Using Zero-inflated Models to Analyze Environmental Data Sets with Many Zeroes

Utilizzo di modelli zero-inflazionati per il trattamento di dati ambientali con eccesso di zeri[1]

Lorena C.M. Viviano, Vito M.R. Muggeo, Gianfranco Lovison
Dipartimento di Scienze Statistiche e Matematiche "S.Vianelli"
Università degli Studi di Palermo - Viale delle Scienze - 90128 Palermo
email: (viviano,vmuggeo,lovison) @dssm.unipa.it

Riassunto: L'analisi di dati di conteggio può essere talvolta complessa a causa di un numero di zeri superiore a quello atteso sotto il modello Poissoniano, che rappresenta l'assunzione standard per la modellazione di questo tipo di dati. Obbiettivo primario della comunicazione è quello di impiegare modelli alternativi a quello di Poisson, che contemplino la possibilità di trattare esplicitamente questo eccesso di zeri, per valutare eventuali differenze in termini di bontà di adattamento e di stima dei parametri regressivi. Vengono discussi modelli *Zero Inflated Posson* (ZIP), *Zero Inflated Negative Binomial* (ZINB) e *Hurdle Poisson* (HP) e applicati a due insiemi di dati ambientali reali con un elevato numero di zeri.

Keywords: Count data; Poisson; Negative Binomial; Zero-Inflated; Hurdle.

1. Zero inflated and Hurdle models: an overview

The Poisson distribution is the probability model usually assumed for count data; however, in many real applications it is likely to observe a number of zeroes greater (zero inflation) or smaller (zero deflation) than that expected under the Poisson model. Such situations can be dealt with through models that accommodate the excess of zeroes, such as *Zero Inflated* and *Hurdle* models. These models are encountered in the econometric, demographic and medical literature, and are characterized by a parametric structure that models the 'zero' and 'non-zero' responses separately. Zorn (1996) justifies this approach in terms of a 'dual regime' data generating process: in the first stage, a presence/absence model determines whether the count is zero or non-zero; in the second stage, a count model governs the actual magnitude of the count. Hence the underlying mixture model is :

$$Pr(Y_i = y_i) = \pi_i f_1(y_i) + (1 - \pi_i) f_2(y_i) \qquad i = 1, 2, \ldots, n \tag{1}$$

where for the i^{th} unit, y_i is the count, π_i is the probability of a zero count in the presence/absence model, $f_1(y_i) = I_{\{0\}}(y_i)$ and $f_2(y_i)$ is the p.d.f. of a count random variable.

All models considered in this paper can be represented using (1), through appropriate choices of π_i and $f_2(y_i)$. In particular, if $\pi_i = 0$ and $f_2(y_i)$ is the p.d.f. of the Poisson distribution, we obtain the standard Poisson model as a (degenerate) sub-case. The distinction between Zero-Inflated models and Hurdle models refers to the form of $f_2(y_i)$: in fact, while in Zero-Inflated models a zero can be contributed by the presence/absence model or by the count model, in Hurdle models a zero can only come from the presence/absence model and therefore a (zero)-truncated distribution caters for counts greater

[1]Lavoro svolto con finanziamento PRIN Cofin MIUR 2004 prot.2004173478_003

than zero. In both cases, the usual choices for $f_2(y_i)$ are the Poisson or Negative Binomial distributions, either in their standard (for Zero-Inflated) or truncated (for Hurdle) form; the Negative Binomial is usually preferred when the counts also exhibit over-dispersion. Therefore possible alternatives to the standard Poisson are Zero Inflated Posson (ZIP), Zero Inflated Negative Binomial (ZINB), Hurdle Poisson (HP) and Hurdle Negative Binomial (HNB) models. In practice, due to the computational difficulties met with the HNB model, we focus on the first three models, whose p.d.f. are as follows:

$$\text{ZIP:} \quad Pr(Y=y_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & \text{for} \quad y_i = 0 \\ (1-\pi_i)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} & \text{for} \quad y_i \geq 0 \end{cases}$$

$$\text{ZINB:} \quad Pr(Y=y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi} & \text{for} \quad y_i = 0 \\ (1-\pi_i)\frac{\Gamma(y_i+\phi)}{\Gamma(\phi)\ y_i!}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i} & \text{for} \quad y_i \geq 0 \end{cases}$$

$$\text{HP:} \quad Pr(Y=y_i) = \begin{cases} \pi_i & \text{for} \quad y_i = 0 \\ \frac{(1-\pi_i)\exp(-\mu_i)\mu_i^{y_i}}{(1-\exp(-\mu_i))y_i!} & \text{for} \quad y_i > 0 \end{cases}$$

