Modeling confidential data via modified hurdle mixed models

Marco Enea, Antonella Plaia, Vincenza Capursi

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche - University of Palermo, Italy

E-mail for correspondence: marco.enea@unipa.it

Abstract: In this work we analyze event count data that show zero inflation with non-ignorable missingness due to confidentiality. The emphasis is not on imputation methods but on the choice of a suitable model. As a motivating example, we analyze a dataset on University student's mobility (SM) in Italy whose records are reported aggregately. In such data, records with less than three moving students are automatically removed. To detect the determinants of SM, similarly to a hurdle mixed model, we estimate two separate models, a binomial mixed model for the "zero" part and a two-truncated negative mixed binomial for the "non-zero" part.

Keywords: GLMM; GAMLSS; truncation; students' mobility.

1 Introduction

The analysis of event count data showing over/under-dispersion and/or too many zero counts has become very common in the literature. Popular modeling approaches are the zero-inflated (Lambert, 1992) or zero-altered (also said hurdle) Poisson regression models (Mullahy, 1986), or their negative binomial counterpart if further overdispersion is present. Additionally, the family of count data models needs adjustments in order to accommodate for non-ignorable missingness. In fact, it is common to deal with official statistics data whose feature is the limited disclosure due to privacy and confidentiality reasons. Masking is the most common technique to protect data. In Italy a common masking procedure is to fix a data truncation threshold at two, that is tables with less than three cell counts cannot be released. In this case, the fitting of the zero-inflated model or the hurdle model can be misleading. In fact, estimates could be biased if the amount and the mechanism of missingness is non-ignorable.
In this work, in order to reduce the amount of bias, we suggest to estimate two models separately, a binomial model for the zero part and a two-truncated negative binomial model for the non-zero part.

As a motivating dataset having the above outlined characteristics, we analyze the 2010/2011 cohort of the Italian freshmen in 58 Italian Universities to detect the determinants of SM within the Italian territory. The data come from the Italian university students' register (ANSU is the Italian acronym). Such data are provided in aggregate form and some information on students' characteristics is available, such as the region of origin and destination. Here the focus is on the moving freshman defined as a student who has moved from a region to another to enroll at the University. Unlikely from other European countries, mobility rates of students in Italy are intrinsically very low. From the dataset, about 173000 resident and about 19000 non-resident students are recorded, but their true number is greater because of the masking mechanism (about 240000 freshmen also including foreign students and private universities). Three regions, Sardegna, Valle d'Aosta and Trentino Alto Adige, were not considered as regions of destination. Two universities are present in Sardegna but these have no observed counts of incoming students greater than 2. The other two regions are so small that not all the study fields are represented. Thus they have been aggregated with their neighboring larger regions Piemonte and Veneto, respectively. There were no observed non-resident students in six Universities i.e. Benevento, Palermo, NapoliSeconda, NapoliParthenope, Catania and Catanzaro. Table 1 describes the covariates used in the analysis.

TABLE 1. Description of the covariates. The first category is the baseline

Variable	Categories/Range	Description
RRes	Abruzzo,...,Veneto	Origin region (20 cat.)
RUni	Abruzzo,...,Veneto	Destination region (17 cat.)
University	Bari,...,Verona	University (58 cat.)
Sex	Female, Male	Gender
HSgrade	100-; 90-99; 80-89; 70-79; 60-69;	High school final grade
HStype	liceo; magistrale; vocational; technical;	High school type
Field	Health6; Health3; Social5; Social5; Scientific5; Scientific3; Humanistic3;	Study field and course duration (in years)
INC.res	-4.5, 4.1	Per-capita average income (centered) of the origin region, in thousands of euros
NU.res	1 - 7	No. of Universities in the region of residence
Censis	-18.3, 16.9	University "quality" score (centered) according to CENSIS-Repubblica

The records are "depicted" by the covariates described above. For each record the number of non-resident students is observed. The count distribution is reported in Figure 1. For graphical reasons, the pick of zeroes representing the resident students is not shown. The skewness of the distribution suggests a non-ignorable missingness for counts one and two.

