Modeling zero-inflated count data when exposure varies: With an application to tumor counts

Gregori Baetschmann and **Rainer Winkelmann***

Department of Economics, University of Zurich, CH-8032 Zurich, Switzerland

Received 16 January 2012; revised 5 July 2013; accepted 29 July 2013

This paper is concerned with the analysis of zero-inflated count data when time of exposure varies. It proposes a modified zero-inflated count data model where the probability of an extra zero is derived from an underlying duration model with Weibull hazard rate. The new model is compared to the standard Poisson model with logit zero inflation in an application to the effect of treatment with thiotepa on the number of new bladder tumors.

Keywords: Complementary log-log link; Exposure; Extra zeros; Poisson regression.

1 Introduction

Count models describe the number of events that occurred during a fixed time period T. In some applications, this time period differs across units of observations, which raises the question how variation in T_i should be accounted for by the model. In standard count data models with exponential mean function, such as the Poisson and the negative binomial regression models, including $\log T_i$ among the regressors allows to test both for proportional changes of the mean function as well as for the absence of exposure effects.

In this paper, we consider instead varying exposure in *zero-inflated count data models* (e.g., Mullahy, 1986; Böhning et al., 1997). Such models are constructed from a binary mixture of an ordinary count model and a distribution with probability mass of one at zero. They account for zero inflation, a situation where the proportion of zeros, conditional on a set of observed characteristics, exceeds that predicted by standard models. The mixing generates two types of zeros, "normal" zeros and "extra" zeros. For example, when counting the number of times a person visits a medical specialist for a particular disease, the extra zero group is made up of people who do not suffer from that disease (when that fact is unobserved by the analyst).

Often, the probability of an extra zero is specified as a logit function, using the same regressors as those determining the conditional mean function of the count model. In this way, continuing the above example, the model allows to distinguish between factors determining disease onset from those determining disease progression, for example, when counting the number of tumors (Joe and Rong, 2005; Hsu, 2007), or the number of decayed teeth (Böhning et al., 1997). Further health-related applications of the zero-inflated Poisson and negative binomial models include Pizer and Prentice (2011), Sari (2009), Sarma and Simpson (2006), Yen et al. (2001), Chang and Trivedi (2003), and Street et al. (1999).

Again, the question arises of how to address the effect of varying exposure time T_i in such models. For example, counting the number of developing tumors using an inflow sample of patients can result in varying exposure if the inflow takes place over an extended period of time whereas the observation

*Corresponding author: e-mail: rainer.winkelmann@econ.uzh.ch, Phone: +41-44-634-22-95, Fax: +41-44-634-49-96

window ends at a particular calendar date that is fixed and identical for everyone in the study. One possible approach is to assume that the probability of an extra zero is time invariant, whereas exposure enters the count part in the usual way. This assumption was made by Lee et al. (2001) in an earlier volume of this journal, who considered exposure effects in the zero-inflated Poisson model but limited exposure effects to a logarithmic offset in the count part of the model.

We argue that this assumption is not very plausible. In many applications, one would expect the probability of an extra zero to decrease with increasing exposure, as does the probability of a normal zero, albeit possibly at a different rate. A natural way of modeling time dependence of extra zeros is in terms of an underlying stochastic process, or single spell duration model, where an extra zero arises if the arrival time of the event (such as onset of disease in the above example) exceeds T_i. In this case, the probability of an extra zero is equal to the survivor function at T_i.

We discuss two alternative specifications, one based on the survivor rate of the log-logistic distribution and one based on the survivor rate of the Weibull distribution, respectively. While the first approach can be implemented in standard zero-inflated count data models by including $\log T_i$ as a regressor in both the logit and in the count part of the model, the second approach leads to an alternative model for zero inflation that has to the best of our knowledge not been considered before.

An interesting aspect of this second approach is that it uses, in the case of the zero-inflated Poisson model, the same probability model for extra zeros and normal zeros, albeit with different parameters. Relatedly, the second approach allows to test whether the hazard rate of the underlying stochastic process is constant, whereas the log-logistic hazard is necessarily nonconstant. We compare both models in an application to the effect of a treatment on the recurrence rate of bladder tumors, based on data from a randomized experiment.

