TECHNOMETRICS, FEBRUARY 1992, VOL. 34, NO. 1

Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing

Diane Lambert

AT&T Bell Laboratories
Murray Hill, NJ 07974

Zero-inflated Poisson (ZIP) regression is a model for count data with excess zeros. It assumes that with probability p the only possible observation is 0, and with probability $1 - p$, a Poisson(λ) random variable is observed. For example, when manufacturing equipment is properly aligned, defects may be nearly impossible. But when it is misaligned, defects may occur according to a Poisson(λ) distribution. Both the probability p of the perfect, zero defect state and the mean number of defects λ in the imperfect state may depend on covariates. Sometimes p and λ are unrelated; other times p is a simple function of λ such as $p = 1/(1 + \lambda^\tau)$ for an unknown constant τ. In either case, ZIP regression models are easy to fit. The maximum likelihood estimates (MLE's) are approximately normal in large samples, and confidence intervals can be constructed by inverting likelihood ratio tests or using the approximate normality of the MLE's. Simulations suggest that the confidence intervals based on likelihood ratio tests are better, however. Finally, ZIP regression models are not only easy to interpret, but they can also lead to more refined data analyses. For example, in an experiment concerning soldering defects on printed wiring boards, two sets of conditions gave about the same mean number of defects, but the perfect state was more likely under one set of conditions and the mean number of defects in the imperfect state was smaller under the other set of conditions; that is, ZIP regression can show not only which conditions give lower mean number of defects but also why the means are lower.

KEY WORDS: EM algorithm; Negative binomial; Overdispersion; Positive Poisson; Quality control.

Standard arguments suggest that, when a reliable manufacturing process is in control, the number of defects on an item should be Poisson distributed. If the Poisson mean is λ, a large sample of n items should have about $ne^{-\lambda}$ items with no defects. Sometimes, however, there are many more items without defects than would be predicted from the numbers of defects on imperfect items (an example is given in Sec. 1), One interpretation is that slight, unobserved changes in the environment cause the process to move randomly back and forth between a perfect state in which defects are extremely rare and an imperfect state in which defects are possible but not inevitable. The transient perfect state, or existence of items that are unusually resistant to defects, increases the number of zeros in the data.

This article describes a new technique, called *zero-inflated Poisson* (ZIP) regression, for handling zero-inflated count data. ZIP models without covariates have been discussed by others (for example, see Cohen 1963; Johnson and Kotz 1969), but here both the probability p of the perfect state and the mean λ of the imperfect state can depend on covariates. In particular, $\log(\lambda)$ and $\text{logit}(p) = \log(p/(1 - p))$ are assumed to be linear functions of some covariates. The same or different covariates might affect p and λ, and p and λ might or might not be functionally related. When p is a decreasing function of λ, the probability of the perfect state and the mean of the imperfect state improve or deteriorate together.

Heilbron (1989) concurrently proposed similar zero-altered Poisson and negative binomial regression models with different parameterizations of p and applied them to data on high-risk behavior in gay men. (Although the models were developed independently, the acronym ZIP is just an apt modification of Heilbron's acronym ZAP for zero-altered Poisson.) He also considered models with an arbitrary probability of 0. Arbitrary zeros are introduced by mixing point mass at 0 with a positive Poisson that assigns no mass to 0 rather than a standard Poisson.

Other authors have previously considered mixing a distribution degenerate at 0 with distributions other than the Poisson or negative binomial. For example, Feuerverger (1979) mixed zeros with a gamma distribution and coupled the probability of 0 with the mean of the gamma to model rainfall data. Farewell (1986) and Meeker (1987) mixed zeros with right-censored continuous distributions to model survival data when some items are indestructible and testing

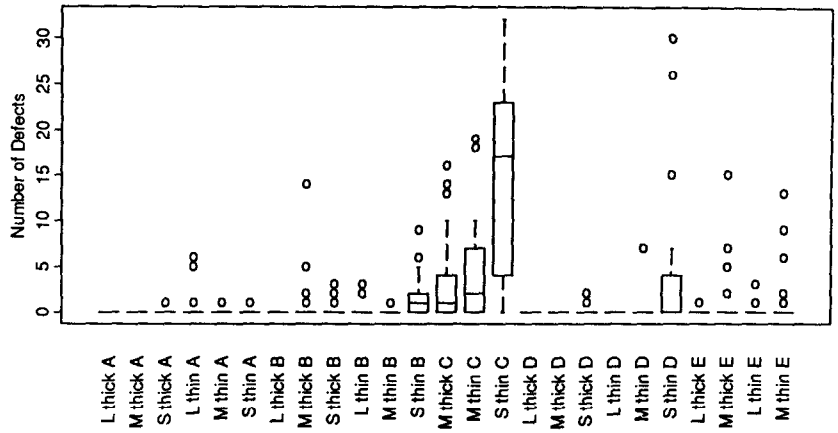

Figure 1. The Data From the Printed Wiring Board Experiment. Each boxplot shows the number of defects on one of the 25 boards in the experiment. The boxes are labeled by the board factors. When the box degenerates to a line, as happens for the leftmost board, which has large openings, thick solder, and mask A, 75% or more of the 27 counts on the board were 0. Counts beyond 1.5 interquartile ranges of the nearest quartile are plotted individually as circles.

is stopped before all items that can fail have failed.

Regression models that mix zeros and Poissons are described in detail in Section 2. Section 3 shows how to compute maximum likelihood estimates (MLE's). Section 4 discusses finite-sample inference in the context of simulations. Section 5 applies ZIP regression to the manufacturing data introduced in Section 1. Section 6 gives conclusions.

1. THE MOTIVATING APPLICATION

Components are mounted on printed wiring boards by soldering their leads onto pads on the board. An experiment at AT&T Bell Laboratories studied five influences on solderability:

Mask—five types of the surface of the board (A–E)

Opening—large (L), medium (M), or small (S) clearances in the mask around the pad

Solder—thick or thin

Pad—nine combinations of different lengths and widths of the pads (A–I)

Panel—1 = first part of the board to be soldered; 2 = second; 3 = last

In the experiment, one combination of mask, solder amount, and opening was applied to a board. Each board was partitioned into three equal panels, and each panel was subdivided into nine areas. A different pad size was used in each area of a panel, and the same pattern of pad sizes was repeated in all three panels. Each area held 24 identical devices with two leads each, giving 48 leads per area. A balanced experiment with 30 boards was designed, but only 25 boards were available for statistical analysis. Thus the factors that vary between boards are unbalanced. Mask is the most unbalanced factor, C is its most unbalanced level, and no board had large openings with mask C or small openings with mask E (see Fig. 1).

In each area, only the total number of leads improperly soldered was recorded, giving 27 responses between 0 and 48 for each of the 25 boards; that is, the response vector $\mathbf{Y}$ has 675 elements between 0 and 48. Out of the 675 areas, 81% had zero defects, 8% had at least five defects, and 5.2% had at least nine. Plainly, most areas had no defects, but those that did have defects often had several, as Figure 1 shows.

Since each lead has only a small probability of not being soldered and there are many leads per area, it is reasonable to fit a log-linear Poisson($\boldsymbol{\lambda}$) model; that is, $\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta}$ for some design matrix $\mathbf{B}$ and coefficients $\boldsymbol{\beta}$. Table 1 summarizes the fit of several such models.

Table 1. Fits of Some Poisson Models for the Number of Defects per Area

Highest terms in the model	*Log-likelihood*	*Residual degrees of freedom*
No interactions	−761.2	657
Solder * opening	−718.0	655
Mask * solder	−719.9	653
Mask * opening	−711.7	651
Mask * solder + opening * solder	−700.4	651
Opening * solder + mask * opening	−663.9	649
Mask * solder + mask * opening	−671.5	647
Mask * solder + opening * solder + mask * opening	−653.0	645
Mask * opening * solder	−638.2	640

NOTE: The main-effects model is $\log(\lambda)$ = panel + pad + mask + opening + solder. Some effects in the mask * opening interaction are not estimable. All models have all main effects; the three-way interaction model also has all two-way interactions.

Even the richest log-linear Poisson model, which has a three-way interaction between mask, opening, and solder, predicts poorly. Although 81% of the areas had no defects, the model predicts that only 71% of the areas will have no defects, since $\sum_{i=1}^{675} P(Y = 0|\hat{\lambda}_i)/675 = .71$. Large counts are also underpredicted. Although 5.2% of the areas had at least nine defects, the model predicts only 3.7% will have at least nine defects. Poisson models with fewer interaction terms are even worse. For example, the richest model for which all coefficients are estimable predicts only 68% zeros and 3.4% counts of at least nine. Modeling the split-plot nature of the experiment in a more complicated, generalized Poisson regression still does not give good predictions, as Section 5.2 shows. In short, there are too many zeros and too many large counts for the data to be Poisson.