where μ_i is the expected value of the model and ϕ^{-1} is the (over-)dispersion parameter. Usually a more realistic context considers vectors of covariates, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ say, to be related to μ_i and π_i through proper link functions in the spirit of Generalized Linear Models: $\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and $\text{logit}(\pi_i) = \boldsymbol{z}_i'\boldsymbol{\gamma}$. Moreover note i)using the same set of covariates serves the purpose of identifying the possibly different roles of the same explanatory variable in each stage; ii)π_i and μ_i can be unrelated or function of each other. The inferential method usually applied to obtain maximum likelihood estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is the EM algorithm and results are based to asymptotic theory of those estimators. Score tests can be useful to compare models. Lambert (1992) introduced ZIP models in a manufactural context, while HP models were introduced by Mullay (1986) and then modified by King (1989); see also Long (1997); Zorn (1996); Ridout et al. (1998); Tu (2002).

2. Comparisons of models

In this section, two environmental real data sets are analyzed. Comparisons in terms of parameter estimates and AIC in particular are carried out for four models: Poisson, ZIP, ZINB and HP. The fitted models in both cases are completely additive and levels of factors are coded as dummy variables through a corner point parameterization.

The first data set refers to a daily time series (1997-1999) data to study the effect of air pollution on health in Palermo. The response is the number of deaths for breathing complications which presents a high number of zeroes ($\approx 41\%$). Some covariates that can influence the response have been included in the final model: Influenza epidemics (binary variable where 0 corresponds to absence of influenza), Month (twelve-levels categorical variable), Temperature (24-hours average in °C) and PM_{10} concentration (moving average of lag 0-3 in μg/m^3). The interest lies in estimating the effect of PM_{10} which is one of the major causes of health problems in air pollution studies.

Table 1 reports estimates from the fitted models. Comparisons can be made at two levels: the former refers to the first stage of ZIP, ZINB and HP models (columns '$P(Y_i = 0)$' in which zero vs non zero outcomes are modelled) and the latter refers to the standard Poisson and the second stage of ZIP, HP and ZINB models. Comparing the same models globally, it is possible to see that the AIC's are close, but the best model is still the standard

Table 1: *Results for the Poisson, ZIP, ZINB, HP models (air pollution data set)*

	Poisson	ZIP		ZINB		HP	
	$P(Y_i \geq 0)$	$P(Y_i = 0)$	$P(Y_i \geq 0)$	$P(Y_i = 0)$	$P(Y_i \geq 0)$	$P(Y_i = 0)$	$P(Y_i > 0)$
`PM10`	.008	-2.457	.007	-5.59	.007	-.019	.005
s.e.	.002	1.782	.002	3.47	.002	.006	.004
p-value	*0*	*.168*	*.005*	*.095*	*.005*	*.002*	*.208*
AIC (n.par.)	2735.30 (16)	2737.78 (32)		2736.34 (33)		2756.42 (32)	

Poisson. Furthermore such conclusion is also confirmed by a modified score test (van den Broek, 1995) comparing a Poisson versus a ZIP model. For such data set the excess of zeroes is really plausible under the standard distribution for count data ($p = 1.00$). In the first stage of HP, estimate of PM_{10} has a negative sign suggesting that the probability of a zero outcome is lower than that of a non zero outcome. Results are very similar for PM_{10} (with a positive effect on expected counts) when comparing the second stage of ZIP, ZINB and the classical Poisson model. It is important to underline the presence of very large standard errors for the coefficients in the inflation equation, especially for HP model. p-value of PM_{10} is not significant in the first stage for ZIP and significant for HP, and the opposite happens in the second stage of the two models. It could be explained by a different way of considering zero counts in the two models and/or by specific features of such dataset, including a modest percentage of zeroes and a low expected value for non-zero counts. However, the findings concerning the health effect of PM_{10} are substantially unchanged.

The second data set is referred to a study of bathing water quality in the district of Palermo. Data are collected in 2001 and are characterized by $n = 1386$ observations. Our goal is to analyze the effect of some covariates (Month, Water Temperature, Oxygen, Sea Condition) on the response variable 'Number of Fecal Streptococcuses' (counts in 100 ml of water), that ranges from 0 to 200 and presents a great percentage of zeroes ($\approx 54\%$). Results from the fitted models are displayed in Table 2 where the `mo-` variables are the dummies relevant to April-September period (data are collected only in the bathing season); the `sea-` variables refer to categories of 'Sea Condition' (respectively 'calm', 'almost wavy', 'wavy') and `temp` and `oxy` stand for the continuous 'Water Temperature' and 'Oxygen'. The van den Broek test suggests that it is advisable to consider a ZIP model instead of the classical Poisson distribution ($p < 0.0001$); this is confirmed also by the AIC value of the Poisson model which is dramatically larger then the AIC of the Zero Inflated models. However among the inflated models, there exists a noticeable improvement in accounting for extra-variability: the ZINB has to be preferred by far, likely due to its capability to catch both excess of zeroes and overdispersion. As regards to parameter estimates, it is worth noting that the sign of coefficients is substantially unchanged among the different models (both logit and log-linear components); however in ignoring the zero-inflation and/or overdispersion the significance is heavily overstated. For instance the significant effect of some variables (actually `mo3`, `mo5`, `sea2` and `temp`) observed in the Poisson model disappears in the ZINB. From a biological standpoint it is worthwhile to stress the role of the months corresponding to beginning and closing of bathing season.

