

Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study

Daniel B. Hall

Department of Statistics, University of Georgia, Athens, Georgia 30602-1952, U.S.A.
email: dhall@stat.uga.edu

SUMMARY. In a 1992 *Technometrics* paper, Lambert (1992, **34**, 1–14) described zero-inflated Poisson (ZIP) regression, a class of models for count data with excess zeros. In a ZIP model, a count response variable is assumed to be distributed as a mixture of a Poisson(λ) distribution and a distribution with point mass of one at zero, with mixing probability p . Both p and λ are allowed to depend on covariates through canonical link generalized linear models. In this paper, we adapt Lambert's methodology to an upper bounded count situation, thereby obtaining a zero-inflated binomial (ZIB) model. In addition, we add to the flexibility of these fixed effects models by incorporating random effects so that, e.g., the within-subject correlation and between-subject heterogeneity typical of repeated measures data can be accommodated. We motivate, develop, and illustrate the methods described here with an example from horticulture, where both upper bounded count (binomial-type) and unbounded count (Poisson-type) data with excess zeros were collected in a repeated measures designed experiment.

KEY WORDS: Excess zeros; EM algorithm; Generalized linear mixed model; Heterogeneity; Mixed effects; Overdispersion; Repeated measures.

1. Introduction

Count data with many zeros are common in a wide variety of disciplines. Recently, a fair amount of statistical methodology has been developed to deal with such data. Ridout, Demétrio, and Hinde (1998) review this literature and cite examples from agriculture, econometrics, manufacturing, patent applications, road safety, species abundance, medicine, use of recreational facilities, and sexual behavior. In the current paper, we consider an example from horticulture and use it to motivate adaptations of Lambert's (1992) zero-inflated Poisson (ZIP) regression models. This example is described in Section 2. We review ZIP regression in Section 3, and we introduce zero-inflated binomial (ZIB) regression models in Section 4. To accommodate the repeated measures features of the example data set, it is useful to incorporate random effects into these models. The resulting mixed versions of the ZIP and ZIB models are introduced in Section 5, including a discussion of maximum likelihood estimation of these models via the EM algorithm. In Section 6, we return to the example data set to illustrate the new methods of this paper. A brief summary is included as Section 7.

2. The Motivating Example

Zero-runoff subirrigation systems are commonly used to irrigate greenhouse crops. In such a system, water and fertilizer solution are made available to the plant through the soil (or other growing medium) and excess is allowed to drain back into a holding tank for reuse. van Iersel, Oetting, and Hall (2000) investigated the use of subirrigation systems to ap-

ply systemic pesticides. These authors report the results of an experiment in which the insecticide imidacloprid was applied to poinsettia plants to control silverleaf whiteflies. The treatments considered in this study included four methods of subirrigated imidacloprid (labeled 0, 1, 2 and 4, corresponding to application following 0, 1, 2, and 4 days, respectively, without water), a standard treatment (H) in which the pesticide is applied via hand watering of the top of the soil, and a control treatment (C) in which no pesticide was applied. These treatments were applied in a randomized complete block design with repeated measures over 12 consecutive weeks. The experimental unit in this study was a trio of poinsettia plants, and 18 such units (54 plants) were randomized to the six treatments in three complete blocks.

Because efficacy of the pesticide is determined by its ability both to kill mature whiteflies and to suppress reproduction, two response variables were measured in this experiment. To measure lethality, at weekly intervals n (mean = 9.5, SD = 1.7), adult whiteflies were placed in clip-on leaf cages attached to one leaf per plant. The number of surviving whiteflies measured 2 days later constitutes the first of the two response variables in this study. To measure reproductive inhibition, the fly cages were removed after the survival count was obtained, but the position of each cage was marked. Three weeks after each survival count was taken, the number of immature whiteflies in the marked location was determined. Thus, number of immature insects forms the second of the two responses in the study. Although the design originally called for 648 observations in a balanced design, one observation was lost on

Table 1
Fits of some fixed effects Poisson regression models for number of immatures measured at the experimental unit level

Highest term in model	Log likelihood	Deviance	Residual degrees of freedom	BIC ^a
No interactions	-1040.48	1458.29	196	-1094.2
Block × trt	-1001.59	1380.51	186	-1082.2
Block × week	-997.37	1372.06	174	-1110.3
Block × trt + block × week	-959.65	1296.63	164	-1099.4
Trt × week	-549.52	476.35	141	-751.1
Block × trt + trt × week	-512.48	402.28	131	-741.0
Block × week + trt × week	-507.72	392.75	119	-768.4
Block × trt + block × week + trt × week	-468.93	315.18	109	-756.5

^a Larger-is-better form.

each of six plants and two observations were lost on one plant, to yield a final data set with $N = 640$ observations.

A natural approach to analyzing these data would be to fit a generalized linear model (GLM) to each response. For example, let $I_{ijk\ell}$ be the number of immatures measured on plant k ($k = 1, \dots, 3$) in treatment i ($i = 1, \dots, 6$) in block j ($j = 1, \dots, 3$) measured at time ℓ ($\ell = 1, \dots, 12$). At first glance, one might expect these data to be modeled appropriately by a log-linear repeated measures analysis of variance (split-plot-type) model, i.e.,

$$\log(\lambda_{ij\cdot\ell}) = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \delta_\ell + (\tau\delta)_{i\ell} + \beta \log(n_{ij\cdot\ell}), \tag{1}$$

where $I_{ij\cdot\ell} = \sum_k I_{ijk\ell}$, the response at the experimental unit level, is assumed to follow a Poisson distribution with mean $\lambda_{ij\cdot\ell}$. In this model, τ_i is the i th treatment effect, α_j is the j th block effect, $(\tau\alpha)_{ij}$ is an interaction effect for treatment i combined with block j , δ_ℓ is the ℓ th week effect, and $(\tau\delta)_{i\ell}$ is an interaction effect for the i th treatment in week ℓ . Here we have also included a covariate, $\log(n_{ij\cdot\ell})$, the natural logarithm of $n_{ij\cdot\ell}$, the total number of insects placed previously upon the three leaves measured in treatment i , block j , and at time ℓ , i.e., $n_{ijk\ell}$ is the number of adult insects from which the $I_{ijk\ell}$ immatures are offspring. For β fixed at one, this final term in model (1) is what is known as an offset in the GLM literature. However, in the models considered in this paper, the coefficient on $\log(n)$ is considered an unknown parameter to be estimated in the model-fitting procedure.

Table 1 summarizes the fit of this model and several other fixed effects log-linear models for I . From Table 1, we see that the minimum deviance model is the no-three-way interaction model in which $(\alpha\delta)_{j\ell}$, a block by week interaction term, is added to (1). According to a model-selection criterion such as Schwarz's (1978) Bayesian information criterion (BIC), this model is overfit; but even this most complex model fits poorly, yielding a deviance nearly three times its residual degrees of freedom. This conclusion is supported by an examination of the quality of the predictions generated by this model. Table 2 presents observed and predicted values for the percentage of the 216 immature insect counts that were equal to 0, 1, ..., 5 or in the ranges 6-10, 11-15, ..., 101-150.

From Table 2, we see that the model fits poorly, particularly for $k = 0$ and $k = 1$. The pattern of the lack of fit here also indicates that the data are not simply overdispersed with respect to the Poisson distribution but instead contain far too many zero counts to be Poisson. The model predicts too few zeros and too many ones. Traditional methods for handling overdispersed Poisson, data such as a quasiliikelihood model in which we allow a nonunity scale parameter ϕ in the Poisson variance, $\text{var}(I) = \phi\lambda$, or a negative binomial model in which $\text{var}(I) = \lambda + \alpha\lambda^2$, account for extra-Poisson variance but cannot compensate for systematic departures from the first moment model as we have here.