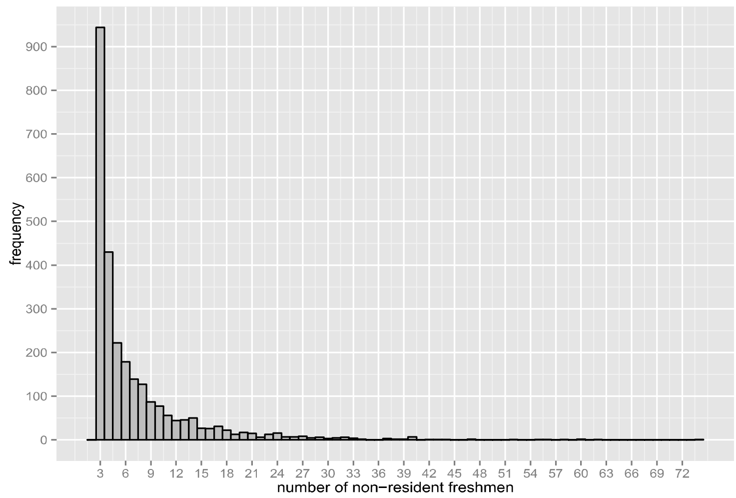

FIGURE 1. Observed non-resident freshmen records.

2 Hurdle mixed model

We first introduce the hurdle negative binomial mixed model and then we introduce the modification we suggest to accommodate for non-ignorable missingness. The choice of the negative binomial rather than the Poisson distribution is due to the likely presence of overdispersion into the non-zero part of the model. Let $Y'_{ijk} \sim NB(\mu_{ijk}, \sigma)$ with type I parameterization as in GAMLSS (Rigby and Stasinopoulos, 2005), with i $(i = 1, ..., n_j)$ indexing students grouped by covariate profiles, j $(j = 1, ..., m_k)$ the universities and k $(k = 1, ..., K)$ the regions, respectively. Under the hurdle model, a new variable Y_{ijk} is introduced whose distribution is:

$$P(Y_{ijk} = 0) = 1 - \pi_{ijk}, \quad (1)$$

$$P(Y_{ijk} = y | y > 0) = \frac{\pi_{ijk} P(Y'_{ijk} = y)}{P(Y'_{ijk} > 0)}, \quad (2)$$

where π_{ijk} is the probability to be non-resident and $P(Y'_{ijk} = y)/P(Y'_{ijk} > 0)$ is the truncated-at-zero NB. In practice Y_{ijk} is zero-altered negative binomial (ZANB) distributed. The hurdle mixed regression model is a two-equation system:

$$\begin{aligned}\text{logit}(\pi_{ijk}) &= \beta_{10} + \text{Sex}_{ijk}\beta_{11} + \text{Field}_{jk}^{\text{T}}\boldsymbol{\beta}_{12} + \text{HStype}_{ijk}^{\text{T}}\boldsymbol{\beta}_{13} + \\ &\quad \text{HSgrade}_{ijk}^{\text{T}}\boldsymbol{\beta}_{14} + \text{RRuni}_{k}^{\text{T}}\boldsymbol{\beta}_{15} + \text{INC.res}_{k}\beta_{16} + \\ &\quad \text{NU.res}_{ijk}\beta_{17} + \text{Censis}_{jk}\beta_{18} + b_{1j}, \qquad (3)\end{aligned}$$

$$\begin{aligned}\log(\mu_{ijk}) &= \log(n_{ijk}) + \beta_{20} + \text{Sex}_{ijk}\beta_{21} + \text{Field}_{jk}^{\text{T}}\boldsymbol{\beta}_{22} + \text{HStype}_{ijk}^{\text{T}}\boldsymbol{\beta}_{23} + \\ &\quad \text{HSgrade}_{ijk}^{\text{T}}\boldsymbol{\beta}_{24} + \text{RRes}_{ijk}^{\text{T}}\boldsymbol{\beta}_{25} + \text{RRuni}_{k}^{\text{T}}\boldsymbol{\beta}_{26} + \text{INC.res}_{k}\beta_{27} + \\ &\quad \text{NU.res}_{ijk}\beta_{28} + \text{Censis}_{jk}\beta_{29} + b_{2j}, \qquad (4)\end{aligned}$$

