Why nematologists should use zero-inflated models when modeling count data. (And how to use them)

August 16, 2016

1 Introduction

Counts of nematodes, abundance measures, and rates within a unit volume of soil are all common count measures of nematodes. The method of counting after elutriation is time consuming but conceptually straightforward for the researcher. In our own work [] we have found count models of nematodes as well as root count measures in vineyards to be inadequate statistical models in most observed data. The count data are often observed with a high prevalence of zeroes, indeed often beyond the frequency of zeroes suggested by the Poisson model itself. In cases with excess zeroes the use of a Poisson model is inadequate to describe the data. Zero-inflated models offer an alternative that accounts for the excess zeroes observed as well as providing additional over dispersion flexibility. Recall, a distribution is called over-dispersed if the variance exceeds the mean, $\sigma^2 > \mu$.

In many cases nematologists, and scientists in general use the Poisson model not because it is appropriate for the observed data but because the scientist is familiar with the modeling approach of calculating sample means. This article is meant to bring the zero-inflated models to the attention of scientists modeling count data and point out the pitfalls of using traditional methods in the presence of excess zero counts. The use of zero inflated models is common in other scientific disciplines such as (cite list[]) and has been seen in the analysis of nematode count data by Murray et al. [14]. Murray mention that the zero-inflation might come not from the lack of nematodes but from the symbiotic relation certain species of nematodes have with the root system, once the nematode embeds in the root system there is no ability to measure the organism. The only time the organism may be measured is when recently-hatched pre-parasitic second stage juveniles (SRKN-J2) are detectable in soil prior to host invasion.

There are readily available software for fitting zero-inflated models. We provide a table of common statistical software to fit zero-inflated models and compare relevant aspects of these software. Moreover, for the scientist who prefers to 'cook their own' statistics we also provide a spreadsheet in an appendix at the publisher's website so that the scientist may perform the calculations on simplified zero-inflated Poisson models themselves. Also the spreadsheet serves as an instructional aid, the scientist may study the formulae and calculations to ensure their understanding of the model and resulting calculations.

Beyond the statistical aspects of zero-inflated models of count data there are very important scientific considerations for using these models, not the least of which is that the failure to use zero-inflated models when the data are in fact zero-inflated leads to bias in the reported rates of nematodes per unit volume of soil. The reported rate estimates under the traditional models will be underestimated and may cause a grower to fail to act when in fact the should make efforts to control their nematode population by perhaps cleansing field tools before moving from one area to another in the field and other mitigation methods beyond methyl bromide outlined by Zasada et al. [20]. Additionally, the scientist might think that as long as they collect a sufficient number of samples the bias from mis-specifying the Poisson model will be negligible, we show this is not the case and in fact the larger the underlying rate of nematodes the larger the error in estimation.

- Models encompass traditional count data models as special cases and account for excess zeros common in nematology.

- Readily available-and often free-software has ready made codes to fit zero-inflated models.
- Using Poisson based models when zero-inflation is present leads to biased estimates and larger mean squared errors.

2 Estimation challenges

The Bias of an estimator is defined as the expected value of the estimator minus the parameter to be estimated. In the case of count data often the researcher wants to know the rate parameter which indicates an average number of nematodes per unit of soil volume. Let the rate parameter be denoted λ and the zero inflation probability be denoted ϕ. If the researchers use the common average of the count data, denoted $\bar{y}$, then under a ZIP distribution the expected value of this estimator is

$$\mathbb{E}(\bar{y}) = (1-\phi)\lambda. \tag{1}$$

Thus the bias of the estimate is

$$\underbrace{\mathbb{E}(\bar{y}) - \lambda}_{=bias} = -\phi\lambda, \tag{2}$$

so that the larger the zero-inflation or the rate parameter, the larger the bias of the estimate.