2. THE MODEL

In ZIP regression, the responses $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ are independent and

$$\begin{aligned} Y_i &\sim 0 && \text{with probability } p_i \\ &\sim \text{Poisson}(\lambda_i) && \text{with probability } 1 - p_i, \end{aligned}$$

so that

$$\begin{aligned} Y_i &= 0 && \text{with probability } p_i + (1 - p_i)e^{-\lambda_i} \\ &= k && \text{with probability } (1 - p_i)e^{-\lambda_i}\lambda_i^k/k!, \\ & && k = 1, 2, \ldots. \end{aligned}$$

Moreover, the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ and $\mathbf{p} = (p_1, \ldots, p_n)'$ satisfy

$$\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta} \quad \text{and}$$
$$\text{logit}(\mathbf{p}) = \log(\mathbf{p}/(1 - \mathbf{p})) = \mathbf{G}\boldsymbol{\gamma} \qquad (1)$$

for covariate matrices $\mathbf{B}$ and $\mathbf{G}$.

The covariates that affect the Poisson mean of the imperfect state may or may not be the same as the covariates that affect the probability of the perfect state. When they are the same and $\boldsymbol{\lambda}$ and $\mathbf{p}$ are not functionally related, $\mathbf{B} = \mathbf{G}$ and ZIP regression requires twice as many parameters as Poisson regression. At the other extreme, when the probability of the perfect state does not depend on the covariates, $\mathbf{G}$ is a column of ones, and ZIP regression requires only one more parameter than Poisson regression.

If the same covariates affect $\mathbf{p}$ and $\boldsymbol{\lambda}$, it is natural to reduce the number of parameters by thinking of $\mathbf{p}$ as a function of $\boldsymbol{\lambda}$. Assuming that the function is known up to a constant nearly halves the number of parameters needed for ZIP regression and may accelerate the computations considerably. [The maximum likelihood computations are no faster than those for least squares, which are $O(k^3)$ in the number of parameters k; for example, see Chambers (1977, p. 144).] In many applications, however, there is little prior information about how $\mathbf{p}$ relates to $\boldsymbol{\lambda}$. If so, a natural parameterization is

$$\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta} \text{ and } \text{logit}(\mathbf{p}) = -\tau\mathbf{B}\boldsymbol{\beta} \qquad (2)$$

for an unknown, real-valued shape parameter τ, which implies that $p_i = (1 + \lambda_i^\tau)^{-1}$. In the terminology of generalized linear models, $\log(\boldsymbol{\lambda})$ and $\text{logit}(\mathbf{p})$ are the natural *links* or transformations that linearize Poisson means and Bernoulli probabilities of success. If the term $\mathbf{B}\boldsymbol{\beta}$ is thought of as stress, the natural links for $\mathbf{p}$ and $\boldsymbol{\lambda}$ are both proportional to stress. ZIP model (2) with logit link for $\mathbf{p}$, log link for $\boldsymbol{\lambda}$, and shape parameter τ will be denoted ZIP(τ).

The logit link for $\mathbf{p}$ is symmetric around .5. Two popular asymmetric links are the log-log link defined by $\log(-\log(\mathbf{p})) = \tau\mathbf{B}\boldsymbol{\beta}$ or, equivalently, $p_i = \exp(-\boldsymbol{\lambda}_i^\tau)$ and the complementary log-log link defined by $\log(-\log(1 - \mathbf{p})) = -\tau\mathbf{B}\boldsymbol{\beta}$ or $p_i = 1 - \exp(-\boldsymbol{\lambda}_i^{-\tau})$. Heilbron (1989) used an additive log-log link defined by $p_i = \exp(-\tau\lambda_i)$ or $\log(-\log(\mathbf{p})) = \mathbf{B}\boldsymbol{\beta} + \log(\tau)$. Linear, instead of proportional or additive, logit links and log-log links could be defined by $\text{logit}(\mathbf{p}) = \log(\alpha) - \tau\mathbf{B}\boldsymbol{\beta}$ and $\log(-\log(\mathbf{p})) = \log(\alpha) + \tau\mathbf{B}\boldsymbol{\beta}$, respectively.

With any of these links, when $\tau > 0$ the perfect state becomes less likely as the imperfect mean increases, and as $\tau \to \infty$ the perfect state becomes impossible if λ_i stays fixed. As $\tau \to 0$ under the additive log-log link and as $\tau \to -\infty$ under the other links, the perfect state becomes certain. Negative τ are not permitted with the additive log-log link, but for the other links with $\tau < 0$, the Poisson mean increases as excess zeros become more likely. This could be relevant in manufacturing applications if setting parameters to improve the fraction of perfect items leads to more defects on imperfect items.

3. MAXIMUM LIKELIHOOD ESTIMATION

The number of parameters that can be estimated in a ZIP regression model depends on the richness of the data. If there are only a few positive counts and $\boldsymbol{\lambda}$ and $\mathbf{p}$ are not functionally related, then only simple models should be considered for $\boldsymbol{\lambda}$. The simulations in Section 4 suggest that the data are adequate to estimate the parameters of a ZIP or ZIP(τ) model if the observed information matrix is nonsingular. This section shows how to compute the MLE's; the observed information matrix is given in the Appendix.

3.1 $\boldsymbol{\lambda}$ and $\mathbf{p}$ Unrelated

When $\boldsymbol{\lambda}$ and $\mathbf{p}$ are not functionally related, the log-likelihood for ZIP regression with the standard parameterization (1) is

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{y_i=0} \log(e^{\mathbf{G}_i\boldsymbol{\gamma}} + \exp(-e^{\mathbf{B}_i\boldsymbol{\beta}})) + \sum_{y_i>0} (y_i\mathbf{B}_i\boldsymbol{\beta} - e^{\mathbf{B}_i\boldsymbol{\beta}}) - \sum_{i=1}^{n} \log(1 + e^{\mathbf{G}_i\boldsymbol{\gamma}}) - \sum_{y_i>0} \log(y_i!), \quad (3)$$

where $\mathbf{G}_i$ and $\mathbf{B}_i$ are the ith rows of $\mathbf{G}$ and $\mathbf{B}$. The sum of exponentials in the first term complicates the maximization of $L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y})$. But suppose we knew which zeros came from the perfect state and which came from the Poisson; that is, suppose we could observe $Z_i = 1$ when Y_i is from the perfect, zero state and $Z_i = 0$ when Y_i is from the Poisson state. Then the log-likelihood with the complete data $(\mathbf{y}, \mathbf{z})$ would be

$$\begin{aligned} L_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{z}) &= \sum_{i=1}^{n} \log(f(z_i|\boldsymbol{\gamma})) + \sum_{i=1}^{n} \log(f(y_i|z_i, \boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} (z_i\mathbf{G}_i\boldsymbol{\gamma} - \log(1 + e^{\mathbf{G}_i\boldsymbol{\gamma}})) + \sum_{i=1}^{n} (1 - z_i)(y_i\mathbf{B}_i\boldsymbol{\beta} - e^{\mathbf{B}_i\boldsymbol{\beta}}) - \sum_{i=1}^{n} (1 - z_i)\log(y_i!) \\ &= L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z}) + L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z}) - \sum_{i=1}^{n} (1 - z_i)\log(y_i!). \end{aligned}$$

This log-likelihood is easy to maximize, because $L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z})$ and $L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z})$ can be maximized separately.

With the EM algorithm, the incomplete log-likelihood (3) is maximized iteratively by alternating between estimating Z_i by its expectation under the current estimates of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ (E step) and then, with the Z_i's fixed at their expected values from the E step, maximizing $L_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{z})$ (M-step) (for example, Dempster, Laird, and Rubin 1977). Once the expected Z_i's converge, the estimated $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ converges, and iteration stops. The estimates from the final iteration are the MLE's $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ for the log-likelihood (3).

In more detail, iteration $(k + 1)$ of the EM algorithm requires three steps.

E Step. Estimate Z_i by its posterior mean $Z_i^{(k)}$ under the current estimates $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$. Here

$$\begin{aligned} Z_i^{(k)} &= P[\text{perfect state}|y_i, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}] \\ &= \frac{P[y_i|\text{perfect state}]\,P[\text{perfect state}]}{P[y_i|\text{perfect state}]\,P[\text{perfect state}] + P[y_i|\text{Poisson}]\,P[\text{Poisson}]} \\ &= (1 + e^{-\mathbf{G}_i\boldsymbol{\gamma}^{(k)} - \exp(\mathbf{B}_i\boldsymbol{\beta}^{(k)})})^{-1} \quad \text{if } y_i = 0 \\ &= 0 \quad \text{if } y_i = 1, 2, \ldots . \end{aligned}$$

M step for $\boldsymbol{\beta}$. Find $\boldsymbol{\beta}^{(k+1)}$ by maximizing $L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{Z}^{(k)})$. Note that $\boldsymbol{\beta}^{(k+1)}$ can be found from a weighted, log-linear Poisson regression with weights $1 - \mathbf{Z}^{(k)}$, as described by McCullagh and Nelder (1989), for example.

