Original Research Paper

Volume-8 | Issue-8 | August-2018 | PRINT ISSN No 2249-555X

Statistics

ZERO TRUNCATED COUNT MODELS WITH AN APPLICATION TO THE DENTAL CARIES DATA

Dr. S. B. Javali* Senior Associate Professor in Statistics, Department of Community Medicine, USM - KLE, International Medical Programme, Belagavi, Karnataka, India *Corresponding Author

Dr. C. M. Math Associate Professor, R. L. Science Institute, College Road, Belagavi, Karnataka, India

ABSTRACT Count response data often exhibit departures from the assumptions of standard Poisson generalized linear models and ordinary least-squares estimation depends on the purpose of that estimation. The propensity of Decayed Missing Filled (DMF) index count data is to contain many zeros and to follow highly skewed distribution or overdispersed. Extra zeros however, violate the variance-mean relationship of the Poisson errors structure. The distribution of DMF counts had nearly 50% of zero values often are not observed. In such case, the appropriate model for modelling positive DMF count data would be the models truncated at zero. This paper examines maximum likelihood regression estimators for DMF count data from truncated samples. Estimators for the Zero Truncated Poisson and Negative Binomial regression models are illustrated and compared with standard Poisson regression and Negative Binomial regression models. Truncated models can provide new insight to caries pattern in an examination of covariates on DMF positive counts. It is anticipated that, Zero truncated models are becoming increasingly useful in recent epidemiological studies of dental caries data with positive counts as an outcome measure.

KEYWORDS : DMF positive count data, Overdispersed, Poisson, Negative Binomial, Zero Truncated Models

INTRODUCTION

In empirical applications, economists and medical scientists are increasingly estimating regression models that are truncated in nature. Studies on the determinants of dental caries by DMF index have frequently used count data models. Count data models are attractive because the dependent variable is a non-negative integer, mutually exclusive and collectively exhaustive. The number of zero counts in some empirical cases exceeds the number that would be expected in applications of the conventional count (Poisson or its variation). Conventional count models fail to account for two different data-generating mechanisms for the zero and strictly positive counts (Mullahy, 1986; Winkelmann, and Zimmermann, 1995; Gurmu and Trivedi, 1996).

Data obtained based on DMF count could have too many zero values. In this case, the generalized linear and zero truncated models are one of the methods used in the modeling the dependent variable having too many zero data. But, nearly 50% of DMF counts had zero values and the DMF count data as whole show overdispersion. In such case, we are considering that portion of data where the DMF>0. However, the appropriate model for the analysis of such data would be the models truncated at zero. The popular models that we use here are truncated Poisson and truncated negative Binomial models.

This article deals with the application of zero-truncated Poisson and Negative Binomial to the positive DMF count data. The zero truncated models have been considered by many authors to analyze positive count data. David & Johnson (1952) and Plackett (1953) have used truncated models in the beginning further by Johnson, Kotz & Kemp (1992). Shaw (1988) extends the Poisson generalized linear model to deal with truncated count data. Alternatively, zero truncated count data can be modelled via the negative binomial generalized linear model, see Gurmu (1991) and Grogger & Carson (1991). Gurmu & Trivedi (1992) present tests for overdispersion in the truncated count model. The truncated Poisson generalized linear model has been applied to adenomatous polyps data by Xie & Aicken (1997). Examples of economic applications of the Truncated Poisson Generalized Linear Model are given in Cameron & Trivedi (1998).

However in many cases, the analyst does not observe the entire distribution of counts. In particular the zeros often are not observed. Consider an example of such a situation. A public dentist wants to administer surveys on public for assessment of caries status and other variables which might be related to that behaviour, deleterious habit and demographic variables. Given these data we seek to construct a model of the number of individuals with dental caries (DMF >0) taken as a function of various variables.

The application of count regression models is receiving much attention from econometricians. (Gourieroux, Monfort, and Trognon, 1984b; Hausman, Hall, and Griliches, 1984; Lee, 1986). These models have seen increasing use in the analysis of outcomes naturally measured as non-negative integers; applications include studies of firms' patenting behaviour (Hausman, Hall, and Griliches, 1984), doctor and hospital visits (Cameron and Trivedi, 1986; Cameron et al., 1988), daily beverage consumption (Mullahy, 1986), incidents of pollution induced illness (Portney and Mullahy, 1986), and daily homicide counts (Grogger, 1990).

