Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Liu and Wang, J Biom Biostat 2019, 10:3

Research Article Open Access

Comparisons of Four Discrete Distributions in Count Regressions Using Elders' Missing Teeth Data

Ying Liu* and Kesheng Wang

Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN, USA

Abstract

Objectives: This present study aimed to select the best count distributions for missing teeth in elders and to investigate the relationship between missing teeth and the predictors.

Materials and methods: Data were extracted from the biennial survey of 2015-2016 the U. S. National Health and Nutrition Examination Survey. Only adults aged 60 years or over who completed oral health examination and demographics interview were included. Descriptive statistics were used to demonstrate the basic information of this studied population. The performances of four different count regression models (Poisson regression, negative binomial regression with linear variance function (NB1), negative binomial regression with quadratic variance function (NB2), and zero-inflated negative binomial regression) were compared through different approaches including the values of model fit test statistics such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), the magnitude of standard errors and a visual graph on the performances of fitted models.

Results: The disparities on missing teeth existed in old adults by poverty and educational level and race/ethnicity. More missing teeth were found in participants who are Blacks (mean=13.89), with less education (<12 years) (mean: 13.11). Significance of t-test for "α" indicated that Poisson distribution is not appropriate for missing teeth due to overdispersion. NB1 is the best model with the smallest AIC and BIC and the smallest standard errors of parameter estimates compared to other three candidate models.

Conclusion: The negative binomial distribution with linear variance function is the best distribution. Due to the fact of missing teeth which ranged from 0 to 28, the caution should be given when we interpreted the fitted model using NB1 as the missing teeth are close to 0 and 28.

Keywords: Count data; Missing teeth; Model fit

Introduction

Count data is widely used in medical research [1]. Examples include the number of seizures, the numbering of hospitalizations, the number of missing teeth in oral health, etc. Traditional linear regression model is not appropriate for count data to investigate the association between outcome variable and its predictors. Poisson regression (log-linear) model is widely used for count data [2]:

$$\ln[E(Y_i / X_i)] = \mu_i = X_i'\beta \quad (1)$$

Where β is a vector of regression coefficient, and x_i is a vector of covariates for subject *i*. In the Poisson regression, the outcome variable Y was assumed drawn from a Poisson distribution and with the density function:

$$f(Y_i / X_i) = \frac{e^{-\mu_i} * \mu_i^{Y_i}}{Y_i!} \quad (2)$$

The main character of Poisson distribution is that the estimated mean value of random variable Y_i is equal to its variance: $\mu_i = E(Y_i) = Var(Y_i)$ $\mu_i = E(Y_i) = Var(Y_i)$. Poisson regression can be viewed as an extension of general linear model that the population means depends on linear function via a nonlinear link function named log link [2].

In the real world, the variance of the observed data usually are much greater than the mean, called over-dispersion, due to heterogeneity and/or exceed zeros, Poisson regression model was criticized for its strict assumption that conditional mean equals to its conditional variance. We cannot correctly interpret the fitted model when the assumption is not held. Negative binomial regression models were used to break through the limitation of Poisson regression [3]. Negative binomial distribution allows various forms of means-variance relationship. A general class of negative binomial models $E(Y_i) = \mu_i$ with, and $Var(Y_i) = \mu_i + \alpha\mu_i^p$, $-\infty < p < \infty$, where α is a constant [4], used to adjust the variance. Negative binomial distribution converges to Poisson distribution when α approaches to 0. In a binomial distribution, we can assume that a sequence of independent Bernoulli trials, and each trial has two possible outcomes called "success" with probability *p* and "failure" with probability 1-*p*. The random variable X is defined the number of success before a predetermined r of failures occurred, that is, X~NB(r, P).

