Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Yao and Liu, J Biomet Biostat 2013, 4:4 DOI: 10.4172/2155-6180.1000172

Research Article Open Access

Semiparametric Analysis of Longitudinal Zero-inflated Count Data with Applications to Instrumental Activities of Daily Living

Ping Yao*[1] and Xiaohong Liu[2]

[1]Public Health and Health Education Program, School of Nursing and Health Studies, Northern Illinois University, Wirtz Hall 260, DeKalb, IL, 60115, USA
[2]Division of Statistics, Northern Illinois University, Dekalb, IL, 60115, USA

Abstract

Background: The instrumental activities of daily living (IADLs) are important index of physical functioning in older adult studies. These count outcomes with a large proportion of zeros are often collected in longitudinal studies. Data were from the Hispanic Established Population for Epidemiological Study of the Elderly (HEPESE), a four wave (seven years) longitudinal study of community-dwelling elderly Mexican-Americans. There were excess zeros IADLs observed during follow-up.

Methods: We present semiparametric zero-inflated Poisson (ZIP) and hurdle model with random effects to evaluate IADLs in the context of excess zeros.

Results: Age, education, household income, marital status, smoking, cognitive functioning and prescription medication use were not significantly associated with IADLs. The counts of IADLs changed with age nonlinearly.

Conclusions: zero-inflated Poisson (ZIP) and hurdle model with random effect fits the IADLs counts with excess zeros in longitudinal studies.

Keywords: Longitudinal; Zero-inflated poisson; Random effect; Instrumental activities of daily living (IADLs)

Introduction

The instrumental activities of daily living (IADLs) are used as measurements of functional status of a person. It is a useful instrument in elderly studies, which let an older adult live independently in a community [1]. The measures of IADLs consist of the following ten items: using telephone without help, driving a car or travel alone, going for groceries/clothes, preparing own meals without help, doing light housework without help, taking medicine without help, handling money without help, doing heavy work around the house, walking up and down stairs without help, and walking half a mile without help. The range of IADLs is from 0-10, with higher scores indicating greater difficulties in activities of daily living. Disabilities on IADLs are good markers to identify individuals at risk of frailty in elderly adults. The study found that elderly women with more than one IADLs were frailer [2]. Several studies investigated the relationship between IADLs and functional health status, showed that the social functioning, health perception and physical functioning were significantly related to IADLs [2-4]. As function health status decline, elderly persons need more dependence on others for assistance. The results of IADLs for nursing administrators help develop health promotion strategies and social policies to improve the independence of elderly people.

Measuring IADLs longitudinally can provide useful information for assessing functional independence among older adults, which can capture the dynamic changes of responses over time. The effect of covariates, such as age may not be linearly related to the link function. Usually the growth curve model with polynomial functions of time is employed to delineate the longitudinal trajectories. However, when the change of IADLs over time is nonlinear curve, the estimates by growth curve model may not capture all the information. The nonparametric methods can be utilized to model potential nonlinear effects of the time on the outcomes. Furthermore, the individual difference in developmental curves observed in the IADLs increases the complex of the model. The counts of IADLs observed in elderly studies often have excess zero values more than what would be expected by a classic Poisson model.

Longitudinal zero-inflated count data are very common in health and medical studies. Cheung [5] used zero-inflated Poisson (ZIP) model to analyze growth failure in children. Dalrymple et al. [6] investigated the sudden infant death syndrome by ZIP and hurdle a model. Rose et al. [7] used ZIP to model vaccine adverse event data. Longitudinal ZIP and hurdle models were proposed to analyze substance abuse disorders data [8]. Zero-inflated count data frequently occur in health utilization, pharmaceutical, epidemiological and medical research [6-12].

This paper is motivated by analysis of IADLs from the Hispanic Established Population for Epidemiological Study of the Elderly (HEPESE), a seven year longitudinal study of community-dwelling elderly Mexican-Americans [13]. IADLs measuring physical functioning status were an important instrument in the HEPESE. An inspection of the records indicates that zeros among the observations from the four waves account for more than 46% of the total.

In this paper, we begin by describing a longitudinal model for zero-inflated count data and then apply zero-inflated Poisson regression (ZIP) and hurdle models to IADL-s from HEPESE data. The utility of two models are compared.