More generally, two common types of sampling schemes are likely to give rise to samples of truncated DMF counts: number of teeth affected with dental caries and examined by dentists only. The underlying statistical similarity between both types of samples is that the observational apparatus potentially becomes active only with the occurrence of some specified (typically one) number of events (Johnson and Kotz, 1969).

Recently, Shaw (1988) has proposed normal and Poisson regression models for the analysis of truncated samples of count data. In this chapter we introduced the two count regression models for non truncated samples and propose estimators based on the truncated Poisson as well as the negative binomial distribution. The choice of models is shown to be important in analyzing truncated samples as an application of the truncated Poisson model to data which fail to meet its stringent moment restrictions that may result in seriously biased and inconsistent parameter estimates. Also we provide an interpretation for the parameters and other statistics estimated from these models. Recently, Shaw (1988) and Grogger and Carson (1991) have proposed Normal, Poisson and Negative Binomial regression models for the analysis of truncated samples of count data.

STANDARD COUNT DATA ESTIMATORS

A number of discrete probability distributions satisfy our requirement of generating nonnegative integers. The simplest one is the one-parameter Poisson distribution. Since many other possible count data distributions represent generalizations of the Poisson, we take it up first. The basic Poisson model can be written as

$$P(X_i = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where there are i = 1, 2... n observations, Xi is the ith observation on the count variable of interest, x = 0, 1,2, ... are the possible values The generalization of the Poisson distribution which is often used to model such overdispersed counts is the negative binomial probability distribution (Johnson and Kotz, 1969; Hausman, Hall and Griliches, 1984; Cameron and Trivedi, 1986). This probability distribution can be written as:

$$P(X_i = x) = F_{NB}(x) = \frac{\Gamma(x+\frac{1}{\alpha})}{\Gamma(x+1)\Gamma(\frac{1}{\alpha})}(\alpha\lambda_i)^{\lambda}[1+\alpha\lambda_i]^{-(x+\frac{1}{\alpha})}$$

where > 0 is a nuisance parameter to be estimated along with . The negative binomial can be derived from a Poisson distribution in which the i are distributed as a gamma random variable. For this reason the negative binomial is sometimes referred to a compound distribution. Other distributions for the i are possible but more difficult to estimate (Hinde, 1982).The first two moments of the negative binomial distribution are given by

$$E(X_i \mid Y_i) = Exp(y_i\beta)$$

and

$$\text{var}(X_i \mid Y_i) = \lambda_i(1+\alpha\lambda_i)$$

So that| var($X_i \mid Y_i$) is greater than $E(X_i \mid Y_i)$

Both the Poisson and Negative Binomial (for given a>0) distributions are members of the linear exponential family of distributions. Quasi-maximum-likelihood methods will therefore generally provide consistent estimates of the correctly specified conditional mean (Wedderburn, 1974; McCullagh, 1983; Gourieroux, Monfort, and Trognon, 1984a) when applied to a random sample from the entire underlying population of interest.

Both the Poisson and Negative Binomial (for given a>0) distributions are members of the linear exponential family of distributions. Quasi-maximum-likelihood methods will therefore generally provide consistent estimates of the correctly specified conditional mean (Wedderburn, 1974; McCullagh, 1983; Gourieroux, Monfort, and Trognon, 1984a) when applied to a random sample from the entire underlying population of interest.

3. TRUNCATED REGRESSION MODELS FOR COUNT DATA

The common statistical structure of truncated estimators follows from the fundamental probability relationship

$$\Pr ob(A/B) = \frac{\Pr ob(A \cap B)}{\Pr ob(B)}$$

In our case, the expression Prob (AB) represents the probability of observing some Xi while Prob (B) represents the probability of being at or above the truncation limit. The term Prob (A|B) represents the probability of observing Xi, given that it exceeds the truncation point. In terms of probability distribution functions i.e. Pr*ob*(*A*|*B*) can be written as

where fk(xi) is the truncated (above k) probability function, f(xi) is the probability function, and F(k) is the distribution function evaluated at k. To derive the maximum-likelihood estimator, a suitable discrete probability function is applied with the relationship for conditional probabilities. We now do this for the Poisson and negative binomial models presented in the previous section, concentrating on the case of k = 0, since this is the case most likely to be encountered in practice. For the Poisson probability function, a model for counts truncated on the left at the value k = 0 can be posited as

$$P(X_i = x \mid X_i > 0) = \frac{e^{-\lambda}\lambda^x}{x!}[1-F_p(0)]^{-1} = \frac{\lambda_i^x}{(\exp(\lambda_y)-1)x!}$$

where x now takes only positive integer values larger than 0.