M step for $\boldsymbol{\gamma}$. Maximize $L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{Z}^{(k)}) = \Sigma_{y_i=0} Z_i^{(k)}\mathbf{G}_i\boldsymbol{\gamma} - \Sigma_{y_i=0} Z_i^{(k)} \log(1 + e^{\mathbf{G}_i\boldsymbol{\gamma}}) - \Sigma_{i=1}^{n}(1 - Z_i^{(k)}) \log(1 + e^{\mathbf{G}_i\boldsymbol{\gamma}})$ as a function of $\boldsymbol{\gamma}$. (The equality holds because $Z_i^{(k)} = 0$ whenever $y_i > 0$.) To do this, suppose that n_0 of the y_i's are 0. Say $y_{i_1}, \ldots, y_{i_{n_0}}$ are 0. Define $\mathbf{y}_*' = (y_1, \ldots, y_n, y_{i_1}, \ldots, y_{i_{n_0}})$, $\mathbf{G}_*' = (\mathbf{G}_1', \ldots, \mathbf{G}_n', \mathbf{G}_{i_1}', \ldots, \mathbf{G}_{i_{n_0}}')$, and $\mathbf{P}_*' = (p_1, \ldots, p_n, p_{i_1}, \ldots, p_{i_{n_0}})$. Define also a diagonal matrix $\mathbf{W}^{(k)}$ with diagonal $\mathbf{w}^{(k)} = (1 - Z_1^{(k)}, \ldots, 1 - Z_n^{(k)}, Z_{i_1}^{(k)}, \ldots, Z_{i_{n_0}}^{(k)})'$. In this notation, $L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{Z}^{(k)}) = \Sigma_{i=1}^{n+n_0} y_{*i}w_i^{(k)}\mathbf{G}_{*i}\boldsymbol{\gamma} - \Sigma_{i=1}^{n+n_0} w_i^{(k)} \log(1 + e^{\mathbf{G}_{*i}\boldsymbol{\gamma}})$, the gradient or score function is $\mathbf{G}_*'\mathbf{W}^{(k)}(\mathbf{y}_* - \mathbf{P}_*) = 0$, and the Hessian or negative of the information matrix is $-\mathbf{G}_*'\mathbf{W}^{(k)}\mathbf{Q}_*\mathbf{G}_*$, where $\mathbf{Q}_*$ is the diagonal matrix with $\mathbf{P}_*(1 - \mathbf{P}_*)$ on the diagonal. These functions are identical to those for a weighted logistic regression with response $\mathbf{y}_*$, covariate matrix $\mathbf{G}_*$, and prior weights $\mathbf{w}^{(k)}$; that is, $\boldsymbol{\gamma}^{(k)}$ can be found by weighted logistic regression.

The EM algorithm converges in this problem (see the Appendix). Moreover, despite its reputation, it converges reasonably quickly because the MLE for the positive Poisson log-likelihood is an excellent guess for $\boldsymbol{\beta}$. The positive Poisson log-likelihood is

$$L_+(\boldsymbol{\beta}; \mathbf{y}_+) = \sum_{y_i>0} (y_i\mathbf{B}_i\boldsymbol{\beta} - e^{\mathbf{B}_i\boldsymbol{\beta}}) - \sum_{y_i>0} \log(1 - e^{-\exp(\mathbf{B}_i\boldsymbol{\beta})}) - \sum_{y_i>0} \log(y_i!),$$

its score equations are

$$\mathbf{B}_+'\left(\mathbf{y}_+ - \frac{e^{\mathbf{B}_+\boldsymbol{\beta}}}{1 - \exp(-e^{\mathbf{B}_+\boldsymbol{\beta}})}\right) = 0,$$

and its Hessian is $\mathbf{B}_+'\mathbf{D}\mathbf{B}_+$, where the subscript + indicates that only the elements or rows corresponding to positive y_i's are used and $\mathbf{D}$ is a diagonal matrix with diagonal

$$\frac{e^{\mathbf{B}_+\boldsymbol{\beta}}(1 - e^{-e^{\mathbf{B}_+\boldsymbol{\beta}}}(1 + e^{\mathbf{B}_+\boldsymbol{\beta}}))}{(1 - e^{-e^{\mathbf{B}_+\boldsymbol{\beta}}})^2}.$$

The score equations can be solved by iteratively reweighted least squares (see Green 1984).

The initial guess for $\boldsymbol{\gamma}$ has been unimportant in the examples and simulations considered to date. One possibility when $\boldsymbol{\gamma}$ includes an intercept is to set all elements other than the intercept to 0 and estimate the intercept by the log odds of the observed average probability of an excess 0. The observed average probability of an excess 0 is

$$\hat{p}_0 = \frac{\#(y_i = 0) - \sum e^{-\exp(\mathbf{B}_i\boldsymbol{\beta}^{(0)})}}{n}.$$

(If the fraction of zeros is smaller than the fitted Poisson models predict, there is no reason to fit a ZIP regression model.)

There are other algorithms, such as the Newton–Raphson, for maximizing the log-likelihood (3). When it converges, the Newton–Raphson algorithm is usually faster than EM. The EM algorithm, however, is simpler to program, especially if weighted Poisson and weighted logistic regression are available. Moreover, the Newton–Raphson algorithm failed for the final ZIP model in Section 5.1. Whatever the algorithm, $\log(1 + \exp(x))$ should be computed carefully in the tails.

3.2 p a Function of λ

The log-likelihood for the ZIP(τ) model with the standard parameterization (2) is, up to a constant,

$$L(\boldsymbol{\beta}, \tau; \mathbf{y}) = \sum_{y_i=0} \log(e^{-\tau\mathbf{B}_i\boldsymbol{\beta}} + \exp(-e^{\mathbf{B}_i\boldsymbol{\beta}})) + \sum_{y_i>0} (y_i\mathbf{B}_i\boldsymbol{\beta} - e^{\mathbf{B}_i\boldsymbol{\beta}}) - \sum_{i=1}^{n} \log(1 + e^{-\tau\mathbf{B}_i\boldsymbol{\beta}}). \quad (4)$$

The EM algorithm is not useful here because $\boldsymbol{\beta}$ and τ cannot be estimated simply even if $\mathbf{Z}$ is known. The Newton–Raphson algorithm, however, has converged in the examples and simulations to date. Initial guesses of $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}_u$ and $\tau^{(0)} = -\text{median}(\hat{\boldsymbol{\gamma}}_u / \hat{\boldsymbol{\beta}}_u)$, where $(\hat{\boldsymbol{\gamma}}_u, \hat{\boldsymbol{\beta}}_u)$ are the ZIP MLE's, perhaps excluding the intercept from the calculation of the median, are often satisfactory. If not, first maximizing over $\boldsymbol{\beta}$ for a few choices of fixed τ_0 and then starting at the $(\hat{\boldsymbol{\beta}}(\tau_0), \tau_0)$ with the largest log-likelihood often succeeds.

4. STANDARD ERRORS AND CONFIDENCE INTERVALS

In large samples, the MLE's $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ for ZIP regression and $(\hat{\boldsymbol{\beta}}, \hat{\tau})$ for ZIP(τ) regression are approximately normal with means $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ and $(\boldsymbol{\beta}, \tau)$ and variances equal to the inverse observed information matrices $\mathbf{I}^{-1}$ and $\mathbf{I}_\tau^{-1}$, respectively. (The formulas for $\mathbf{I}$ and $\mathbf{I}_\tau$ are given in the Appendix.) Thus, for large enough samples, the MLE's and regular functions of the MLE's, such as the probability of a defect and the mean number of defects, are nearly unbiased.

Normal-theory, large-sample confidence intervals are easy to construct, but they assume that the log-likelihood is approximately quadratic near the MLE. When it is not quadratic a few standard errors from the MLE, normal-theory confidence intervals can mislead. Likelihood ratio confidence intervals are more difficult to compute but are often more trustworthy. A two-sided $(1 - \alpha)100\%$ likelihood ratio confidence interval for β_1 in ZIP regression, for example, is found by computing the set of β_1 for which $2(L(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) - \max_{\boldsymbol{\gamma}, \boldsymbol{\beta}_{-1}} L((\beta_1, \boldsymbol{\beta}_{-1}), \boldsymbol{\gamma})) < \chi^2_{\alpha,1}$ where $\chi^2_{\alpha,1}$ is the upper α quantile of a χ^2 distribution with 1 df and $\boldsymbol{\beta}_{-1}$ is the parameter vector $\boldsymbol{\beta}$ without its first element.

4.1 The Simulation Experiments

To explore whether the asymptotic results pertain to finite samples, three simulations with 2,000 runs each were carried out in S (Becker, Chambers, and Wilks 1988). In the first experiment, $n = 25$, in the second $n = 50$, and in the third $n = 100$. Throughout, there is one covariate $\mathbf{x}$ taking on n uniformly spaced values between 0 and 1, $\boldsymbol{\gamma} = (-1.5, 2)$, and $\boldsymbol{\beta} = (1.5, -2)$, so $\tau = 1$. With these choices, the Poisson means $\boldsymbol{\lambda}$ range from .6 to 4.5 with a median of 1.6, and, on average, 50% of the responses y_i are 0 and 23% of the zeros are Poisson. The response $\mathbf{y}$ was obtained by first generating a uniform (0, 1) random vector $\mathbf{U}$ of length n and then assigning $y_i = 0$ if $U_i \le p_i$ and $y_i \sim \text{Poisson}(\lambda_i)$, otherwise.