A similar problem is encountered when analyzing the other study outcome, number of live insects, L . A natural approach to handling the data on this response is to assume a binomial distribution for the number of live insects and then fit a logistic regression model of a similar form to (1); i.e., we assume $L_{ij\cdot\ell} \sim \text{binomial}(n_{ij\cdot\ell}, \pi_{ij\cdot\ell})$, where $n_{ij\cdot\ell}$ is the total number of insects at risk of being killed by the pesticide at time ℓ on experimental unit ij , $\pi_{ij\cdot\ell}$ is the common survival probability

Table 2
Observed values and predictions for the percentage of counts equal to k based on the no-three-way interaction model from Table 1

k	Observed	Predicted	Difference
0	35.19	29.04	6.15
1	2.78	8.83	-6.05
2	5.56	6.05	-0.50
3	3.24	4.04	-0.80
4	4.63	2.94	1.69
5	2.31	2.41	-0.10
6-10	11.11	9.76	1.35
11-15	5.56	7.21	-1.66
16-20	3.70	5.05	-1.35
21-25	3.70	2.94	0.76
26-50	8.80	7.87	0.92
51-750	7.87	7.43	0.44
76-100	3.24	3.86	-0.62
101-150	2.31	2.38	-0.07

Table 3

Observed values and predictions for the percentage of live insects equal to k based on the minimum deviance binomial logistic regression model

k	Observed	Predicted	Difference
0	25.46	18.92	6.55
1	10.19	13.21	-3.03
2	8.33	10.90	-2.57
3	6.94	8.85	-1.91
4	7.87	7.13	0.74
5	4.17	5.72	-1.55
6-10	15.27	15.10	0.17
11-15	4.63	3.41	1.22
16-20	3.70	2.18	1.52
21-25	4.17	5.13	-0.96
26-30	6.48	8.19	-1.71
31-35	2.78	1.16	1.62

ty for each of those $n_{ij,\ell}$ insects, and we model π as $\text{logit}(\pi_{ij,\ell}) = \eta_{ij,\ell}$, where $\eta_{ij,\ell}$ is a linear predictor involving main effects and interactions among the factors trt (treatment), block, and week. Again, models of this form fit poorly. The minimum deviance unsaturated model was, again, the no-three-way interaction model, which yielded a deviance of 438.43 on 110 residual degrees of freedom. Predictions from this model are summarized in Table 3. As in the Poisson case, this model underpredicts the percentage of zero counts and overpredicts ones, twos, and threes. The pattern of these prediction errors once again suggests an overabundance of zeros as the reason for lack of fit here.

The abundance of zeros in the response variables of this study rules out one tempting, simple alternative to the fixed effects GLMs considered thus far, i.e., to transform the response variables to normality and fit mixed effects models. For example, a repeated measures analysis of variance model similar in form to the right-hand side of (1) could be fit to the mean of a suitably chosen power-family (say) transformation of the response, with block effects assumed random and the treatment by block effects replaced by a random whole-plot error term. Given that more than one third of the $I_{ij,\ell}$'s are equal to zero, the problem with this approach is obvious: while a transformation may normalize the distribution of the nonzero responses, no transformation has the ability to spread out the zeros. After transformation, the very high frequency of zeros in the data will simply be replaced by an equally high frequency of the transformation of zero. The situation is nearly as bad for $L_{ij,\ell}$, the number of live insects quantified at the experimental unit level, for which over a quarter of the observations are zero, and even worse if an analysis is attempted for the plant-specific counts, I_{ijkl} and L_{ijkl} . At the plant level, over 50% of the observations of L are equal to zero and just under 50% of the observations of I are equal to zero.

Even without a normalizing transformation, a mixed model approach can be attempted by utilizing generalized linear mixed models (GLMMs). However, an extremely high frequency at zero combined with substantial frequencies for nonzero counts is no more a feature of the discrete data distributions such as the binomial and Poisson on which GLMMs are

based than it is a feature of the normal distribution on which linear mixed models are based. Therefore, GLMMs can be expected to provide limited improvements in fit. Instead, we pursue an approach based on models that explicitly account for a high frequency at zero by mixing discrete distributions with the degenerate distribution with point mass of one at zero. As will be described in subsequent sections, such models, especially when adapted to accommodate random as well as fixed effects, provide substantial improvements in quality of fit over GLM and GLMM alternatives.

3. ZIP Regression

An examination of the data and consideration of the results of the previous section suggest that the treatments under consideration in this experiment may affect the leaves of the poinsettia plant in two distinct ways, producing in some leaves a complete suppression of insect reproduction and survival and allowing in other leaves a (possibly reduced) suitability for reproduction and survival; i.e., the process generating the data has two states, a zero-state from which only zero values are observed and a Poisson (in the case of response variable I) or binomial (in the case of L) state from which all of the nonzero values and a few of the zero values are observed. Mixtures of the Poisson distribution with zero have been considered by several authors, including Yip (1988), who described a number of insects per leaf application of such a mixture, and Lambert (1992) and Heilbron (1989, 1994), who presented regression models for counts based on mixtures of zero and various distributions. Here we adopt the zero-inflated-Poisson, or ZIP, regression models introduced by Lambert (1992) for the analysis of our response I . We also introduce zero-inflated-binomial, or ZIB, models for the analysis of L .

For a vector of responses $\mathbf{Y} = (Y_1, \dots, Y_N)^T$, Lambert (1992) discusses the ZIP regression model, in which

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i; \\ \text{Poisson}(\lambda_i), & \text{with probability } 1 - p_i. \end{cases} \quad (2)$$

In addition, the parameters $\mathbf{p} = (p_1, \dots, p_N)^T$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N)^T$ are modeled via canonical link GLMs as $\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta}$ and $\text{logit}(\mathbf{p}) = \mathbf{G}\boldsymbol{\gamma}$ for design matrices \mathbf{B} and \mathbf{G} . Although this model consists of two distinct parts, the model components must be fit simultaneously. Heilbron (1994) proposes an alternative two-part model that allows the two model components to be fit separately to the subset of positive responses and to the entire response vector dichotomized as an indicator variable for whether or not the response is positive. Although the separate fitting feature of Heilbron's model is a convenience that can be useful for model selection and for selecting a mixing distribution other than the Poisson, we prefer Lambert's model for the present application because of its interpretability and because of the suitability of the Poisson distribution in this case.

In Lambert's ZIP model, the design matrices \mathbf{G} and \mathbf{B} contain potentially different sets of experimental factor and covariate effects that pertain to the probability of the zero state (corresponding to the pesticide being fully effective) and the Poisson mean in the nonzero state (pesticide not fully effective), respectively. Therefore, the $\boldsymbol{\gamma}$'s have interpretations in terms of a covariate or factor level's effect on the probability of the pesticide being fully effective and the $\boldsymbol{\beta}$'s have interpretations in terms of the effect on the mean number of immature

insects produced when the pesticide is less than fully effective. Both sets of parameters, β and γ , pertain to the mean number of immature insects produced marginally, i.e., unconditional on the pesticide's effectiveness state. A simple consequence of the ZIP model (2) is that, marginally, $E(Y_i) = (1 - p_i)\lambda_i$, so that the mean number of immatures under the experimental conditions under which Y_i is generated depends on γ through p_i and on β through λ_i . Therefore, for a covariate that occurs in both \mathbf{G} and \mathbf{B} , hypothesis tests, say, on that covariate's γ and β values are scientifically meaningful when conducted separately or simultaneously.