The truncated probability function differs from the standard probability function by the factor [1 – FP (0)] -1. Since Fp(0) < 1, multiplication of the standard probabilities by [1 - Fp(0)] -1 inflates them, accounting for the unobserved zeros.

AN EMPIRICAL APPLICATION TO DMF COUNT DATA

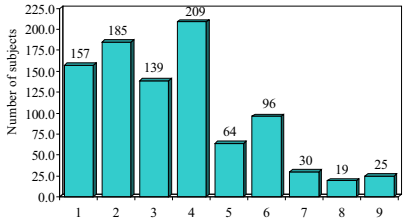
In this section, we illustrate the use of various count data models to estimate the dental caries by DMF index data on a random sample of 1760 individuals were collected. 924 (52.50%) individuals have dental caries and 836 (47.50 %) individuals do not have dental caries. This means that, the DMF count data is overdispersed and excess zeros in the distribution and considering the part of the count data, where the DMF>0.

The parameter estimates of zero truncated count data from standard and truncated forms of the Poisson and negative binomial models by a large number of possible predictor variables are included. These independent covariates are age as a continuous, socio-economic status (as a continuous), family size (as a continuous), dietary habits (1, if non-vegetarian), frequency of sweet consumption (=1 if greater than 2, 0=otherwise), oral hygiene habits (tooth brush/finger=1, datum/others=2), frequency of brushing (=1 if less than twice a day, 2=otherwise), mouth rinsing habit (yes=1, no=0), smoking habit (yes=1, no=0) and chewing habit (yes=1, no=0). The variables used are those suggested by Zero Truncated regression models and generalized linear models to DMF count data. The following tables presents, the parameter estimates by standard Poisson, Standard Negative Binomial, Zero truncated Poisson and Zero truncated Negative Binomial models and the results are explained in the preceeeding section.

COMPARISON OF MODEL RESULTS

In this section, we illustrate the use of various count models to estimate the DMF counts on a random sample of 1760 individuals by considering 924 individuals with DMF>0. It means that, an individual must have at least one count. The maximum value of DMF counts is 9 respectively. The nature of distribution of truncated DMF counts of 924 individuals is presented in the Figure 1. The mean DMF counts is 3.53 2.01.

Figure 1: The nature of distribution of DMF positive counts (n=924)



COMPARISON OF NORMAL, POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS TO ZERO TRUNCATED DMFT COUNT DATA

Table 1 presents parameter estimates and their standard errors by Normal, Poisson and Negative Binomial regression models to zero truncated DMF count data. Out of the total of 11 covariates, seven covariates are significantly associated with truncated DMF counts. In which, The DMF counts are significantly positively associated with all covariates namely, age (in years), dietary habits, smoking habit and alcohol habit. These significant covariates exhibited positive regression coefficients, indicating that they are likely to increase the higher DMF count. However, the covariates like Socio-economic status, Frequency of brushing and Mouth rinsing habit are significantly negatively associated with DMF counts. But other covariates are not significantly associated with DMF counts (p>0.05). Table 1 also shows that, the regression coefficients and their standard errors estimated from Poisson and Negative Binomial models and they are quite smaller and similar in magnitude compared to regression coefficients of Normal regression model.

In the additional model fitting information, the log likelihood for the Normal regression, Poisson regression and Negative Binomial models are -1805.2145, -1774.8566 and -1774.8563 respectively. Based on the log likelihood, the Poisson regression and Negative Binomial regression models fits better compared to Normal regression model. The Poisson and Negative Binomial model log likelihood values are similar and indicate a reasonably better fit than the Normal regression model to the truncated DMF counts.