ZIP and ZIP(τ) MLE's and likelihood ratio intervals were found by the Newton–Raphson algorithm using the S function ms, which is based on an algorithm of Gay (1983). (The function ms is included in the 1991 version of S and was described by Chambers and Hastie [1992].) ZIP regression always converged when started from the positive Poisson MLE and one EM step for $\boldsymbol{\gamma}$, even for $n = 25$. For ZIP(τ) regression, the Newton–Raphson algorithm always converged from the ZIP MLE's for $n = 100$. But for $n = 50$, these starting values failed in 15 runs. In 14 of those runs, $|\hat{\gamma}_2| > 1{,}600$, and in the remaining run, $\hat{\gamma}_2 = -64$. (This is not as alarming as it might appear. Standard logistic regression coefficients are infinite when the x_i's whose responses are 0 and the x_i's whose responses are 1 do not overlap. Likewise, ZIP regression breaks down when the set of zeros that are classified as perfect and the set of zeros that are classified as Poisson come from nonoverlapping x_i's.) In the 15 runs with outlying $\hat{\gamma}_2$'s, ZIP(τ) MLE's

were found by the Newton–Raphson algorithm when the log-likelihood was first maximized over $\boldsymbol{\beta}$ for fixed τ, as described in Section 3. For $n = 25$, the Newton–Raphson algorithm failed to find ZIP(τ) MLE's when started from the ZIP MLE's in 180 runs, but it succeeded in these runs when τ was arbitrarily initialized to 10. In short, the ZIP and ZIP(τ) regressions were not difficult to fit.

4.2 Simulated Properties of the MLE's

Table 2 summarizes the finite-sample properties of the estimated coefficients. For all n, the mean $\hat{\boldsymbol{\beta}}$ is close to the true $\boldsymbol{\beta}$ for both ZIP and ZIP(τ) regressions. For $n = 50$, however, the mean $\hat{\boldsymbol{\gamma}}$ is far from $\boldsymbol{\gamma}$, and for $n = 25$ the mean $\hat{\boldsymbol{\gamma}}$ is off by a factor of 4,500. This is not surprising, because even standard logistic regression MLE's have infinite bias in finite samples. ZIP regressions are, of course, harder to fit than standard logistic regressions because it is not known which of the zeros are perfect. What is encouraging, however, is that $\boldsymbol{\beta}$ can be estimated even when $\boldsymbol{\gamma}$ cannot. Moreover, if we require that the observed information matrix $\mathbf{I}$ be nonsingular, then the bias in $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ decreases dramatically, especially for $n = 25$.

Looking at only runs with nonsingular observed information, the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated from observed information are reasonably close to the sample standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$ (within 8% for ZIP regression and $n = 25$, within 14% for ZIP(τ) regression and $n = 25$, and less than 4% otherwise). The results for $\hat{\gamma}_1$ are slightly worse, and those for $\hat{\gamma}_2$ and $\hat{\tau}$ are noticeably worse (about 50% too small for $n \leq 50$); that is, the precision in the estimated slope of the relationship between the covariate and the log odds of the probability of a perfect 0 is hard to estimate well. As often happens, standard deviations based on expected information seriously underestimate the variability in the MLE's and should not be trusted. Standard deviations based on observed information are also optimistic but to a much lesser extent.

Table 2. Behavior of ZIP and ZIP(τ) Coefficients As Estimated From 2,000 Simulated Trials

	ZIP				ZIP(τ)		
	β_1	β_2	γ_1	γ_2	β_1	β_2	τ
				Mean			
$n = 25$	1.476	−2.118	−6,449.	9,370.	1.466	−2.026	32,736.
$n = 50$	1.497	−2.090	−4.565	−4.027	1.489	−2.026	1.192
$n = 100$	1.496	−2.037	−1.555	1.876	1.491	−2.004	1.088
Asymptotic	1.500	−2.000	−1.500	2.000	1.500	−2.000	1.000
			*Mean omitting runs with singular observed information**				
$n = 25$	1.483	−2.075	−1.834	1.629	1.477	−2.037	1.495
$n = 50$	1.497	−2.082	−1.587	1.514	1.489	−2.025	1.191
				Median			
$n = 25$	1.504	−2.084	−1.607	2.091	1.497	−2.012	1.126
$n = 50$	1.513	−2.037	−1.481	1.939	1.505	−2.020	1.059
$n = 100$	1.505	−2.025	−1.506	2.000	1.502	−1.996	1.034
Asymptotic	1.500	−2.000	−1.500	2.000	1.500	−2.000	1.000
				Standard deviation			
$n = 25$	.332	1.055	103,573.	158,075.	.314	.756	210,328.
$n = 50$	.218	.704	70.125	159.99	.203	.462	1.028
$n = 100$	.155	.493	.604	1.742	.145	.319	.466
			Standard deviation omitting runs with singular observed information				
$n = 25$	.323	1.021	2.021	5.130	.303	.749	2.228
$n = 50$	.219	.701	1.096	3.931	.203	.461	1.027
			Mean of the standard deviations estimated from nonsingular observed information				
$n = 25$	.299	.950	1.871	3.906	.285	.657	1.544
$n = 50$	.214	.677	.928	2.315	.202	.454	.669
$n = 100$	.153	.485	.594	1.354	.143	.319	.412
			Standard deviation estimated from expected information				
$n = 25$	.288	.927	1.053	2.170	.254	.602	.672
$n = 50$	.210	.667	.756	1.560	.184	.434	.479
$n = 100$	.151	.476	.539	1.112	.132	.310	.341

*For $n = 100$, observed information was never singular. ZIP observed information was singular in 17 runs for $n = 50$ and in 160 runs for $n = 25$; ZIP(τ) observed information was singular in one run for $n = 50$ and in 62 runs for $n = 25$.

4.3 Simulated Behavior of Confidence Intervals and Tests

Table 3 shows that 95% normal-theory confidence intervals for γ_2 are unreliable for n as large as 100 and one-sided normal-theory error rates for τ are obviously asymmetric. Fortunately, even for n as small as 25, ZIP and ZIP(τ) 95% likelihood ratio confidence intervals perform well when only runs with nonsingular observed information are used (see Table 3). For $n = 25$, 6.3% of the ZIP 95% intervals for β_2 and 5.5% of the 85% intervals for γ_2 did not cover the true values; 5.6% of the ZIP(τ) 95% intervals for β_2 and 5.6% of the 95% intervals for τ did not cover the true values. For $n = 50$, the corresponding error rates were 5.3% and 5.5% for ZIP and 5.1% and 4.5% for ZIP(τ) regression. The errors were not concentrated on one side of the confidence intervals. Moreover, quantile-quantile plots (not given here) showed that the χ^2_1 distribution is appropriate for testing the null hypothesis that the ZIP(τ) model is correct. In particular, for $n = 25$ twice the difference of the ZIP and ZIP(τ) log-likelihoods is approximately χ^2_1 up until at least the .99 quantile, and the approximation improves with n.

4.4 Simulated Behavior of Properties of the Observable Y

A different question is how well are the parameters of the unconditional distribution of the observable y estimated? For example, how good is the estimate $\exp(\mathbf{X}\hat{\boldsymbol{\beta}})/(1 + \exp(\mathbf{X}\hat{\boldsymbol{\gamma}}))$ of the mean of $\mathbf{Y}$? The left half of Figure 2 shows properties of the relative bias $[\hat{E}(Y|x) - E(Y|x)]/E(Y|x)$ plotted against $E(Y|x)$ for x between 0 and 1. In Figure 2, all 2,000 simulation runs are used, including those with singular observed information, because predictions from models with poorly determined coefficients can be valid. The top left plot shows that the average relative bias is much better for $n = 50$ or 100 than it is for $n = 25$, but even for $n = 15$ the relative bias lies only between -3.8% and 2.5%. For all n, $E(Y|x)$ is, on average, overestimated near $x = 0$ and underestimated near $x = 1$. Loosely speaking $\hat{E}(Y|x)$ shrinks toward $n^{-1}\sum_{i=1}^{n}\hat{E}(Y|x_i)$. Figure 2 also shows that the relative bias in $\hat{E}(Y|x)$ is even smaller for ZIP(τ) regression than for ZIP regression, except for x near 0.

The right half of Figure 2 shows properties of the relative bias of the estimated probability that $Y = 0$; that is, it shows $[\hat{P}(Y = 0|x) - P(Y = 0|x)]/P(Y = 0|x)$ for x between 0 and 1. When ZIP(τ) regression is appropriate, it is better to use it instead of ZIP regression even if n is as large as 100. For $n = 100$, the average ZIP relative bias is between -1.7% and 4.8% and the average ZIP(τ) relative bias is between $-.5\%$ and 1.3%. For $n = 25$, the relative biases are -17.9% and 19.6% for ZIP regression and 1.6% and 7.1% for ZIP(τ) regression.

To summarize, these simulations with one covariate for both $\boldsymbol{\lambda}$ and $\mathbf{p}$ are encouraging. The ZIP and ZIP(τ) regressions were not difficult to compute, and as long as inference was applied only when the observed information matrix was nonsingular, estimated coefficients, standard errors based on observed information, likelihood ratio confidence intervals, and estimated properties of Y could be trusted.