***Corresponding author:** Ping Yao, Public Health and Health Education Program, School of Nursing and Health Studies, Northern Illinois University, Wirtz Hall 260, DeKalb, IL, 60115, USA, Tel: 815-753-0853; Fax: 815-753-5406, E-mail: pyao@niu.edu

Received July 24, 2013; **Accepted** August 28, 2013; **Published** August 30, 2013

Citation: Yao P, Liu X (2013) Semiparametric Analysis of Longitudinal Zero-inflated Count Data with Applications to Instrumental Activities of Daily Living. J Biomet Biostat 4: 172. doi:10.4172/2155-6180.1000172

Statistical Models and Estimation

The zero-inflated Poisson model

The ZIP model includes two parts of parameters: one for zero inflated measures and the other for repeated Poisson counts.

Let response vectors $Y=(Y_1^T,...,Y_N^T)^T$, where $Y_i=(Y_{i1}^T,...,Y_{iT}^T)^T$ and the response Y_{ij} denote the count for i^{th} subject at time j, i=1,…, N, j=1,…, T. The probability of an excess zero is denoted by π_{ij}, $0 \le \pi_{ij} \le 1$. Here we adopt ZIP regression models introduced by Lambert [14]

$$P\left(Y_{ij}=y_{ij}|\pi_{ij},\theta\right)=\begin{cases}\pi_{ij}+\left(1-\pi_{ij}\right)e^{-\mu_{ij}}, & y_{ij}=0;\\ \dfrac{\left(1-\pi_{ij}\right)e^{-\mu_{ij}}\mu_{ij}^{\,y_{ij}}}{y_{ij}!}, & y_{ij}>0,\end{cases} \quad (1)$$

Let $u_i=(u_{i1,...,},u_{iT})^T$ and $\pi_i=(\pi_{i1,...,},\pi_{iT})^T$ are assumed to be log-linear and logistic regression models with the parameter θ. Both logit (π_{ij}) and log (μ_{ij}) are assumed to depend on a non-linear function of independent variables. The covariates in these two parts can be different.

First, let describe the trajectory in zero-inflation based on a logistic regression model,

$$\text{logit}\,(\pi_{ij})=x_{ij}^{'}\beta_{\pi}+v_i+g(t_{ij}), \quad (2)$$

Where t_{ij} is the time, x_{ij} are vectors of covariates for ith subject, β_{π} are the corresponding regression coefficients. We assume that vi are independent normal distribution with mean 0 and variance δ_v^2. The function $g(t_{ij})$ is the baseline function for the nonlinear time effects and fitted with the following quadratic spline basis (to simplify the presentation, *ij* is dropped from t):

$$g(t)=\alpha_0+\alpha_1 t+\alpha_2 t^2+\sum_{k=1}^{K}\theta_k\,(t-t_k)^2+\,, \quad (3)$$

Where the + indicates the positive value with k knots that divide the range of time into segments.

This part of ZIP model is to estimate the initial probability and the change in the probability of zero inflation.

Next, the second part of ZIP model estimates Poisson counts over time,

$$\log\,(\mu_{ij})=m_{ij}^{'}\beta_p+u_i+h(t_{ij}) \quad (4)$$

Where t_{ij} is the time, where m_{ij} are vectors of covariates for i_{th} subject, β_p are the corresponding regression coefficients. We assume that μi are independent normal distribution with mean 0 and variance δ_v^2. The function $h(t_{ij})$ is the same format as $g(t_{ij})$. It is assumed that covariance between μ_i and v_i is $\delta_{\mu v}^2$.

Then the log-likelihood for ZIP model with random effects is

$$l(\theta;y)=\sum_{i=1}^{N}\log\int\int_{-\infty}^{+\infty}\prod_{j=1}^{T}P(Y_{ij}=y_{ij}\,|v_i,\mu_i)\,\phi(v_i,\mu_i)\,dv_i d\mu_i\,,\ \text{Where}$$

$$P(Y_{ij}=y_{ij}\,|\theta,(v_i,\mu_i)=\left[\pi_{ij}+(1-\pi_{ij})e^{-\mu_{ij}}\right]^{\Delta_{ij}}\left[\frac{(1-\pi_{ij})e^{-\mu_{ij}}\mu_{ij}^{\,y_{ij}})}{y_{ij}!}\right]^{1-\Delta_{ij}}$$

Here, θ denotes all parameters, ϕ denotes the standard bivariate normal probability density function, and Δ_{ij} is an indicator, with value 1 if y_{ij}=0 and 0 if y_{ij}>0.