COMPARISON OF NORMAL, ZERO TRUNCATED POISSON AND ZERO TRUNCATED NEGATIVE BINOMIAL REGRESSION MODELS TO DMFT COUNT DATA

Table 1 presents parameter estimates and their standard errors by Normal, Poisson and Negative Binomial regression models to zero truncated DMF count data. Out of the total of 11 covariates, seven covariates are significantly associated with truncated DMF counts. In which, the DMF counts are significantly positively associated with all covariates namely, age (in years), dietary habits, smoking habit and alcohol habit. These significant covariates exhibited positive regression coefficients, indicating that they are likely to increase the higher DMF count. However, the covariates like Socio-economic status, Frequency of brushing and Mouth rinsing habit are significantly negatively associated with DMF counts. But other covariates are not significantly associated with DMF counts (p>0.05). Table 2 also shows

that, the regression coefficients and their standard errors estimated from truncated Poisson and truncated Negative Binomial models and they are quite similar in magnitude compared to regression coefficients of Poisson and Negative Binomial regression models.

In the additional model fitting information, the log likelihood for the zero truncated Poisson regression and zero truncated Negative Binomial models are -1728.7061 and -1728.7054 respectively. Based on the log likelihood, the both zero truncated Poisson and zero truncated Negative Binomial models fits better compared to Normal regression, Poisson and Negative Binomial regression models. Based on these findings, the zero truncated Poisson and zero truncated Negative Binomial models which seem to be equally better fit to the truncated DMF counts than the Normal, Poisson and Negative Binomial regression models.

Table 1: Parameter estimates from Normal, Poisson and Negative Binomial regression models to zero truncated DMF count data

DMF count	Normal regression model	Poisson regression model	Negative Binomial regression model						
	Estimate	SE of estimate	z-value	Estimate	SE of estimate	z-value	Estimate	SE of estimate	z-value
Constant	1.2906	0.9798	0.1880	0.4301	0.3399	0.2060	0.4301	0.3399	0.2060
Education	-0.4463	0.0536	0.0001*	-0.1168	0.0156	0.0001*	-0.1168	0.0156	0.0001*
Age (in years)	0.0121	0.0040	0.0020*	0.0033	0.0012	0.0040*	0.0033	0.0012	0.0040*
Socio-economic status	-0.0380	0.0466	0.4150	-0.0012	0.0139	0.9330	-0.0012	0.0139	0.9330
Dietary habits	0.3446	0.0918	0.0001*	0.0905	0.0270	0.0010*	0.0905	0.0270	0.0010*
Frequency of sweet consumption	0.0024	0.0445	0.9570	-0.0009	0.0138	0.9480	-0.0009	0.0138	0.9480
Oral hygiene habits	0.1396	0.0835	0.0950	0.0311	0.0242	0.2000	0.0311	0.0242	0.2000
Frequency of brushing	-0.3101	0.1676	0.0650	-0.1246	0.0522	0.0170*	-0.1246	0.0522	0.0170*
Mouth rinsing habit	-0.2648	0.0414	0.0001*	-0.0696	0.0122	0.0001*	-0.0696	0.0122	0.0001*
Smoking habit	0.7692	0.1472	0.0001*	0.2138	0.0461	0.0001*	0.2138	0.0461	0.0001*
Chewing habit	0.1937	0.2786	0.4870	0.1038	0.0998	0.2980	0.1038	0.0998	0.2980
Alcohol habit	1.4584	0.2719	0.0001*	0.4704	0.0975	0.0001*	0.4704	0.0975	0.0001*
Log likelihood				-1774.8566			-1774.8563		
R2	0.2480			0.0673			0.0652		

Significant at 5% level of significance (p<0.05)

Table 2: Parameter estimates from Normal, truncated Poisson and Negative Binomial regression models to zero truncated DMF count data

