International Journal of Medical and Health Sciences

Journal Home Page: http://www.ijmhs.net ISSN:2277-4505

Original Article

Factors Affecting Number of Somatic Symptoms among Pregnant Women: A Multimodal Regression Approach

Adhin Bhaskar[*1], Kandavel Thennarasu[2], Mariamma Philip[3], Supraja Thirumalai Ananthanpillai[4], Geetha Desai[5], Prabha Satish Chandra[6]

[1]Scientist B, Dept. of Statistics, ICMR-NIRT, #1, Mayor Sathiyamoorthy road, Chetpet, Chennai-600 031,

[2]Professor, [3]Associate professor, Dept. of Biostatistics, NIMHANS, Hosur Road, Bengaluru -29, PhD

[4]Student, Dept. of Psychiatric Social Work, NIMHANS, Hosur Road, Bengaluru -29,

[5&6]Professor, Dept. of Psychiatry, NIMHANS, Hosur Road, Bnegaluru-29.

INTRODUCTION

Background: Multimodality, the occurrence of multiple modes in the distribution of the data is a less-discussed issue in the area of count regression models. **Objectives**: To assess the suitability of Hermite regression model to predict the factors affecting the number of symptoms among pregnant women. **Methods:** This study is focused on comparing the performance of various count regression models when the dependent variable is both overdispersed and multimodal. The data is obtained from a community based prospective study of anxiety, depression during pregnancy and its relationship to pregnancy outcomes. Poisson, negative binomial, Hermite and generalized Hermite regression models were fitted to find the relationship between the variables. The models were compared using fit indices along with the estimates and standard errors. Distribution of randomized quantile residuals was also assessed to determine the goodness of fit of the models. **Results**: Based on the values of fit indices and tests, Hermite regression was chosen as the best to establish the relationship between the response variable, number of somatic symptoms and the predictors. The model identified parity, stress and depression as the factors affecting number of somatic symptoms in pregnant women. **Conclusions**: The Poisson and negative binomial model may not accommodate multimodality as they are framed based on unimodal distributions. The Hermite regression approach is an ideal approach for count data, as it can handle both overdispersion as well as multimodality.

KEY WORDS: Count data, Hermite regression, multimodality, negative binomial regression, Poisson regression.

INTRODUCTION

Researchers in biological, behavioural science, epidemiology, etc. are likely to gather information on different variables as a continuum rather than a categorical one. Often, the research questions on outcome involve count variables1,2,3,4,5. Essentially, a count variable reflects the number of occurrences of an event, generally over a period of time and it always assumes it's values to be either a zero or a positive integer (i.e. 0, 1, 2, 3, etc.).

Poisson and Negative Binomial (NB) are the two regression approaches commonly used for count data. The NB regression is ideal when the count data is observed with high variability called over dispersion. Over dispersion may occur due to several reasons[1] such as effect of omitted covariates, dependency between the measurements, etc.

Apart from overdispersion count variables may also exhibit Multimodality, described as the occurrence of multiple peaks or local maxima[2]. Hermite regression is a generalized form of Poisson regression that can handle both overdispersion and multimodality[3].

This study is aimed to compare and assess the suitability of Hermite regression approach with the other count regression models to find out the variables associated with number of somatic symptoms experienced by pregnant women.

MATERIALS AND METHODS

Poisson regression

The bench mark model for count data, Poisson regression fits the data under Generalized Linear Modeling (GLM) framework assuming, a Poisson distribution for the response variable[4]. The choice of Poisson distribution restricts the response variable to take only nonnegative values which most suits a count variable. The probability mass function of Poisson distribution is expressed as,

$$p(Y=y)=\frac{e^{-\lambda}\lambda^{y}}{y!} \quad ; \begin{matrix} y=0,1,2,\ldots \\ E(y)=V(y)=\lambda \end{matrix} \qquad (1)$$

The Poisson regression models the mean count as function of covariates,

$$E(y_i\,|x_i\,)=\lambda_i=\exp(x_i'\,\beta)=e^{\sum_{j=1}^{k}\beta_j x_{ji}} \quad ; \text{i=1,2,3…,n; j=1,2,3,…,k} \qquad (2)$$

The exponential link function restricts the expected count to be positive always. The Poisson regression can be written in linear from by taking $ln(\lambda_i)$,

$$\ln(\lambda_i)=x_i{}'\beta=\sum_{j=1}^{k}\beta_j\,x_{ij} \qquad (3)$$

Inheriting the properties of Poisson distribution, The Poisson regression model assumes that the conditional mean is equal to the conditional variance i.e., $E(y_i|x_i)=V(y_i|x_i)$. This assumption is called equidispersion. The adherence of Poisson regression to the equidispersion assumption limits it's application in real life situations as the count data is usually overdispersed, i.e., $E(y_i|x_i)<V(y_i|x_i)$. Overdispersion causes the Poisson regression to underestimate the standard errors of regression coefficients, which results in overestimating the significance of the predictors[5].