Table 2: *Results for the Poisson, ZIP, ZINB, HP models (bathing water quality data set)*

	Poisson	ZIP		ZINB		HP	
	$P(Y_i \geq 0)$	$P(Y_i = 0)$	$P(Y_i \geq 0)$	$P(Y_i = 0)$	$P(Y_i \geq 0)$	$P(Y_i = 0)$	$P(Y_i > 0)$
mo2(s.e.)	1.15(.04)	-.79(.21)	.76(.04)	-.79(.25)	.81(.21)	-.79(.21)	.76(.04)
p-value	*.00*	*.00*	*.00*	*.00*	*.00*	*.00*	*.00*
mo3(s.e.)	.16(.05)	-.35(.28)	.12(.05)	-.45(.35)	.04(.27)	-.35(.28)	.12(.05)
p-value	*.00*	*.22*	*.02*	*.20*	*.87*	*.21*	*.02*
mo4(s.e.)	.08(.06)	-.31(.37)	.08(.06)	-.49(.47)	-.21(.39)	-.31(.37)	.08(.06)
p-value	*.22*	*.41*	*.20*	*.30*	*.60*	*.41*	*.20*
mo5(s.e.)	.69(.06)	-1.45(.38)	.21(.06)	-1.92(.53)	.01(.39)	-1.45(.38)	.21(.06)
p-value	*.00*	*.00*	*.00*	*.00*	*.987*	*.00*	*.00*
mo6(s.e.)	1.06(.05)	-.79(.33)	.78(.06)	-.84(.42)	.70(.35)	-.79(.33)	.78(.06)
p-value	*.00*	*.02*	*.00*	*.05*	*.05*	*.01*	*.00*
sea2(s.e.)	-.18(.02)	.27(.12)	.001(.02)	.30(.15)	.006(.13)	.27(.12)	.001(.02)
p-value	*.00*	*.03*	*.95*	*.05*	*.96*	*.03*	*.95*
sea3(s.e.)	.03(.03)	-.60(.19)	-.20(.03)	-.88(.29)	-.31(.18)	-.60(.19)	-.20(.03)
p-value	*.22*	*.00*	*.00*	*.00*	*.08*	*.00*	*.00*
temp(s.e.)	-.03(.01)	.03(.04)	-.03(.01)	.05(.05)	-.000(.04)	.03(.04)	-.03(.01)
p-value	*.00*	*.36*	*.00*	*.27*	*.99*	*.36*	*.00*
oxy(s.e.)	-.06(.002)	.02(.01)	-.04(.18)	.02(.01)	-.04(.01)	.02(.01)	-.04(.001)
p-value	*.00*	*.02*	*.00*	*.05*	*.00*	*.02*	*.00*
AIC (n.par.)	37701 (10)	21467.14 (20)		6792.54 (21)		21459.24 (20)	

3. Conclusions

Count data with zero mass need particular care and should be properly modelled. Unlike the seeming excess of zeroes, given the covariates, sometimes the standard Poisson suffices. Otherwise wrong conclusions can be reached and different models (ZIP, ZINB, HP) should be considered. In the mortality data set, the classical Poisson model is still the best choice, while in the second data set ZINB is preferable. Possible drawback in employing these alternative models is the difficulty of using standard software as computational aspects are often non-negligible. Our analysis were conducted in R employing two libraries (*zeroinfl - hurdle*) created by S. Jackman (*http://pscl/standford.edu/content.html*). Other possible functions are *yipp, zipbipp, zipoissonX* (vgam), *zicounts* and *fmr* (gnlm) created by J. Lindsey (*http://www.luc.ac.be/* $\sim$ *jlindsey/rcode*).

References

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

Long, J. (1997). *Regression models for categorical and limited dependent variables*, Sage.

Mullay, J. (1986). Specifications and testing of some modified count data model, *Journal of Econometrics* **33**, 341–365.

Ridout, M., Demetrio, C. and Hinde, J. (1998). Models for count data with many zeros, *Proceedings of the XIX*th *International Biometric Conference* pp. 179–192.

Tu, W. (2002). *Encyclopedia of Environmetrics (Zero-inflated data)*, Vol. 4, Wiley.

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

Zorn, C. (1996). Evaluating zero-inflated and Hurdle Poisson specifications, *Midwest Political Science Association* **18-20 april**, 1–16.