2 Methods

2.1 Zero-inflated count models and exposure

The zero-inflated Poisson model with covariates but ignoring exposure can be written as (see, e.g., Böhning et al., 1997)

$$\Pr(y_i|x_i, z_i) = \begin{cases} \omega(z_i) + (1-\omega(z_i))\exp(-\lambda(x_i)) & \text{for } y_i = 0 \\ (1-\omega(z_i))\dfrac{\exp(-\lambda(x_i))\lambda(x_i)^{y_i}}{y_i!} & \text{for } y_i = 1,\ 2,\ 3,\ \ldots, \end{cases} \tag{1}$$

where y_i is a count-valued random variable, $0 \le \omega_i \le 1$ is a zero inflation parameter, and x_i and z_i are covariates that can be disjunct, overlapping, or identical. The standard zero-inflated Poisson model is obtained setting $\lambda(x_i) = \exp(x_i'\beta)$ and $\omega(z_i) = \exp(z_i'\gamma)/[1+\exp(z_i'\gamma)]$. For a sample of n independent observations, the parameters of the model can be estimated either by maximum likelihood or by exploiting moment conditions (e.g., using nonlinear least squares or Poisson pseudomaximum likelihood, see Staub and Winkelmann, 2013). In order to account for exposure effects in the rate of the count process, let

$$\lambda(x_i, T_i) = \exp(x_i'\beta + \delta \log T_i)\,, \tag{2}$$

where δ is identified if there is variation in T_i across observation units. Proportionality is obtained for $\delta = 1$, an assumption that has been imposed *a priori* by Lee et al. (2001). If $\delta = 0$, there are no exposure effects.

2.2 Extra zeros and exposure

We approach the issue of time dependence of extra zeros from the viewpoint of an underlying stochastic process, or single spell duration model, where an extra zero arises if the arrival time of the event exceeds

T_i and the probability of an extra zero is equal to the survivor function at T_i. For example, adding logarithmic exposure as a regressor to the logit part of a standard zero-inflated Poisson model amounts to the assumption that the time to event has a log-logistic distribution. For $\xi \leq 0$, the log-logistic survivor function and thus the probability of an extra zero is given by

$$\omega(z_i, T_i) = \exp(z_i'\gamma + \xi \log T_i)/[1 + \exp(z_i'\gamma + \xi \log T_i)] = [1 + T_i^{-\xi} \exp(-z_i'\gamma)]^{-1}. \tag{3}$$

In this model, extra zeros decrease with increasing exposure unless $\xi = 0$. Numerous patterns of duration dependence are possible, including nonmonotonic hazard rates (see Johnson et al., 1994).

However, the log-logistic distribution does not allow for a constant hazard rate and thus treats extra zeros differently from "normal" zeros that, under the maintained assumptions, are generated from a Poisson process. To address this discrepancy, we suggest using an alternative formulation for the probability of an extra zero based on a complementary log–log (cloglog) model. Specifically, if events occur randomly over time with constant hazard rate $\exp(z_i'\gamma)$ and T_i is the elapsed time, then

$$S(T_i; z) = \exp(-\exp(z_i'\gamma)T_i) \tag{4}$$

is the survivor rate at time T_i. As before, the probability of an extra zero can be modeled in terms of the survivor rate for the single spell event. Suppose we do not want to impose the constant hazard assumption but rather put it to test. In this case, we can use the survivor rate of the Weibull distribution, whereby

$$\omega(z_i, T_i) = \exp(-\exp(z_i'\gamma)T_i^{\xi}) = \exp(-\exp(z_i'\gamma + \xi \log T_i)). \tag{5}$$

Therefore, the zero-inflated part effectively uses a proportional hazard model for the event "termination of the extra-zero (or "perfect") state by time T_i". A positive ξ means that an increase in exposure reduces the survivor rate and therefore the probability of an extra zero. For $\xi = 1$, we obtain a constant hazard rate. For $\xi > 1$, duration dependence is positive. If $\xi = 0$, the implied hazard rate is zero and extra zeros are time invariant and hence truly "strategic". γ has the usual interpretation for proportional hazard models: $\exp(\gamma)$ gives the hazard ratio for a unit increase in the associated regressor.