where b_{1j} and b_{2j} are the random effects of the universities. It is assumed that $b_{1j} \sim N(0, \sigma_{b_1}^2)$, $b_{2j} \sim N(0, \sigma_{b_2}^2)$. Equation (3) uses a binomial mixed model for the zero part, whereas a zero-truncated negative binomial mixed model is employed in (4) to model the non-zero part. The offset $\log(n_{ijk})$ implies we are modeling the ratio between the number of incoming students and the number n_{ijk} of "local" students in the kth region. Estimation in (3) and (4) is usually performed via maximum likelihood. For non-longitudinal data, b_{1j} and b_{2j} are independent and the log-likelihood decomposes into the sum of the log-likelihood pertaining to the binomial part and the log-likelihood of the truncated negative binomial part. Thus both log-likelihoods can be maximized separately (Molas and Lasaffre, 2010). However, maximizing the log-likelihood under the zero-truncated negative binomial using the SM data can be misleading, since the missing data appear to be nonignorable. Since missing counts belong to the non-zero part (2), this leads to underestimate π_{ijk} and, in addition, the truncated-at-zero negative binomial does not fit well. To reduce the bias, at least for the non-zero part, we fit (4) using a two-truncated negative binomial.
Of course, that implies there is not an underlying "true" probability distribution of Y_{ijk}, and two analyses must be carried out separately, accordingly. On the other hand, the advantage is that the relationship (4) between the linear predictor and the expected value of the untruncated NB does not change. The expected value of the two-truncated distribution is $E(Y'_{ijk} = y|y > 2) = \mu_{ijk} + \mu_{ijk}(1 + 2\sigma)\frac{P(Y'_{ijk}=y|\mu_{ijk},\sigma)}{1-P(Y'_{ijk}\leq 2|\mu_{ijk},\sigma)}$. We adopt a GAMLSS framework to estimate (4).

3 Application to the SM dataset

Tables 2 and 3 show the fixed-effects estimates from the binomial mixed model and the two-truncated negative binomial mixed model, respectively. Due to lack of space we prefer to report results regarding just the fixed effects. Table 2 shows that the odds ratios of the non-resident students as opposed to the resident ones are larger in the wealthier North-Central regions. These odds ratios become larger for regions of origin with a smaller number of universities, and " more" larger if they are considered "prestigious" (Censis). In general non-resident students have a high school diploma (liceo) while is not statistically significative the high school grade. They mainly attend five/six-year degree courses in the social and health fields. These results seem to be consistent with the literature (Bruno and Genovese, 2010).

TABLE 2. Fixed effects from the binomial mixed model.

	Estimate	Std. Error	Pr(>|z|)
Intercept	-1.71	0.79	0.03
SexMale	-0.32	0.08	0.00
HSgrade			
60_69	0.03	0.14	0.85
70_79	0.23	0.14	0.10
80_89	0.15	0.14	0.29
90_99	-0.26	0.16	0.10
HStype			
Magistrale	-1.08	0.14	0.00
Vocational	-1.29	0.19	0.00
Technical	-0.78	0.10	0.00
Field			
Health3	-0.62	0.19	0.00
Scientific5	-1.47	0.30	0.00
Scientific3	0.03	0.16	0.86
Social5	1.04	0.19	0.00
Social3	-0.12	0.16	0.44
Humanistic3	-0.18	0.17	0.28
RRuni			
BASILICATA	-3.11	1.56	0.05
CALABRIA	-6.99	1.17	0.00
CAMPANIA	-6.51	1.05	0.00
EMILIA	5.01	1.04	0.00
FRIULI	4.29	1.25	0.00
LAZIO	1.68	0.97	0.09
LIGURIA	2.47	1.57	0.12
LOMBARDIA	3.85	0.94	0.00
MARCHE	0.92	1.06	0.39
MOLISE	-1.46	1.59	0.36
PIEMONTE	1.86	1.11	0.10
PUGLIA	-5.10	1.05	0.00
SICILIA	-5.10	1.05	0.00
TOSCANA	1.78	1.10	0.11
UMBRIA	0.28	1.56	0.86
VENETO	3.21	1.01	0.00
INC.res	-1.18	0.01	0.00
NU.res	-0.29	0.01	0.00
Censis	0.11	0.04	0.00

