International Biometric Conference
Cape Town, December 1998

Models for count data with many zeros

Martin Ridout
Horticulture Research International-East Malling,
West Malling, Kent ME19 6BJ, UK.
email: `Martin.Ridout@hri.ac.uk`

Clarice G.B. Demétrio
DME/ESALQ, University of São Paulo, 13418-900 Piracicaba SP, Brazil
email: `clarice@carpa.ciagri.usp.br`

John Hinde
School of Mathematical Sciences, Laver Building, University of Exeter,
North Park Road, Exeter EX4 4QE, UK.
email: `J.P.Hinde@exeter.ac.uk`

SUMMARY

We consider the problem of modelling count data with excess zeros and review some possible models. Aspects of model fitting and inference are considered. An example from horticultural research is used for illustration.

1. Introduction

Poisson regression models provide a standard framework for the analysis of count data. In practice, however, count data are often overdispersed relative to the Poisson distribution. One frequent manifestation of overdispersion is that the incidence of zero counts is greater than expected for the Poisson distribution and this is of interest because zero counts frequently have special status. For example, in counting disease lesions on plants, a plant may have no lesions either because it is resistant to the disease, or simply because no disease spores have landed on it. This is the distinction between *structural zeros*, which are inevitable, and *sampling zeros*, which occur by chance.

In recent years there has been considerable interest in models for count data that allow for excess zeros, particularly in the econometric literature. These models complement more conventional models for overdispersion that concentrate on modelling the variance-mean relationship correctly. Application areas are diverse and have included manufacturing defects (Lambert, 1992), patent applications (Crepon & Duguet, 1997), road safety (Miaou, 1994), species abundance (Welsh *et al.*, 1996; Faddy, 1998), medical consultations

Key words: Count data; EM algorithm; Hurdle models; Negative Binomial; Poisson; Zero inflation.

(Gurmu, 1997), use of recreational facilities (Gurmu & Trivedi, 1996; Shonkwiler & Shaw, 1996) and sexual behaviour (Heilbron, 1994).

In this paper we review and compare these methods with particular focus on potential applications in agricultural and horticultural research. Section 2 provides a survey of models that have been proposed. Sections 3 and 4 briefly discuss some aspects of model fitting and inference and Section 5 looks at biological examples.

2. A survey of models for count data with excess zeros

We shall consider excess zeros particularly in relation to the Poisson distribution, but the term may be used in conjunction with any discrete distribution to indicate that there are more zeros than would be expected on the basis of the non-zero counts. Of course it is also possible for there to be fewer zero counts than expected, but this is much less common in practice.

In discussing different models, it is helpful to have a particular example in mind, and for this we consider the number of roots, Y, produced by a plant cutting during a period in a propagation environment. We emphasize, however, that this example is purely illustrative; not all of the mechanisms that we discuss are intended to be realistic biologically. The baseline model is $Y \sim \text{Poisson}(\mu)$.

2.1 Mixed Poisson distributions

Cuttings vary, for example in their basal diameter and in the position on the stockplant from which they were taken. When these factors are not explicitly taken into account, we may expect the Poisson parameter to vary from cutting to cutting, leading to a mixed Poisson distribution. In particular, if the Poisson parameter is μV, where V is a random variable with expected value one and variance α, representing the unobserved heterogeneity, then $E(Y) = \mu$ and

$$\text{var}(Y) = \mu + \alpha\mu^2. \tag{1}$$

Feller (1943) and Mullahy (1997) have shown that the probability of zero in a mixed Poisson distribution is greater than the probability of zero in an ordinary Poisson distribution with the same mean.

Mixed Poisson distributions have been used widely to model overdispersed data; see Hinde & Demétrio (1998) for a recent review. The most commonly used distribution is the negative binomial but other distributions may be more appropriate for modelling data with excess zeros, because, unlike the negative binomial, they can have more than one mode, including a mode at zero. Examples include the Neyman Type A and Pólya-Aeppli distributions. In mixed Poisson regression models, covariates are usually introduced via a log-linear model for μ, as in the standard Poisson model. Extended models can also be considered that allow the degree of dispersion to depend on covariates (Hinde & Demétrio, 1998).