As described in Lambert's (1992) paper, the ZIP model (2) can be fit using maximum likelihood via the EM algorithm. The log likelihood for regression parameters γ and β based on all of the data is given by

$$\ell(\gamma, \beta; \mathbf{y}) = \sum_{i=1}^N \left\{ u_i \log \left[e^{\mathbf{G}_i \gamma} + \exp \left(-e^{\mathbf{B}_i \beta} \right) \right] + (1 - u_i) \left(y_i \mathbf{B}_i \beta - e^{\mathbf{B}_i \beta} \right) - \log \left(1 + e^{\mathbf{G}_i \gamma} \right) - (1 - u_i) \log(y_i!) \right\},$$

where \mathbf{G}_i and \mathbf{B}_i are the i th rows of \mathbf{G} and \mathbf{B} and $u_i = 1$ if $y_i = 0$ and $u_i = 0$ otherwise. The missing data in this problem is a vector of indicator variables $\mathbf{z} = (z_1, \dots, z_N)^T$, where $z_i = 1$ when Y_i is from the zero state and $z_i = 0$ when Y_i is from the Poisson state. The complete-data log likelihood is then $\ell_c(\gamma, \beta; \mathbf{y}, \mathbf{z}) = \ell_c(\gamma; \mathbf{y}, \mathbf{z}) + \ell(\beta; \mathbf{y}, \mathbf{z})$, where $\ell_c(\gamma; \mathbf{y}, \mathbf{z}) = \sum_i [z_i \mathbf{G}_i \gamma - \log(1 + e^{\mathbf{G}_i \gamma})]$ and $\ell(\beta; \mathbf{y}, \mathbf{z}) = \sum_i (1 - z_i) [y_i \mathbf{B}_i \beta - e^{\mathbf{B}_i \beta} - \log(y_i!)]$. Notice that $\ell_c(\gamma, \beta; \mathbf{y}, \mathbf{z})$ has a particularly convenient form for the EM algorithm; i.e., $\ell_c(\gamma, \beta; \mathbf{y}, \mathbf{z})$ is linear in \mathbf{z} , so that at iteration $(r + 1)$ of the algorithm, the E step consists of replacing \mathbf{z} by its conditional expectation given \mathbf{y} , $\gamma^{(r)}$, and $\beta^{(r)}$. This conditional expectation is easily calculated as

$$z_i^{(r)} = \begin{cases} [1 + \exp(-\mathbf{G}_i \gamma^{(r)} - e^{\mathbf{B}_i \beta^{(r)}})]^{-1} & \text{if } y_i = 0; \\ 0 & \text{if } y_i > 0. \end{cases}$$

With this substitution, $\ell_c(\gamma, \beta; \mathbf{y}, \mathbf{z}^{(r)})$ is easily maximized with respect to γ and β because it is equal to the sum of the log likelihood for an unweighted binomial logistic regression of $\mathbf{z}^{(r)}$ on \mathbf{G} (a term not involving β) and the log likelihood for a weighted Poisson log-linear regression of \mathbf{y} on \mathbf{B} , with weights $\mathbf{1}_{N \times 1} - \mathbf{z}^{(r)}$ (a term not involving γ). In Lambert's (1992) paper, the M step for γ is accomplished via an equivalent procedure that involves fitting a weighted logistic regression based on an augmented data set. We find this approach unnecessarily complicated and prefer solving the equivalent unweighted logistic regression problem.

Asymptotic variance-covariance matrices for the parameter estimates can be estimated using the inverse of the observed Fisher information matrix, and inference can be performed using likelihood ratio tests and confidence intervals. Based on Lambert's (1992) simulation results, it appears that normal theory (Wald) tests and confidence intervals work well for β but are unreliable for γ for values of N as large as 100. Various other issues associated with these models, including choice of starting values, model interpretation, asymptotic distribu-

tion of ZIP parameter estimators, and convergence of the EM algorithm, are discussed in Lambert's (1992) paper.

4. ZIB Regression

The number of live adult insects, $L_{ijk\ell}$, measured on plant k at occasion ℓ under experimental conditions i, j is a count bounded between zero and $n_{ijk\ell}$. The presence of an upper bound, $n_{ijk\ell}$, on this response suggests a model based on the binomial rather than the Poisson distribution. Lambert's ZIP regression techniques are easily adapted to yield a zero-inflated binomial model.

We again denote our response vector as $\mathbf{Y} = (Y_1, \dots, Y_N)^T$, where now we assume

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{binomial}(n_i, \pi_i), & \text{with probability } 1 - p_i. \end{cases} \quad (3)$$

This model implies

$$Y_i = \begin{cases} 0, & \text{with probability } p_i + (1 - p_i)(1 - \pi_i)^{n_i}; \\ k, & \text{with probability } (1 - p_i) \binom{n_i}{k} \pi_i^k (1 - \pi_i)^{n_i - k}, \\ & k = 1, \dots, n_i, \end{cases}$$

which leads to moments $E(Y_i) = (1 - p_i)n_i\pi_i$ and $\text{var}(Y_i) = (1 - p_i)n_i\pi_i[1 - \pi_i(1 - p_i n_i)]$. The parameters $\mathbf{p} = (p_1, \dots, p_N)^T$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^T$ are modeled via logit link GLMs as $\text{logit}(\boldsymbol{\pi}) = \mathbf{B}\beta$ and $\text{logit}(\mathbf{p}) = \mathbf{G}\gamma$. The log likelihood for the ZIB model is

$$\begin{aligned} \ell(\gamma, \beta; \mathbf{y}) &= \sum_{i=1}^N \left\{ u_i \log \left[e^{\mathbf{G}_i \gamma} + \left(1 + e^{\mathbf{B}_i \beta} \right)^{-n_i} \right] - \log \left(1 + e^{\mathbf{G}_i \gamma} \right) + (1 - u_i) \right. \\ &\quad \left. \times \left[y_i \mathbf{B}_i \beta - n_i \log \left(1 + e^{\mathbf{B}_i \beta} \right) + \log \binom{n_i}{y_i} \right] \right\}. \end{aligned} \quad (4)$$

ML estimates of γ and β can be obtained via the EM algorithm in a manner similar to that described above for the ZIP model. Define the unobserved random variable $Z_i = 1$ when Y_i is generated from the zero state and $Z_i = 0$ when Y_i comes from the binomial state. If we could observe $\mathbf{z} = (z_1, \dots, z_N)^T$, then the complete-data (\mathbf{y}, \mathbf{z}) log likelihood would be

$$\begin{aligned} \ell_c(\gamma, \beta; \mathbf{y}, \mathbf{z}) &= \log \prod_{i=1}^N \Pr(Y_i = y_i, Z_i = z_i) \\ &= \sum_{i=1}^N \left[z_i \mathbf{G}_i \gamma - \log \left(1 + e^{\mathbf{G}_i \gamma} \right) \right] \\ &\quad + \sum_{i=1}^N (1 - z_i) \left[y_i \mathbf{B}_i \beta - n_i \log \left(1 + e^{\mathbf{B}_i \beta} \right) + \log \binom{n_i}{y_i} \right] \\ &= \ell_c(\gamma; \mathbf{y}, \mathbf{z}) + \ell_c(\beta; \mathbf{y}, \mathbf{z}). \end{aligned}$$

The EM algorithm can be used to maximize $\ell(\gamma, \beta; \mathbf{y})$ by alternating between an E step, in which the missing data, \mathbf{z} , is estimated by its expectation under the current estimates of

(γ, β) , and a maximization step, in which $\ell_c(\gamma, \beta)$ evaluated at the current (fixed) estimate of \mathbf{z} is maximized with respect to both γ and β . As in the ZIP regression model, this procedure is particularly convenient because $\ell_c(\gamma, \beta; \mathbf{y}, \mathbf{z})$ is linear in \mathbf{z} and also splits into a sum of two exponential family (in this case, binomial) log likelihoods, each of which depends on only one of the regression parameters γ and β .