One common approach to handle excess zeroes in count data is to add 1 to each observed count. If the counts are indeed Poisson distributed, the resultant random variable is a size-biased Poisson [3, 2]. Regardless of the underlying distribution the rate estimate will now be biased by 1. To remove this bias you may safely subtract 1 from the average. Note this does not solve the problem of excess zeroes. Another common method especially when analyzing abundance data (e.g. [13]) is to condition on the count being greater than zero which is equivalent to ignoring the zero counts. Under a ZIP distribution one case show that this distribution $\Pr(Y = y|y > 0)$ has a zero-truncated Poisson distribution. For more on the statistical properties of this distribution see Cohen [7] and Singh [16]. Unfortunately the sample mean is not an unbiased estimate of the rate parameter λ in the zero-truncated case either. Moreover, some authors prefer to use a transformed version of the count with a 1 added, some examples are $log(y+1)$ [12, 5] or $\sqrt{2(y+1)}$ [1, 19]. Both lead to biased estimators of the underlying rate. The second transform which uses a squared root, has roots in the variance stabilization literature [11] and would be reasonable without the addition of a 1 only if the underlying counts are Poisson distributed.

One common measure of the closeness of an estimator to a parameter is the mean squared error (MSE). The MSE is defined as the expected value of the square of the difference between the estimator and the true value of the parameter. Formally, write

$$MSE(\bar{y}, \lambda) = \mathbb{E}(\bar{y} - \lambda)^2, \tag{3}$$

For the average as an estimator of the rate the MSE is

$$MSE(\bar{y}, \lambda) = \phi^2\lambda^2 + \frac{\lambda(1-\phi) + \lambda^2(1-\phi)\phi}{n}. \tag{4}$$

This last equation tells us that even with an infinite number of samples ($n \to \infty$) unless the zero-inflation is non-existent, there will be a bias proportional to the rate and the traditional methods will be suboptimal to use in place of estimators of zero-inflated data generating processes. The preceding portion of this section has discussed the statistical estimation of rates for data which are truly zero-inflated, in practice we never know for certain whether the data is truly zero-inflated or not and we must rely on statistical tests to make this determination. The next section discusses the methods to use to make this determination as well as the statistical nuances of these procedures.

3 Model determination: Poisson or ZIP?

There are two commonly used methods to determine whether the count data you are modeling is from a zero-inflated distribution or not: the likelihood ratio (LR) test and the Vuong test [17]. There is a fair amount of disagreement within the scholarly community about when to use the Vuong test or the LR test [18]. Notwithstanding these disputes, the Vuong test is actually a class of tests that are appropriate when comparing two zero-inflated regression models that may or may not be nested. We will not detail the Vuong test for reasons of scope but the interested reader is referred to the paper by Vuong (cited above)) as well as the two econometrics textbooks [4, 9]. The likelihood ratio test method is appropriate when comparing a Poisson model against a zero-inflated model. We will first discuss the likelihood ratio test and then the Vuong test.

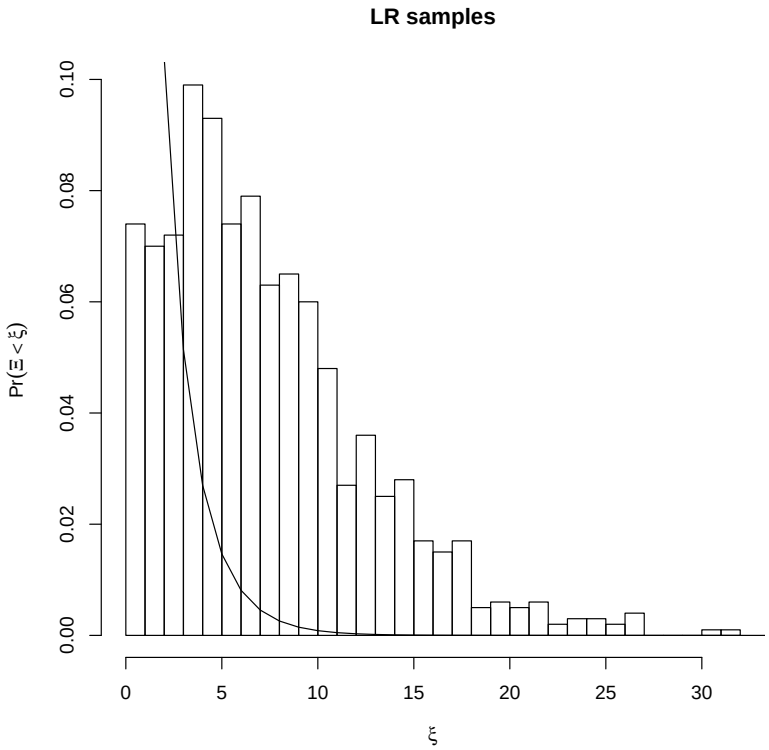
The likelihood ratio (LR) test statistic is constructed by taking twice the logarithm of the ratio of the likelihood under the unconstrained model to the likelihood under the model constrained by the null hypothesis. The standard results in statistics, applying a second order taylor series and taking limits, implies that this statistic will be asymptotically distributed as a chi-squared distribution. In our case of making a determination between a model with and one without zero-inflation the likelihood ratio test will not converge to the standard distribution in the statistical literature because the zero-inflation parameter lies on the boundary ($\phi = 0$) of the parameter space [6, 10, 15]. The case when the likelihood ratio statistic has a parameter on the boundary of the parameter space is a well studied and understood problem within statistics that occurs in other models as well, such as a mixed effect model when you want to test whether one of the variance components is 0 or not [8]. The traditional results state that the distribution of the LR test statistic has a point mass mixture distribution with probability one-half at the zero point and then other half is spread over a chi-squared distribution with a single degree of freedom. Formally twice the log-likelihood random variable, denoted ξ is distributed