2.2 Zero-modified distributions

An extreme form of mixture distribution arises when a proportion ω of the cuttings are unable to root, and for the remainder the Poisson parameter takes the fixed value λ. Then Y has a *zero-inflated Poisson* (ZIP) distribution, given by

$$\Pr(Y=y)=\begin{cases}\omega+(1-\omega)\exp(-\lambda) & y=0\\ (1-\omega)\exp(-\lambda)\lambda^y/\,y! & y>0\end{cases} \tag{2}$$

It is possible for ω in equation (2) to assume negative values, giving a *zero-deflated* distribution, although the distribution can no longer arise as a mixture distribution. Zero-deflated data seldom arise in practice, however, and we shall assume $0 \le \omega < 1$ in this paper. Zero-inflated forms of other count distributions, such as the negative binomial, can be defined similarly. Gupta, Gupta & Tripathi (1996), for example, investigated the zero-inflated form of the generalized Poisson distribution. For the zero-inflated Poisson distribution

$$\begin{aligned} E(Y) &= (1-\omega)\lambda=\mu,\\ \operatorname{var}(Y) &= \mu+\left(\frac{\omega}{1-\omega}\right)\mu^2. \end{aligned}$$

The second of these equations has the same form as equation (1).
Mullahy (1986), Heilbron (1989, 1994) and Lambert (1992) pioneered the use of regression models based on the ZIP distribution. Lambert (1992) considered models in which

$$\log(\lambda)=X\boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{\omega}{1-\omega}\right)=Z\boldsymbol{\gamma}$$

where X and Z are matrices of covariates and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of parameters. The two sets of covariates may or may not coincide. When they do coincide, more parsimonious models may be developed by supposing that the two linear predictors are related in some way. Perhaps the simplest such model, which Lambert refers to as the ZIP(τ) model, has

$$\log(\lambda)=X\boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{\omega}{1-\omega}\right)=\tau X\boldsymbol{\beta}$$

where τ is a scalar parameter. This implies that $\omega=\left(1+\lambda^{-\tau}\right)^{-1}$. A great variety of alternative models can be generated by using different link functions for λ and/or ω. Greene (1994) gives details of analogous zero-inflated negative binomial regression models.
Shonkwiler & Shaw (1996) proposed a generalization of the ZIP distribution in which there is an underlying bivariate Poisson process giving rise to unobservable count variables Y^* and D, with marginal distributions Poisson($\lambda+\zeta$)

and Poisson($\theta + \zeta$) respectively, where ζ is the covariance between Y^* and D. The observed count, Y, is zero if either Y^* or D is zero, and is equal to Y^* otherwise. The distribution of Y reduces to the ZIP distribution, with $\omega = \exp(-\theta)$, when $\zeta = 0$.

Crepon & Duguet (1997) considered a somewhat similar model in which there are unobserved variables Y^* and C, and again the observed count, Y, is zero if either Y^* is zero or $C < 0$, and is equal to Y^* otherwise. In their model there are latent variables U and V that are normally distributed with zero mean, and which may be correlated. Conditional on $U = u$, Y^* has a Poisson distribution with mean $\exp(X\boldsymbol{\beta} + u)$ whereas $C = Z\boldsymbol{\gamma} + V$. This reduces to a standard ZIP model, with probit link function for ω, when the distribution of U is degenerate with U always taking the value zero.

2.3 Hurdle models

In practice, propagation experiments are often analysed by considering separately the proportion of cuttings that rooted and the mean number of roots per rooted cutting. This approach recognises the possibility that the mechanisms that determine whether or not a cutting roots at all may differ from the mechanisms that determine how many roots are produced by cuttings that do root. To develop models in this framework we need to specify

(a) the probability, π_0, that a cutting fails to root;

(b) a distribution (defined on the positive integers) for the number of roots when a cutting does root.