Two major forms of negative binomial distribution were known as linear (NB1) and quadratic (NB2) negative binomial distribution given by *p*=1 and *p*=2, respectively [5,6]. The outcome variable Y_i given X_i is distributed as a negative binomial and its density function is denoted as:

$$f(Y_i \mid \mathbf{X}_i) = \frac{\Gamma(Y_i + \alpha^{-1})}{\Gamma(Y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{Y_i} \quad (3)$$

Zero-inflated count models were proposed by the fact of excess zeros in the real-life data [7,8]. In this study, we limited our description to zero-inflated negative binomial distribution. Zero-inflated negative

***Corresponding author:** Ying Liu, Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN, USA, Tel: 4234394477; E-mail: liuy09@etsu.edu

Received February 08, 2019; **Accepted** March 29, 2019; **Published** April 05, 2019

Citation: Liu Y, Wang K (2019) Comparisons of Four Discrete Distributions in Count Regressions Using Elders' Missing Teeth Data. J Biom Biostat 10: 427.

binomial regression is a way of modeling the data that both excess zeros and overdispersion exist. Zero-inflated regression, therefore, can be viewed a mixture of two statistics processes. For an i[th] observation, process one is used with a Bernoulli probability $_i$ which generate zero counts, and process two is used with probability 1-ϕ_i which generates a negative binomial model, that is,

$$y_i \sim \begin{cases} 0 & with\ probability\ \phi_i \\ g(y_i \mid \mathbf{x}_i)\ with\ probability\ 1-\phi_i \end{cases} \quad (4)$$

Then the probability of $(Y_i|X_{i,} Z_i)$ is expressed as:

$$P(Y_i \mid \mathbf{x}_i, \mathbf{z}_i) = \begin{cases} \phi(\mathbf{z}_i'\gamma) + \{1-\phi(\mathbf{z}_i'\gamma)g(0 \mid \mathbf{x}_i) & if\ y_i = 0 \\ \{1-\phi(\mathbf{z}_i'\gamma)g(y_i \mid \mathbf{x}_i) & if\ y_i. > 0 \end{cases} \quad (5)$$

Where Z_i and X_i are covariate matrix, γ is the vector of zero-inflated coefficient to be estimated in the model building. The function to the probabilityϕ_i is the zero-inflated link function, it can be a logit link or probit link function. The function to the probability of 1-ϕ_i can be a Poisson or negative binomial distribution.

Missing teeth is a very common problem which may result in a great embarrassment for people, especially for old adults. Lost teeth bring problems in daily life including difficulty eating or chewing, speech problems and shifting of adjacent teeth. The goal of this study is to select the best count distributions for missing teeth in old adults and to investigate the relationship between missing teeth and the predictors.

Materials and Method

Data used in the present study are from survey of 2015-2016 which are subset of the series of National Health and Nutrition Examination Surveys (NHANES) conducted by the National Center for Health Statistics (NCHS). NHANES survey datasets were nationally representative information using a stratified, multistage design. Full details of the survey methods used can be found at http://www.cdc.gov/nchs/nhanes.htm. For the current study, only subjects who were ≥ 60 years old and had completed an oral examination were included. The oral health examination data were released in April 2018.

Ethics statement

This study was conducted in accordance with the latest version of the principles of the Declaration of Helsinki. The NCHS Research Ethics Review Board (ERB) approved the study (NCHS IRB/ERB Protocol #2011-17); further ethical approval for the use of NHANES data is not required and the data are is available on the internet.

Demographic and socio-economics (SES) variables

Five socio-demographic variables were included in data analysis: (1) age at screening, (2) gender, (3) race/ethnicity including Hispanic, non-Hispanic white, non-Hispanic black and other race, (4) poverty income ratio (PIR) which is the ratio of family income to the Federal Poverty Threshold (FTP), adjusted for family size and composition, and (5) education which reflects the highest grade or level of school completed by the participant was used in the data analysis as being <12 years, 12 years and >12 years.

Oral examination variables

The oral health of participants were examined by dentists (D.D.D/D.M.D.) licensed in at least one state in the U.S. The examinations were conducted in the mobile examination centers (MECs), and oral health data were recorded directly on the computerized data form. Third molars were excluded in this study and number of missing teeth is ranged from 0 to 28. The quality of data is assessed by internal quality control and it is acceptable.

