Comparison of methods for modelling a count outcome with excess zeros: an application to Activities of Daily Living (ADL-s)

Paola Zaninotto, Emanuela Falaschetti

Title: Comparison of methods for modelling a count outcome with excess zeros: an application to Activities of Daily Living (ADL-s)

Paola Zaninotto[1] and Emanuela Falaschetti[1]

[1] Department of Epidemiology & Public Health, UCL,

Corresponding author:
Paola Zaninotto
Department of Epidemiology & Public Health, UCL
1-19 Torrington Place
WC1E 7HB
London, UK
tel +44(0)20 7679 1668
fax +44(0)20 7813 0280
e-mail: p.zaninotto@ucl.ac.uk

Word count: 2,890

Abstract

Background: count outcomes are commonly encountered in many epidemiology applications, and are often characterized by a large proportion of zeros. While linear or logistic regression models have often been used to analyse count outcomes, the resulting estimates are likely to be inefficient, inconsistent or biased.

Methods: data are from wave1 of the English Longitudinal Study of Ageing (ELSA). The main outcome measure is number of difficulties (ranging from 0 to 6) with 'Activities of Daily Living (ADL-s)' such as dressing, walk across a room, bathing, eating, getting in and out of bed and using the toilet. We fitted four regressions models specifically developed for count outcomes: Poisson, NB, ZIP and ZINB. We then compared these models using the Likelihood Ratio test of overdispersion, the Vuong test, and graphical methods. **Results**: The plots of predictions showed that overall, the ZINB model fit best. Although the ZINB and the ZIP models showed similar fit the LR test provided strong evidence that the ZINB improves the fit over the ZIP. Increasing number of difficulties with ADL-s was associated with fair/poor self-reported health, limiting longstanding illness and physical inactivity. The probability of not having any difficulty with ADL-s decreases with a limiting longstanding illness, increasing age, no education, fair/poor self-reported health and with not living with the partner.

Conclusion: models specifically developed for count outcomes with excess zeros such as ZINB can provide better insights into the investigation of the factors associated with the number of difficulties with Activities of Daily Living.

Key word: count outcome, activities of daily living, old age.

INTRODUCTION

Many studies have addressed the issue of functional dependence and its risk factors among older adults [1-8]. A measure of functional dependence widely used in epidemiology studies of older adults is the difficulties to perform Activities of Daily Living (ADL-s) such as bathing, dressing, walk across a room, eating, getting in or out of bed and using a toilet [9]. From these items a variable is derived to assess the number of difficulties in doing ADL-s, which ranges from 0 (to indicate no difficulties) to 6 (to indicate difficulties with all 6 activities of daily living). This outcome has been used as a continuous variable, or as a dichotomous variable for comparison between those reporting zero difficulties with ADL-s and those reporting 1 to 6 difficulties with ADL-s . [5-8] However, using this variable as a continuous outcome in a linear regression would not be appropriate and the categorisation of the variable to be used in logistic regression may result in loss of information.

Count outcomes are commonly encountered in many epidemiology applications, and are often characterized by a large proportion of zeros. While linear or logistic regression models have often been used to analyse count outcomes, the resulting estimates are likely to be inefficient, inconsistent or biased. [10, 11] Several models belonging to the family of Generalised Linear Models are available for performing regressions with count outcomes [12]. Such models are the Poisson and negative binomial (NB) as well as zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) specifically developed for count outcomes with excess zeros and dispersion. [10-21] In recent years there has been increasing interest in the use of these models in research studies, and few examples are available in epidemiology and public health literature. [22-27]

The aim of this paper is to examine and compare the utility of the Poisson, negative binomial (NB), ZIP and ZINB regressions for modelling a count outcome, with particular focus on application in epidemiology and public health research. For this purpose we use data on the number of difficulty with ADL-s from a national sample of older adults, participants of the first wave of the English Longitudinal Study of Ageing (ELSA).