In the econometric literature, this type of model is called a *hurdle model* (Mullahy, 1986), where $\pi_+ = 1 - \pi_0$ is the probability of clearing the "hurdle" and generating a non-zero count. Another name is *two-part model* (Heilbron, 1994). Typically the distribution in part (b) is the zero-truncated form of a standard discrete distribution such as the Poisson or negative binomial distribution, though distributions defined directly on the positive integers, such as the logarithmic distribution could also be considered. If the truncated Poisson distribution is used then the distribution of Y is

$$\Pr(Y = y) = \begin{cases} \pi_0 & y = 0 \\ \dfrac{(1-\pi_0)e^{-\lambda}\lambda^y}{(1-e^{-\lambda})\, y!} & y > 0 \end{cases}$$

This is just a reparameterisation of the zero-inflated Poisson distribution, given by equation (2) with $\pi_0 = \omega + (1-\omega)e^{-\lambda}$. However, in regression contexts different parameters (π_0 or ω) are modelled and the hurdle and zero-inflated models are no longer equivalent.

When the same covariates affect π_0 and λ, it is useful to consider a model that involves the complementary-log-log link function for π_+ and the log link function for λ, say

$$\log(\lambda) = X\boldsymbol{\beta} \quad \text{and} \ \log[-\log(1-\pi_+)] = X\boldsymbol{\gamma}$$

because this reduces to the standard Poisson model when $\boldsymbol{\beta} = \boldsymbol{\gamma}$. This model was proposed originally by Mullahy (1986), and is an example of what Heilbron (1994) calls a compatible model, because the probability of a zero count is compatible with the distribution of positive counts when the two linear predictors are equal. Heilbron (1994) and Mullahy (1986) discuss other compatible models.

Whilst compatible models have the desirable property of reducing to a standard count data model when the parameters are suitably constrained, other, non-compatible, models may be more useful empirically. For example, in an application involving health care utilization, Gurmu (1998) used a generalized logistic model for the probability of a non-zero count in conjunction with a zero-truncated negative binomial distribution for the non-zero counts.

Most applications of hurdle models have assumed that the linear predictor for π_+ is unrelated to the linear predictor for μ, and this has computational advantages which will be discussed in Section 3. However, it seems a rather restrictive assumption in practice. For example, if only a small proportion of cuttings succeed in rooting under certain experimental conditions, these cuttings usually have fewer roots than rooted cuttings from more successful treatments. Analogously to Lambert's ZIP(τ) model, we may consider, for example, a Poisson hurdle model with

$$\log(\lambda) = X\boldsymbol{\beta} \quad \text{and} \quad \log[-\log(1-\pi_+)] = \tau X\boldsymbol{\beta} \ .$$

2.4 Semi-parametric hurdle models

If a standard Poisson model is fitted to data that are overdispersed, then under fairly general conditions the parameter estimates remain consistent and reasonably efficient. Standard errors of parameter estimates will be underestimated, but use of a simple heterogeneity adjustment (McCullagh & Nelder, 1989, Section 6.2.3) can correct for this. This is a robust approach, which, unlike maximum likelihood estimation, does not require strong distributional assumptions.

Unfortunately, for hurdle models (Gurmu, 1997), as for zero-truncated models (Grogger & Carson, 1991), misspecification of the underlying distribution leads to inconsistent parameter estimates. Grogger & Carson (1991), for example, fitted zero-truncated Poisson models to data simulated from zero-truncated negative binomial distributions, and found biases of up to 30% in the estimated parameters. The source of this inconsistency is the fact that the mean of a zero-truncated distribution depends on the form of the zero probability. For example, if a Poisson distribution and a negative binomial distribution with the same mean are truncated at zero, the means of the truncated distributions will differ. This is not simply because the variance of the distributions differs; the same is true of, for example, a negative binomial distribution and a Neyman Type A distribution that have the same mean and variance.

To provide a more robust approach, Gurmu (1997), developed a semi-parametric hurdle model. Consider again a mixed Poisson model in which the distribution of Y, conditional on V, is Poisson$(V\mu)$ and V has a distribution with probability density function $h(v)$ and $E(V) = 1$. Provided that the distribution of V satisfies some smoothness conditions, the density function $h(v)$ has an infinite Laguerre series expansion. If the expansion is truncated after a finite number of terms and re-normalised to give a proper density function, it is possible to derive a complicated, but explicit, expression for the distribution of Y, thus allowing maximum likelihood estimation of parameters. The method is semi-parametric insofar as the choice of the number of terms at which the expansion is truncated is data-driven, for example using the Akaike information criterion. Using a single term from the expansion is equivalent to assuming a gamma distribution for V and hence a negative binomial distribution for Y. Gurmu (1997) gives an example involving health care utilization.