To the best of our knowledge, such a modified zero-inflated count data model, where the binary model for extra zeros is based on a cloglog link rather than a logit link, has not been considered in the literature so far. Estimation of the model parameters is quite straightforward. Based on a sample of n independent observations on (y_i, x_i, z_i), $i = 1, \ldots n$, with exposure $T_1, T_2, \ldots, T_n$, respectively, the log-likelihood function of the Poisson-cloglog model for zero-inflated count data can be written as

$$\begin{aligned} \log L(\beta, \delta, \gamma, \xi) = &\sum_{y_i=0} \ln[\exp(-\mu_i) + \exp(-\lambda_i) - \exp(-\mu_i - \lambda_i)] \\ &+ \sum_{y_i>0} \ln[1 - \exp(-\mu_i)] - \lambda_i + y_i \ln \lambda_i, \end{aligned} \tag{6}$$

where $\lambda_i = \exp(x_i'\beta + \delta \log T_i)$ and $\mu_i = \exp(z_i'\gamma + \xi \log T_i)$. The EM algorithm has been shown to work well in this kind of problem, but straight Newton–Raphson maximization is possible as well. For testing, it should be noted that neither do zero-inflated models nest their standard parent models, nor do the logit and cloglog specifications for the extra zeros have a nested structure. Tests can be based on procedures developed for non-nested models, as discussed, for example, in Vuong (1989).

2.3 Further aspects

While zero-inflated models allow for a conditional variance of the count dependent variable that exceeds its mean, the implied degree of *overdispersion* might be insufficient to account for the full

amount of variation in the data. In such a situation, one can use a mixture of a zero mass point with a negative binomial distribution (that arises, for instance, if the Poisson rate is mixed with a gamma distribution) and add logarithmic exposure terms as before.

Once the full model is specified and estimated, one can construct counterfactual outcome distributions for alternative time periods. For example, one can predict annual rates in cases where actual observations are for shorter periods of time. With exposure-dependent zero inflation, the mean does not move proportionally with time. Differentiating

$$\mathrm{E}(y_i|x_i, z_i, T_i) = (1 - \omega(z_i, T_i))\lambda(x_i, T_i) \tag{7}$$

with respect to T_i yields

$$\frac{\partial \mathrm{E}(y_i|x_i, z_i, T_i)}{\partial T_i} = -\frac{\partial \omega(z_i, T_i)}{\partial T_i}\lambda(x_i, T_i) + (1 - \omega(z_i, T_i))\frac{\partial \lambda(x_i, T_i)}{\partial T_i}. \tag{8}$$

The effect of exposure time is proportional only if the effect in the parent model is proportional to exposure ($\partial\lambda(x_i, T_i)/\partial T_i = \lambda(x_i)$) and the probability of an extra zero is independent of time (the zeros are purely strategic). If extra zeros decrease with increasing exposure, the first term on the right-hand side of (8) is positive, and the expected value of such a zero-inflated count model increases overproportionally with increasing exposure.

3 Application: The recurrence of bladder tumors

In this section, we apply the proposed methodology to a study of tumor recurrence in a sample of 85 US veterans who underwent an initial surgery to remove bladder tumors. The data were obtained from the data repository of the *Journal of the Royal Statistical Society*. A detailed description of the study can be found in Byar (1980). Earlier uses of the data, focussing on different methodological aspects, include Sun and Wei (2000), Zhao and Sun (2011), and Wellner and Zhang (2000).

The dataset records pretreatment differences in the severity of disease progression for the 85 patients. The number of initial tumors (INITNR) varied between 1 and 8; the tumor size (SIZE) was classified using an integer scale between 1 and 7. Following surgery, patients were randomized to either placebo (TREATMENT = 0, $n = 47$) or treatment with thiotepa (TREATMENT = 1, $n = 38$). Patients were reexamined during subsequent clinic visits, and at each occasion, the number of new tumors was recorded, and they were removed by surgical procedure. Our dependent variable COUNT is the total number of new tumors recorded over the entire observed patient record time.