In more detail, the EM algorithm begins with starting values $(\gamma^{(0)}, \beta^{(0)})$ and proceeds iteratively. At iteration $r + 1$, we have the following three steps:

(1) *E step.* Estimate Z_i by its conditional mean $Z_i^{(r)} = E(Z_i | y_i, \gamma^{(r)}, \beta^{(r)})$ under current estimates of the regression parameters. This expectation is given by

$$\begin{aligned} Z_i^{(r)} &= \Pr(\text{zero state} | y_i, \gamma^{(r)}, \beta^{(r)}) \\ &= \Pr(y_i | \text{zero state}) \Pr(\text{zero state}) \\ &\quad \div \left[\Pr(y_i | \text{zero state}) \Pr(\text{zero state}) \right. \\ &\quad \left. + \Pr(y_i | \text{binomial}) \Pr(\text{binomial}) \right] \\ &= \begin{cases} \left[1 + e^{-\mathbf{G}_i \gamma^{(r)}} \left(1 + e^{\mathbf{B}_i \beta^{(r)}} \right)^{-n_i} \right]^{-1}, & \text{if } y_i = 0; \\ 0, & \text{if } y_i > 0. \end{cases} \end{aligned}$$

(2) *M step for γ .* Find $\gamma^{(r+1)}$ by maximizing $\ell_c(\gamma; \mathbf{y}, \mathbf{Z}^{(r)})$. This can be accomplished by performing an unweighted binomial logistic regression of $\mathbf{Z}^{(r)}$ on design matrix \mathbf{G} using a binomial denominator of one for each observation.

(3) *M step for β .* Find $\beta^{(r+1)}$ by maximizing $\ell_c(\beta; \mathbf{y}, \mathbf{Z}^{(r)}) = \sum_i (1 - Z_i^{(r)}) [y_i \mathbf{B}_i \beta - n_i \log(1 + e^{\mathbf{B}_i \beta}) + \log \binom{n_i}{y_i}]$. This can be done via a weighted logistic regression with weights $(1 - Z_i^{(r)})$, $i = 1, \dots, N$, and binomial error distribution with denominators n_1, \dots, n_N .

Good starting values for β in the EM algorithm can be obtained by maximizing the positive part binomial log likelihood as

$$\begin{aligned} \ell_+(\beta; \mathbf{y}_+) &= \sum_{i=1}^N (1 - u_i) \left\{ y_i \mathbf{B}_i \beta - n_i \log(1 + e^{\mathbf{B}_i \beta}) \right. \\ &\quad \left. - \log \left[1 - \left(1 + e^{\mathbf{B}_i \beta} \right)^{-n_i} \right] + \log \binom{n_i}{y_i} \right\}. \end{aligned}$$

Here, \mathbf{y}_+ is the vector of positive elements of \mathbf{y} and \mathbf{B}_+ is the matrix consisting of the rows of \mathbf{B} corresponding to positive elements of \mathbf{y} . Maximization of ℓ_+ can be done using iteratively reweighted least squares (see Green, 1984). The score equations based on ℓ_+ are given by

$$\sum_{i=1}^N \mathbf{B}_{+i}^T \left\{ y_i - n_i \pi_i [1 - (1 - \pi_i)^{n_i}]^{-1} \right\} = \mathbf{0},$$

and its Hessian is

$$-\sum_{i=1}^N \mathbf{B}_{+i}^T \mathbf{B}_{+i} n_i \pi_i (1 - \pi_i)$$

$$\times \left[1 - \frac{(1 - \pi_i)^{n_i - 1} \{ n_i \pi_i [1 - (1 - \pi_i)^{n_i}]^{-1} - (1 - \pi_i) \}}{1 - (1 - \pi_i)^{n_i}} \right].$$

As in the ZIP model, the starting value chosen for γ has been less important in our experience. Following Lambert (1992), we suggest an initial value for the intercept in γ equal to the observed average probability of an excess zero, or $\hat{p}_0 = [\#(y_i = 0) - \sum_{i=1}^N (1 - \pi_i)^{n_i}] / N$, and an initial value of zero for the remaining elements of γ .

Convergence of the EM algorithm in this problem follows from arguments similar to those given by Lambert (1992, Appendix A.1). Alternative algorithms, such as Newton–Raphson or Newton–Raphson with Fisher scoring, for maximizing the log likelihood (equation (4)) can be used in this problem. However, the EM algorithm is simpler to program, especially if GLM fitting routines are available. In addition, the EM algorithm and its extensions (e.g., Monte Carlo EM; McCulloch, 1997) have been useful for fitting models with random effects, and we make such use of the algorithm in Section 5. Our experience regarding the convergence of the EM and Newton–Raphson algorithms for fitting the zero-inflated models discussed in this paper is in agreement with Lambert’s summarizing statement that “In short, the ZIP . . . regressions were not difficult to fit” (Lambert, 1992, p. 6).

5. ZIP and ZIB Regression with Random Effects

The best ZIP and ZIB models that we fit to the whitefly data model the data much more closely than corresponding generalized linear models (as in Section 2). For example, based on BIC, the best fitting ZIB model that we fit to the number of live insects response was

$$\text{logit}(p) = \mu + \text{trt} + \text{block} + \text{week}$$

and

$$\text{logit}(\pi) = \mu + \text{trt} + \text{block} + \text{week} + \text{trt} \times \text{block} + \text{trt} \times \text{week}.$$

This model yielded a maximized log likelihood value of -851.6 on 537 residual degrees of freedom and a BIC of -1184.4 . More importantly, it fit the data much better than any fixed-effects logistic regression model. Table 4 displays observed and predicted counts from this model. A comparison with the

Table 4
Observed values and predictions for the percentage of counts equal to k based on the fitted ZIB model

k	Observed	Predicted	Difference
0	52.97	53.20	-0.23
1	8.91	7.08	1.82
2	7.81	6.52	1.30
3	5.63	5.64	-0.01
4	2.66	4.45	-1.79
5	1.25	3.44	-2.19
6	4.38	3.04	1.34
7	2.81	2.69	0.12
8	2.34	3.24	-0.89
9	1.25	4.09	-2.84
10	7.97	5.38	2.58
11	1.25	0.96	0.29
12	0.63	0.27	0.35
13	0.16	0.02	0.14

results of Table 3 reveals a substantial improvement when we allow a mixture of the binomial distribution with zero.

However, these fixed effects models are based on the assumption of independence among the responses. In a repeated measures design such as that of the whitefly data, such an assumption is clearly violated. While there may be independence from plant to plant, there almost certainly is correlation among repeated observations on the same plant. Neither the ZIP nor ZIB models make allowances for this dependence, nor do they separate plant-to-plant heterogeneity from the residual variance (at the experimental unit level) remaining after accounting for design factors. Such inadequacies in the model can seriously affect the validity of statistical inference.

Recently, many researchers have incorporated random effects into a wide variety of regression models to account for correlated responses and multiple sources of variance. We propose this approach in the zero-inflated Poisson and binomial models discussed earlier in the paper. In particular, we consider models in which a random intercept is added to the exponential family portion of the model.

5.1 ZIP Regression with a Random Intercept

Suppose our response vector \mathbf{Y} contains data from K independent clusters so that $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_K^T)^T$, where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT_i})^T$. We assume that, conditional on a random effect b_i ,

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } p_{ij}; \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } (1 - p_{ij}), \end{cases}$$

where we model $\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{iT_i})^T$ and $\mathbf{p}_i = (p_{i1}, \dots, p_{iT_i})^T$ with log-linear and logistic regression models

$$\begin{aligned} \log(\boldsymbol{\lambda}_i) &= \mathbf{B}_i\boldsymbol{\beta} + \sigma b_i \quad \text{and} \\ \text{logit}(\mathbf{p}_i) &= \mathbf{G}_i\boldsymbol{\gamma}, \quad i = 1, \dots, K. \end{aligned}$$

Here, $\mathbf{B} = (\mathbf{B}_1^T, \dots, \mathbf{B}_K^T)^T$ and $\mathbf{G} = (\mathbf{G}_1^T, \dots, \mathbf{G}_K^T)^T$ are design matrices, and we assume b_1, \dots, b_K are independent standard normal random variables.