The two-truncated negative binomial mixed model in Table 3 is fitted on data including only the non-resident students. From the estimates it results that, after Sicily, the region of destination Abruzzo (at the baseline) is the one with the higher average ratio between non-resident and resident students. However one should consider that, in this model, Sicily is represented only by the University of Messina, which takes students almost exclusively from the neighboring region Calabria, since the universities of Palermo and Catania have not non-resident students. In conclusion, the proposed approach seems to work well in presence of non-ignorable missing data via modeling a truncated distribution. Further developments of this approach may lead to construct imputation procedures which take into account the distribution following an EM-like algorithm.

TABLE 3. Estimates from the two-truncated negative binomial mixed model.

	Est.	S.E.	P(>|t|)		Est.	S.E.	P(>|t|)
Intercept	-2.42	0.33	0.00	RRuni			
SexMale	-0.01	0.05	0.77	BASILICATA	-0.65	0.38	0.09
HSgrade				CALABRIA	-0.97	0.50	0.05
60_69	-0.24	0.09	0.01	CAMPANIA	-2.20	0.24	0.00
70_79	-0.25	0.08	0.00	EMILIA	-1.40	0.10	0.00
80_89	-0.24	0.08	0.00	FRIULI	-0.85	0.17	0.00
90_99	-0.18	0.09	0.05	LAZIO	-1.25	0.09	0.00
INC.res	-0.38	0.14	0.01	LIGURIA	-2.24	0.21	0.00
Censis	0.05	0.00	0.00	LOMBARDIA	-2.54	0.12	0.00
RRes				MARCHE	-1.18	0.16	0.00
BASILICATA	-1.08	0.18	0.00	MOLISE	0.36	0.30	0.23
CALABRIA	-0.64	0.27	0.02	PIEMONTE	-1.60	0.16	0.00
CAMPANIA	-0.67	0.36	0.07	PUGLIA	-0.70	0.19	0.00
EMILIA	2.90	0.88	0.00	SICILIA	1.00	0.16	0.00
FRIULI	2.08	0.79	0.01	TOSCANA	-1.56	0.11	0.00
LAZIO	1.81	0.58	0.00	UMBRIA	-1.76	0.16	0.00
LIGURIA	1.57	0.76	0.04	VENETO	-1.57	0.14	0.00
LOMBARDIA	2.34	0.74	0.00	HStype			
MARCHE	1.66	0.49	0.00	Magistrale	0.40	0.10	0.00
MOLISE	-0.77	0.17	0.00	Vocational	0.52	0.18	0.00
PIEMONTE	3.15	0.71	0.00	Technical	0.01	0.06	0.83
PUGLIA	-0.15	0.26	0.57	Field			
SARDEGNA	-1.43	0.26	0.00	Health3	0.23	0.12	0.05
SICILIA	-0.56	0.28	0.05	Scientific5	-0.50	0.31	0.10
TOSCANA	1.32	0.67	0.05	Scientific3	-0.43	0.08	0.00
TRENTINO	1.04	0.85	0.22	Social5	-0.27	0.12	0.02
UMBRIA	-0.08	0.48	0.87	Social3	-0.42	0.09	0.00
VDAOSTA	1.73	0.89	0.05	Humanistic3	-0.30	0.10	0.00
VENETO	2.58	0.64	0.00				

References

Bruno, G., Genovese, A. (2010). A spatial interaction model for the representation of the mobility of University students on the Italian territory. *Netw Spat Econ*, DOI 10.1007/s11067-010-9142-7

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

Molas, M., Lesaffre, E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in medicine*, **29**, 3294 − 3310.

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

Rigby, R.A., Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*,**54(3)**, 507 − 554.