The number and timing of follow-up visits vary quite a bit among study participants. The variation in exposure is rooted in the study design: subjects were recruited into the study over a five-year period (between November 1971 and August 1976, see Byar, 1980), while the follow-up period was limited by construction to the end of 1976. Thus, while some patients were observed over a period of 53 months, three participants had only one recorded follow-up visit that took place after a month. Since the variation in exposure is mainly due to the timing of inflow into the study (rather than to treatment status or severity of disease), one can reasonably treat it as exogenous.

The first two columns of Table 1 provide supporting evidence for exogeneity of exposure. They display estimated parameters from log-linear Poisson regressions of exposure (in months) on TREATMENT and on the pretreatment measures INITNR and SIZE. There is no effect of TREATMENT on overall exposure time. Similarly, neither SIZE nor INITNR have an effect on the duration of exposure. Hence, it is justifiable to proceed with the models developed earlier in the paper and treat the between-subject variation in exposure as exogenous.

A second aspect of the study is that there was quite a bit of variation in the frequency of follow-up visits, even keeping exposure constant. The overall numbers range from 1 to 38 visits. Indeed, it has been found by earlier research that the number of visits depended on treatment (Sun and Wei, 2000).

Table 1 Parameter estimates for exposure and number of visits (n=85) (Poisson model, robust standard errors in parentheses).

	Exposure (months)		Number of visits	
TREATMENT	−0.037	−0.032	0.444	0.454
	(0.102)	(0.103)	(0.157)	(0.159)
SIZE		0.026		0.049
		(0.032)		(0.055)
INITNR		0.0003		−0.003
		(0.029)		(0.046)
Pval (χ^2 test)	0.7165	0.8380	0.0049	0.0427

Table 2 Frequency of new tumors for treatment and control group.

	Frequency										
	0	1	2	3	5	6	8	9	10	11+	Total
Control	18	5	1	3	2	1	2	1	2	12	47
Treatment	20	3	6	1	0	2	3	0	0	3	38
Total	38	8	7	4	2	3	5	1	2	15	85

Patients in the thiotepa group tended to visit the clinic centers more often than those in the control group, since the thiotepa treatment required such visits in order to instill the substance into the bladder. Results in columns 3 and 4 of Table 1 confirm this relationship. For our purposes, however, this source of endogeneity does not play any role, since we focus on the total number of newly detected tumors rather than on their timing.

The distribution of the total number of tumors for thiotepa and placebo treatments for the 85 patients is shown in Table 2. The fraction of zeros is 53% for the treatment group, and 38% for the control group. Overall, the thiotepa treatment appears to have a tumor reducing effect. Applications of zero-inflated Poisson regression models allow to control for the two pretreatment variables, SIZE and INITNR, as well as exposure, and to disentangle two channels of the treatment effect: the effect of thiotepa on extra zeros versus the effect on the mean of the count distribution. Formally, we consider the following model:

Extra zeros:

$$\omega_i = \exp[-\exp(\gamma_0 + \gamma_1 TREATMENT + \gamma_2 SIZE + \gamma_3 INITNR + \xi \log(Time))] \tag{9}$$

Count model:

$$\lambda_i = \exp[\beta_0 + \beta_1 TREATMENT + \beta_2 SIZE + \beta_3 INITNR + \delta \log(Time)]. \tag{10}$$

The estimation results for the zero-inflated Poisson model with complementary log–log link (ZIP-cloglog) are shown in the fourth column of Table 3. We see that the treatment effect is attributed equally to the two processes: treatment with thiotepa lowers the hazard rate of the Poisson process by about $\exp(-0.53) - 1 = 41\%$. An almost identical estimate is obtained for the hazard of the extra zero process, although it is not statistically significant in a t-test at conventional significance levels. A lower hazard in the two parts of the model means that the probability of extra and normal zeros both increase due to treatment. If a single-index Poisson model is estimated instead (column 1 of Table 3),

Table 3 Parameter estimates for number of tumors ($n = 85$) (standard errors in parentheses).