2.5 Birth process models

A different approach is to suppose that the emergence of roots during the propagation period follows a pure birth process with

$$\Pr\{Y(t+\delta t) = y+1 \,|\, Y(t) = y\} = \lambda_y\, \delta t + o(\delta t).$$

If $\lambda_y = \lambda$ is independent of y then the distribution of the number of roots present at the end of the propagation period (which may be taken as time $t = 1$) is Poisson(λ). Alternatively, if the sequence λ_y increases linearly with y then the distribution is negative binomial. More generally, any count distribution can be obtained by suitable choice of the sequence λ_y, (Faddy, 1997). In particular, the sequence

$$\lambda_y = \begin{cases} \lambda_0 & y = 0 \\ \lambda_1 & y > 0 \end{cases}$$

with $\lambda_1 > \lambda_0$ may be useful for data with excess zeros. This would imply that the rate at which new roots are formed increases after the first root has formed. For regression models, λ_0 and λ_1 can be related to covariates via log-linear models.

2.6 Threshold models

When the number of different non-zero observations, say m, is small, Saei *et al.* (1996) suggest the use of *threshold models.* The basic idea is to assume the existence of a continuous latent variable V such that if V lies in the interval $(\theta_{y-1}, \theta_y]$ then the response $Y = y$ is observed. The cumulative distribution of Y is given by $\Pr(Y \leq y) = G(\theta_y)$ where G is the cumulative distribution function of the latent variable V and it is modelled by the threshold parameters θ_k, $k = -1, 0, \ldots, m$ (with $\theta_{-1} = -\infty$ and $\theta_m = \infty$). The model is extended to the regression situation by assuming that the linear predictor $\eta = X\boldsymbol{\beta}$ simply moves all thresholds up or down by the same amount, that is $\Pr(Y \leq y) = G(\theta_y - \eta)$. This is the standard ordinal regression model discussed by McCullagh (1980). Various common choices for G are the normal, logistic or extreme-value cumulative distributions. This model makes no specific assumption about the form of the probability distribution of the response Y and so may be particularly useful in situations where the non-zero part of the data is not easily modelled. The model also has the property that the covariate effects for the zero and non-zero counts are the same. Saei and McGilchrist (1997) extend this model to include random effects in the linear predictor.

3. Model fitting

Most of the models that we have discussed involve a full specification of the distribution of counts, and maximum likelihood methods are therefore appropriate for parameter estimation. The EM algorithm is a natural contender for zero-inflated models (Lambert, 1992) by formulating the model in terms of an unobserved binary indicator W of whether the observation is a structural or sampling zero. In the M-step the $\boldsymbol{\beta}$ parameter vector is estimated from a weighted fit of the standard distribution for Y and the $\boldsymbol{\gamma}$ parameter vector is obtained by fitting a binary regression model to the current estimate of the indicator W. The E-step simply involves updating the estimate of W by the expected value of its conditional distribution given Y and the current estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Although this is very simple to implement convergence can be rather slow and direct maximization of the likelihood may be preferred. Greene (1994) reports that gradient methods work well in practice, and gives details of the computations. This approach is implemented in the LIMDEP package (http://www.limdep.com) which has facilities for fitting various types of zero-inflated Poisson and negative binomial models. GLIM macros and a Genstat procedure implementing Lambert's EM algorithm are available from the third author.

As we noted earlier, most applications of hurdle models have assumed that the linear predictor for π_+ is unrelated to the linear predictor for λ. In this case the likelihood function factorises and the two sets of parameters can be estimated from separate analyses of the proportion of non-zero counts and the positive counts. Zero-truncated Poisson models for the positive counts are generalized linear models, and Demétrio & Ridout (1994) provide GLIM macros. Grogger & Carson (1991) discuss the fitting of zero-truncated negative binomial models. Both types of model can be fitted in LIMDEP, and also in the free package COUNT (http://gking.harvard.edu/stats.shtml). The

COUNT package also includes a specific module for fitting hurdle models. Simple threshold models can be fitted using standard ordinal regression modelling procedures in, for example, Genstat, GLIM, Minitab, SAS or LIMDEP. Including random effects in these models leads to a generalized linear mixed model. Saei and McGilchrist (1997) discuss the fitting of such models using approximate maximum likelihood, or REML, estimation and suitable software is available in SAS and Genstat.