Table 3. Error Rates for 95% Two-Sided ZIP and ZIP(τ) Confidence Intervals From 2,000 Simulated Trials

	ZIP				ZIP(τ)			
	β_2		γ_2		β_2		γ_2	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Normal theory intervals, all runs								
$n = 25$	.0365	.0445	.001	.003	.042	.017	.000	.047
$n = 50$	.025	.037	.011	.0035	.029	.0205	.001	.0415
$n = 100$	.028	.0245	.0255	.0075	.0365	.018	.001	.0345
Likelihood ratio intervals, all runs								
$n = 25$	.0325	.0455	.051	.042	.0365	.033	.0655	.0185
$n = 50$	.0215	.034	.026	.035	.0245	.0255	.025	.020
$n = 100$	.025	.024	.0305	.0275	.031	.0215	.028	.0235
Normal theory intervals, only runs with nonsingular observed information								
$n = 25$	.040	.048	.001	.003	.043	.018	.000	.049
$n = 50$	.025	.037	.011	.004	.029	.021	.001	.042
Likelihood ratio intervals, only runs with nonsingular observed information								
$n = 25$	.031	.032	.026	.029	.027	.029	.037	.019
$n = 50$	.022	.031	.025	.030	.025	.026	.025	.020

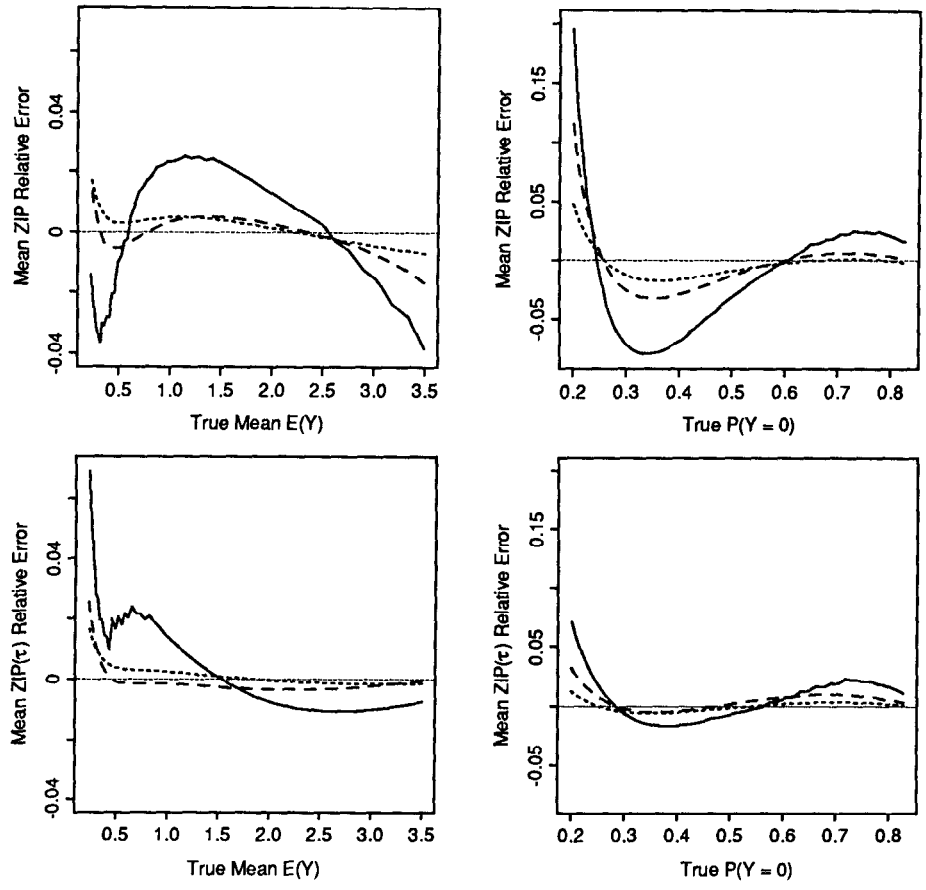

Figure 2. Bias in the Estimated Mean of Y and Estimated Probability of a Zero for 2,000 Simulated Trials. The left plots describe $\hat{E}(Y)/E(Y) - 1$; the right plots show $\hat{P}(Y = 0)/P(Y = 0) - 1$. The solid line represents n = 25, the dashed line represents n = 50, and the dotted line represents n = 100.

5. THE MOTIVATING APPLICATION REVISITED

5.1 ZIP Regression

Tables 4 and 5 summarize the fit of several ZIP models to the data described in Section 1. In Table 4, the probability **p** of the perfect state is constant; in Table 5 it depends on the factors. The simple ZIP models in Table 4 are significantly better than the Poisson models in Table 1, but the models in Table 5 are even better.

The ZIP model in Table 5 with the largest log-likelihood can be represented as

$$\begin{aligned}\log(\hat{\lambda}) &= \text{panel} + \text{pad} + \text{mask} \\ &\quad + \text{opening} + \text{solder} + \text{mask} * \text{solder} \\ &\quad + \text{opening} * \text{solder} \\ \log\left(\frac{\hat{p}}{1-\hat{p}}\right) &= \text{panel} + \text{pad} + \text{mask} + \text{opening}. \qquad (5)\end{aligned}$$

Although there are many parameters in the ZIP regression model (5), 24 for $\boldsymbol{\lambda}$ and 17 for **p**, the observed information matrix is nonsingular. The estimated coefficients are shown in Figure 3 for the full rank parameterization that arbitrarily sets the first level of each main effect to 0. The $\hat{p}_i$'s for the 675 design points range from .0004 to .9998, with a median of .932 and quartiles of .643 and .988, reinforcing the conclusion that the probability of the perfect state varies with the factors. The estimated Pois-

Table 4. Fits of ZIP Regression Models With Constant Probability of the Perfect State

Highest term in the model for log(λ)	*Log-likelihood*	*Residual degrees of freedom*
No interactions	−664.9	656
Opening * solder	−630.6	654
Mask * solder	−630.7	652
Mask * opening	−622.8	650
Mask * solder + opening * solder	−614.1	650
Opening * solder + mask * opening	−592.9	648
Mask * solder + mask * opening	−594.8	646
Mask * solder + opening * solder + mask * opening	−582.1	644
Mask * opening * solder	−567.1	639

NOTE: Some effects in the mask * opening interaction are not estimable. All models have all main effects; the three-way interaction model also has all two-way interactions.

Table 5. Fits of ZIP Regression Models With Logit(p) = Panel + Pad + Opening + Mask

Highest term in the model for log(λ)	Log-likelihood	Residual degrees of freedom
No interactions	−563.3	640
Opening * solder	−524.0	638
Mask * solder	−537.9	637
Mask * opening	−524.3	634
Mask * solder + opening * solder	−511.2	634

NOTE: The model for p excludes solder, because including it increases the log-likelihood by less than 1.0 whenever λ includes at least one interaction. Observed information is singular when the mask * opening interaction is included in the model for λ, even if a full rank design matrix is used. All models for λ have all main effects.

son means range from .02 to 41.8 with a median of 1.42 and quartiles of .38 and 3.14.

ZIP regression can be difficult to interpret when the covariates affect λ, p, and the mean number of defects $E(Y) = \mu = (1 - p)\lambda$ differently. Fortunately, for these data the effects on p and λ, and hence the effect on μ, generally agree, as Figure 3 shows. In Figure 3, levels of a factor that are near the center of the figure are better, in the sense of producing fewer defects; that is, levels near the top of the plot for log($\boldsymbol{\lambda}$) give large Poisson means; levels near the bottom of the plot for logit($\mathbf{p}$) give few perfect zones. For example, pad g is better than pad b, because pad g gives a higher p and lower λ. In contrast, opening M has a complex effect on quality. It gives a high λ when combined with thick solder but has little effect on p.

It is hard to predict whether $E(Y) = \mu$ improves or degrades when a level of a factor increases both p and λ. Similarly, it is hard to predict how μ changes when $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ change, because no transformation of μ is linear in both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. (The same difficulty arises in interpreting fixed coefficients in log-linear models with random effects.) One way to address such questions is to average the estimated means $\hat{\mu}_i$ over all design points that share the same level of a factor and then compare the averages. For example, the average for mask C, or *marginal mean* of C, is defined by

$$\frac{\sum_{\text{design points } i \text{ with mask } C} \hat{\mu}_i}{\#(\text{design points } i \text{ with mask } C)} = \frac{\sum_{\text{design points } i \text{ with mask } C} (1 - \hat{p}_i)\hat{\lambda}_i}{\#(\text{design points } i \text{ with mask } C)}.$$

Figure 4 shows the marginal means for the terms in Model (5) under the full factorial design of 810 design points. Using the actual unbalanced experimental design instead would misrepresent the unbalanced factors. For example, exactly one board (27 counts) with mask C has small openings and thin solder, which is a "bad" combination, but no board with mask C has large openings and thick solder, which is a "good" combination. Under the unbalanced design the marginal mean for mask C is 7.47; under the balanced design it is only 4.48.