METHODS

The sample

The data are from the first wave of the English Longitudinal Study of Ageing (ELSA), a panel study where the same individuals are followed and re-interviewed every two years. The technical details of this study and the results of primary analyses have been published elsewhere [28, 29] and are also available at the web site of the Institute of Fiscal Studies (http://www.ifs.org.uk/elsa/report.htm). Briefly the ELSA sample was drawn from people who had taken part in the Health Survey for England (HSE) in 1998, 2000 or 2001 and were born before March 1952. The HSE samples are selected to be representative of people living in private households in England. A total of 11,392 eligible sample members took part in wave1 (2002-2003), giving a response rate of 67%.

Participants gave their informed consent to take part in the study. Ethical approval for ELSA was given by the London Multi-centre Research Ethics Committee (MREC/04/2/006).

Measures

Information on difficulty with Activities of Daily Living (ADL-s) was collected during the interview using the health-assessment questionnaire. [6] The original scale of ADL-s was developed by Katz *et al*,. [9]

Respondents were shown a card and the following text was read to them: ‘Here are a few everyday activities. Please tell me if you have any difficulties with these because of a physical, mental, emotional or memory problem. Exclude any difficulties they expect to last less than three months. Because of a health or memory problem, do you have any difficulty in doing any of the activities on this card?’.

The following six ADL-s items were listed in the card:

1. Dressing, including putting on shoes and socks
2. Walk across a room
3. Bathing or showering
4. Eating, such as cutting up food
5. Getting out of bed
6. Using toilet, including getting up or down

From 11,220 valid answers to this question a variable is derived to count the number of difficulty with ADL-s (range 0 to 6).

Variables included in the models as potential predictors for ADL-s were age, sex, marital status, education level, smoking status, weekly alcohol consumption, physical activity, limiting longstanding illness and self-reported general health. Apart from age, all the variables were entered into the models as dummy.

Analysis

Poisson, negative binomial, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regressions models for count outcomes were fitted to the data in order to examine and compare which model gives the best fit for the data.

The Poisson distribution is a probability distribution for non-negative integers. Let *y* be a random variable which indicates the number of times an event occurred and μ the expected count. The Poisson distribution specifies the relationship between the expected count μ and the probability of observing any observed count *y*:

$$\Pr(\mathrm{Y} = \mathrm{y} \mid \mu) = \frac{\mathrm{e}^{-\mu}\mu^{\mathrm{y}}}{\mathrm{y}!} \qquad \text{for y} = 0, 1, 2,\ldots \qquad (1)$$

where μ is the mean and the variance of the distribution, which is known as *equidispersion*. [17] The mean (μ) of the distribution strongly controls the shape of the distribution. When the mean is small the most common count is zero. For example, if

μ =0.5, then the probability of obtaining a count of zero is 0.61, a count of one is 0.33, a count of two is 0.12, and as the mean increases, the probability of a zero count becomes less than zero. The distribution is therefore highly skewed towards the lowest value, zero. The Poisson regression model can be represented as:

$$\Pr(\mathrm{Y_i} = \mathrm{y_i} \mid \mathrm{X_i}) = \frac{\mathrm{e}^{-\mu_\mathrm{i}}\mu_\mathrm{i}^{\mathrm{y_i}}}{\mathrm{y_i}!} \qquad (2)$$

where $\mathrm{Y_i} \mid \mathbf{X}_\mathrm{i} \rightarrow P(\mu_i)$ following a Poisson distribution; and where

$$\mu_i = \mathrm{E}(\mathrm{Y_i} \mid \mathbf{X}_\mathrm{i}) = \exp\{\beta_0 + \beta_1 \mathrm{X_{i1}} + \beta_2 \mathrm{X_{i2}} + \ldots + \beta_p \mathrm{X_{ip}}\} \qquad (3)$$

$\mathrm{Y_i}$ is the count for the $\mathrm{i^{th}}$ subject, $\mathbf{X}_\mathrm{i} = \{1 + \mathrm{X_{i1}} + \mathrm{X_{i2}} + \ldots + \mathrm{X_{ip}}\}$ is a vector of covariates plus 1 and $\beta_0 \ldots \beta_p$ the regression coefficients to be estimated.

One problem with Poisson regression is over-dispersion, which means that the assumption of equality between the mean and the variance is not often adapted to observed data, as the variance often exceeds the mean. [17] This leads to underestimation of the standard errors of the regression estimates, confidence intervals that are too narrow, and p-values that are too small.