This likelihood function can be fitted by a Newton-Raphson or quasi-Newton algorithm. Because there are two random effects in the function, the proposed likelihood (5) needs integrate random effect first and then get the maximum likelihood estimation. We apply Gaussian quadrature numerical integration techniques [15].

The hurdle model

The hurdle model consists of a mixture of a point mass at zero and a truncated Poisson model for non-zero counts. The hurdle model has been used in healthcare utilization analysis [11] and vaccine adverse study [7]. The model is represented as follows

$$P\left(Y_{ij}=y_{ij}|\pi_{ij},\theta\right)=\begin{cases}\pi_{ij}, & y_{ij}=0;\\ \dfrac{\left(1-\pi_{ij}\right)e^{-\mu_{ij}}\mu_{ij}^{\,y_{ij}}}{(1-e^{-\mu_{ij}})y_{ij}!}, & y_{ij}>0,\end{cases} \quad (6)$$

where both logit (π_{ij}) and log (μ_{ij}) are modeled as Equation (2) and (4) in the ZIP, respectively. The likelihood function of the hurdle model also has the same form as Equation (5) with

$$P\left(Y_{ij}=y_{ij}|\theta,v_i,u_i\right)=\left[\pi_{ij}\right]^{\Delta_{ij}}\left[\frac{\left(1-\pi_{ij}\right)e^{-\mu_{ij}}\mu_{ij}^{\,y_{ij}}}{y_{ij}!(1-e^{-\mu_{ij}})}\right]^{1-\Delta_{ij}}$$

Model comparison and goodness-of-fits statistics

The ZIP and hurdle model with random effects both are nonlinear mixed effect modes, thus SAS PROC NLMIXED procedure is used to estimate the parameters [15]. Meanwhile a standard Poisson model with random effects is fitted as the base model to compare with the above models.

Application and Results

Data and measures

This longitudinal study comprises four waves of follow-up: 1993-1994, 1995-1996, 1998-1999 and 2000-2001. The sample consists of 3050 Mexican-Americans at baseline, aged 65 years and older, residing in the five of the southwestern states of Arizona, California, Colorado, New Mexico, and Texas.

Variables

Repeated measures of IADLs: The outcome variable of interest is total number of IADLs. The measures of IADLs are available in wave

IADLs	0	1	2	3	4	5	6	7	8	9	10
Wave one	1424	463	276	180	146	117	86	102	60	63	128
	46.77%	15.21%	9.06%	5.91%	4.79%	3.84%	2.82%	3.35%	1.97%	2.07%	4.20%
Wave two	1175	302	182	169	84	88	83	71	53	58	172
	48.22%	12.39%	7.47%	6.93%	3.45%	3.61%	3.41%	2.91%	2.17%	2.38%	7.06%
Wave three	959	186	124	111	80	74	70	79	50	56	182
	48.66%	9.44%	6.29%	5.63%	4.06%	3.75%	3.55%	4.01%	2.54%	2.84%	9.23%
Wave four	812	146	88	113	69	71	64	57	50	62	141
	48.54%	8.73%	5.26%	6.75%	4.12%	4.24%	3.83%	3.41%	2.99%	3.71%	8.43%

Table 1: Frequency of IADLs at wave one, two, three and four.

Parameter	ZIP model			Hurdle model		
	Estimate	Std error	p-value	Estimate	Std error	p-value
The zero component						
Sex	-0.22	0.12	0.06	-0.15	0.10	0.16
Years of education	0.02	0.02	0.21	0.02	0.10	0.25
Marital status	-0.06	0.12	0.60	-0.01	0.10	0.93
Household income	-0.07	0.05	0.10	-0.07	0.04	0.08
Smoking	0.02	0.17	0.89	-0.05	0.15	0.75
Drug	0.11	0.12	0.35	0.13	0.10	0.20
MMSE	0.004	0.01	0.73	0.01	0.01	0.36
δ_{μ}^2	3.75	0.55	<0.0001	3.29	0.43	<0.0001
The Poisson component						
Sex	-0.05	0.04	0.68	-0.02	0.04	0.67
Years of education	-0.002	0.01	0.69	-0.002	0.01	0.63
Marital status	-0.05	0.04	0.20	-0.05	0.04	0.22
Household income	0.01	0.01	0.41	0.01	0.01	0.51
Smoking	0.12	0.05	0.03	0.10	0.05	0.05
Drug	-0.02	0.04	0.52	-0.02	0.03	0.51
MMSE	0.002	0.004	0.64	0.001	0.004	0.67
δ_{μ}^2	0.44	0.04	<0.0001	0.44	0.04	<0.0001
$\delta_{\mu v}^2$	-0.88	0.07	<0.0001	-1.08	0.06	<0.0001
Fit Statistics						
AIC	31490			31481		
BIC	31640			31637		