Together, Figures 3 and 4 describe the roles of the factors in detail. For example, pad e improves the mean number of defects because it improves the probability of the perfect state and the mean in the imperfect state. Conditions Athin and Bthick have about the same marginal means, but Athin is more often in the perfect state and Bthick has fewer defects in the imperfect state. Small openings are less likely to be in the perfect state than large openings. But

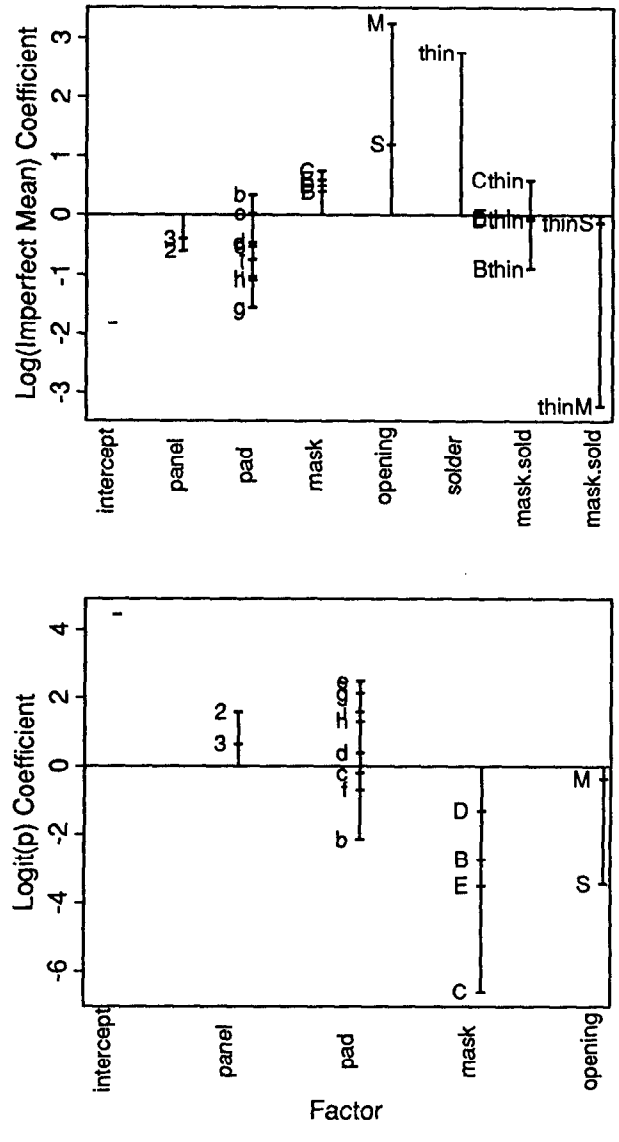

Figure 3. ZIP Regression Effects for the Printed Wiring Board Data. The top half of the figure shows the effects on log(λ). The bottom half shows the effect on logit(p). Effects not shown were set to 0 to obtain a full rank parameterization.

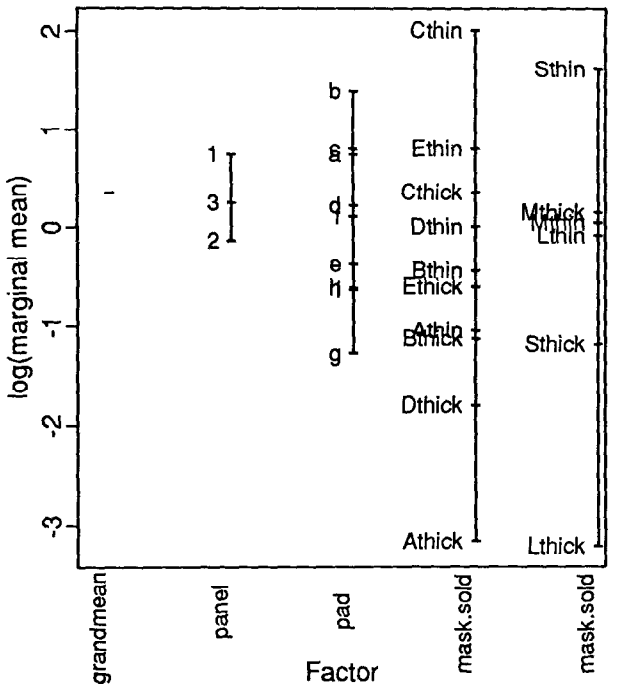

Figure 4. Estimated Marginal Means Under ZIP Regression in the Manufacturing Experiment. Each point averages the estimated means $(1 - p_i)\lambda_i$ with the specified level of the factor under the full factorial design.

small openings with thick solder give fewer defects in the imperfect state than large openings with thin solder. Overall, small openings with thick solder have smaller marginal means.

In short, Figures 3 and 4 show not only which levels give lower mean numbers of defects but also why the means are lower. With plots like these, ZIP regression is nearly as easy to interpret as standard Poisson regression.

5.2 ZIP Regression Compared to Negative Binomial Regression

One could ask whether the excess zeros can be explained more simply as random variation between boards. Expanding the Poisson model by including a random-board factor is not enough to model these data satisfactorily, however. To see this, suppose that counts from different areas j of the same board i are Poisson$(\lambda_{ij}R_i)$, where R_i is a random effect shared by all areas of a board, R_i has a gamma(α, α) distribution, and $\log(\lambda_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$. The likelihood for the 675 observations is then

$$\prod_{i=1}^{25} L(\boldsymbol{\beta}, \alpha | y_{i1}, \ldots, y_{i27})$$

$$= \prod_{i=1}^{25} \int \prod_{j=1}^{27} f(y_{ij}|R_i,\beta) f(R_i|\alpha) dR_i$$

$$= \prod_i \left(\frac{\Gamma(\alpha + \sum_j Y_{ij} - 1)\alpha^\alpha}{\Gamma(\alpha) \prod y_{ij}!} \cdot \frac{\exp(\sum_j y_{ij}\mathbf{X}_{ij}\boldsymbol{\beta})}{(\alpha + \sum_j \exp(\mathbf{X}_{ij}\boldsymbol{\beta}))^{\alpha + \Sigma_j y_{ij}}} \right),$$

which is a product of negative binomials. The MLE $(\hat{\boldsymbol{\beta}}, \hat{\alpha})$ can be found by maximizing the log-likelihood for a set of fixed α and then maximizing over α.

If all terms in a model are required to be estimable, the negative binomial log-likelihood is maximized when λ has the same form as the fitted imperfect mean (5): $\log(\hat{\lambda})$ = panel + pad + mask + opening + solder + mask * solder + opening * solder and $\hat{\alpha} = 1.38$. The maximum log-likelihood is -674.2, which is significantly higher than the Poisson log-likelihood (-700.4) for the same model of the mean. Thus random variability between boards is important if the Poisson model is used.

The best negative binomial model does not predict as well as the ZIP model (5), however. To see this, first estimate probabilities for the negative binomial model at the 675 design points by taking the number of defects Y_{ij} in area j on board i to be Poisson$(\hat{R}_i\hat{\lambda}_{ij})$, where $\hat{R}_i$ is the estimated posterior mean $(\hat{\alpha} + \Sigma_j Y_{ij})/(\hat{\alpha} + \Sigma_j \hat{\lambda}_{ij})$ of R_i. The 25 $\hat{R}_i$'s range from .1 to 2.5 with a mean of 1.00 and a standard deviation of .60. Figure 5 shows that the negative binomial model underestimates the probability of a 0 and the probability of at least nine defects and overestimates the probability of one, two, or three

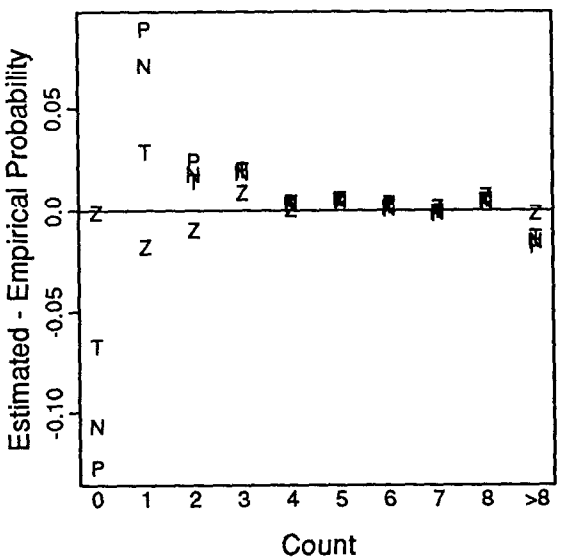

Figure 5. Estimated Probabilities Compared to Empirical Probabilities for the Manufacturing Experiment. In each case, $\Sigma\hat{P}(Y_i = k)/675 - \#(Y_i = k)/675$ is plotted against $\#(Y_i = k)/675$. P denotes Poisson probabilities, Z denotes ZIP, N denotes negative binomial, and T denotes ZIP(τ). All of these probabilities have Model (5) for log(λ).

defects almost as badly as the pure Poisson model does. Additionally, although ZIP regression does not include a parameter for random board variability, the ZIP estimates of the mean total number of defects on a board, $\sum_{j=1}^{27} (1 - \hat{p}_{ij})\hat{\lambda}_{ij}$, are as good as the negative binomial mean estimates $\sum_{j=1}^{27} \hat{\lambda}_{ij}$. The 25 residuals observed − expected for the total number of defects per board have mean 7.6 under the negative binomial and mean −.3 under the ZIP and quartiles −3.8 and 4.2 under the negative binomial and −2.9 and 1.9 under the ZIP. Thus the negative binomial model is not as good as the ZIP regression model for these data, and simply increasing the variability in the Poisson distribution does not necessarily accommodate excess zeros. Of course, inflating a negative binomial model with "perfect zeros" might provide an even better model for the printed-wiring-board data than ZIP regression does. Such a model was not successfully fit to these data, however.