The negative binomial (NB) regression model is an alternative to the Poisson model when the count data are over-dispersed in relation to the mean. [30, 31] The NB regression adds an error ε to equation (3) as follow:

$$\mu_i = E(Y_i \mid \mathbf{X}_i) = \exp\{ \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \} \qquad (4)$$

where ε_i is assumed to be uncorrelated with $\mathbf{X}_i$ and exp{ ε_i } has a gamma distribution with mean 1 and variance α^2. It follows that:

$$\Pr(Y_i = y_i \mid \mu_i, \varepsilon_i) = \frac{e^{-\mu_i \exp\{\varepsilon_i\}} \mu_i^{y_i}}{y_i!} \qquad (5)$$

If α=0 then the NB (5) reduces to the Poisson (2). [15,17] Therefore the inappropriateness of the Poisson regression relative to the NB regression is determined by the statistical significance of the α parameter.

In addition to dispersion, data often display a greater number of zero observations than expected from the Poisson model, this problem is known as “excess zeros”. [17]

Zero-inflated count models provide a way of modelling dispersion and the excess zeros by changing the mean structure to allow zeros to be generated by two distinct processes. [13, 21] Zero-inflated count models assume that there are two latent groups: the first group is the group of ‘zeros’ and the second is the group of ‘non-zeros’. The first group generates only zeros while the second is either a Poisson or a negative binomial distribution. In the first situation we will have a ZIP regression represented as:

$$\Pr(Y_i = y_i \mid X_i) = \begin{cases} p_i + (1 - p_i) e^{-\mu_i} & y_i = 0 \\ \dfrac{(1 - p_i) e^{-\mu_i} \mu_i^{y_i}}{y_i!} & y_i \geq 1 \end{cases} \qquad (6)$$

where p_i is the probability of being an extra zero, which is determined, in our application, by a logistic model. Substituting the negative binomial for the Poisson yields the zero-inflated negative binomial (ZINB).

All analyses were run using Stata version 10. We first looked at how the probability distributions, which underpin the four models, fit the observed data; to do so, we have run the Poisson, NB, ZIP and ZINB regressions without any independent variables and compared graphically the predicted proportions (from models with intercept only) with the observed proportions from the ADL-s score. The Poisson, NB, ZIP and ZINB regressions with covariates were then fitted to the data. Because the NB reduces to the Poisson when $\alpha=0$ these models can be compared using the Likelihood Ratio test for overdispersion available in Stata, to test the $H_0:\alpha=0$. Similarly because the ZINB and ZIP are nested, [15] it is possible to use the Likelihood Ratio test for overdispersion, which is easy to compute in Stata. [15, p:285] Comparisons of the Poisson with the ZIP and comparisons of the NB with the ZINB involve non-nested comparisons which can be assessed using the Vuong test for non-nested models, [32] available in Stata. In addition, to obtain graphical illustration of fit, we plotted the observed minus predicted probabilities across models, adjusted for covariates. The same set of covariates was used to fit both the logistic part and the intensity part of the ZIP and ZINB models.

The results from the best model were then compared to results from a logistic regression (with same predictors), in which ADL-s was analysed as a dichotomous variable for comparison between those reporting 1 to 6 difficulties and those reporting no difficulties with ADL-s.

Statement of Ethics: We certify that all applicable institutional and governmental regulations concerning the ethical use of human participants were followed during this research. The survey was approved by the appropriate Research Ethics Committee.

RESULTS

Table 1 presents the distribution of the outcome variable and predictor variables. Of the 11,220 respondents, 79% report no difficulties with ADL-s.

Figure 1 shows how the probability distributions, which underpin the four models, fit the observed proportions from the ADL-s score. The Poisson model was a poor fit, while the predicted values of the NB model were very close to the observed. The ZIP and ZINB models gave a good prediction of zeros. In fact, 85% of the observed zeros were predicted by both models. The ZINB model produced the best fit for the entire range of the ADL-s values.

The likelihood test for overdispersion comparing the NB to the Poisson, which tests H_0:α=0, yields a statistic of 948.67. The estimate of α is 1.1 (s.e. 0.06) which is significantly different from 0, therefore the NB model is favored over the Poisson.