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T, \sigma)^T$ be the combined parameter vector. The log likelihood for the ZIP model with random intercept is

$$\ell(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i=1}^K \log \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{T_i} \Pr(Y_{ij} = y_{ij} \mid b_i) \right] \phi(b_i) db_i, \tag{5}$$

where

$$\begin{aligned} \Pr(Y_{ij} = y_{ij} \mid b_i) &= \left[p_{ij} + (1 - p_{ij})e^{-\lambda_{ij}} \right]^{u_{ij}} \left[\frac{(1 - p_{ij})e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!} \right]^{1-u_{ij}} \\ &= \left(1 + e^{\mathbf{G}_{ij}\boldsymbol{\gamma}} \right)^{-1} \\ &\times \left\{ u_{ij} \left[\exp(\mathbf{G}_{ij}\boldsymbol{\gamma}) + \exp\left(-e^{\mathbf{B}_{ij}\boldsymbol{\beta} + \sigma b_i}\right) \right] \right. \\ &\quad \left. + (1 - u_{ij}) \frac{\exp\left[y_{ij}(\mathbf{B}_{ij}\boldsymbol{\beta} + \sigma b_i) - e^{\mathbf{B}_{ij}\boldsymbol{\beta} + \sigma b_i}\right]}{y_{ij}!} \right\}. \end{aligned}$$

Here, ϕ denotes the standard normal probability density function, and $u_{ij} = 1$ if $y_{ij} = 0$ and $u_{ij} = 0$ otherwise.

Maximization of (5) with respect to $\boldsymbol{\psi}$ is complicated by the integration with respect to b_i . Several authors (e.g., Hinde, 1982; Anderson and Aitkin, 1985) have dealt with the same challenge in the context of a GLM with random intercept by employing the EM algorithm with Gaussian quadrature. This is a natural approach to employ here because we are already using the EM algorithm in the fixed effects version of the model. By regarding both the state of the process (zero state versus Poisson state) and the random effects as missing data, we can use the EM algorithm to more conveniently maximize (5).

Let $Z_{ij} = 1$ when Y_{ij} comes from the zero state and $Z_{ij} = 0$ when Y_{ij} comes from the $\text{Poisson}(\lambda_{ij})$ state. The complete-data log likelihood is

$$\begin{aligned} \ell_c(\boldsymbol{\psi}; \mathbf{y}, \mathbf{z}, \mathbf{b}) &= \log f(\mathbf{b}; \boldsymbol{\psi}) + \log f(\mathbf{y}, \mathbf{z} \mid \mathbf{b}; \boldsymbol{\psi}) \\ &= \sum_{i=1}^K \log \phi(b_i) \\ &\quad + \sum_{i=1}^K \sum_{j=1}^{T_i} \left\{ \left[z_{ij} \mathbf{G}_{ij}\boldsymbol{\gamma} - \log\left(1 + e^{\mathbf{G}_{ij}\boldsymbol{\gamma}}\right) \right] \right. \\ &\quad \left. + (1 - z_{ij}) \right. \\ &\quad \left. \times \left[y_{ij}(\mathbf{B}_{ij}\boldsymbol{\beta} + \sigma b_i) - \exp(\mathbf{B}_{ij}\boldsymbol{\beta} + \sigma b_i) \right. \right. \\ &\quad \left. \left. - \log(y_{ij}!) \right] \right\}. \end{aligned}$$

The $(r + 1)$ th iteration of the EM algorithm consists of the following three steps:

(1) *E step.* The E step requires the calculation of $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}) = E(\log f(\mathbf{y}, \mathbf{z}, \mathbf{b}; \boldsymbol{\psi}) \mid \mathbf{y}, \boldsymbol{\psi}^{(r)})$, where here the expectation is with respect to the joint distribution of \mathbf{z}, \mathbf{b} given \mathbf{y} and $\boldsymbol{\psi}^{(r)}$, the parameter estimate based on the r th iteration. This expectation can be taken in two steps,

$$\begin{aligned} Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}) &= E \left[E(\log f(\mathbf{y}, \mathbf{z}, \mathbf{b} \mid \boldsymbol{\psi}) \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\psi}^{(r)}) \mid \mathbf{y}, \boldsymbol{\psi}^{(r)} \right]. \end{aligned}$$

The inner expectation is with respect to \mathbf{z} only and, since $\log f(\mathbf{y}, \mathbf{z}, \mathbf{b} \mid \boldsymbol{\psi})$ is linear with respect to \mathbf{z} , this expectation becomes $\log f(\mathbf{y}, \mathbf{Z}^{(r)}, \mathbf{b} \mid \boldsymbol{\psi})$, where $\mathbf{Z}^{(r)}$ contains elements

$$\begin{aligned} Z_{ij}^{(r)} &= E \left(Z_{ij} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\psi}^{(r)} \right) \\ &= \begin{cases} 0, & \text{if } y_{ij} > 0; \\ \left[1 + \exp\left(-\mathbf{G}_{ij}\boldsymbol{\gamma}^{(r)} - e^{\mathbf{B}_{ij}\boldsymbol{\beta}^{(r)} + \sigma^{(r)} b_i}\right) \right]^{-1}, & \text{if } y_{ij} = 0. \end{cases} \end{aligned}$$

Note that $Z_{ij}^{(r)}$ depends on b_i , so we will emphasize this dependence by writing $Z_{ij}^{(r)}(b_i)$. To complete the E step, we now need to take the expectation with respect to the distribution of $\mathbf{b} \mid \mathbf{y}, \boldsymbol{\psi}^{(r)}$. Dropping terms that don't involve $\boldsymbol{\psi}$ and are therefore irrelevant in the M step, it follows that

$Q(\psi | \psi^{(r)})$ equals

$$\sum_{i=1}^K \sum_{j=1}^{T_i} \int_{-\infty}^{+\infty} \left\{ \left[Z_{ij}^{(r)}(b_i) \mathbf{G}_{ij} \gamma - \log(1 + e^{\mathbf{G}_{ij} \gamma}) \right] + \left[1 - Z_{ij}^{(r)}(b_i) \right] \times [y_{ij}(\mathbf{B}_{ij} \beta + \sigma b_i) - \exp(\mathbf{B}_{ij} \beta + \sigma b_i)] \right\} \times f(\mathbf{y}_i; \psi^{(r)} | b_i) \phi(b_i) db_i \div \left[\int_{-\infty}^{+\infty} f(\mathbf{y}_i; \psi^{(r)} | b_i) \phi(b_i) db_i \right].$$

Here we have used the fact that

$$f(b_i; \psi^{(r)} | \mathbf{y}_i) = \frac{f(\mathbf{y}_i; \psi^{(r)} | b_i) \phi(b_i)}{\int f(\mathbf{y}_i; \psi^{(r)} | b_i) \phi(b_i) db_i}.$$

Using m -point Gaussian quadrature to approximate these integrals, we have

$$Q(\psi | \psi^{(r)}) \approx \sum_{i,j} \left\{ \frac{\sum_{\ell=1}^m [Z_{ij}^{(r)}(b_\ell) \mathbf{G}_{ij} \gamma - \log(1 + e^{\mathbf{G}_{ij} \gamma})]}{\sum_{\ell=1}^m f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell} \times f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell + \frac{\sum_{\ell=1}^m [1 - Z_{ij}^{(r)}(b_\ell)]}{\sum_{\ell=1}^m f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell} \times [y_{ij}(\mathbf{B}_{ij} \beta + \sigma b_\ell) - \exp(\mathbf{B}_{ij} \beta + \sigma b_\ell)] \times f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell \right\},$$

where b_ℓ are quadrature points and w_ℓ the associated weights.