Statistical analysis

Descriptive statistics were used to summarize the demographics and status of the missing teeth of the sample participants based on an oral examination. We calculated the proportions by gender, race/ethnicity and educational level. The mean and standard deviation of age and family income ratio were also calculated. The average missing teeth and its standard deviation were accordingly calculated. Based on properties of the response variables (number of missing teeth), count regression was used to investigate the relationship between the number of missing teeth and its predictors: age, gender, race/ethnicity, family poverty level and educational level. To obtain better model, four different count regression models were considered: Poisson regression, negative binomial regression with linear and quadratic variances; zero-inflated negative binomial regression with linear variance.

In this study, SAS COUNTREG procedure was used to execute data analyses for count response variable. Nonlinear Newton-Raphson optimization technique was used for iterative minimization. Akaike's Information Criterion (AIC) and Schwarz's Bayesian information criterion (BIC) are estimators of the relative quality of model (s) given a dataset and were used to compare a set of candidate count regression models in this study [9,10]. Smaller values of these two criteria indicate better models. The formulas for AIC and BIC are denoted as the below:

$$AIC = 2k - 2\ln(\hat{L}) \quad (6)$$

$$BIC = \ln(n)K - 2\ln(\hat{L}) \quad (7)$$

Where

$\hat{L}$: The maximized value of the likelihood function of the fitted model;

n: The number of data points in observed data; and

k: The number of parameters estimated by the model.

Scaled deviance (Value of Deviance/df) is an estimate of dispersion parameter in the Poisson distribution. Aplha was used to measure the dispersion in the negative binomial distribution and zero-inflated negative binomial distribution, where a t test was applied to test the hypothesis H_0: α=0 and p<0.05 indicates the evidence of significant overdispersion. Variable selection used a rigorous approach - penalized likelihood method which consider model complexity [11,12]. Data management and analyses were performed with the Statistical Analysis System (SAS, version 9.4; SAS Institute Inc., Cary, NC, USA). A *p*-value<0.05 was considered statistically significant.

Results

Demographics and socio-economic characteristics

A total of 1567 participants were included in this study with mean age of 69.84 (sd=6.87). Demographics and socio-economic characteristics were showed in Table 1. Non-Hispanic white accounted for 40% followed by Hispanic (31%) and Black (19%). About 48% percent of old adults received college education (>12 years education). The average missing teeth among old adults was 10.6 (sd=9.89). The larger number of missing teeth were observed in old Black adults (mean=13.89, sd=10.11) followed by individuals from other races (mean=9.99, sd=10.28). A gradient of missing teeth was seen by educational level, that is, the individuals whose education years were less than twelve years (mean=13.11, sd=10.29) have more missing teeth

than individuals with 12 years education (mean=11.50, sd=9.75) and more than 12 years education (mean=8.60, sd=9.25). Correlation was found between missing teeth and age and family poverty level with correlation coefficient 0.22 and -0.31, respectively.

Results from four different count regression model

To obtain a better model, we did the following three steps: we firstly fit Poisson regression model. Penalized likelihood method indicated gender could not significantly contribute to the model (data not shown). Age, race, poverty level and educational level are significant in these four models. The estimates and their standard errors of remained predictors were shown in Table 2. In the fitted Poisson regression model, the scaled deviance (value=8.3) is much greater than 1 which indicated a large variability existed in the fitted model and the fitted model probably unmatched the assumption of Poisson distribution: the mean and variance of the model are identical. The large scaled deviance also indicated Poisson regression model is not adequate to describe the relationship between missing teeth and its predictors since there is a greater variability among missing teeth than that would be expected for Poisson distribution. In next step, two negative binomial distributions with different variance format were used: linear and quadratic forms. α were far from zero and the *p*-values of t test for α showed a strong sign that negative binomial distribution is preferred to Poisson distribution. Consider the adequacy of the fitted models in this study, both AIC and BIC were compared between two negative binomial distributions, the negative binomial distribution with linear function of variance was preferred with smaller AIC and BIC (Table 2). Three are 209 old adults (≥ 60 years old) had no missing teeth, that is, there are 209 zeros in this study. In step three, we fit a zero-inflated negative binomial distribution with a linear variance format, where we thought age is the variable which contributes to increase missing teeth. However, both AIC and BIC of the fitted zero-inflated model are greater than in the negative binomial distribution with linear function. Therefore, the best distribution considered in this study is negative binomial distribution with a linear variance function. It is worth noting that the standard error of parameter estimates was the smallest in NB1 model than Poisson regression, NB2 and zero-inflated NB1 models. Therefore, the final equation is denoted as:

