosage level. For these two genotypes, it appears that perhaps an alternative parameterization to

the one suggested in Table 2 might be more descriptive, such as some kind of "cut-point" parameterization.

A 95% interval estimation of the LD50, the dosage required to achieve a mortality probability of 0.5, can be achieved for each genotype using either the asymptotic confidence intervals via the delta method or the Fieller intervals (Faraggi, Izikson, and Reiser 2003). For each genotype, the predicted mortality probability can be expressed as a function of dosage d:

$$\hat{p} = (1 + \exp(-\hat{\gamma_0} - \hat{\gamma_1} d))^{-1}, \tag{23}$$

where

$$\hat{\gamma} = A\hat{\phi} \tag{24}$$

for some genotype-specific matrix A. Then using z as a critical value from the standard normal distribution, the general LDp can be expressed as

$$LD_p = \frac{1}{\hat{\gamma}_1^2}\left(\log\left(\frac{p}{1-p}\right) - \hat{\gamma}_0\right). \tag{25}$$

By the delta method,

$$Var[LD_p] \approx \frac{1}{\hat{\gamma}}(1 \quad LD_p)\Sigma(1 \quad LD_p)', \tag{26}$$

where $\Sigma = Var[\hat{\gamma}] = AVA'$. Then the approximate $(1-\alpha)100\%$ confidence interval for the LD_p is

$$LD_p \pm z_{1-\alpha/2}\sqrt{Var[LD_p]}. \tag{27}$$

Using the same Σ as above, when $p = 0.5$ the non-symmetric $(1-\alpha)100\%$ Fieller interval is

$$\begin{aligned} &LD_p + \frac{C}{1-C}\left(LD_p + \frac{\Sigma_{1,2}}{\Sigma_{2,2}}\right) \\ &\pm\frac{z_{1-\alpha/2}}{\hat{\gamma}_1(1-C)}\sqrt{\Sigma_{1,1} + 2LD_p\Sigma_{1,2} + LD_p^2\Sigma_{2,2} - C\left(\Sigma_{1,1} - \frac{\Sigma_{1,2}^2}{\Sigma_{2,2}}\right)}, \end{aligned} \tag{28}$$

where $C = z_{1-\alpha/2}\Sigma_{2,2}/\hat{\gamma}_1$. In order for the Fieller intervals to be obtained, it is necessary that $C > 1$, which is equivalent to having the γ_1 term be statistically significant at the α level for the corresponding genotype.

Table 3 gives the genotype-specific models (the $\hat{\gamma}$ and corresponding Σ) as well as the LDp and corresponding intervals. Note that when the dosage effect is "nonsignificant" ($t < 1.96$), the Fieller approach gives no interval. It is important to note that "nonsignificance" can occur here due to large variance, as seen in the Σ matrices for genotypes -/A and +/B in

Table 3. In general, the interval estimates for the LD50 tend to agree with the graphical representation in Figure 1. However, the confidence interval for the -/A genotype's LD50 is much wider than the other genotypes', a result of the high variance associated with the dosage effect for that genotype, which further suggests an alternative parameterization (such as a cut-point model) for this genotype.

As seen in Figure 1, for the three genotypes -/A, +/B, and +/A the observed dosages did not occur at levels where the dose-response curve changes most rapidly. In addition, the observed dosages were insufficiently high to capture the rate parameter (the γ_1 term in equation 23) for genotype +/A. This lack of "support" for the portion of greatest change in the dose-response curve can lead to high variance of the estimates of dosage effect, resulting in the "non-significance" of the dosage effect estimates for these three genotypes. Another possible reason for the non-significance of dosage effect in the three genotypes -/A, +/B, and +/A is the non-monotonicity of the ratio of genotype survivors to total beetles at each dosage level (see Table 1). As dosage increases, this ratio is expected to decrease in a monotone fashion, but for these three genotypes it does not.