*AIC=32235, BIC=32324 from standard Poisson regression model

Table 2: Parameter estimates of ZIP and hurdle model on the IADLs data.

one, two, three and four. The discrete number of IADLs indicates the amount of difficult items in activities of daily living. The IADLs frequency distribution (Table 1) illustrates that there are excessive zeros at each wave. There are 46%-49% of older adults reported zero difficulties.

The demographic and health behavior variables are controlled by age, education, household income, marital status, smoking, prescription medication use and cognitive functioning. These covariates were chosen because they are potentially associated with elderly independence. Age, education (number of years of formal schooling), and cognitive function (indicated by the Mini-Mental State Exam total score) were all entered as continuous variables. Household income was categorized as <$5,000, $5,000-$9,999, $10,000-$14,999, $15,000-$19,999 and ≥ $20,000, with the latter as the reference group. Marital status was dichotomized as unmarried (consisting of separated, divorced widowed and never married people) and currently married, with married people as the reference group. Smoking status was dichotomized into current smokers and nonsmokers, with nonsmokers as the reference. Prescription medication use was dichotomized into any or none, with no prescription medications as the reference group.

Table 1 indicates that there are zero-inflated IADL-s counts from four wave interviews: of the total IADL-s, 46.77% zero counts at wave one, 48.22% zero counts at wave two, 48.66% zero counts at wave three, and 48.54% zero counts at wave four. Therefore the ZIP and hurdle models are necessary to fit this zero-inflated data.

Table 2 lists the estimated regression coefficients, standard errors, p-value and goodness-of-fit test results. The estimates in the zero and Poisson components produced by the two models are very close, whereas standard errors in the hurdle model are smaller than ZIP model. Given the significance level at 0.05, both models can't identify significant association between IADLs and education, household income, marital status, smoking, and prescription medication use, except for that smoking in the Poisson component of ZIP model had positive association with IADLs counts.

Both models indicate that there are significant individual difference (the random effect) in the zero and Poisson component, and the individual difference in the zero components is much bigger than the Poisson component. The variance of random effect in hurdle model is smaller than ZIP model. The zero component and Poisson component are negatively correlated with correlation coefficient 0.69 (Table 2).

Figures 1 and 2 show the spline functions of age, which explains the fixed effect on the zero and Poisson component. The spline functions in the Poisson component generated by two models are very similar. The severity of disability increased rapidly from age around 70 to 90 years, stays at high level, and deceased after 100. The spline functions in the zero component from two models are also very close. The curve indicates that the probability of without IADLs decreased with age increased from 67 to 90, then stayed at low level. After 90, the probability of zero count of IADLs increased.

The longitudinal development trajectories of IADLs under the ZIP and hurdle models are very similar, so only trajectories of IADLs under the ZIP were presented here. The values of AIC and BIC from ZIP and hurdle models are also very close, which indicates both models are good fit for zero-inflated IADLs data. The corresponding values of AIC and BIC from standard Poisson regression model are much larger than from ZIP and hurdle models, which indicates ZIP and hurdle models are better fit (Table 2).

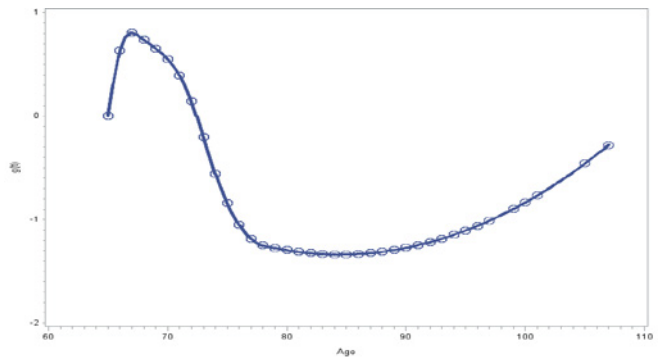

Figure 1: Regression spline: the zero component.