$$E(missingteeth/X) = \exp(0.48 + 0.0288 * age + 0.04301(if\ race = black) + 0.0560 * (if\ race = Hispanic) - 0.0521 * (if\ race = other) - 0.1664 * poverty\ ratio + 0.1547(if\ edu < 12) + 0.1800(if\ edu = 12)$$

	n (%) or mean (sd)	Missing teeth mean (sd)
Gender		
Male	795 (50.73)	10.84 (9.97)
Female	772 (49.27)	10.36 (9.80)
Race		
Non-Hispanic Black	289 (18.84)	13.89 (10.11)
Hispanic	486 (31.01)	9.93 (8.91)
Non-Hispanic White	627 (40.01)	9.14 (9.56)
Other race	165 (10.53)	9.99 (10.28)
Education		
<12	272 (30.12)	13.11 (10.29)
12	349 (22.27)	11.50 (9.75)
>12	746 (47.61)	8.60 (9.25)
Age (years)	69.84 (6.87)	0.22[a]
Family poverty ratio	2.31 (1.55)	-0.31[a]
[a]The correlation coefficient with number of missing teeth		

Table 1: Demographics of American adults from NHANES, 2015-2016.

Comparisons between observed and predicted probability

In this session, we compared the performances of four different models by comparing the sample probability distribution of the data to the average probability distribution predicted from four fitted models. The sample probability distribution and average probability distributions from fitted models were summarized in Table 3 and demonstrated in Figure 1. Poisson regression had lowest probability at zero missing teeth (0.34%) which is much smaller than the observed probability by 13%. NB2 had the smallest difference with observed probability at zero missing teeth, followed by zero-inflated NB1 and NB1. The average probability from Poisson regression kept much higher than observed probability with range 7 to 15 missing teeth. NB1, NB2 and zero-inflated NB1 have similar performance, and their average probabilities were getting closer to the observed probability from 6 to 27 missing teeth. All four models showed much lower probability than

	Model 1	Model 2	Model 3	Model 4
	Estimate (S. E)	Estimate (S. E)	Estimate (S. E)	Estimate (S. E)
Age	0.0283 (0.0012)***	0.0288 (0.0033)***	0.0312 (0.0040)***	-0.0289 (0.0103)a**
Race				
Black	0.3323 (0.0212)***	0.4301 (0.0613)***	0.4181 (0.0759)***	0.02496 (0.0690)****
Hispanic	-0.1034 (0.0219)***	0.0503 (0.0624)	-0.0554 (0.0693)	-0.2230 (0.0625)***
Other race	-0.0815 (0.0293)**	0.0560 (0.0830)	-0.0051 (0.0930)	-0.0824 (0.0882)
White	(Reference)			
Poverty level	-0.1840 (0.0293)***	-0.1664 (0.0172)***	-0.1984 (0.0187)***	-0.1842 (0.0180)***
Education				
<12	0.2092 (0.0202)***	0.1547 (0.0594)**	0.2371 (0.0673)***	0.2746 (0.0627)***
12	0.1461 (0.0205)***	0.1800 (0.0569)**	0.1721 (0.0689)*	0.1428 (0.0642)***
>12	(Reference)			
Alpha	N/A	9.7604 (0.4782)***	0.9818 (0.0408)***	0.7628 (0.0443)***
Scaled Deviance	8.2287	N/A	N/A	N/A
AIC	18207	10380	10442	10460
BIC	18250	10428	10490	10514
*p<0.05; **p<0.01; ***p<0.001 Model 1: Poisson regression. Model 2: Negative binomial regression with linear variance. Model 3: Negative binomial regression with quadratic variance. Model 4: Zero-inflated negative binomial regression with linear variance.				

Table 2: Parameter estimates from three different types of count regression models.