Test of overdispersion

Equidispersion, the key assumption of Poisson regression can be tested by employing a test for overdispersion[6]. The test can be performed by fitting the Poisson regression. H_0: $\alpha = 0$ (equidispersion) versus H_1:$\alpha > 0$ can be performed after fitting an auxiliary ordinary least square regression on the dependent variable generated using, $\frac{(y_i-\hat{\lambda}_i)^2-y_i}{\hat{\lambda}_i}$. The regression equation can be written as,

$$\frac{(y_i-\hat{\lambda}_i)^2-y_i}{\hat{\lambda}_i}=\alpha\hat{\lambda}_i+e_i \qquad (4)$$

where, e_i is an error term and $\hat{\lambda}_i=\exp(x_i{}'\beta)$ is the expected value obtained from the Poisson fit. A positive α indicates the overdispersion.

Negative binomial regression

The Negative Binomial (NB) regression is an alternative approach to model count data when the Poisson regression fails due to overdispersion. The NB regression can handle the overdispersion occurring due to unobserved heterogeneity and dependency between the events[7]. In NB regression, the unobserved heterogeneity is introduced through a random error function, e_i. The NB model can be expressed by modifying equation (2) as,

$$E(y_i\,|x_i\,)=\tilde{\lambda}_i=\exp(x_i'\,\beta+e_i)$$

$$\tilde{\lambda}_i=\exp(x_i'\,\beta)\exp(e_i)=\exp(x_i'\,\beta)\delta_i \qquad (5)$$

where, $exp(e_i)=\delta_i$ is assumed to follow gamma distribution with mean 1 and variance $\alpha={}^{1}/_{\nu_i}$. Hence, the NB distribution can be expressed as a mixture of Poisson and gamma distribution. The NB probability mass function can be written as

$$p(y_i\,|x_i\,)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{y_i} \quad ; y_i=0,\,1,\,2,\ldots \qquad (6)$$

The NB distribution has the mean and variance as λ_i and $\lambda_i+\alpha{\lambda_i}^2$ respectively. NB regression is flexible to account the over dispersion in count data. However, it may not be compatible for multimodal count data. Regression models based on Hermite distribution are capable to account over dispersion as well as multimodality[4].

Hermite regression

Hermite distribution is essentially an extended form of Poisson distribution. The distribution got the name because the probability and factorial moments can be derived from a modified Hermite polynomial equation. The Hermite distribution is derived in several ways. It is a special case of Poisson binomial distribution. It can also be derived as a sum of two correlated Poisson variables or as a sum of an ordinary Poisson variable and an independent 'doublet' Poisson variable[8]. The probability mass function can be written as,

$$p(y_i=k)=\exp\{-(a_1+a_2\}\sum_{j=0}^{[k/2]}\frac{{a_1}^{k-2j}{a_2}^{2j}}{(k-2j)!\,j!} \quad ; k=0,1,2,\ldots. \qquad (7)$$

where, [k/2] is the integer part of k/2 and a_1 and a_2 are the parameters of the distribution which are non-negative. The mean and variance of the distribution is defined as a_1+2a_2 and a_1+4a_2.

Gupta and Jain introduced the generalized version of the Hermite distribution by explaining the distribution of $Y=Y_1+mY_m$, where Y_1 and Y_m are two independent Poisson random variables with mean a_1 and a_m respectively and m is a positive integer[9]. The probability mass function of generalized Hermite distribution is defined as,

$$p(y=k)=\begin{cases} e^{-a_1-a_m} & ;\ k=0 \\ e^{-a_1-a_m}\sum_{j=0}^{[k/m]}\frac{(a_m)^j(a_1)^{k-mj}}{j!(k-mj)!} & ;\ k=1,2,3\ldots \end{cases} \qquad (8)$$

The mean and variance of the generalized Hermite distribution is defined as a_1+ma_m and $a_1+m^2a_m$ respectively. The generalized Hermite distribution converges to a Hermite distribution when m=2.
In hermit regression the distribution of the response variable is assumed to follow a Hermite distribution. Similar to Poisson regression, Gile has related the mean function to an exponential of linear combination of covariates i.e., $\mu_i = \exp(x_i'\beta)$ [9]. The parameters of the Hermite regression can be obtained using maximum likelihood estimation technique[10].