	Poisson (1)	Negbin (2)	ZIP-logit (3)	ZIP-cloglog (4)	ZINB-cloglog (5)
γ_1 (trt)			0.712	−0.523	−0.162
			(0.508)	(0.333)	(0.437)
γ_2 (size)			−0.147	0.106	0.161
			(0.160)	(0.114)	(0.175)
γ_3 (init)			−0.447	0.299	0.276
			(0.152)	(0.093)	(0.159)
ξ (log time)			−0.809	0.577	0.649
			(0.310)	(0.234)	(0.479)
β_1 (trt)	−0.788	−1.216	−0.534	−0.534	−1.116
	(0.318)	(0.367)	(0.282)	(0.282)	(0.436)
β_2 (size)	−0.026	−0.020	−0.065	−0.065	−0.119
	(0.095)	(0.119)	(0.076)	(0.076)	(0.126)
β_3 (init)	0.255	0.382	0.132	0.132	0.246
	(0.070)	(0.091)	(0.060)	(0.060)	(0.097)
δ (log time)	0.706	0.837	0.345	0.346	0.256
	(0.270)	(0.237)	(0.267)	(0.268)	(0.393)
σ^2		2.341			0.942
		(0.485)			(0.378)
Log-likelihood	−380.4	−193.8	−261.9	−261.7	−189.5

Note: The probability of an extra zero is modeled as $\omega(z_i, T_i) = [1 + \exp(-z_i'\gamma - \xi \log T_i)]^{-1}$ in (3), and as $\omega(z_i, T_i) = \exp(-\exp(z_i'\gamma + \xi \log T_i))$ in (4) and (5).

then the overall, combined treatment effect is correspondingly larger. The expected number of counts is reduced by −55% due to treatment.

Returning to the ZIP-cloglog model, the estimated exposure effect in the cloglog part is 0.577, with standard error 0.234. Hence, we can reject the null hypothesis of no exposure effect, that is, genuinely strategic extra zeros, as well as the null of proportionality, the latter at the 10% level of significance. A 10% increase in exposure increases the hazard of leaving the zero state by about 6%. For example, if the probability of an extra zero is 20%, this effect translates into a 1.8 percentage point reduction in extra zeros.

Table 3 includes, for comparison, results for two further models, first the Poisson model with logit-type zero inflation (ZIP-logit), and second the negative binomial model with cloglog zero inflation (ZINB-cloglog). The ZIP-logit model gives almost identical results for the count part, whereas the estimated zero inflation parameters differ in sign: this is a consequence of the log-logistic/logit parameterization defined in Eq. (3) where the survivor function increases in $z_i'\gamma$. However, effect sizes and p-values tend to be similar to that of the cloglog specification. The logit estimate of the exposure effect is −0.081; for example, increasing exposure by 10% when the baseline probability of an extra zero is 20% reduces the probability of an extra zero by $0.2 \times (1 - 0.2) \times -0.081 \times 0.1 = -1.3$ percentage points.

The final column 5 of Table 3 shows the ZINB-cloglog results. While the log-likelihood improves substantially relative to the ZIP-cloglog, it is interesting to note that much of the improvement is already achieved by a standard negative binomial model (see column 2 of Table 3, $\log L = -193.8$), while allowing for additional zero inflation only leads to a relatively small further improvement that is statistically insignificant if judged by a Vuong test. However, the predicted probability of extra zeros

is 34.5% and thus substantial even in the ZINB-cloglog model. By contrast, the predicted probability of a normal zero is only 11.3% . In the ZIP-cloglog, almost all zeros are predicted to be extra zeros, 44.5% versus 0.2% .