The ZIP regression yielded a log-likelihood of -7215.37; while the ZINB yielded a log-likelihood of -7189.80 with an estimate of α equals to 0.25 (s.e. 0.05).

The ZIP and ZINB models are nested so they can be compared by using the likelihood test for overdispersion to test H_0:α=0, which yields a statistic of 51.1 which provide evidence for preferring the ZINB over the ZIP.

The Poisson and ZIP and also the NB and ZINB are not nested; therefore the models cannot be compared using the Log-likelihood test. We used the Vuong test for non-nested models. The Vuong test for the ZIP vs the Poisson yielded a z equals to 14.85, which supported the ZIP model over the Poisson (p<0.0001); similarly, the Vuong test supported the ZINB model over NB (p<0.0001).

Table 1 Sample characteristics

Number of difficulties with ADL-s	
0	79.2%
1	10.3%
2	4.9%
3	2.6%
4	1.7%
5	0.9%
6	0.4%
Mean Age of participants	65.2
(s.e.)	(0.10)
Females	54.5%
No educational qualification	42.6%
Not living with the partner	31.3%
Fair or poor health	27.1%
Limiting longstanding illness	35.0%
Current smoker	17.8%
Current drinker	28.1%
Physically inactive	65.7%

Finally, Figure 2 shows the observed minus the predicted probabilities at each count, for each model. Points above 0 on the y-axis indicate more observed count than predicted; while those below 0 indicate more predicted counts than observed. From the graph it is clear that the Poisson regression does not predict well the average number of zeros. The NB model is a substantial improvement over the Poisson; however, it underestimates the proportions at 1. Both, the ZIP and ZINB models fit the data quite well; however, the ZIP predicts more 1s and less 2s and 3s than the ZINB. Based on formal tests and graphical methods, we prefer the ZINB modelling approach.

Results from the ZINB model are shown in table 2. The first three columns report respectively the raw coefficients, p-values and coefficients for the factor change in the odds (for unit increase in the dependent variable) of being in the group without any difficulty with ADL-s, modelled with a logit model. The last three columns report the raw coefficients, p-values and the factor change in the expected count for those who have one and more difficulty with ADL-s, modelled with a Negative Binomial. Increasing age, not having educational qualification, not living with the partner, having a limiting longstanding illness, reporting fair or poor health were all associated with decrease odds of reporting zero difficulty with ADL-s. Among those with one and more difficulty with ADL-s, having a limiting longstanding illness, reporting fair or poor health and being physically inactive increase the expected rate of reporting difficulties with ADL-s, while drinking alcohol decrease the expected rate of reporting difficulty with ADL-s.

Table 2 ZINB regression coefficients for the number of difficulties with ADL-s

	Logit portion[a]			Negative Binomial portion		
	b	**p-value**	**exp(b)**	**b**	**p-value**	**exp(b)**
	(s.e.)			**(s.e.)**		
Age	-0.054	0.000	0.95	-0.001	0.713	1.00
	(0.002)			(0.006)		
Females	0.112	0.281	1.12	0.024	0.630	1.02
	(0.049)			(0.104)		
No educational qualification	-0.210	0.041	0.81	0.027	0.573	1.03
	(0.049)			(0.103)		
Not living with the partner	-0.327	0.003	0.72	0.013	0.801	1.01
	(0.051)			(0.112)		
Fair or poor health	-1.116	0.000	0.33	0.355	0.000	1.43
	(0.058)			(0.124)		
Limiting longstanding illness	-1.498	0.000	0.22	0.903	0.000	2.47
	(0.082)			(0.132)		
Current smoker	-0.060	0.631	0.94	-0.019	0.747	0.98
	(0.059)			(0.124)		
Current drinker	-0.050	0.687	0.95	-0.133	0.026	0.88
	(0.060)			(0.124)		
Physically inactive	-0.177	0.314	0.84	0.584	0.000	1.79
	(0.109)			(0.175)		
	value		**(95% CI)**	**s.e.**		
α	0.251		(0.167, 0.375)	0.05		

[a] **Models the probability of no difficulty with ADL-s**

Results from a logistic regression (Table 3) comparing those with 1 to 6 difficulties with ADL-s those with 0 difficulties show that increasing age, not having educational qualification, not living with the partner, having a limiting longstanding illness, reporting fair or poor health and being physically inactive were all associated with increased odds of reporting one and more ADL-s.