$$\xi \stackrel{\text{d}}{=} \frac{\chi_0^2}{2} + \frac{\chi_1^2}{2}, \tag{5}$$

where the zero degree of freedom chi-squared random variable is interpreted as a point mass at zero. To illustrate this phenomena to the reader we generate Figure 3 which displays the results of $10,000$ random draws from a zero-inflated Poisson distribution with $\phi = 0.36788$ and rate parameter $\lambda = 1$. We form the LR test statistic by comparing a Poisson likelihood and a zero-inflated Poisson likelihood and taking twice the logarithm of the difference of the two. The bold line displays the theoretical distribution which contains a point mass at zero and a chi-squared density on the positive axis. This has implications for the nematologist who is testing whether to use a ZIP model versus a Poisson model, the test statistic p-value may be calculated exactly by calculating one-half the upper tail of a chi-squared distribution with 1 degree of freedom, unless the LR test statistic is zero in which case the p-value is 1. Table 3 gives the critical values of the distribution for various common type 1 error rates.

Likelihood ratio test	
P-value	Critical value ξ
0.001	9.549536
0.01	5.411894
0.025	3.841459
0.05	2.705543
0.1	1.642374
0.25	0.454936

Table 1: The critical values of the likelihood ratio test.



4 An algorithm for estimating ϕ and λ with no covariates

In this section we detail an algorithm for estimating the two parameters of a ZIP model that is readily programmable within a spreadsheet with basic arithmetic calculations. Although the algorithm contains a while loop that is potentially infinite, in our experience the number of iterations is typically no more than 5-10. Moreover, the calculations required in each iteration are readily programmed so that the researcher may apply the technique to various data via copy and paste to mimic the iterations of the while loop. The algorithm follows:

Algorithm 1 ZIP EM algorithm

1: **procedure** ZIP EM($\phi^0, \lambda^0, \varepsilon, \vec{y}, \vec{w}$)
2: **for** $i \leftarrow 1, n$ **do**
3: **if** $y_i = 0$ **then**
4: $w_i \leftarrow (1 + \exp[-\lambda_i^0 - \log(\phi_i^0/(1-\phi_i^0))])^{-1}$
5: **else**
6: $w_i \leftarrow 0$
7: $t \leftarrow 1$
8: **while** $\varepsilon > |\phi^t - \phi^{t-1}|$ and $\varepsilon > |\lambda^t - \lambda^{t-1}|$ **do**
9: $\phi^t \leftarrow \sum_{i=1}^n w_i/n$
10:
11: $\lambda^t \leftarrow \sum_{i=1}^n (1-w_i)y_i/(\sum_{i=1}^n 1 - w_i)$
12:
13: **for** $i \leftarrow 1, n$ **do**
14: **if** $y_i = 0$ **then**
15: $w_i \leftarrow (1 + \exp[-\lambda_i^t - \log(\phi_i^t/(1-\phi_i^t))])^{-1}$
16: **else**
17: $w_i \leftarrow 0$
18: $t \leftarrow t + 1$
19:
20: **return** ϕ^t, λ^t ▷ The EM algorithm (MLE) estimates
21:

The Algorithmic description is a formal way of stating the algorithm which in essence involves iterating the calculations on lines 9 and 11 and update after an initialization. For initial values of ϕ^0, λ^0, and $\vec{z}$ I suggest random uniform numbers between 0 and 1 for all three. Note that all final values output by the algorithm (e.g z_i and ϕ^t) will be between 0 and 1 except for λ^t which may be any positive value. Although a while loop is formally used in the algorithm often iterating for $t = 10$ is sufficient for reasonable settings of numerical accuracy.