Test for multimodality
Haritgan's *dip* test can be used to test the significance of multimodality in a variable[11]. The test is performed by comparing the empirical distribution function and a unimodal distribution function. Let $F_n(x)$ be the empirical distribution function an $H(x)$ be a the closest unimodal distribution function respect to the empirical distribution,

The DIP statistic = $sup|F_n(x)-H(x)|$ (9)

The test uses uniform distribution as the reference distribution to compute the p value. A significant *dip* test indicates the presence of more than one mode.

Analysis of residuals
Pearson and deviance residual are the most common standardized residuals for count regression model. As count variable are observed in a very limited range, Pearson and deviance residual produces overlapping residuals which might distort the information in residual plots[12]. Hence, in this study Randomized Quantile Residuals (RQR) are used to evaluate the goodness of fit of the models. The RQR are obtained by inverting the fitted distribution function and calculating the corresponding standard normal quantile for each observation. RQR for continuous distribution function is defined as,

$q_i = \Phi^{-1}(F(y_i|\widehat{\theta_i}))$ (10)

where, Φ^{-1} is the quantile function of a standard normal distribution. For discrete variables, a uniform random component is introduced in order to avoid the overlapping of residual values. The RQR for discrete distribution function is defined as,

$q_i = \Phi^{-1}(u_i)$ (11)

where, u_i is random value from a uniform distribution with interval $[F(y_i - 1|\widehat{\theta_i}), F(y_i|\widehat{\theta_i})]$. The RQR follows a standard normal distribution if the model is correctly specified[13]. Hence the test of normality of RQR can be used as a technique to assess the goodness of fit of the model. Due to the randomness involved in the calculation, RQR were computed 1000 times for each model and the average of summary statistics and p value of normality test were taken.

Data
The data for comparing the performance of the models was obtained from the PRAMMS - Prospective Assessment of Maternal Mental Health Study, a community based prospective study on anxiety, depression and stress during pregnancy and its relationship to pregnancy outcomes. The data was collected from pregnant women attending antenatal clinic at the government referral hospital, Bengaluru, India during 2014 and November 2015. Informed consent was obtained from each participants and study was approved by the institute ethical committee. The number of somatic symptoms experienced by women during the first trimester, a count variable was considered as the response variable for fitting the models. The variable was composed of 25 symptoms such as headache, palpitations, weakness of mind, lack of sleep, nausea, etc.

RESULTS

The final dataset used for analysis contained information on 490 pregnant women. The response variable, the number of somatic symptoms was observed with a mean count of 5.50 and standard deviation 3.29 ranging from 0 to 25 (fig. 1). The average age of the subjects was 22.99 (s. d. = 3.40) years. The Socio Economic Status (SES) of 80.41% (n = 394) women was above poverty line (table 1). Among the pregnant women 78 (15.92%) had history of abortion and 219 (44.69%) were primiparous.