Table 3 Results from a logistic regression for the risk of having one to six difficulties with ADL-s

	b **(s.e.)**	**p-value**	**exp(b)**
Age	0.035 (0.003)	0.000	1.04
Females	-0.035 (0.058)	0.542	0.97
No educational qualification	0.180 (0.059)	0.002	1.20
Not living with the partner	0.260 (0.061)	0.000	1.30
Fair or poor health	1.032 (0.061)	0.000	2.81
Limiting longstanding illness	1.772 (0.063)	0.000	5.88
Current smoker	0.027 (0.073)	0.711	1.03
Current drinker	-0.098 (0.067)	0.142	0.91
Physically inactive	0.682 (0.075)	0.000	1.98

DISCUSSION

This study investigated the utility of the Poisson, negative binomial (NB), ZIP and ZINB regressions for modelling the number of difficulty with Activities of Daily Living (ADL-s) in a national sample of older adults, participants of the first wave of the English Longitudinal Study of Ageing (ELSA).

There is a growing literature on models specifically developed for count outcomes, and several examples are available in epidemiology and public health research. [22-27] However, this is the first study to bring the utility of these approaches to model a count measure of physical functioning, ADL-s, widely used in epidemiology.

We found that 79% of respondents in our study reported no difficulties with ADL-s. Poisson regression is often used for count data; however, we have shown that when the observed counts exhibit more variability than what is predicted by the Poisson (known as overdispersion), like in our outcome variable, is better to use a different model. To allow for overdispersion we used the NB regression model. The NB provided a substantial improvement over the Poisson. To account for both dispersion and the excess zeros the ZIP and ZINB were fitted to the data. The ZIP and ZINB provided a similar fit; however, we found that the ZINB modelling approach yielded the best fit. The choice of the ZINB however, cannot be made strictly on the basis of model fitting. [15, 17-18] In our application it seems reasonable to assume that there is a separate population of zeros formed of those who did not experience any difficulty in performing ADL-s, as results of being less limited in their physical functioning or simply healthier people.

We have also performed logistic regression using the dichotomous variable of ADL-s (comparing 0”difficulty with ADL-s versus 1”1 and more difficulty with ADL-s). Results from logistic regression were consistent with those found from the ZINB regression, the exception was only for physical activity. In the logistic regression, being physically inactive was associated with increased odds of having 1 to 6 difficulties with ADL-s. In the ZINB model, being physically inactive was not significantly associated with the odds of remaining without difficulties for those with zero difficulties (Always zero group) but it was positively associated with the odds of having increasing number of difficulties for those with difficulties (Not always zero group). This information is not available with logistic regression. Also by using a logistic regression is not possible to consider the range of difficulties with ADL-s experienced; while the ZINB model offers the advantage of modelling difficulties with ADL-s scores on a continuum instead of the dichotomized outcome used in logistic regression. Therefore the ZINB model allows assessing the association of the exposure variables with disease severity and not just the presence or absence of disease.[13]

The finding that being a current drinker is associated with a reduction in the number of difficulty with ADL-s may be part explained by the cross-sectional nature of the data and in part by the insufficient information available about drinking frequency. Respondents were asked to report whether they drink twice a day or more, whether they drink less frequently than twice a day or whether they abstain completely from drinking. Detailed information about the number of daily alcohol consumption would have provided a better understanding of the relationship between this variable and ADL-s. Future research could investigate the utility of the ZIP and ZINB regressions for modelling the number of difficulty with ADL-s using longitudinal data.

In regression analysis of count data, independent variables are often modelled by their linear effects under the assumption of log-linearity. However, the validity of such an assumption is difficult to test due to the lack of readily accessible testing procedures.