(2) *M step for γ .* Notice that, as in the EM algorithm for the fixed effects ZIP model, $Q(\psi | \psi^{(r)})$ decomposes into the sum of a term involving only γ and a second term involving only β . Therefore, we maximize $Q(\psi | \psi^{(r)})$ with respect to γ by maximizing the first term. This maximization can be done via a weighted logistic regression of $Z_{ij}^{(r)}(b_\ell)$, $i = 1, \dots, K$, $j = 1, \dots, T_i$, $\ell = 1, \dots, m$, on $\mathbf{G} \otimes \mathbf{1}_{m \times 1}$ with weights $f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell / g_i^{(r)}$, where $g_i^{(r)} = \sum_{\ell=1}^m f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell$, i.e., we perform a weighted logistic regression with an $Nm \times 1$ response vector $(Z_{11}^{(r)}(b_1), \dots, Z_{11}^{(r)}(b_m), Z_{12}^{(r)}(b_1), \dots, Z_{12}^{(r)}(b_m), \dots, Z_{KT_K}^{(r)}(b_m))^T$, a matrix of explanatory variables equal to the matrix obtained by repeating each row of \mathbf{G} m times, and weight

$$f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell / g_i^{(r)} = [\prod_{j=1}^{T_i} f(y_{ij}; \psi^{(r)} | b_\ell)] w_\ell / g_i^{(r)}$$

(constant over index j) corresponding to the (i, j, ℓ) th response.

(3) *M step for β, σ .* Maximization of the second term in $Q(\psi | \psi^{(r)})$ with respect to β and σ can be done simultaneously. Define $\mathbf{B}^* = [(\mathbf{B} \otimes \mathbf{1}_m), (\mathbf{1}_N \otimes (b_1, \dots, b_m))^T]$, $\beta^* = (\beta^T, \sigma)^T$. Maximization with respect to β^* can be

accomplished by fitting a weighted log-linear regression of $\mathbf{y} \otimes \mathbf{1}_{m \times 1}$ on \mathbf{B}^* with weights

$$[1 - Z_{ij}^{(r)}(b_\ell)] f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell / g_i^{(r)},$$

$i = 1, \dots, K$, $j = 1, \dots, T_i$, $\ell = 1, \dots, m$.

5.2 ZIB Regression with a Random Intercept

In the ZIB regression model with random intercept, we assume that, conditional on a random effect b_i ,

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } p_{ij}; \\ \text{binomial}(n_{ij}, \pi_{ij}), & \text{with probability } (1 - p_{ij}), \end{cases}$$

where we model $\pi_i = (\pi_{i1}, \dots, \pi_{iT_i})^T$ and \mathbf{p}_i using logistic regression models

$$\begin{aligned} \text{logit}(\pi_i) &= \mathbf{B}_i \beta + \sigma b_i \quad \text{and} \\ \text{logit}(\mathbf{p}_i) &= \mathbf{G}_i \gamma, \quad i = 1, \dots, K. \end{aligned}$$

Again, we assume b_1, \dots, b_K are independent standard normal random variables. The log likelihood for this model is as in (5) but now where

$$\begin{aligned} \Pr(Y_{ij} = y_{ij} | b_i) &= [p_{ij} + (1 - p_{ij})(1 - \pi_{ij})^{n_{ij}}]^{u_{ij}} \\ &\times \left[(1 - p_{ij}) \binom{n_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{n_{ij} - y_{ij}} \right]^{1 - u_{ij}} \\ &= (1 + e^{\mathbf{G}_{ij} \gamma})^{-1} \\ &\times \left\{ u_{ij} \left[\exp(\mathbf{G}_{ij} \gamma) + (1 + e^{\mathbf{B}_{ij} \beta + \sigma b_i})^{-n_{ij}} \right] \right. \\ &\quad \left. + (1 - u_{ij}) \binom{n_{ij}}{y_{ij}} e^{y_{ij}(\mathbf{B}_{ij} \beta + \sigma b_i)} \right. \\ &\quad \left. \times (1 + e^{\mathbf{B}_{ij} \beta + \sigma b_i})^{-n_{ij}} \right\}. \end{aligned}$$

As in the ZIP mixed model, the complete-data log likelihood can be constructed for use in the EM algorithm as

$$\begin{aligned} \ell_c(\psi; \mathbf{y}, \mathbf{z}, \mathbf{b}) &= \sum_{i=1}^K \log \phi(b_i) \\ &+ \sum_{i=1}^K \sum_{j=1}^{T_i} \left\{ \left[z_{ij} \mathbf{G}_{ij} \gamma - \log(1 + e^{\mathbf{G}_{ij} \gamma}) \right] \right. \\ &\quad \left. + (1 - z_{ij}) \right. \\ &\quad \times \left[y_{ij}(\mathbf{B}_{ij} \beta + \sigma b_i) \right. \\ &\quad \left. - n_{ij} \log(1 + e^{\mathbf{B}_{ij} \beta + \sigma b_i}) \right. \\ &\quad \left. \left. + \log \binom{n_{ij}}{y_{ij}} \right] \right\}. \end{aligned}$$

The $(r + 1)$ th iteration of the EM algorithm consists of the following three steps:

(1) *E step*. In the E step, $Q(\psi | \psi^{(r)}) = E[\log f(\mathbf{y}, \mathbf{Z}^{(r)}, \mathbf{b} | \psi)]$, where $\mathbf{Z}^{(r)}$ contains elements

$$Z_{ij}^{(r)}(b_i) = \begin{cases} 0, & \text{if } y_{ij} > 0 \\ \left[1 + e^{-\mathbf{G}_{ij}\gamma^{(r)}} \left(1 + e^{\mathbf{B}_{ij}\beta^{(r)} + \sigma^{(r)}b_i} \right)^{-n_{ij}} \right]^{-1}, & \\ \text{if } y_{ij} = 0. \end{cases}$$

Taking the expectation with respect to the distribution of $\mathbf{b} | \mathbf{y}, \psi^{(r)}$ and dropping irrelevant terms, we obtain

$$\begin{aligned} Q(\psi | \psi^{(r)}) &= \sum_{i=1}^K \sum_{j=1}^{T_i} \int_{-\infty}^{+\infty} \left\{ \left[Z_{ij}^{(r)}(b_i) \mathbf{G}_{ij} \gamma - \log \left(1 + e^{\mathbf{G}_{ij} \gamma} \right) \right] \right. \\ &\quad + \left[1 - Z_{ij}^{(r)}(b_i) \right] \\ &\quad \times \left[y_{ij} (\mathbf{B}_{ij} \beta + \sigma b_i) \right. \\ &\quad \left. \left. - n_{ij} \log \left(1 + e^{\mathbf{B}_{ij} \beta + \sigma b_i} \right) \right] \right\} \\ &\quad \times f(\mathbf{y}_i; \psi^{(r)} | b_i) \phi(b_i) db_i \\ &\quad \div \left[\int_{-\infty}^{+\infty} f(\mathbf{y}_i; \psi^{(r)} | b_i) \phi(b_i) db_i \right], \end{aligned}$$

which, using Gaussian quadrature, is approximately equal to

$$\begin{aligned} \sum_{i,j} \left\{ \frac{\sum_{\ell=1}^m \left[Z_{ij}^{(r)}(b_\ell) \mathbf{G}_{ij} \gamma - \log \left(1 + e^{\mathbf{G}_{ij} \gamma} \right) \right]}{\sum_{\ell=1}^m f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell} \right. \\ \times f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell \\ + \frac{\sum_{\ell=1}^m \left[1 - Z_{ij}^{(r)}(b_\ell) \right]}{\sum_{\ell=1}^m f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell} \\ \times \left[y_{ij} (\mathbf{B}_{ij} \beta + \sigma b_\ell) - n_{ij} \log \left(1 + e^{\mathbf{B}_{ij} \beta + \sigma b_\ell} \right) \right] \\ \left. \times f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell \right\}. \end{aligned}$$

(2) *M step for γ* . Maximization of $Q(\psi | \psi^{(r)})$ with respect to γ can be done exactly as in Section 5.1.