In zero-inflated models, the overall treatment effect is the combination of two effects, the effect on the mean of the count distribution and the effect on the extra zeros (see Eq. (8)). One should not interpret these in isolation. In the present case, the extra zeros treatment effect is smaller in the ZINB than in the ZIP model, thereby offsetting the larger estimated treatment effect on the mean of the count distribution in the ZINB (-1.116 vs. -0.534 in the ZIP). The attribution of the treatment effect to the two parts of the model differs between the ZIP and the ZINB, likely because the ZINB introduces two sources of heterogeneity, one related to strategic zeros and one related to the intensity of the process, whereas the ZIP only allows for the former. As mentioned before, a formal test rejects the ZIP against the ZINB and thus underlines the relevance of accounting for such additional heterogeneity in this particular application.

4 Concluding remarks

This paper discussed the problem of modeling count data with extra zeros when time of exposure varies. Our approach builds on the class of so-called zero-inflated count data models, a mixture of an ordinary count model and a distribution that is degenerate at zero. We generalize earlier research by Lee et al. (2001) where exposure effects were excluded from the zero-inflated part of the model. This exclusion makes a very strong assumption on the nature of the extra zeros, and we argue that, in many situations, it is more reasonable to allow for exposure effects among extra zeros as well. Under the assumption that extra zeros are generated by a separate Poisson-like process, one should use a cloglog link to parameterize the probability of an extra zero. As in the Poisson process, this allows to adjust for varying period at risk in a theory consistent way, by including the logarithm of exposure time as an offset or as an additional control variable in the inflated part of the model. The extension to a negative binomial model with complementary log–log zero inflation is straightforward.

Acknowledgements We are grateful to two anonymous referees and the editor for helpful comments and suggestions.

Conflict of interest
The authors have declared no conflict of interest.

References

Böhning, D., Dietz, E. and Schlattmann, P. (1997). Zero-inflated count models and their applications in public health and social science. In: Rost, J. and Langeheine, R. (Eds.), *Application of Latent Trait and Latent Class Model in Social Sciences*. Wasemann, Münster, pp. 333–344.

Byar, D. P. (1980). The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In: Pavone-Macaluso, M., Smith, P. H. and Edsmyr, F. (Eds.), *Bladder Tumors and Other Topics in Urological Oncology*. Plenum, New York, pp. 363–370.

Chang, F.-R. and Trivedi, P. K. (2003). Economics of self-medication: theory and evidence. *Health Economics* **12**, 721–739.

Hsu, C. H. (2007). A weighted zero-inflated Poisson model for estimation of recurrence of adenomas. *Statistical Methods in Medical Research* **16**, 155–166.

Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal* **47**, 219–229.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Volume 1. Wiley, New York.

Jones, A. (2007). *Applied Econometrics for Health Economists: A Practical Guide* (2nd edn.). Radcliffe Publishing, Milton Keynes.

Lee, A. H., Wang, K. and Yau, K. K. W. (2001). Analysis of zero-inflated Poisson incorporating extent of exposure. *Biometrical Journal* **43**, 963–975.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**, 1–19.

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

Pizer, S. D. and Prentice, J. C. (2011). Time is money: outpatient waiting times and health insurance choices of elderly veterans in the United States. *Journal of Health Economics* **30**, 626–636.

Sari, N. (2009). Physical inactivity and its impact on healthcare utilization. *Health Economics* **18**, 885–901.

Sarma, S. and Simpson, W. (2006). A microeconometric analysis of Canadian health care utilization. *Health Economics* **15**, 219–239.

Staub, K. E. and Winkelmann, R. (2013). Consistent estimation of zero-inflated count models. *Health Economics* **22**, 673–686.

Street, A., Jones, A. and Furuta, A. (1999). Cost sharing and pharmaceutical utilisation and expenditure in Russia. *Journal of Health Economics* **18**, 459–472.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society*, Series B **60**, 293–302.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics* **28**, 779–814.

Yen, S. T., Tang, C.-H. and Su, S.-J. B. (2001). Demand for traditional medicine in Taiwan: a mixed Gaussian-Poisson model approach. *Health Economics* **10**, 221–232.

Zhao, X. and Sun, J. (2011). Nonparametric comparison for panel count data with unequal observation processes. *Biometrics* **67**, 770–779.