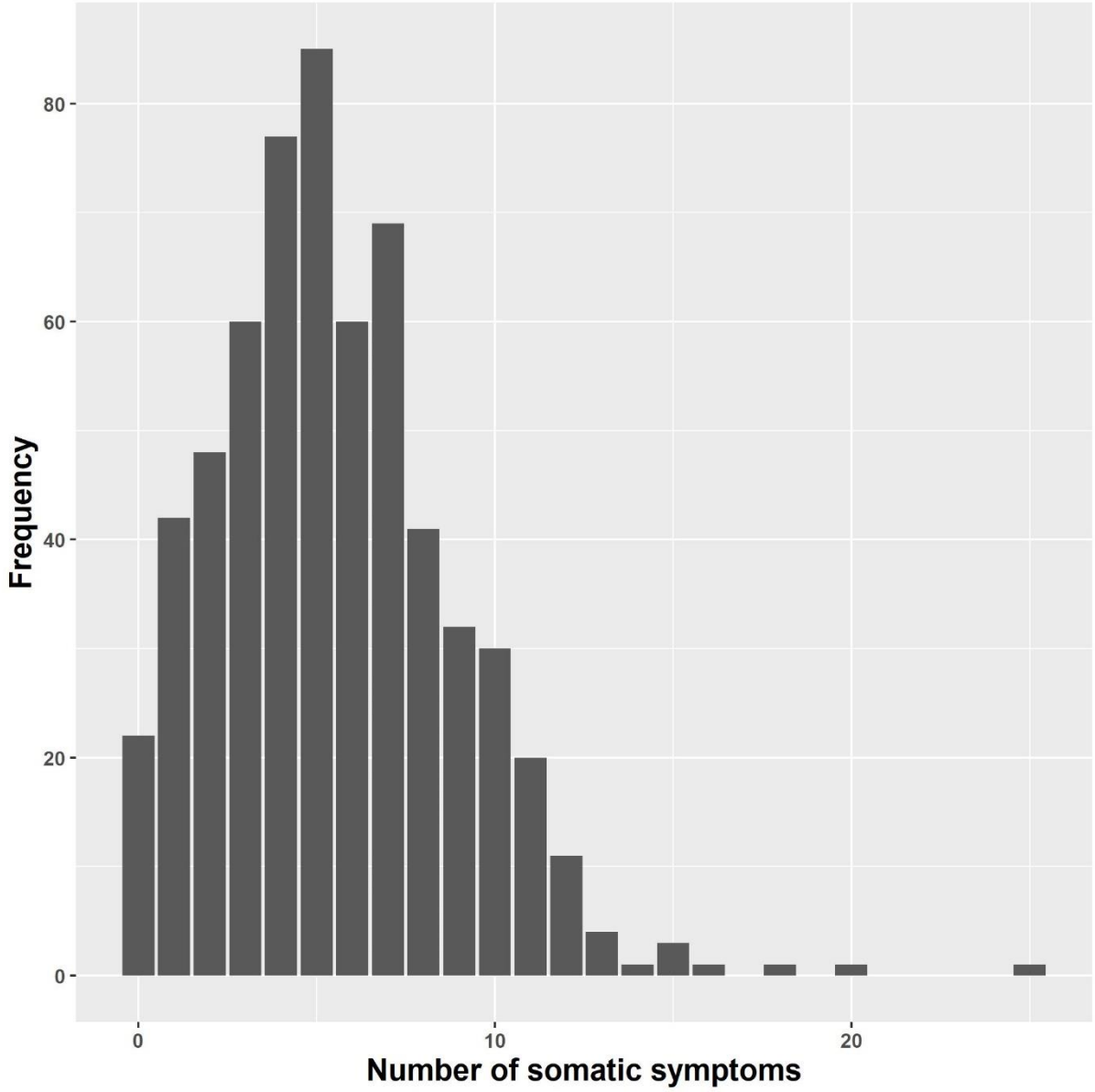
Table 1: Socio demographic and clinical profile of participants

Variable		**n (%)**
Age in years*		23.06 (3.44)
Education	Above secondary	137 (27. 96)
	Up to secondary	353 (72.04)
SES	Medium or upper	183 (37.35)
	Low	211 (43.06)
	BPL†	96 (19.59)
Parity	Two or more	29 (5.92)
	One	242 (49.39)
	Zero	219 (44.69)
Abortion History	Yes	78 (15.92)
	No	412 (84.08)

Depression	Yes	34 (6.94)
	No	456 (93.06)
Stress score*		12.79 (3.44)
Mid arm circumference (c.m.)		25.07 (3.23)

* Mean(s.d.), † Below Poverty Line

Figure 1: Distribution of the number of somatic symptoms



Univariate analyses were carried out to select the predictors to construct the multivariable models. Poisson regression was fitted to the data as first attempt of modelling count data (table 2). The equidispersion assumption was tested using an auxiliary regression based test assuming overdispersion under the alternative hypothesis. The significance of test (t statistic = 6.223, p value = <0.001) indicated that the data was overdispersed. Hence, NB regression was fitted in order to account the extra Poisson variability in the data.

Due to over dispersion, the standard errors of the estimates of the Poisson regression were observed to be underestimated (table 2). The smaller standard errors inflated the values of z statistic, which resulted in smaller p values. However, the significant factors predating number of somatic symptoms were same for both models. The history of abortion was found to be marginally significant (p value = 0.067) in Poisson regression. Even in the presence of over dispersion the estimated parameters of the Poisson model were similar to the estimates of NB regression for most of the variables.