To conclude we have shown that models specifically developed for count outcomes with excess zeros such as ZIP and ZINB can provide better insights into the investigation of the factors associated with the number of difficulties with Activities of Daily Living (ADL-s). When dealing with the number of difficulties with ADL-s we recommend the use of the ZIP and ZINB models rather dichotomizing the outcome which yields to loss of information with respect to the association of exposure with disease severity.

What is already known on this subject?

Counting outcomes have often been used as continuous variables, transformed to induce normality, or categorized which can yield to biased estimates or loss of information. Alternatives models such as Poisson, negative binomial (NB), zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) have been proposed which better suited to counting processes.

What does this study add?

This paper provides a useful methodology to analyse a count outcome measures like Activities of Daily Living (ADL-s) in epidemiological studies. We recommend the use of the ZIP or ZINB models which can provide better insights into the investigation of the factors associated with the number of difficulties with ADL-s.

Acknowledgements: The data were made available through the UK Data Archive. Elsa was developed by a team of researchers based at University College of London, the National Centre for Social Research and the Institute for Fiscal Studies. The data were collected by the National Centre for Social Research.

Competing interests: None.

Funding: The funding is provided by the National Institute of Aging in the United States, and a consortium of UK Government departments coordinated by the Office for National Statistics. The developers and funders of ELSA and the Archive do not bear any responsibility for the analyses or interpretations presented here.

Licence Statement

Figures Legend

Figure 1 Predicted proportions from intercept-only Poisson, NB, ZIP and ZINB models compared with the observed proportions from ADL-s score

Figure 2 Observed minus predicted probabilities for four models

References

1. Boult C, Kane RL, Louis TA, Boult L, & McCaffrey, D. Chronic conditions that lead to functional limitation in the elderly. *J Gerontol.*, 1994 ; **49**:M28-M36.

2. Gill TM, Williams CS, Richardson ED, Tinetti ME. Impairments in physical performance and cognitive status as predisposing factors for functional dependence among nondisabled older persons. *J Gerontol.A Biol.Sci Med.Sci*, 1996; **51**: M283-M288.

3. Stuck AE, Walthert JM, Nikolaus T, Bula CJ, Hohmann C, Beck JC. Risk factors for functional status decline in community-living elderly people: a systematic literature review. *Soc.Sci Med.*, 1999; **48**: 445-469.

4. Tinetti ME, Inouye SK, Gill TM, Doucette JT. Shared risk factors for falls, incontinence, and functional dependence. Unifying the approach to geriatric syndromes. *JAMA*, 1995; **273**: 1348-1353.

5. Turvey CL, Schultz SK, Klein D. M. Alcohol use and health outcomes in the oldest old. *Subst.Abuse Treat.Prev.Policy*, 2006; **1**: 8.

6. Steel N, Huppert F, McWilliams B, Melzer D. Physical and Cognitive Function. In: *Health, wealth and lifestyles of the older population in England. The 2002 English Longitudinal Study of Ageing*, Marmot M, Banks J, Blundell R, Lessof C, Nazroo J, (Eds) London: The Institute for Fiscal Studies; 2003 p. 249-301.

7. Aijanseppa S, Notkola IL, Tijhuis M, van SW, Kromhout D, Nissinen A. Physical functioning in elderly Europeans: 10 year changes in the north and south: the HALE project. *J Epidemiol.Community Health*, 2005; **59**: 413-419.

8. HRS (2007) Health In: *Growing Older in America: The Health and Retirement Study* http://hrsonline.isr.umich.edu//docs/databook/HRS_Text_WEB_Ch1.pdf (accessed 06/05/2008)

9. Katz S, Ford AB, Moskowitz RE, Jackson BA, Jaffee MW, Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. Jou*rnal of the American Medical Association*, 1963; **185**: 914-919

10. McCullagh P & Nelder JA. *Generalized linear models*. London : Chapman and Hall 1989.

11. Agresti A. *An introduction to categorical data analysis*. New York; Chichester: Wiley 1996.

12. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics,* 2000; **56**: 1030-1039.

13. Lambert D. Zero-Inflated Poisson Regression, with An Application to Defects in Manufacturing. *Technometrics*, 1992; **34**: 1-14.

14. Mullahy J. Specification and Testing of Some Modified Count Data Models. Journal of Econometrics, 1986; **33**: 341-365.