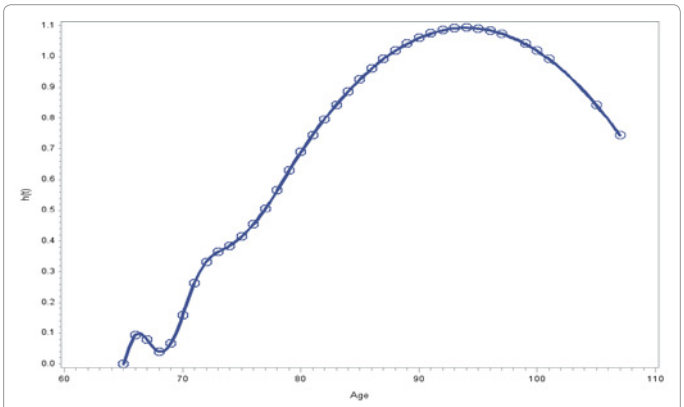

Figure 2: Regression spline: the Poisson component.

For IADLs data, the hurdle model is more appropriate than ZIP model. When all the elderly Mexican-Americans started at early stage, they were relatively healthy without any help. Once they crossed the hurdle to get helps, more and more dependence with aging.

Discussion

In this article, we have illustrated the use of semi-parametric longitudinal model for zero-inflated IADLs count data in health study. The ZIP and hurdle with nonlinear time effect models are used to capture the dynamic profile of IADLs in elderly Mexican American population. The results from both models indicated the probability of being independent for elderly Mexican American adults decreased and needed more help with age. The results of IADLs for nursing administrators help develop interventions to maximize functional independence among elderly people.

There are two limitations for this paper. First, dependent variable IADLs was self-reported by elderly people thorough face to face interview, instead of performance-based IADLs assessment, thus there were potential biases with cognitive status among elderly people. Second, there were death and unknown drop-out during the four wave follow-up, so only available observations are used for longitudinal analysis by assuming responses are missing completely at random.

Acknowledgments

The authors thank Arlene Keddie from Northern Illinois University for providing the data of HEPESE.

References

1. Yana S, Paula GW, Matthew LK, Emilie F, Jonathan B (2010) Instrumental Activities of Daily Living Among Community-Dwelling Older Adults: Personality Associations With Self-Report, Performance, and Awareness of Functional Difficulties. J Gerontol B Psychol Sci Soc Sci 65: 542-550.
2. Nourhashémi F, Andrieu S, Gillette-Guyonnet S, Vellas B, Albarède JL, et al. (2001) Instrumental activities of daily living as a potential marker of frailty: a study of 7364 community-dwelling elderly women (the EPIDOS study). J Gerontol A Biol Sci Med Sci 56: M448-453.
3. Reuben DB, Rubenstein LV, Hirsch SH, Hays RD (1992) Value of functional status as a predictor of mortality: results of a prospective study. Am J Med 93: 663-669.
4. Scott WK, Macera CA, Cornman CB, Sharpe PA (1997) Functional health status as a predictor of mortality in men and women over 65. J Clin Epidemiol 50: 291-296.
5. Cheung YB (2002) Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med 21: 1461-1469.
6. Dalrymple ML, Hudson IL, Ford RPK (2003) Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. Comput Stat Data Anal 41: 491-504.
7. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD (2006) On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J Biopharm Stat 16: 463-481.
8. Buu A, Li R, Tan X, Zucker RA (2012) Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Stat Med 31: 4074-4086.
9. Lee AH, Wang K, Scott JA, Yau KK, McLachlan GJ (2006) Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. Stat Methods Med Res 15: 47-61.
10. Baughman AL (2007) Mixture model framework facilitates understanding of zero-inflated and hurdle models for count data. J Biopharm Stat 17: 943-946.
11. Wang K, Yau KK, Lee AH (2002) A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. Comput Methods Programs Biomed 68: 195-203.
12. Yau KKW, Wang K, Lee AH (2003) Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. Biometrical Journal 45: 437-452.
13. Keddie AM, Peek MK, Markides KS (2005) Variation in the associations of education, occupation, income, and assets with functional limitations in older Mexican Americans. Ann Epidemiol 15: 579-589.
14. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34: 1-14.
15. SAS Institute Inc. SAS/STAT User's Guide, Version 9.2: Software Manual. Cary, NC: SAS Institute Inc, 2002-2008.

Citation: Yao P, Liu X (2013) Semiparametric Analysis of Longitudinal Zero-inflated Count Data with Applications to Instrumental Activities of Daily Living. J Biomet Biostat 4: 172. doi:10.4172/2155-6180.1000172