Because the +/A genotype has been observed to be the most resistant, its results deserve a final comment. It has been established by practical experience that there is a synergistic effect between the two markers (rp6.79, rp5.11) defining the six genotypes. This synergy or interaction causes the +/A genotype to be highly resistant, much more so than could be attributed to solely additive effects of the two markers. Based on the parameterization in Table 2, the null hypothesis of additive marker effects is $H_0 : g_{0,2} = g_{0,3} = \beta_{0,2} = \beta_{0,3} = 0$, or $H_0 : a'\phi = 0$ for a specific vector a. With these data, the value of the test statistic is $t = 0.14$, and the null hypothesis is not rejected. One possible reason for this failure to statistically detect the known interaction is the greater missingness of data for this +/A genotype – higher numbers of beetles were allocated to the higher dose levels with the sole purpose of observing surviving +/A beetles at those dose levels (see Table 1). The extent to which this greater missingness of data, the previously mentioned non-monotonicity, and the lack of "support" limit the analysis will be the subject of future work involving extensive simulations and alternative parameterizations.

4 Summary

To control costs, it is often desirable in agricultural settings to eliminate a pest such as the lesser grain borer using the lowest practical effective dose of a particular fumigant such as phosphine. In the application presented here, of interest was how beetle genotype affected resistance to phosphine. However, due to the high expense involved, only surviving beetles were genotyped. Due to the missing data, then, a simple dose-response model could not adequately address these data. The two important factors that allowed a useful application of the EM algorithm to these data were the missingness mechanism (specifically, genotype was "missing at random") and the availability of marginal information on the missing categorical

covariate (genotype was observed at the zero dosage level). The "direct calculation" approach (Oakes 1999) was found to be the most straightforward method for estimating the variance-covariance matrix of the final estimates from the EM algorithm.

The greater degree of data missingness for the most resistant genotype, along with non-monotonicity and lack of dosage "support" in two other genotypes, introduced greater variance for the corresponding parameter estimates, causing the statistical tests to fail to detect the marker synergy observed in practice. The main biological conclusion drawn from this analysis would be a recommendation to consider the dosages along the portions of greatest change in the dose-response curves of Figure 1 for each genotype. Additional study of other dosage levels (more "support") would add greater accuracy and resolution to the dose-response curve. This will be the subject of future work involving extensive simulations.

Alternative parameterizations (of mortality probability in terms of genotype and dosage effects) are possible, and a cut-point model might be reasonable for two of the genotypes. The methods described in this paper were implemented using both R (for the EM implementation) and Maple (for the symbolic differentiation required for the variance-covariance estimation). The adaptation of these methods to other similar data sets should be reasonably straightforward, and the R code and Maple worksheet are available from the authors.

Acknowledgments

Purdue University's Statistical Consulting Service supported J.R.S. during the early stages of this project. The experimental portion of the project was conducted by D.I.S. and colleagues, funded by the Grains Research and Development Corporation (GRS23), the Australian Research Council (00/ARC098G), the Queensland Department of Primary Industries, and Grainco Australia.

References

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Faraggi, D., P. Izikson, and B. Reiser (2003). Confidence intervals for the 50 percent response dose. *Statistics in Medicine 22*, 1977–1988.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York, NY.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association 85*(411), 765–769.

Meng, X.-L. and D. B. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the Americal Statistical Association 86*(416), 899–909.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B 61*(2), 479–482.

Schlipalius, D. I., Q. Cheng, P. E. B. Reilly, P. J. Collins, and P. R. Ebert (2002). Genetic linkage analysis of the lesser grain borer *Rhyzopertha dominica* identifies two loci that confer high-level resistance to the fumigant phosphine. *Genetics 161*, 773–782.

Thisted, R. A. (1988). *Elements of Statistical Computing*. Chapman and Hall, New York, NY.

Table 1: Distribution of numbers of surviving beetles at various genotypes and dose levels.

Phosphine Dosage (mg/L)	Total Receiving Dosage	Total Deaths	Total Survivors	Observed Survivors at Genotype					
				-/B	-/H	-/A	+/B	+/H	+/A
0	98	0	98	31	27	10	6	20	4
0.003	100	16	84	18	26	10	6	20	4
0.004	100	68	32	10	4	3	4	7	4
0.005	100	78	22	1	4	7	2	6	2
0.01	100	77	23	0	1	9	8	5	0
0.05	300	270	30	0	0	0	5	20	5
0.1	400	383	17	0	0	0	0	10	7
0.2	750	740	10	0	0	0	0	0	10
0.3	500	490	10	0	0	0	0	0	10
0.4	500	492	8	0	0	0	0	0	8
1	7,850	7,806	44	0	0	0	0	0	44
	10,798	10,420	378						

Table 2: Possible parameterization of $p_{i,j}$, the probability of death at dosage level j for genotype i, with the logit link $\log \frac{p_{i,j}}{1-p_{i,j}} = \eta_{i,j}$ in terms of the vector $\vartheta = (\mu, g_0, g_2, g_3, g_{0,2}, g_{0,3}, \alpha, \beta_0, \beta_2, \beta_3, \beta_{0,2}, \beta_{0,3})$. Here, d_j corresponds to dosage level j.