Table 2: Poisson and negative binomial model fit to number of somatic symptoms

Variable		Poisson			Negative binomial		
		Estimate	S. E.	p value	Estimate	S. E.	p value
Age in years		-0.006	0.007	0.338	-0.006	0.009	0.509
Education	Above secondary*	-0.067	0.045	0.131	-0.079	0.06	0.191
Socio Economic Status	Medium or upper	-0.058	0.055	0.695	-0.049	0.075	0.509
	Low	0.021	0.053	0.293	0.031	0.072	0.673
	BPL	Reference	-		Reference	-	
Parity	two or more	-0.113	0.086	0.186	-0.141	0.12	0.237
	One	-0.261	0.044	<0.001	-0.264	0.06	<0.001
	Zero	Reference	-		Reference	-	
Abortion History	Yes	0.096	0.053	0.069	0.102	0.073	0.161
Stress score		0.035	0.005	<0.001	0.036	0.008	<0.001
Mid arm circumference (c.m.)		0.009	0.006	0.137	0.01	0.009	0.265
Depression	Yes	0.242	0.076	0.001	0.244	0.111	0.028
Dispersion parameter		-	-		6.54	1.000	

*Ref. category-up to secondary, BPL-Below Poverty Line,

The response variable was observed to have multiple peaks (figure 1). Hence, Hartigan's *dip* test was performed suspecting multimodality in the distribution of response variable. The significance of the test (test statistic = 0.077, p value = <0.001) indicated that the population from where the response variable was taken is at least bimodal. Hence, suspecting a poor fit of NB regression for the multimodal response variable, Hermite and generalized Hermite regression model were attempted (table 3).
Even in the presence of multimodality, significant variables identified by NB model were similar to the significant predictors obtained from Hermite regression fit. The estimated coefficients and standard errors of Hermite and generalized Hermite regressions were similar (table 3). The models identified parity, stress and depression as the significant predictors of the response variable.

Likelihood based fit indices based fit such as Akaike Information Citerion (AIC) and Bayesian Information Criterion (BIC) were used to compare the relative fit of the models (table 4). Models with smaller fit indices indicates the better fit of the model for the data. The lower value of the fit indices (AIC = 2458 and BIC = 2515) indicates a superior fit of the Hermite regression model over the others, supporting the detection of multimodality and overdispersion in the response variable. Generalized Hermite regression was found to be inferior to Hermite regression model. AIC indicated that both multimodal regression models had better fit over the Poisson and NB models. The values of BIC were similar for NB and the multimodal models. Whereas, Poison model had substantially larger BIC (2606) values. To check the absolute goodness of fit of the models, RQR were computed and assessed the distributional properties (table 4). Only Hermite and generalized Hermite models had standard normal RQR, indicated good fit of the models to the data. The RQR of Poisson regression was normal but the standard deviation (1.330) was larger. Whereas the NB regression failed to have normal residuals (p = 0.014). Due to the incidence of overdispersion and multimodality Poisson regression gave the poor fit among the fitted models.

Table 3: Hermite and generalized Hermite model fit to number of somatic symptoms

Variable		**Hermite**			**Generalized Hermite**		
		Estimate	**S. E.**	**p value**	**Estimate**	**S. E.**	**p value**
Age in years		-0.003	0.008	0.729	-0.004	0.009	0.675
Education*	Above secondary	-0.065	0.057	0.258	-0.067	0.061	0.271
Socio Economic Status$	Medium or upper	-0.074	0.070	0.294	-0.070	0.075	0.355
	Low	-0.011	0.068	0.866	0.001	0.073	0.991
	BPL	Reference	-		Reference	-	
Parity	2 or more	-0.160	0.112	0.152	-0.127	0.118	0.284
	1	-0.274	0.057	<0.001	-0.273	0.061	<0.001
	Zero	Reference	-		Reference	-	

Abortion History	Yes	0.089	0.068	0.189	0.128	0.073	0.076
Stress score		0.035	0.007	<0.001	0.031	0.007	<0.001
Mid arm circumference (c.m.)		0.011	0.008	0.176	0.007	0.009	0.364
Depression	Yes	0.247	0.098	0.011	0.278	0.105	0.008

*Ref. category-up to secondary, BPL-Below Poverty Line

Table 4: Comparison of model fit

Model fit	Fit index		Randomized Quantile Residuals*			
	AIC	BIC	Mean	S. D.	Test of normality	
					W	p value
Poisson	2560	2606	0.057	1.330	0.996	0.350
Negative binomial	2465	2516	-0.007	1.024	0.855	0.014
Hermite	2458	2515	-0.006	1.037	0.997	0.464
Generalized Hermite	2460	2517	0.011	0.953	0.996	0.411

*average of 1000 residuals

Similar to the Poisson regression model, the Hermite regression and the generalized Hermite regression were modelled using a log link function. As the coefficients of count regression models are interpreted generally as Incidence Rate Ratio [7], we are interpreting the exponential of coefficients of the Hermite regression in terms of Incidence Rate Ratio (IRR).