DMFT count	Truncated Poisson regression model	Truncated Negative Binomial regression model				
	Estimate	SE of estimate	p-value	Estimate	SE of estimate	p-value
Constant	0.1606	0.3926	0.6820	0.1606	0.3926	0.6820
Education	-0.1314	0.0166	0.0001*	-0.1314	0.0166	0.0001*
Age (in years)	0.0038	0.0012	0.0020*	0.0038	0.0012	0.0020*
Socio-economic status	0.0017	0.0148	0.9070	0.0017	0.0148	0.9070
Dietary habits	0.1018	0.0287	0.0001*	0.1018	0.0287	0.0001*
Frequency of sweet consumption	-0.0016	0.0149	0.9130	-0.0016	0.0149	0.9130
Oral hygiene habits	0.0324	0.0256	0.2050	0.0324	0.0256	0.2050
Frequency of brushing	-0.1563	0.0569	0.0060*	-0.1563	0.0569	0.0060*
Mouth rinsing habit	-0.0781	0.0130	0.0001*	-0.0781	0.0130	0.0001*
Smoking habit	0.2438	0.0499	0.0001*	0.2438	0.0499	0.0001*
Chewing habit	0.1512	0.1170	0.1960	0.1512	0.1170	0.1960
Alcohol habit	0.5759	0.1145	0.0001*	0.5759	0.1145	0.0001*
Log likelihood	-1728.7061	-1728.7054				
R2	0.0782	0.0683				

Significant at 5% level of significance (p<0.05)

COMPARISON OF MODELS

In this section, we analyze and compare the fitting performances of five regression count models to the truncated DMF data set in terms of log likelihood procedure.

Table 3: Model fitting information of Normal, Poisson, Negative Binomial, Zero Truncated Poisson and Zero Truncated Negative Binomial regression models to DMF count data.

Regression models	Information	DMF Data
Normal	Log likelihood	-1805.2145
Standard Poisson	Log likelihood	-1774.8566
Standard Negative Binomial	Log likelihood	-1774.8563
Zero-Truncated Poisson	Log likelihood	-1728.7061
Zero Truncated Negative Binomial	Log likelihood	-1728.7054

The table 3 presents the model fitting information on Normal, Poisson, Negative Binomial, zero Truncated Poisson and zero Truncated Negative Binomial regression models to truncated DMFT count data. The log likelihood of the Normal, Poisson, Negative Binomial, zero truncated Poisson and zero truncated Negative Binomial regression models to DMFT count data are -1805.2145, -1774.8566, -1774.8563, -1728.7061 and -1728.7054 respectively. Based on these log-likelihoods, zero truncated poisson binomial and zero truncated negative binomial are equivalently better fits to the truncated DMFT count data followed by the standard negative binomial is a next better fit, third best fit is standard poisson followed by Normal regression model. The standard Poisson and Negative Binomial models remarkably not better fits to the truncated DMF count data. However among the truncated models, there exists a noticeable improvement in accounting for extra-variability: the truncated Negative Binomial regression model have to be preferred by far, likely due to its capability to catch overdispersed truncated DMF count data.

DISCUSSIONS AND CONCLUSIONS

Our motivation in using the truncated count data models presented here was the feeling that there are a large number of potential applications for such models. Currently we are using them to predict the dental caries by DMF index among a population of individuals with at least one DMF>0. Applications from the fields of dental epidemiology and medicine also seem natural.

Our results showed that regression coefficients can be substantially biased when overdispersion is not accounted for and the mean number of counts is relatively low; that is, when many zeros would be expected in a non-truncated sample. The results of the analysis pointed out the serious consequences for inference that may arise when overdispersion

is neglected, even when mean counts are large. Some of the specification tests of the Poisson versus the negative binomial in the non truncated case can be easily extended to cover the truncated case. Given that only one tail of the distribution is observable, special attention should be paid to deriving tests with particular power against overdispersion in the upper tail. The finite sample performance of all these tests would need to be assessed. Our experience suggests that the Poisson mean-variance equality restriction is rarely appropriate.

The situation is further complicated by a mixing of two types of zeros: those who individuals without dental caries (DMF =0). If the observed zeros are a mixture of the two types, the researcher may be better off simply using the appropriate truncated count model on the positive counts, as inclusion of a large number of structural zeros will severely bias the regression coefficients for the caries (DMF) process of interest.