Missing teeth	Frequency	Observed (%)	M1 Estimate (%)	M2 Estimate (%)	M3 Estimate (%)	M4 Estimate (%)
0	209	13.34	0.3412	16.8360	12.2950	14.4310
1	118	7.53	1.3651	9.6070	10.2390	7.1770
2	93	5.93	3.1443	7.5750	8.7130	6.9530
3	100	6.38	5.1156	6.7470	7.7470	6.7880
4	97	6.19	7.4352	5.8860	6.8080	6.3090
5	62	3.96	6.3723	5.2830	5.6620	5.3840
6	70	4.47	7.3093	4.8100	5.0500	5.0130
7	71	4.53	8.0358	4.3510	4.5180	4.6540
8	54	3.45	8.5366	3.9480	4.0360	4.1200
9	47	3	8.6659	3.8210	3.6920	3.8280
10	43	2.74	8.5483	3.4520	3.3350	3.5320
11	35	2.23	8.5900	3.2370	3.0810	3.2490
12	21	1.34	7.3810	3.0590	2.7770	2.9740
13	21	1.34	6.8591	2.8340	2.5390	2.7970
14	29	1.85	5.0317	2.4770	2.2910	2.5210
15	22	1.4	5.4327	2.4680	2.1880	2.3750
16	21	1.34	3.9017	2.0790	1.8750	2.0840
17	26	1.66	3.6758	2.1170	1.8640	2.0250
18	21	1.34	2.4190	1.8310	1.5790	1.7370
19	37	2.36	2.3971	1.7320	1.5530	1.6930
20	35	2.23	2.5199	1.6520	1.4240	1.5180
21	20	1.28	1.6759	1.5120	1.3250	1.4920
22	25	1.6	2.3513	1.5950	1.3090	1.3390
23	15	0.96	0.9609	1.1650	1.0220	1.1220
24	16	1.02	1.0172	1.1400	0.9600	1.0340
25	7	0.45	1.0934	1.3350	1.1020	1.1760
26	9	0.57	0.2633	0.8980	0.7860	0.8740
27	8	0.51	1.2340	1.1740	0.9940	0.9850
28	235	15	0.5101	0.9850	0.8790	0.8820

M1: Poisson regression.
M2: Negative binomial regression with linear variance.
M3: Negative binomial regression with quadratic variance.
M4: Zero-inflated negative binomial regression with linear variance.

Table 3: Frequency and percentage of missing teeth from observed and four models.

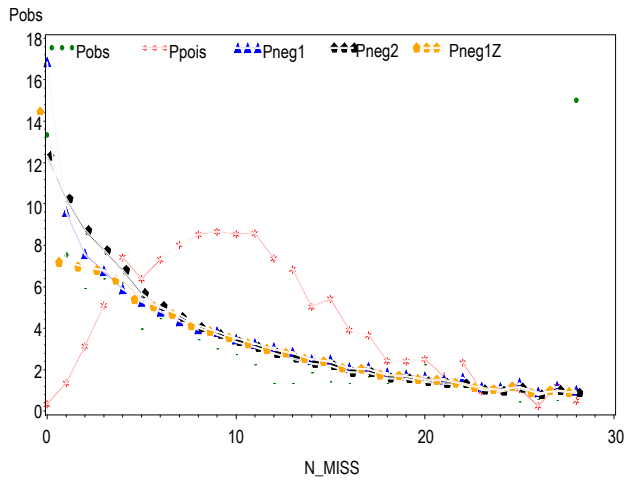

Figure 1: Plot of percent vs. number of missing teeth in American elders. Pobs: Observed percent; Ppois: Estimate percent from Poisson regression; Pneg1: Estimate percent from negative binomial regression with linear variance; Pneg2: Estimate percent from negative binomial regression with quadratic variance.

observed probability when individuals had 28 missing teeth (Table 3 and Figure 1).