The Hermite regression identified parity, stress score and history of depression as the significant predictors of number of somatic symptoms experienced by pregnant women (table 5). Women with depression experienced 26% more symptoms compared to women who were not depressed. Women who had given birth once (parity = 1) experienced 24% less events compared to women who have not given birth at least once. However, more than one birth (parity = two or more) did not have any significant effect on the number of somatic symptoms. Though the history of abortion was not significant, women with history of abortion experienced 11% more symptoms compared to women who did not have history of abortion.

Table 5: Hermite regression model fit

Variable		**IRR ((95% C. I.)**
Age in years		0.995 (0.980, 1.011)
Education	Above secondary	0.934(0.836, 1.045)
	Up to secondary	-
Socio Economic Status	Medium or upper	0.937(0.817, 1.075)
	Low	1.015(0.888, 1.160)
	BPL	-
Parity	Two or more	0.889 (0.716, 1.103)
	One	0.763(0.682, 0.853)
	Zero	-
Abortion History	Yes	1.110 (0.971, 1.268)
	No	-
Stress score		1.037 (1.023, 1.051)
Mid arm circumference (c.m.)		1.009(0.993, 1.025)
Depression	Yes	1.261 (1.041, 1.528)
	No	-

BPL- Below Poverty Line

DISCUSSION

The count variables are often observed in biomedical research as an outcome of interest. Count variable shows varying properties such as, overdispersion, zero inflation, zero truncation at various situations. Multimodality is also a concern while modelling count variables. However, only handful of literatures have discussed the multimodality in count data modelling. We have found with the illustrative example that the Hermite regression approach is appropriate when the data is tested multimodal and overdispersed. The Hermite regression was emerged as the most plausible fitting model among the four fitted models. The better fit of Hermite regression over generalized Hermite regression was might be due to the less severe overdispersion and multimodality.

The results were also quite similar between the Poisson and NB regression models as the overdispersion was not so severe. The multimodal model regression models might show substantial difference in estimates and standard errors from the standard count models if fitted to a data with high multimodality. The hermit regression is easy to use as its modeling is similar to other basic count regression models. Moreover the estimates can be obtained using maximum likelihood method of estimation. However the prevalence of multimodality in count variable is not really known.

ACKNOWLEDGMENTS

We acknowledge the funding for the PRAMMS project by the Indian Council of Medical Research, Grant Number 7/7/01PSRH/12-RCH.

REFERENCES

1. Xekalaki E. On the distribution theory of over-dispersion. J Stat Distrib Appl 2014;1:1–22.
2. Pfingsthorn M, Birk A. Simultaneous localization and mapping with multimodal probability distributions. Int J Rob Res 2013;32:143–171.
3. Giles DE. Hermite regression analysis of multi-modal count data. Econ Bull 2010;30:2936–2945.
4. Nelder JA, Wedderburn RWM. Generalized linear models. J R Statist Soc A 1972;13517213:370–384.
5. Palmer A, Losilla JM, Vives J, et al. Overdispersion in the Poisson regression model. Meth Eur J Res Meth Behav Soc Sci 2007;3:89–99.
6.Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the Poisson model. J Econom 1990;46:347–364.
7.Long JS. Regression models for categorical and limited dependent variables. Thousand Oaks: Sage Publications; 1997.
8.Kemp BYCD, Kemp AW. Some properties of the ' Hermite ' distribution. Biometrika 1965;52:381–394.
9.Gupta RP, JAIN GC. A generalized Hermite distribution and its properties. Siam J Appl Math 1974;27:359–363.
10.Moriña D, Higueras M, Puig P, et al. Generalized {H}ermite distribution modelling with the {R} package Hermite . R Journal 2015;7:263–274.
11.Hartigan JA, Hartigan PM. The Dip test of unimodality. Ann Stat 1985;13:70–84.
12.Feng C, Sadeghpour A, Li L. Randomized quantile residuals : An omnibus model diagnostic tool with unified reference distribution. Cornel Univ Libr 2017;1–33.
13.Dunn PK, Smyth GK. Randomized Quantile Residuals. J Comput Graph Stat 1996;5:236–244

*Corresponding author: Adhin Bhaskar
E-Mail: adhinb6001@gmail.com