The application addressed in this chapter involves the estimation of normal Poisson, Negative Binomial, zero truncated Poisson and zero truncated Negative Binomial models to predict the dental caries by DMF indices independently. Since count data frequently exhibit over-dispersion even after truncation at zero, an obvious methodology is to use a model that can accommodate over-dispersion and zero-truncation. We also consider the zero truncated Poisson and zero truncated Negative Binomial models for over-dispersion situation. The zero truncated Poisson and zero truncated Negative Binomial models are alternative to the Poisson and Negative Binomial models respectively when there is a situation of zero truncation.

For this reason, we apply the zero truncated Poisson and zero truncated Negative Binomial models over Poisson and Negative Binomial models for modeling over-dispersed DMF count data with positive integers. Based on results, for DMF count data, the zero truncated negative binomial model is good fit over the standard negative binomial, the truncated Poisson model is better fit over the standard Poisson model. But, the zero truncated negative binomial models is good statistical fit compared to zero truncated Poisson for modeling the DMF count data.

REFERENCES

1. Mullahy, J. (1986): “Specification and testing of some modified count data models”, Journal of Econometrics, 33, 341-365.
2. Winkelmann, R., and K.F. Zimmermann (1993): “Count Data Models for Demographic Data.”Mathematical Population Studies 4: 205-221.
3. Gurmu, S., and P. Trivedi (1992): “Overdispersion Tests for Truncated Poisson Regression Models.” Journal of Econometrics 54: 347-370.
4. Gurmu, S., and P. Trivedi (1996): “Excess Zeros in Count Models for Recreational Trips”. Journal of Business and Economic Statistics 14: 469-477.
5. David, F.N. & Johnson, N.L. (1952): “The truncated Poisson”. Biometrics 8, 275 - 285.
6. Plackett, R.L. (1953): “The truncated Poisson distributions”. Biometrics 9, 485 - 488.
7. Johnson, N.L., Kotz, S. & Kemp, A.W. (1992): Univariate discrete distributions. Second Ed. John Wiley & Sons, Inc.
8. Shaw, D. (1988): “One-site samples- regression problems of non-negative integers”, truncation, and endogenous stratification. J. Econometrics 37, 211 - 223.
9. McCullagh, P. & Nelder, J.A. (1989): “Generalized linear models”. Second Ed. London: Chapman and Hall
10. McDonald, J., and R. Moffitt (1980): 'The uses of Tobit analysis', Review of Economics and Statistics, 62, 318-321.
11. Gurmu, S (1991): “Tests for Detecting Overdispersion in the Positive Poisson Regression”. Journal of Business and Economic Statistics 9: 215-222.
12. Grogger, J.T., and R.T. Carson (1991): “Models for Truncated Counts.” Journal of Applied Econometrics 6: 225-238.
13. Xie, T. & Aicken, M. (1997): “A truncated Poisson regression model with applications to occurrence of adenomatous polyps”. Statistics in Medicine 16, 1845 - 1857
14. Cameron, A.C. & Trivedi, P. (1998): “Regression analysis of count data”. Cambridge University Press, Cambridge.
15. Gourieroux, C., A. Monfort and A. Trognon (1984b): “Pseudo maximum likelihood methods: applications to Poisson models”, Econometrica, 52, 701-720.
16. Hausman, J., B. Hall and Z. Griliches (1984): “'Econometric models for count data with an application to the patents-R&D relationship”, Econometrica, 52, 909-938.
17. Lee, L. F. (1986): “Specification tests for Poisson regression models”, International Economic Review, 27, 689-706.
18. Cameron, A. C., and P. K. Trivedi (1986): “Econometric models based on count data: comparisons and application of some estimators and tests”, Journal of Applied Econometrics, 1, 29-53.
19. Cameron, A. C., P. K. Trivedi, F. Milne and J. Piggott (1988): “A micro-econometric model of the demand for health care and health insurance in Australia”, Review of Economic Studies, 55, 85-106.
20. Portney, P. R., and J. Mullahy (1986): “Urban air quality and acute respiratory illness”, Journal of Urban Economics, 20, 21 -38.
21. Grogger, J. (1990): “The deterrent effect of capital punishment: an analysis of daily homicide counts”, Journal of the American Statistical Association, 85, 295-303.
22. Johnson, N. L., and S. Kotz (1969): “Discrete Distributions”, Wiley, New York.
23. Grogger J. T., and Carson R. T. (1991): “Models for Truncated Counts”. Journal of Econometrics, 6, 225-238.