4. Inferential aspects

Given a baseline model, score tests can be useful in determining whether a more complex model is appropriate, without the need to fit the more complex model. Van den Broek (1995) gives a score test for comparing a standard Poisson model with a ZIP model. It is assumed that ω, the proportion of excess zeros in the ZIP model, does not depend on covariates. This test is discussed also by Mullahy (1986, 1997). We have developed a score test for comparing a ZIP model with a zero-inflated negative binomial model, the details of which are too lengthy to give here and will be published elsewhere. When there are no covariates this is equivalent to a score test for comparing the zero-truncated Poisson distribution with the zero-truncated negative binomial distribution (Gurmu, 1991).

Comparisons between alternative models, for example between a zero-inflated model and a hurdle model, are often comparisons of non-nested models. Several applications have used the test statistic of Vuong (1989) in this context, following the suggestion of Greene (1994). Alternatives have included use of the Akaike information criterion (e.g. Miaou, 1994) and an approach based on embedding the alternative models in an artificial compound model (Crepon & Duguet, 1997).

5. Biological examples

Various authors have considered the ZIP distribution as a possible model for biological count data. A recent entomological example is Desouhant *et al.* (1998) who found that the distribution gave a good fit to 25 out of 31 data sets involving the chestnut weevil. However, there appear to have been few biological applications of zero-inflated regression models. Exceptions are the work of Heilbron (1989, 1994) on sexual behaviour in relation to AIDS risk, and Welsh *et al.* (1996) who modelled data on the abundance of Leadbeater's possum. The possum data were also analysed by Faddy (1998) using various types of birth process model, including the model discussed in Section 2.5. Welsh *et al.* (1996) also used a hurdle model as did Ridout & Demétrio (1992). Saei & McGilchrist (1997) apply the threshold model with random effects to data on the use of chemotherapy across the counties of Washington State.

Table 1 provides an additional data set. The data are the number of roots produced by 270 micropropagated shoots of the columnar apple cultivar *Trajan*. During the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultured on media containing different concentrations of the cytokinin BAP, in growth cabinets with an 8 or 16 hour photoperiod. The full experimental background is given by Marin *et al.* (1993). A striking feature of the data is that although almost all shoots produced under the 8 hour photoperiod rooted, only about half of those produced under the 16 hour photoperiod did.

Fitting a sequence of Poisson regression models, with factors for the two different photoperiods and the four different levels of BAP, there is a very large photoperiod effect and marginal evidence of a significant interaction. However, the residual deviance for the full interaction model is 813.0 on 262 df, giving strong evidence of lack of fit or overdispersion. The large numbers of zeros for the 16 hour photoperiod are an obvious problem for the Poisson fit and a score test for zero-inflation has a value of 7.54 on 1 df, which is highly significant when compared with the asymptotic $N(0,1)$ reference distribution. From the model fitting results in Table 2 we can obtain an equivalent likelihood ratio test with a value of 218.9 on 1 df. The large discrepancy between the square of the score test statistic and the likelihood ratio test statistic is rather surprising, but may be related to the fact that the score test is against a ZIP model with constant ω and here the zero-inflation is very different for the two photoperiods. Van den Broek (1995) notes that the asymptotic distribution may be a poor approximation to the true distribution of the score test statistic when the mean count is high and there are few zeros. A score test for negative binomial overdispersion has a value of 12.22 (the likelihood ratio test has the value 157.3 on 1 df), which is also highly significant.