(3) *M step for β, σ* . Maximization of $Q(\psi | \psi^{(r)})$ with respect to β^* can be accomplished by fitting a weighted logistic regression of $\mathbf{y} \otimes \mathbf{1}_{m \times 1}$ on \mathbf{B}^* with weights $[1 - Z_{ij}^{(r)}(b_\ell)] f(\mathbf{y}_i; \psi^{(r)} | b_\ell) w_\ell / g_i^{(r)}$, $i = 1, \dots, K, j = 1, \dots, T_i, \ell = 1, \dots, m$, and binomial denominators $(n_{11}, \dots, n_{1T_1}, n_{21}, \dots, n_{KT_K})^T \otimes \mathbf{1}_{m \times 1}$.

In both the ZIP mixed and ZIB mixed models, an estimate of the asymptotic variance-covariance matrix of the MLE $\hat{\gamma}$ can be obtained by inverting the observed Fisher information matrix evaluated at $\hat{\gamma}$. For the models fit in this paper, this information matrix was obtained numerically. Based on Lambert's (1992) results, Wald tests and confidence intervals based on the asymptotic normality of $\hat{\gamma}$ may perform poorly in ZIP mixed and ZIB mixed models unless the sample size is quite large. Instead, we follow Lambert (1992) in recommend-

ing likelihood ratio tests and confidence intervals as the basis of inference when N is moderate to small.

6. Examples

6.1 Whitefly Data

Based on the experimental design, a reasonable starting point for selecting a ZIP mixed model for the number of immature insects per leaf data is to adapt the repeated measures analysis of variance model given in equation (1). Informally, we write the model for the Poisson-state mean as

$$\begin{aligned} \log(\lambda) &= \mu + \text{plant} + \text{block} + \text{trt} + \text{week} + \text{trt} \times \text{block} \\ &\quad + \text{trt} \times \text{week} + \log(n), \end{aligned} \tag{6}$$

where μ is a fixed intercept, plant is a random plant-specific effect, $\log(n)$ represents a covariate corresponding to the natural logarithm of the number of adult insects placed on the leaf prior to measurement of the response, and the other terms are fixed effects corresponding to factors and two-way interactions among the factors. We fit several models in which the relationship given in (6) is assumed combined with various choices for the linear predictor associated with $\text{logit}(p)$. Based on BIC, we selected the model with the covariate $\log(n)$ plus all main effects in the linear predictor for $\text{logit}(p)$. This model yielded a maximum log likelihood of -1219.3 and a BIC of -1561.8 on 534 residual degrees of freedom.

For comparison, we fit the ZIP model with the same design matrices \mathbf{B} and \mathbf{G} but no random effects. This model yielded a maximum log likelihood value of -1238.4 on 535 residual degrees of freedom. Notice that inclusion of a random plant effect has resulted in a significantly improved fit. Under $H_0: \sigma = 0$, two times the difference in the maximum log likelihood values for the ZIP and ZIP mixed versions of the model has an asymptotic distribution which is a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$. Noting that H_0 places the identifiable parameter σ^2 on the boundary of its parameter space, this result follows from the work of Stram and Lee (1994) and authors referenced therein. Thus an approximate p -value for H_0 is $(1/2)\text{Pr}[\chi^2(1) > 38.2] < .0001$.

As mentioned in Section 2, an alternative class of models worthy of consideration for these data is the class of GLMMs. Lambert (1992) compared her ZIP regression model with a Poisson-gamma (negative binomial) GLMM and demonstrated the superiority of her approach for the data set that she analyzed in that paper. We consider a similar Poisson-gamma model for response variable I in the whitefly data set. In this model, it is assumed that counts from different weeks j on the same plant i are $\text{Poisson}(\lambda_{ij} R_i)$, where R_i is assumed to follow a $\text{gamma}(\alpha, \alpha)$ distribution and $\log(\lambda_{ij}) = \mathbf{X}_{ij} \beta$. We also explored a Poisson-normal model of the same form except we assumed $R_i \sim N(0, \sigma_R^2)$. Restricting attention to identifiable models in these classes, the best fitting Poisson-gamma and Poisson-normal models that we were able to fit to the whitefly data were of the form given in (6), although certain of the design matrix columns corresponding to two-way interaction effects were eliminated in each case for identifiability.

These models fit better than the corresponding Poisson models (e.g., the negative binomial model gave a maximum log likelihood of -1583.1 versus -1650.8 for the Poisson model). However, they do not fit nearly as well as the ZIP and ZIP mixed models described above. In Figure 1 (cf., Lambert,

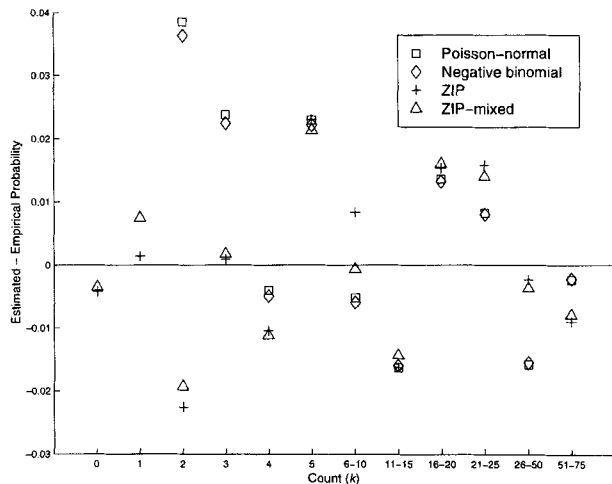


Figure 1. Errors of estimation for the proportion of responses (I) equal to k , for various k . Values for the Poisson-normal and negative binomial GLMMs are respectively, $-.205$ and $-.198$ for $k = 0$ and $.141$ and $.140$ for $k = 1$.

1992, Figure 5), we plot $\sum_{i=1}^N \hat{P}r(Y_i = k)/N - \#(Y_i = k)/N$ versus $\#(Y_i = k)/N$ for the ZIP and ZIP mixed models and the best fitting GLMMs; i.e., we plot the difference between the estimated and the observed proportion of the responses equal to k for $k = 0, 1, \dots$. From Figure 1, it appears that the ZIP and ZIP mixed models fit about equally well according to this criterion. However, for small k , the GLMMs do a very poor job of fitting the observed proportion of counts to k .

While the superior fit of the ZIP mixed model versus the ZIP model is not obvious from Figure 1, it is easily apparent based on an examination of the residuals from these models. We computed the 54 residuals (observed minus expected) for the total number of immature insects counted over the 12-week experiment on each plant. Based on the ZIP model, these raw residuals had quartiles -7.53 , $-.29$, and 3.86 and standard deviation 25.49. Based on the analogous ZIP mixed model, these residuals had quartiles -3.52 , 0.036 , and 3.05 and standard deviation 15.74.

For the number of live adults response, we fit several models with

$$\begin{aligned} \text{logit}(\pi) = & \mu + \text{plant} + \text{block} + \text{trt} + \text{week} + \text{trt} \times \text{block} \\ & + \text{trt} \times \text{week}, \end{aligned} \quad (7)$$

combined with various specifications for $\text{logit}(p)$. Again, to fit these models, m (the number of abscissas used for Gaussian quadrature) was taken to be nine. As for I , the model that was found to maximize BIC has only main effects in the linear predictor for $\text{logit}(p)$. This model yielded a maximum log likelihood of -839.6 on 536 residual degrees of freedom and a BIC of -1175.6 . By comparison, the ZIB version of this model gave a maximum log likelihood value of -851.6 and a BIC of -1184.4 . Again, the mixed effects version of the model fits our data set substantially better than the ordinary ZIB model.