5.3 ZIP (τ) Regression

Since the factor levels in the fitted model (5) that improve λ generally tend to improve p as well, it is reasonable to fit ZIP(τ) regressions to the printed-wiring-board data. Table 6 summarizes the fits of several ZIP(τ) models. The model with mask * solder and opening * solder interactions has significantly higher likelihood than the others and is the only one considered here.

The best ZIP(τ) regression, which has $\hat{\tau} = .81$, fits much better than the Poisson regressions in Table 1 but not as well as the ZIP regression (5). First, the ZIP(τ) model underestimates the probability of a 0 and the probability of at least nine defects, as Figure 5 shows. Indeed, the ZIP(τ) regression is little better than the Poisson for predicting large counts. It is much better than the Poisson at predicting zero, one, or two defects, however. Second, the Pearson residuals, defined by $(Y_i - \hat{E}(Y_i))/[\hat{\text{var}}(Y_i)]^{1/2}$, are generally worse for ZIP(τ) regression than for ZIP regression, as Figure 6 shows. For 16 out of 25 boards, most of the ZIP(τ) absolute residuals exceed the ZIP absolute residuals. For four other boards, the ZIP and ZIP(τ) absolute residuals are nearly the same.

Table 6. Fits of ZIP(τ) Regression Models

Highest term in the model for log(λ)	*Log-likelihood*	*Residual degrees of freedom*
No interactions	−643.2	656
Opening * solder	−609.4	654
Mask * solder	−611.2	652
Mask * opening	−602.4	650
Mask * solder + opening * solder	−595.0	650

NOTE: All models have all main effects. Not all mask * opening coefficients are estimable.

Most of the estimated effects on λ are similar for ZIP(τ) and unconstrained ZIP regression. There are some differences, however. For example, boards with small openings and thick solder cannot be distinguished from boards with large openings and thin solder in the ZIP(τ) model, but they can be distinguished in the ZIP model (large openings with thin solder are worse). Moreover, the marginal mean for Cthick is better than the marginal mean for Ethin under the ZIP model but not under the ZIP(τ) model.

Although ZIP(τ) regression does not fit the printed wiring board data as well, it might be favored over the ZIP regression because it has fewer parameters,

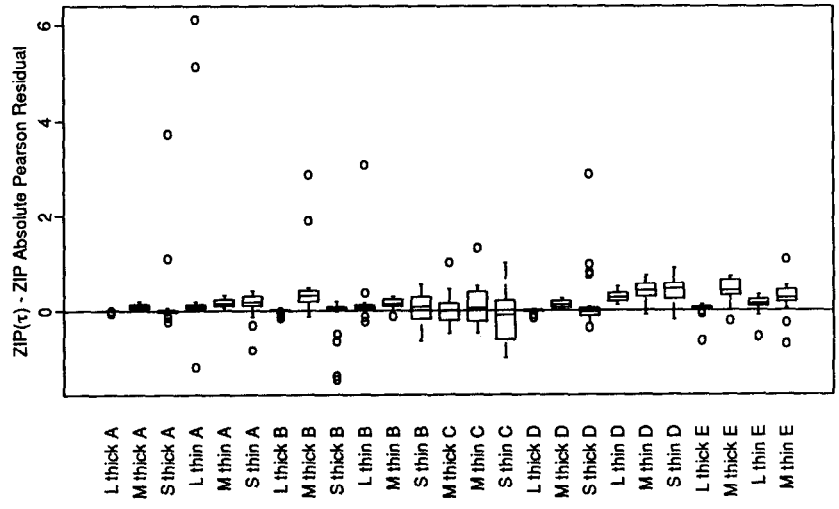

Figure 6. Difference in Absolute Pearson Residuals Under the ZIP(τ) and ZIP Regressions for the 25 Printed Wiring Boards. Positive values imply that the absolute ZIP(τ) residual is larger than the absolute ZIP residual. Thus boxes that are positioned above the y = 0 line favor the ZIP regression model over the ZIP(τ) regression model.

fits better than the Poisson, and can be easier to interpret, since the same coefficients describe the imperfect mean, the probability of the perfect state, and the mean of Y. But in this application, ZIP regression was preferred for two reasons. First, separating the effects on the perfect state from the effects on the imperfect mean made more sense to the engineers than coupling the parameters of the two models. More importantly, ZIP regression gave results that agreed more closely with the engineers' prior experience. When the same manufacturing experiment had been run earlier with older practices and different equipment at a different factory, few counts of 0 had been observed. A Poisson regression fit the earlier data well. When the same kind of Poisson regression model was fit to the later data, however, the effects of some of the factors changed. But when the Poisson model was expanded to the ZIP, the estimated effects on the *imperfect* mean agreed well with the estimated effects from the earlier experiment; that is, the two experiments led to the same conclusions about large numbers of defects, but the earlier experiment was never in a state that produced very low levels of defects and the later one was. This simple conclusion was so appealing that ZIP regression was preferred over ZIP(τ) regression.

6. CONCLUSIONS

ZIP regression is a practical way to model count data with both zeros and large counts. It is straightforward to fit and not difficult to interpret, especially with the plots introduced here. Expressing the zero-inflation term p as a function of λ reduces the number of parameters and simplifies interpretation. Both kinds of parameterization have been fit without difficulty to simulated and real data in this article. Furthermore, ZIP and ZIP(τ) MLE's are asymptotically normal, twice the differences of log-likelihoods under nested hypotheses are asymptotically χ^2, and simulations suggest that log-likelihood ratio confidence intervals are better than normal-theory confidence intervals.

ZIP regression inflates the number of zeros by mixing point mass at 0 with a Poisson distribution. Plainly, the number of zeros in other discrete exponential family distributions can also be inflated by mixing with point mass at 0. The natural parameterization $\eta(\mu)$ of the mean μ of the discrete distribution would be linear in the coefficients $\boldsymbol{\beta}$ of some covariates $\mathbf{B}$, and the logit (or log-log or complementary log-log) link of the probability p of the perfect state would be linear in the coefficients $\boldsymbol{\gamma}$ of covariates $\mathbf{G}$. When the probability of the perfect state depends on $\eta(\mu)$, taking logit(p), or $-\log(-\log(p))$, or $\log(-\log(1-p))$ equal to $-\tau\mathbf{B}\boldsymbol{\beta}$, $\mathbf{B}\boldsymbol{\beta} - \log(\tau)$, or $\alpha - \tau\mathbf{B}\boldsymbol{\beta}$ should make the computations tractable. Mixing zeros with nonexponential family discrete distributions, such as negative binomial with unknown shape, however, may be computationally more difficult.

ACKNOWLEDGMENTS

Many thanks are owed Don Rudy, who provided not only the data but also patient explanations about printed wiring boards and an attentive audience for new ways to model the data, and John Chambers, Anne Freeny, Jim Landwehr, Scott Van der Weil, and the referees and associate editor whose thoughtful comments led to a better article.

APPENDIX: SOME DETAILS

A.1 CONVERGENCE OF THE EM ALGORITHM

The EM algorithm is monotone in the sense that the log-likelihood $L(\boldsymbol{\theta})$ (3) satisfies $L(\boldsymbol{\theta}^{(k+1)}) \geq L(\boldsymbol{\theta}^{(k)})$, where $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ (for example, Dempster et al. 1977). Because $L(\boldsymbol{\theta})$ is bounded above, monotonicity implies that the sequence of log-likelihoods generated by the EM iterations converges, although not necessarily to a stationary point of L. The proof of that depends on the smoothness of $Q(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = E[L_c(\boldsymbol{\theta}^* | y, z) | y, \boldsymbol{\theta}]$.

Conditional on y_i, the unobserved indicator Z_i of the perfect state is identically 0 if $y_i > 0$ and Bernoulli(r_i) if $y_i = 0$, where r_i is the posterior probability of the perfect state:

$$r_i = \frac{e^{\mathbf{G}_i\boldsymbol{\gamma} + \lambda_i}}{1 + e^{\mathbf{G}_i\boldsymbol{\gamma} + \lambda_i}} \quad \text{if } y_i = 0$$

$$= 0 \quad \text{if } y_i = 1, 2, \ldots .$$

Therefore,

$$Q(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\mathbf{G}_i\boldsymbol{\gamma}^* e^{\mathbf{G}_i\boldsymbol{\gamma} + \exp(\mathbf{B}_i\boldsymbol{\beta}^*)} + y_i\mathbf{B}_i\boldsymbol{\beta}^* - e^{\mathbf{B}_i\boldsymbol{\beta}^*} - \log(y_i!)}{1 + e^{\mathbf{G}_i\boldsymbol{\gamma} + \exp(\mathbf{B}_i\boldsymbol{\beta}^*)}},$$

which is continuous in $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$. Using theorem 2 of Wu (1983), this continuity guarantees that the sequence of EM log-likelihoods converges to a stationary point of $L(\boldsymbol{\theta})$. Moreover, since the first derivative of $Q(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}^*$ is continuous in $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$, it follows that $\boldsymbol{\theta}^{(k)}$ converges to a local maximizer of $L(\boldsymbol{\theta})$ (corollary 1, Wu 1983).