Comparing the fitted models in Table 2 we see that we need to take account of different degrees of overdispersion and zero-inflation over the two photoperiods, however, there is no evidence that these vary over the hormone levels. With the full interaction model for λ the best fit (smallest AIC and BIC) is obtained for the zero-inflated negative binomial with both ω and α depending on photoperiod. In this model, the parameter estimates of ω and α are close to zero for the 8 hour photoperiod, although they are not exactly zero, reflecting the evidence of overdispersion from the low BAP concentration and zero-inflation from the high BAP concentration. However, a model where α is constant fits almost as well, showing that once we have taken account of the zero-inflation there is only a small degree of additional overdispersion.

Using the zero inflated negative binomial model (ZINB) some simplification of the model for λ is possible. Among the well fitting models (smallest AIC) is one with a linear trend over the log concentration of BAP with different slopes for each photoperiod. For the 8 hour photoperiod the linear trend coefficient is small and positive, 0.064 (s.e. 0.033), while for the 16 hour period it is larger and negative, -0.119 (s.e. 0.068), indicating a significant difference between the two photoperiods.

Basing model selection on BIC, we are led to the even simpler model where

all of the parameters depend only upon photoperiod. Again there is little evidence that the overdispersion parameter α is different for the two groups. However, it is clear that there is additional overdisperion from that accounted for by the simple ZIP model. There is also strong evidence of the photoperiod effect on both λ and ω. These analyses support the conclusion of Marin *et al.* (1993) that in Trajan, as with other columnar apple varieties that have been tested, there is little effect of BAP concentration. This contrasts with conventional apple varieties where BAP has a strong deleterious effect.
Because of limitations of space, we have restricted our attention here to zero-inflated models. A more extensive analyses of these data, using some of the alternative models discussed in Section 2 will be presented elsewhere.

Acknowledgments

We are grateful to FAPESP and The Royal Society for funding exchange visits which have made this collaborative work possible.

References

Crepon, B. and E. Duguet (1997). Research and development, competition and innovation – pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity. *Journal of Econometrics*, **79**, 355–378.

Demétrio, C. and M. Ridout (1994). Letter to the Editor: Coping with extra Poisson variability in the analysis of factors influencing vaginal ring expulsions. *Statistics in Medicine*, **13**, 873–874.

Desouhant, E., D. Debouzie, and F. Menu (1998). Oviposition pattern of phytophagous insects: on the importance of host population heterogeneity. *Oecologia*, **114**, 382–388.

Faddy, M. (1997). Extended Poisson process modelling and analysis of count data. *Biometrical Journal*, **39**, 431–440.

Faddy, M. (1998). Stochastic models for analysis of species abundance data. In D. J. Fletcher, L. Kavalieris, and B. F. Manly (Eds.), *Statistics in Ecology and Environmental Monitoring 2: Decision Making and Risk Assessment in Biology*, pp. 33–40. University of Otago Press.

Feller, W. (1943). On a general class of "contagious" distributions. *Annals of Mathematical Statistics*, **16**, 319–329.

Greene, W. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, Department of Economics, New York University.

Grogger, J. and R. Carson (1991). Models for truncated counts. *Journal of Applied Econometrics*, **6**, 225–238.

Gupta, P., R. Gupta, and R. Tripathi (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis*, **23**, 207–218.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business & Economic Statistics*, **9**, 215–222.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics*, **12**, 225–242.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, **58**, 263–268.

Gurmu, S. and P. Trivedi (1996). Excess zeros in count models for recreational trips. *Journal of Business & Economic Statistics*, **14**, 469–477.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. SIMS Technical Report 9, Department of Epidemiology and Biostatistics, University of California, San Francisco.

Heilbron, D. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, **36**, 531–547.

Hinde, J. and C. Demétrio (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Marin, J., O. Jones, and W. Hadlow (1993). Micropropagation of columnar apple trees. *Journal of Horticultural Science*, **68**, 289–297.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, **42**, 109–142.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (Second ed.). Chapman & Hall.

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections. Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, **26**, 471–482.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, **12**, 337–350.

Ridout, M. and C. Demétrio (1992). Generalized linear models for positive count data. *Revista de Matemática e Estatística*, **10**, 139–148.

Saei, A. and C. McGilchrist (1997). Random threshold models applied to inflated zero class data. *Australian Journal of Statistics*, **39**, 5–16.

Saei, A., J. Ward, and C. McGilchrist (1996). Threshold models in a methadone program-evaluation. *Statistics in Medicine*, **15**, 2253–2260.