A spreadsheet to accompany the manuscript is available to illustrate the necessary calculations as well as provide a working template for the nematologist. The spreadsheet is titled `phi_lambda_calc.xlsx` and is available in the supplementary material on the journal website.

5 Discussion

In this brief research note we discussed the zero-inflated distributions and several statistical and scientific implications of nematodes using the standard statistical techniques when modeling nematode abundance with count data. We argue that all nematologists should be familiar with zero-inflated models and use these methods in place of the standard classical statistics to estimate rates of nematodes in soil volume data. For the researcher to determine whether a Poisson model and a ZIP model should be preferred, we provided both likelihood ratio and the Vuong tests. The Vuong test should be used when trying to determine between two competing ZIP models with non-nested covariate structure. Moreover we stressed the computational aspects of these methods as well as the statistical nuances of the methods. Finally we provided an algorithm to estimate the rate of nematodes under a zero-inflated model and the zero-inflation probability component as well as provided a spreadsheet with calculations ready programmed for the researcher to modify as needed for their nematode abundance data.

References

[1] Francis J Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[2] Richard Arratia and Larry Goldstein. Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent? *arXiv preprint arXiv:1007.3910*, 2010.

[3] Richard Arratia, Larry Goldstein, and Fred Kochman. Size bias for one and all. *arXiv preprint arXiv:1308.2729*, 2013.

[4] A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[5] M Centinari, JE Vanden Heuvel, M Goebel, MS Smith, and TL Bauerle. Root-zone management practices impact above and belowground growth in cabernet franc grapevines. *Australian Journal of Grape and Wine Research*, 22(1):137–148, 2016.

[6] Herman Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, pages 573–578, 1954.

[7] A Clifford Cohen. Estimating the parameter in a conditional poisson distribution. *Biometrics*, 16(2):203–211, 1960.

[8] Ciprian M Crainiceanu and David Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185, 2004.

[9] Russell Davidson and James G MacKinnon. *Estimation and inference in econometrics*. Cambridge Univ Press, 1995.

[10] Ziding Feng and Charles E McCulloch. Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, 13(4):325–332, 1992.

[11] Murray F Freeman and John W Tukey. Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, pages 607–611, 1950.

[12] Amanda D Howland, R Paul Schreiner, and Inga A Zasada. Spatial distribution of plant-parasitic nematodes in semi-arid vitis vinifera vineyards in washington. *Journal of nematology*, 46(4):321, 2014.

[13] Ganpati B Jagdale, Ted Holladay, PM Brannen, WO Cline, Paula Agudelo, AP Nyczepir, and JP Noe. Incidence and pathogenicity of plant-parasitic nematodes associated with blueberry (vaccinium spp.) replant disease in georgia and north carolina. *Journal of nematology*, 45(2):92, 2013.

[14] Leigh Murray, Stephen H Thomas, Jill Schroeder, Scott Kreider, Zhining Ou, JM Trojan, and C Fiore. Modeling the root-knot nematode/nutsedge pest complex: Perspectives from weed science, nematology and statistics. 2011.

[15] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[16] Jagbir Singh. A characterization of positive poisson distribution and its statistical application. *SIAM Journal on Applied Mathematics*, 34(3):545–548, 1978.

[17] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

[18] Paul Wilson. The misuse of the vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127:51–53, 2015.

[19] Guan Yu. Variance stabilizing transformations of poisson, binomial and negative binomial distributions. *Statistics & Probability Letters*, 79(14):1621–1629, 2009.

[20] Inga A Zasada, John M Halbrendt, Nancy Kokalis-Burelle, James LaMondia, Michael V McKenry, and Joe W Noling. Managing nematodes without methyl bromide. *Annual review of phytopathology*, 48:311–328, 2010.