A.2 ASYMPTOTIC DISTRIBUTION OF $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ AND THE ZIP LOG-LIKELIHOOD

Define $\mathbf{q} = 1 - \mathbf{p}$ and $\mathbf{s} = 1 - \mathbf{r}$, and let $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$, $\hat{\mathbf{r}}$, and $\hat{\mathbf{s}}$ be the analogous quantities with the MLE's substituted for the true parameter values. Then the

observed information matrix corresponding to the ZIP log-likelihood (3) is

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{I}_2 \\ \mathbf{I}_3 & \mathbf{I}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{G}' & 0 \\ 0 & \mathbf{B}' \end{bmatrix} \begin{bmatrix} \mathbf{D}_{\hat{\gamma},\hat{\gamma}} & \mathbf{D}_{\hat{\gamma},\hat{\beta}} \\ \mathbf{D}_{\hat{\gamma},\hat{\beta}} & \mathbf{D}_{\hat{\beta},\hat{\beta}} \end{bmatrix} \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{B} \end{bmatrix},$$

where $\mathbf{D}_{\hat{\gamma},\hat{\gamma}}$ is diagonal with elements $\hat{\mathbf{q}} - \hat{\mathbf{s}}$, $\mathbf{D}_{\hat{\gamma},\hat{\beta}}$ is diagonal with elements $-\hat{\boldsymbol{\lambda}}\hat{\mathbf{s}}$, and $\mathbf{D}_{\hat{\beta},\hat{\beta}}$ is diagonal with elements $\hat{\boldsymbol{\lambda}}(1 - \hat{\mathbf{r}})(1 - \hat{\boldsymbol{\lambda}}\hat{\mathbf{r}})$. Expected information is

$$\mathbf{i}_{\gamma,\beta} = \begin{bmatrix} \mathbf{G}' & 0 \\ 0 & \mathbf{B}' \end{bmatrix} \begin{bmatrix} \mathbf{d}_{\gamma,\gamma} & \mathbf{d}_{\gamma,\beta} \\ \mathbf{d}_{\gamma,\beta} & \mathbf{d}_{\beta,\beta} \end{bmatrix} \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{B} \end{bmatrix},$$

where $\mathbf{d}_{\gamma,\gamma}$ is diagonal with elements $\mathbf{p}(\mathbf{r} - \mathbf{p})$, $\mathbf{d}_{\gamma,\beta}$ is diagonal with elements $-\boldsymbol{\lambda}\mathbf{p}(1 - r)$, and $\mathbf{d}_{\beta,\beta}$ is diagonal with elements $\boldsymbol{\lambda}(1 - \mathbf{p}) - \boldsymbol{\lambda}^2\mathbf{p}(1 - \mathbf{r})$.

If $n^{-1}\mathbf{i}_{\gamma,\beta}$ has a positive definite limit, then, as in the work of McCullagh (1983),

$$n^{1/2} \begin{bmatrix} \hat{\gamma} - \gamma \\ \hat{\beta} - \beta \end{bmatrix} \sim \text{asymptotically normal}(0, ni^{-1}_{\gamma,\beta}).$$

Observed information can be substituted for expected information in the asymptotic distribution. If $(\hat{\gamma}_0, \hat{\beta}_0)$ maximizes the log-likelihood (3) under a null hypothesis H_0 of dimension q_0 and $(\hat{\gamma}, \hat{\beta})$ maximizes the log-likelihood (3) under a nested alternative hypothesis H of dimension $q > q_0$, then

$$2\left[L(\hat{\gamma}, \hat{\beta}) - L(\hat{\gamma}_0, \hat{\beta}_0)\right] \sim \text{asymptotically } \chi^2_{q-q_0}$$

(for example, see McCullagh 1983). Therefore, twice the difference of nested ZIP log-likelihood under nested alternative and null hypotheses is asymptotically chi-squared.

A.3 ASYMPTOTIC DISTRIBUTION OF $(\hat{\beta}, \hat{\tau})$ AND THE ZIP(τ) LOG-LIKELIHOOD

Define $\mathbf{p}$, $\mathbf{r}$, $\mathbf{q}$, and $\mathbf{s}$ as before, but with $\mathbf{B}$ substituted for $\mathbf{G}$ and $-\tau\boldsymbol{\beta}$ substituted for γ. Then the observed information matrix corresponding to the ZIP(τ) log-likelihood (4) is

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{I}_2 \\ \mathbf{I}_3 & \mathbf{I}_4 \end{bmatrix},$$

where $\mathbf{I}_1 = \mathbf{B}'\mathbf{D}_1\mathbf{B}$ and $\mathbf{D}_1$ is diagonal with elements $\hat{\tau}^2(\hat{\mathbf{q}} - \hat{\mathbf{s}}) + \hat{\boldsymbol{\lambda}}(1 - \hat{\mathbf{r}})(1 - \hat{\boldsymbol{\lambda}}\hat{\mathbf{r}}) + 2\hat{\tau}\hat{\boldsymbol{\lambda}}\hat{\mathbf{s}}$, $\mathbf{I}_2 = \mathbf{B}'\mathbf{D}_2$, $\mathbf{D}_2$ is diagonal with elements $\hat{\mathbf{r}} - \hat{\mathbf{p}} + \hat{\tau}\mathbf{B}\hat{\boldsymbol{\beta}}(\hat{\mathbf{q}} - \hat{\mathbf{s}}) + \hat{\boldsymbol{\lambda}}\mathbf{B}\hat{\boldsymbol{\beta}}\hat{\mathbf{s}}$, and $\mathbf{I}_4 = (\mathbf{B}\hat{\boldsymbol{\beta}})'\mathbf{B}\hat{\boldsymbol{\beta}}(\hat{\mathbf{q}} - \hat{\mathbf{s}})$. Expected information is

$$\mathbf{i} = \begin{bmatrix} \mathbf{i}_1 & \mathbf{i}_2 \\ \mathbf{i}_3 & \mathbf{i}_4 \end{bmatrix},$$

where $\mathbf{i}_1 = \mathbf{B}'\mathbf{D}_1\mathbf{B}$ and $\mathbf{D}_1$ is diagonal with elements $\tau^2(\mathbf{p}(\mathbf{r} - \mathbf{p}) - \mathbf{r}(1 - \mathbf{r})) + \boldsymbol{\lambda}(1 - \mathbf{p}) - \boldsymbol{\lambda}\mathbf{p}(1 - \mathbf{r}) + 2\tau\boldsymbol{\lambda}\mathbf{p}(1 - \mathbf{r})$, $\hat{\mathbf{I}}_2 = \mathbf{B}'\mathbf{D}_2$ and $\mathbf{D}_2$ is diagonal with elements $\tau\mathbf{B}\boldsymbol{\beta}\mathbf{p}(\mathbf{R} - \mathbf{p}) + \boldsymbol{\lambda}\mathbf{B}\boldsymbol{\beta}\mathbf{p}(1 - \mathbf{R})$, and $\hat{\mathbf{I}}_4 = (\mathbf{B}\boldsymbol{\beta})'\mathbf{B}\boldsymbol{\beta}\mathbf{p}(\mathbf{r} - \mathbf{p})$.

If $n^{-1}\mathbf{i}_{\beta,\tau}$ has a positive-definite limit, then

$$n^{1/2} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\tau} - \tau \end{pmatrix} \sim \text{asymptotically normal}(0, n\tilde{\mathbf{i}}^{-1}_{\beta,\tau}).$$

Observed information can be substituted for expected information in the asymptotic distribution. If $(\hat{\boldsymbol{\beta}}_0, \hat{\tau}_0)$ maximizes the log-likelihood (4) under a null hypothesis H_0 of dimension q_0 and $(\hat{\boldsymbol{\beta}}, \hat{\tau})$ maximizes the log-likelihood (4) under a nested alternative hypothesis H of dimension $q > q_0$, then $2[L(\hat{\boldsymbol{\beta}}, \hat{\tau}) - L(\hat{\boldsymbol{\beta}}_0, \hat{\tau}_0)] \sim$ asymptotically $\chi^2_{q-q_0}$.

[*Received April 1990. Revised July 1991.*]

REFERENCES

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Chambers, J. M. (1977), *Computational Methods for Data Analysis*, New York: John Wiley.

Chambers, J. M., and Hastie, T. (eds.) (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cohen, A. C. (1963), "Estimation in Mixtures of Discrete Distributions," in *Proceedings of the International Symposium on Discrete Distributions*, Montreal, pp. 373–378.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Farewell, V. T. (1986), "Mixture Models in Survival Analysis: Are They Worth the Risk?" *Canadian Journal of Statistics*, 14, 257–262.

Feuerverger, A. (1979), "On Some Methods of Analysis for Weather Experiments," *Biometrika*, 66, 665–668.

Gay, D. M. (1983), "ALGORITHM 611—Subroutines for Unconstrained Minimization Using a Model/Trust-Region Approach," *ACM Transactions on Mathematical Software*, 9, 503–524.

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*, Ser. B, 46, 149–192.

Heilbron, D. C. (1989), "Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data," unpublished technical report, University of California, San Francisco, Dept. of Epidemiology and Biostatistics.

Johnson, N. L., and Kotz, S. (1969), *Distributions in Statistics: Discrete Distributions*, Boston: Houghton Mifflin.

McCullagh, P. (1983), "Quasi-likelihood Functions," *The Annals of Statistics*, 11, 59–67.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

Meeker, W. Q. (1987), "Limited Failure Population Life Tests: Application to Integrated Circuit Reliability," *Technometrics*, 29, 51–65.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.