Genotype	Parameterization											
-/B:	$\eta_{1,j}$	$=$	$(\mu$				$)$	$+$	$(\alpha$			$)\, d_j$
-/H:	$\eta_{2,j}$	$=$	$(\mu$		$+\, g_2$		$)$	$+$	$(\alpha$		$+\, \beta_2$	$)\, d_j$
-/A:	$\eta_{3,j}$	$=$	$(\mu$		$+\, g_3$		$)$	$+$	$(\alpha$		$+\, \beta_3$	$)\, d_j$
+/B:	$\eta_{4,j}$	$=$	$(\mu$	$+\, g_0$			$)$	$+$	$(\alpha$	$+\, \beta_0$		$)\, d_j$
+/H:	$\eta_{5,j}$	$=$	$(\mu$	$+\, g_0$	$+\, g_2$	$+\, g_{0,2}$	$)$	$+$	$(\alpha$	$+\, \beta_0$	$+\, \beta_2$	$+\, \beta_{0,2}$ $)\, d_j$
+/A:	$\eta_{6,j}$	$=$	$(\mu$	$+\, g_0$	$+\, g_3$	$+\, g_{0,3}$	$)$	$+$	$(\alpha$	$+\, \beta_0$	$+\, \beta_3$	$+\, \beta_{0,3}$ $)\, d_j$

Table 3: Summary of results. $\hat{P}_i$ is the final EM estimate of the proportion of each genotype in the population. $\hat{\gamma}$ is the vector of genotype-specific parameter estimates to represent mortality probability in terms of dosage, as in equation 23. Σ is the estimated covariance matrix of $\hat{\gamma}$. The test statistic $t = \hat{\gamma}_1/\sqrt{\Sigma_{2,2}}$ tests for significance of dosage effect. The LD50 point and 95% interval estimates are also given for each genotype. Note that when the dosage effect is "nonsignificant" ($t < 1.96$), the Fieller approach gives no interval.

					LD50		
Genotype	$\hat{P}_i$	$\hat{\gamma}$	Σ	t	point	95% Confidence	95% Fieller
-/B	0.362	$\begin{pmatrix} -8.40 \\ 2411.99 \end{pmatrix}$	$\begin{pmatrix} 5.95 & -1452.99 \\ -1452.99 & 363470.80 \end{pmatrix}$	4.00	0.0035	(0.0031, 0.0039)	(0.0028, 0.0038)
-/H	0.366	$\begin{pmatrix} -4.36 \\ 1325.36 \end{pmatrix}$	$\begin{pmatrix} 1.05 & -261.05 \\ -261.05 & 70774.20 \end{pmatrix}$	4.98	0.0033	(0.0028, 0.0038)	(0.0027, 0.0037)
-/A	0.079	$\begin{pmatrix} -28.43 \\ 981.77 \end{pmatrix}$	$\begin{pmatrix} 12501105.30 & -93828.16 \\ -93828.16 & 58208692.65 \end{pmatrix}$	0.13	0.0290	(-7.1862,7.2441)	NA
+/B	0.053	$\begin{pmatrix} -24.78 \\ 511.68 \end{pmatrix}$	$\begin{pmatrix} 80149.44 & -1602979.00 \\ -1602979.00 & 32059577.00 \end{pmatrix}$	0.09	0.0484	(0.0135, 0.0833)	NA
+/H	0.118	$\begin{pmatrix} -2.98 \\ 44.93 \end{pmatrix}$	$\begin{pmatrix} 1.36 & -11.50 \\ -11.50 & 114.51 \end{pmatrix}$	4.20	0.0664	(0.0407, 0.0921)	(0.0256, 0.0873)
+/A	0.021	$\begin{pmatrix} -2.82 \\ 3.82 \end{pmatrix}$	$\begin{pmatrix} 10.08 & -8.85 \\ -8.85 & 7.88 \end{pmatrix}$	1.36	0.7382	(0.1407, 1.3356)	NA