15. Long JS. *Regression models for categorical and limited dependent variables*. Thousand Oaks; London: Sage 1997.

16. Long JS., Freese, J., & Stata Corporation. *Regression models for categorical dependent variables using Stata*. College Station, Texas : Stata Corporation 2003.

17. Cameron AC & Trivedi PK. *Regression analysis of count data.* Cambridge, UK; New York, NY, USA : Cambridge University Press 1998.

18. Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. *Stat.Med.*, 2002; **21**: 1461-1469.

19. Afifi AA, Kotlerman JB, Ettner SL, Cowan M. Methods for improving regression analysis for skewed continuous or counted responses. *Annu.Rev.Public Health*, 2007; **28**: 95-111.

20. Ridout M, Demetrio CGB, Hinde J. Models for count data with many zeros. *International Biometric Conference*, Cape Town 2008.

http://www.kent.ac.uk/IMS/personal/msr/webfiles/zip/ibc_fin.pdf (accessed 26/11/2008).

21. Greene WH Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. 1994 Working Paper no.EC-94-10, Department of Economics, Stern School of business, New York University.

22. Lewsey JD & Thomson WM. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology* 2004; **32**:183-9.

23. Wong EL, Roddy RE, Tucker H, Tamoufe U, Ryan K, Ngampoua F. Use of male condoms during and after randomized, controlled trial participation in Cameroon. *Sexually Transmitted Diseases* 2005; **32**:300-307.

24. Bulsara MK, Holman CDJ, Davis EA, Jones TW. Evaluating risk factors associated with severe hypoglycaemia in epidemiology studies- what method should we use? *Diabetic Medicine* 2004; **21**:914-919.

25. Qin X, Ivan JN, Ravishanker N. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention* 2004; **36**:183-91.

26. Slymen DJ, Ayala GX, Arredondo EM, Elder JP. A demonstration of modeling count data with an application to physical activity. *Epidemiol Perspect Innov*. 2006; **21**:1-9.

27. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *J Biopharm Stat*. 2006; **16**: 463-81.

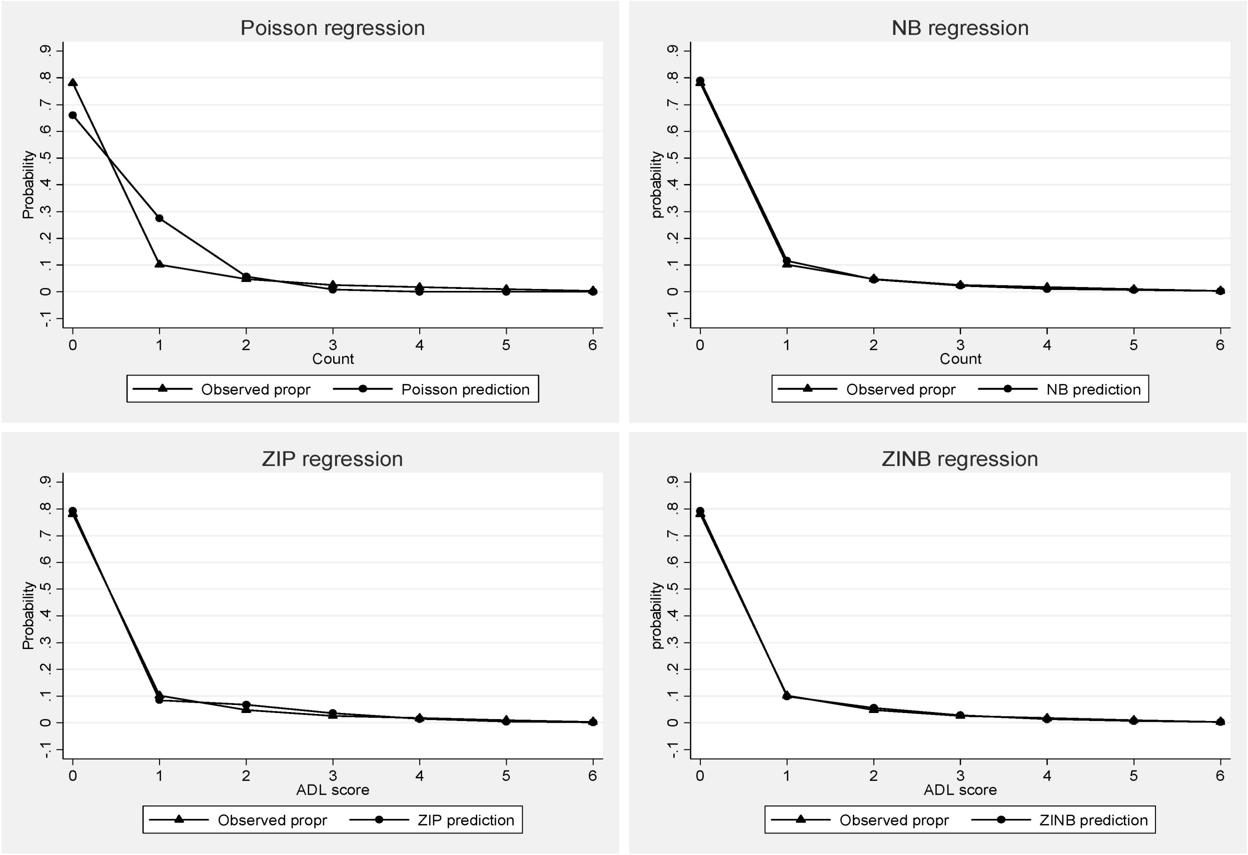
28. Marmot MG. Institute for Fiscal Studies. *Health, wealth and lifestyles of the older population in England : the 2002 English Longitudinal Study of Ageing*. London : Institute for Fiscal Studies 2003.