Detailed results of analyses of the whitefly data set based on ZIP and ZIB models can be found in van Iersel et al.

(2000). Results based on ZIP mixed and ZIB mixed models are qualitatively the same. Briefly, for I , contrasts between the control treatment and all active treatments and between the standard active treatment (H) and the subirrigation treatments (0, 1, 2, 4) were highly significant when performed on β and γ or on β and γ simultaneously. All effects were in the expected direction, with active treatments suppressing whitefly reproduction, most effectively when subirrigation was used to deliver the pesticide. A significant disorderly interaction between week and the subirrigation treatments prevented a marginal comparison between the 0, 1, 2, and 4 treatments, but the treatment by week profile plot of the marginal mean number of immatures suggested a trend toward greater effectiveness for longer delays between last watering and application of the pesticide.

For L , contrasts between control and active treatments were again highly significant in the expected direction when performed on β and γ separately and simultaneously. The hand-watering treatment was significantly less effective than the subirrigation treatments as measured by Wald tests for the appropriate 1 d.f. contrast on β and for the 2 d.f. contrast performed simultaneously on β and γ . However, this contrast was not significant when tested on γ alone. Finally, a significant (in β), highly disorderly interaction between week and the subirrigation treatments was found in the marginal mean profile plot of L , obscuring any trend in the effectiveness of the subirrigation treatments across different lengths in the delay since last watering.

The results presented here and in van Iersel et al. (2000) regarding specific treatment comparisons are based on Wald tests of contrasts. In this example, the Wald approach has a substantial advantage over likelihood ratio-based inference in terms of computational simplicity. With the large sample size of the whitefly experiment ($N = 640$), we expect these inferences to be reasonably accurate. However, we reiterate that a likelihood ratio approach is preferable for moderate- to small-sized data sets.

6.2 Lambert's Printed Wiring Board Data

Lambert (1992) illustrates her ZIP regression methodology using an example from manufacturing. A split-plot experiment was conducted to determine the influence of five factors on the frequency of soldering errors in the manufacture of printed wiring boards. The five factors were mask (five levels), opening (three levels), solder (two levels), pad (nine levels), and panel (three levels). Twenty-five wiring boards were used in the experiment, where each board was divided into three panels and each panel was further divided into nine areas receiving different pad types. For details of the experimental design and a full description of the experimental factors, see Lambert (1992). The response variable was the number of soldering errors made in each of the 675 board areas.

Lambert provides excellent motivation for ZIP regression for these data (81% of the responses were zero) and demonstrates that the final ZIP model she obtains fits the data much better than alternative models based on ordinary Poisson regression and negative binomial regression (i.e., Poisson-gamma regression, as described in Section 6.1). However, due to the split-plot nature of the experimental design for these data, it is natural to introduce random effects

Table 5

Fits of models including mask \times solder and opening \times solder interactions and all main effects to printed wiring board data

Method	Log likelihood	d.f.	BIC
Poisson	-700.4	651	1557.2
Negative binomial	-674.2	650	1511.3
ZIP ^a	-511.2	634	1289.5
ZIP mixed ^a	-505.3	633	1284.2

^a Model for logit(p) includes all main effects except solder.

for the whole plots, which are the wiring boards here. Such an approach works poorly when the conditional distribution is Poisson (e.g., the negative binomial model fit by Lambert). But when the conditional distribution is zero-inflated Poisson, it is possible to obtain a significantly better fitting model than Lambert's ZIP model at the cost of only one additional degree of freedom.

In Table 5, we present the maximum log likelihood and BIC values for the model in which the design matrix associated with $\log(\lambda)$ contains main effects and two-way interactions between mask and solder and between opening and solder. The ZIP version of this model was chosen for analysis of the printed wiring board data by Lambert (1992). Again, the inclusion of a random board effect results in a significantly improved fit. An approximate p -value for $H_0: \sigma = 0$ is $(1/2) \Pr[\chi^2(1) > 11.8] = .0003$. Residuals from the ZIP mixed model are also smaller in magnitude than for the corresponding ZIP model. Lambert (1992) reports that the 25 raw residuals computed as observed minus expected for the number of defects per board have a mean of $-.3$ for the ZIP model with quartiles of -2.9 and 1.9 . For the corresponding ZIP mixed model, these residuals have mean $-.2$ with quartiles $-.6$ and $.5$. In addition, the quality of the estimates for the proportion of responses equal to k , $k = 0, 1, \dots$ (as in Tables 2-4), is slightly better for the ZIP mixed model.

7. Summary

In this article, we have adapted Lambert's zero-inflated Poisson regression to the situation in which the response is an upper-bounded count. The resulting zero-inflated binomial regression models can be fit using the EM algorithm in a manner similar to that described in Lambert's (1992) paper. In addition, we introduce random effects into the portion of the ZIP and ZIB models that describes the dependence of the non-zero state mean on covariates. The resulting ZIP mixed and ZIB mixed models can also be fit with the EM algorithm with Gaussian quadrature. These mixed models are useful for modeling sources of heterogeneity and dependence in zero-inflated count data. These models have been demonstrated to fit better than corresponding fixed-effects models with zero inflation (ZIP and ZIB models) and mixed effects models without zero inflation (GLMMs) for the repeated measures and split-plot data sets considered in this paper.

RÉSUMÉ

Dans un papier de *Technometrics* de 1992, Lambert (1992) a décrit une régression de Poisson à inflation nulle (ZIP), une classe de modèles pour données de comptage avec un excès de zéros. Dans un modèle ZIP, la variable réponse de comptage

est considérée comme étant le mélange d'une distribution de Poisson (λ) et d'une distribution ponctuelle avec une masse de 1 au point zéro, avec une probabilité de mélange p . A la fois p et λ sont autorisées à dépendre des covariables au travers de modèles linéaires généralisés à liens canoniques. Dans ce papier, nous adoptons la méthodologie de Lambert à une situation de comptage borné vers le haut, obtenant ainsi un modèle binomial à inflation nulle (ZIB). De plus, nous ajoutons de la flexibilité à ces modèles à effets fixes en incorporant des effets aléatoires, de façon (par exemple) que la corrélation intra-sujet et l'hétérogénéité inter sujets, typiques des mesures répétées, soient prises en compte. Nous motivons, développons et illustrons les méthodes décrites ici avec un exemple d'horticulture, où à la fois des données de comptage bornées (binomiales) et non bornées (Poisson) avec un excès de zéros ont été recueillies dans une expérimentation en mesures répétées.

REFERENCES

- Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary responses: interviewer variability. *Journal of the Royal Statistical Society, Series B* **47**, 203-210.
- Green, P. J. (1984). Iteratively reweighted least-squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, Series B* **46**, 149-192.
- Heilbron, D. C. (1989). *Generalized linear models for altered zero probabilities and overdispersion in count data*. Technical Report, University of California, Department of Epidemiology and Biostatistics, San Francisco.
- Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36**, 531-547.
- Hinde, J. (1982). Compound Poisson regression models. In *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, R. Gilchrist (ed), 109-121. New York: Springer-Verlag.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14.
- McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162-170.
- Ridout, M., Demétrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros. Invited paper presented at the Nineteenth International Biometric Conference, Capetown, South Africa.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.
- Stram, D. O. and Lee, J. W. (1994). Variance component testing in the longitudinal mixed effect model. *Biometrics* **50**, 1171-1177.
- van Iersel, M., Oetting, R., and Hall, D. B. (2000). Imidicloprid applications by subirrigation for control of silverleaf whitefly on poinsettia. *Journal of Economic Entomology*, in press.
- Yip, P. (1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. *Australian Journal of Statistics* **30**, 299-306.

Received September 1999. Revised March 2000.

Accepted April 2000.