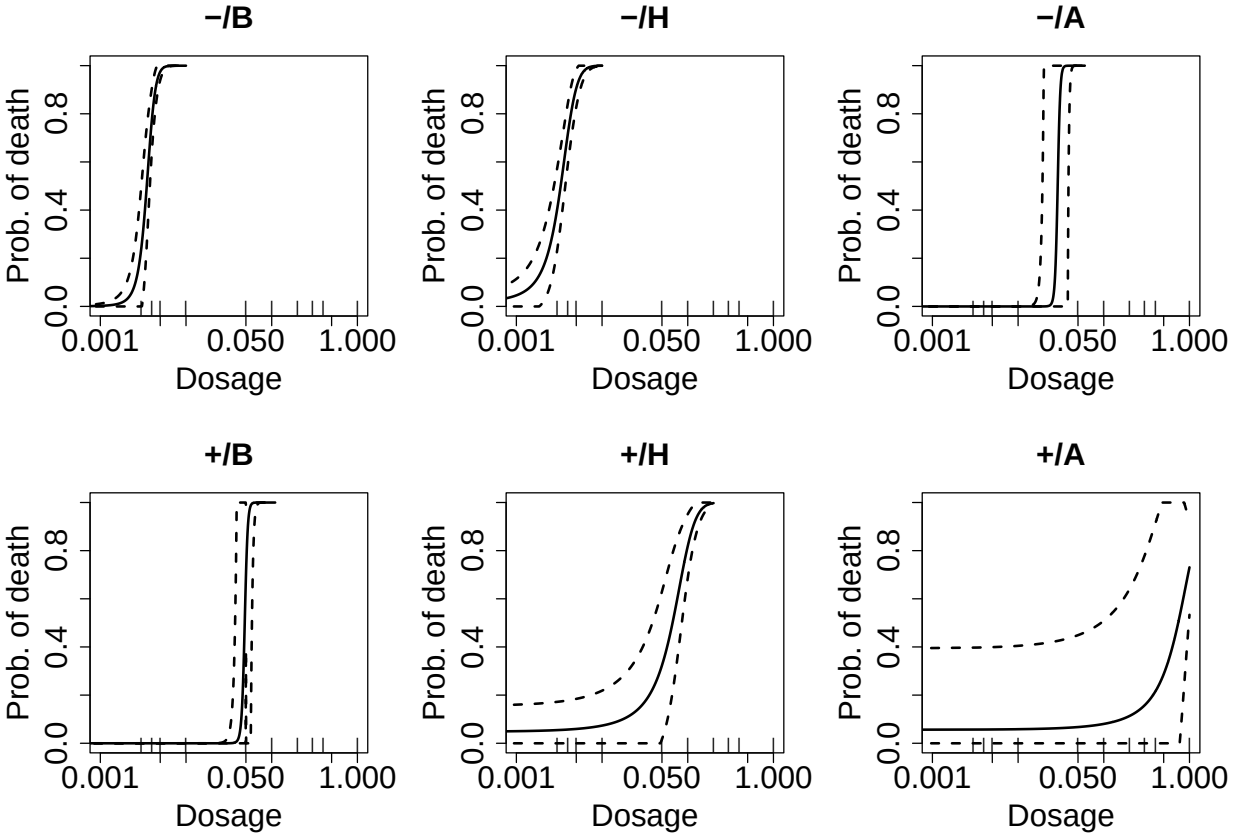

Figure 1: Estimated dose-response curves, with pointwise approximate 95% confidence intervals. The small hash marks above the dosage axis indicate dosage levels observed in the experiment. Because the dosage axis is on the log scale here, the zero dose hash mark does not appear. The apparent erratic behavior of the lower bound for the +/B genotype at dosage 0.05 mg/L is due to the dramatic jump in the response curve at that dosage. The curves for genotypes -/A and +/B suggest that alternative parameterizations may be considered, such as a cut-point model.