Discussion and Conclusion

In this study, the overall missing teeth (10.6) among old adults were lower than in 2005-06 and 2007-08. We showed the missing teeth by different demographics and seriocomic status. The average missing teeth are similar in males (10.84) and females (10.36). The disparities of missing teeth were observed by educational and poverty level and races. More specifically, the individuals with higher level of SES have fewer missing teeth compared to the individuals with lower level of SES. The individuals who have college level education had fewest missing teeth (8.6) followed by those who have high school diploma. The negative correlation coefficient (-0.31) between missing teeth and family poverty ratio indicated that the poor individuals are more likely to have more missing teeth than non-poor individuals. Black old adults have more missing teeth (13.89) compared to their white peers (9.14) and other races including Hispanics. Generally, people would have more missing teeth as they are getting older even for the old adults aged over 60 years [13]. In this present study, a positive correlation coefficient was found between missing teeth and age (0.22).

We identified four predictors considered in this study are significantly associated with number of missing teeth in old adults aged ≥ 60 years. By comparing four proposed models in terms of the model fit statistics, standard errors of estimates, we finally found NB1 is the best among these four candidates. Furthermore, we also compared the performance estimated probabilities from four models with observed

probabilities in a visualized graph, NB1 is still the winner even though the BN2 is the most commonly known and utilized [5]. The final model means the better model among candidate models rather than the perfect model for the observed data in real life.

In the fact of that, missing teeth are ranged from 0 to 28 rather than non-negative unbounded range. That is, an individual has 28 maximum missing teeth. It is not surprising there is a large difference between the estimated probability and the observed probability, and the caution should be given for the interpretation. As we noticed this final model has higher probability than observed probability when missing teeth are less than 3. Therefore, we should caution when the fitted model needs to be explained, especially at small number of missing teeth (<3). The real phenomenon in the old adults is that there are relative large proportions with 0-3, and 28 missing teeth, we speculated that made the data more dispersion. McCullagh and Nelder [2] and Cameron and Trivedi [14] stated that the negative binomial distribution may produce a very similar result to the an overdispersion Poisson distribution given a modest overdispersion. The very different results were observed from Poisson regression and three types of negative binomial regressions indicated that the overdispersion is severe in this study.

In this study, our final model (NB1) was used for older adults which is not appropriate for young kids or young adults based on the fact that very few missing teeth existed in young people. We speculated that zero-inflated count regressions are more appropriate which merit further investigation. With property of cross-sectional study, the final model can be only used to investigate the association relationship rather than a causal relationship.

Acknowledgement

The authors thank the U.S. CDC/NCHS for providing the NHANES 2015-16 data.

Declaration

The authors declare that they have no conflicts of interest.

References

1. Glynn RJ, Buring JE (1996) Ways of Measuring Rates of Recurrent Events. Education and Debate, BMJ 312: 364-367.
2. McCullagh P, Nelder JA (1989) Generalized Linear Models (2ndedn), Monographs on Statistics & Applied Probability, Chapman & Hall/CRC.
3. Lawless JF (1987) Negative binomial and mixed Poisson regression. Can J Stat 15: 209-225.
4. Cameron AC, Trivedi PK (1986) Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. J Appl Econ 1: 29-53.
5. Lord D, Washington SP, Ivan JN (2005) Poisson, Poisson-Gamma and Zero Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. Accid Anal Prev 37: 35-46.
6. Maher MJ, Summersgill I (1996) A Comprehensive Methodology for the Fitting Predictive Accident Models. Accid Anal Prev 28: 281-296.
7. Lambert D (1992) Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing. Technometrics 34: 1-14.
8. Greene WH (1994) Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. NYU Working Paper No. EC-94-10.
9. Kass RE, Wasserman L (1995) A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. J Am Stat Assoc 90: 928-934.
10. Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95: 631-636.
11. Preminger A, Wettstein D (2005) Using the penalized likelihood method for model selection with nuisance parameter present only the alternative: an application to switching regression models. J Time Ser Anal 26: 715-741.
12. Cole SR, Chu H, Greenland S (2014) Maximum likelihood, profile likelihood, and penalized likelihood: A Primer. Am J Epidemiol 179: 252-260.
13. Dye BA, Weatherspoon DJ, Mitnik LG (2019) Tooth loss among older adults according to poverty status in the Uinted States from 1999 through 2004 and 2009 through 2014. J Am Dent Assoc 150: 9-23.
14. Cameron AC, Trivedi PK (1998) Regression analysis of count data (2ndedn), Cambridge University Press.