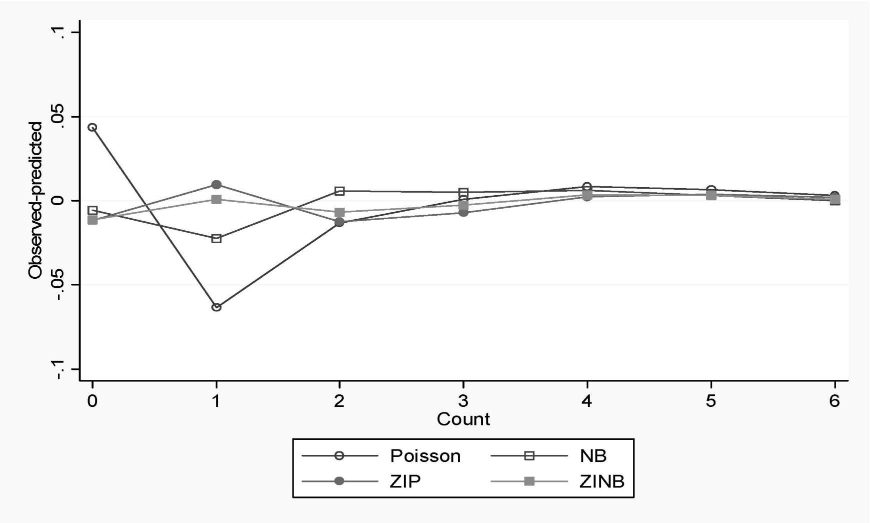
29. Banks J, Breeze E, Lessof C, Nazroo J. *Retirement, health and relationships of the older population in England : the 2004 English Longitudinal Study of Ageing.* London, The Institute for Fiscal Studies 2006.

30. Miaou SP. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions *Accid.Anal.Prev.*, 1994; **26**: 471-482.

31. Poch M. & Mannering F. Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering-Asce*, 1996; **122**: 105-113.

32. Vuong QH. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica,* 1989; **57**: 307-333.


Poisson regression
Probability
-.1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
0
1
2
3
4
5
6
Count
Observed propr
Poisson prediction
NB regression
probability
Count
Observed propr
NB prediction
ZIP regression
Probability
ADL score
Observed propr
ZIP prediction
ZINB regression
probability
ADL score
Observed propr
ZINB prediction


.1
.05
0
-.05
-.1
Observed-predicted
0
1
2
3
4
5
6
Count
Poisson
NB
ZIP
ZINB