Shonkwiler, J. and W. Shaw (1996). Hurdle count-data models in recreation demand analysis. *Journal of Agricultural and Resource Economics*, **21**, 210–219.

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**, 738–743.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–334.

Welsh, A., R. Cunningham, C. Donnelly, and D. Lindenmayer (1996). Modeling the abundance of rare species – statistical-models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

Table 1: Frequency distributions of the number of roots produced by 270 shoots of the apple cultivar *Trajan*, classified by the experimental conditions (BAP concentration and photoperiod) under which the shoots were reared. The table shows the number of shoots that produced $0, 1, \ldots, 12$ roots. Counts that exceeded 12 are shown individually.

	Photoperiod							
	8				16			
BAP (μM)	2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.6
No. of roots								
0	0	0	0	2	15	16	12	19
1	3	0	0	0	0	2	3	2
2	2	3	1	0	2	1	2	2
3	3	0	2	2	2	1	1	4
4	6	1	4	2	1	2	2	3
5	3	0	4	5	2	1	2	1
6	2	3	4	5	1	2	3	4
7	2	7	4	4	0	0	1	3
8	3	3	7	8	1	1	0	0
9	1	5	5	3	3	0	2	2
10	2	3	4	4	1	3	0	0
11	1	4	1	4	1	0	1	0
12	0	0	2	0	1	1	1	0
>12	13,17	13	14,14	14				
No. of shoots	30	30	40	40	30	30	30	40
Mean	5.8	7.8	7.5	7.2	3.3	2.7	3.1	2.5
Variance	14.1	7.6	8.5	8.8	16.6	14.8	13.5	8.5
Overdispersion index	1.42	-0.03	0.13	0.22	4.06	4.40	3.31	2.47

$$\text{Overdispersion index} = \frac{\text{variance - mean}}{\text{mean}}$$

Table 2: Results of fitting various models to the data from Table 1. The λ-model is for the parameter of the basic distribution, the ω-model is for the zero-inflation parameter and the α-model is for the negative binomial overdispersion parameter as in equation (1). In describing the models P is a two level factor for photoperiod, H denotes a four level factor for the BAP levels and Lin(H) is a linear trend over the levels of H, i.e. on the log concentration scale for BAP.

	Models						
Description	λ	ω	α	$-2\log\mathrm{L}$	df	AIC	BIC
Poisson	H*P	0	0	1556.9	262	1572.9	1601.7
	P	0	0	1571.9	268	1575.9	1583.1
Neg-Bin	H*P	0	const	1399.6	261	1417.6	1450.0
	H*P	0	P	1264.6	260	1284.6	1320.6
	H*P	0	H*P	1254.8	254	1286.8	1344.4
	Lin(H)*P	0	P	1270.1	264	1282.1	1303.7
	P	0	P	1272.4	266	1280.4	1294.8
	P	0	const	1403.9	267	1409.9	1420.7
ZIP	H*P	const	0	1338.0	261	1356.0	1388.4
	H*P	P	0	1244.5	260	1264.5	1300.5
	H*P	H*P	0	1238.2	254	1270.2	1327.8
	Lin(H)*P	P	0	1250.2	264	1262.2	1283.8
	P	P	0	1261.3	266	1269.3	1283.7
	P	const	0	1355.2	267	1361.2	1372.0
ZINB	H*P	const	const	1324.8	260	1344.8	1380.8
	H*P	P	const	1232.5	259	1254.5	1294.1
	H*P	P	P	1226.3	258	1250.3	1293.5
	H*P	H*P	H*P	1205.6	246	1253.6	1340.0
	Lin(H)*P	P	P	1231.0	262	1247.0	1275.8
	P	P	P	1237.7	264	1249.7	1271.3
	P	P	const	1243.9	265	1253.9	1271.9
	P	const	const	1336.5	266	1344.5	1358.9
	const	P	const	1257.8	266	1265.8	1280.2

$\mathrm{AIC} = -2\log\mathrm{L} + 2$ (number of fitted parameters)

$\mathrm{BIC} = -2\log\mathrm{L} + \log n$ (number of fitted parameters)