Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online): 2229-5666
Vol. 12, Issue 2 (2019): 51-60

MODELS FOR ANALYZING OVER-DISPERSED HURDLE NEGATIVE BINOMIAL REGRESSION MODEL: AN APPLICATION TO MANUFACTURED CIGARETTE USE

***Sujan Rudra and Soma Chowdhury Biswas**
Department of Statistics, University of Chittagong, Chittagong, Bangladesh
E Mail:sujan1rudra@gmail.com; soma.stat@cu.ac.bd

Received March 13, 2019
Modified May 08, 2019
Accepted October 13, 2019

Abstract

Our main aim is to identify the factors that influence the use of manufactured cigarettes among tobacco users especially those whose age is above fifteen. Among the tobacco users, a large portion of adult does not take manufactured cigarettes but take other tobacco. As a result, we need to construct a model that can handle the existence of excess zero counts and the over-dispersed phenomenon. Motivated by these facts, in this paper, we propose to apply the Hurdle Negative Binomial (HNB) regression model to discover the relationships between uses of manufactured cigarettes and social factors. The data were found to have excess zeros (35%); moreover, the variance is 47.122, which is much higher than its mean 5.933. With excess zeros and high variability of non-zero outcomes, the HNB model was found to be better fitted.

Key Words: Manufactured Cigarettes, Tobacco, Over Dispersed, Hurdle Negative Binomial, Bangladesh.

1. Introduction

Hurdle model is used in the presence of excess zero as well as for over dispersion in the model. There are two processes in this model: the first process is governed by a point binomial distribution that generates structural zeros and the later one is governed by a truncated Poisson distribution (PD) that generates positive counts. As a result, it is possible to study the positive counts in the count part as well as the zero counts the zero part separately, which is the major advantage over other models (Especially, zero-inflated models). Hurdle model is a “finite mixture generated by combining the zeros generated by one density with the zeros and positives generated by a second zero-truncated density separately . . .” (Cameron and Trivedi, 1998). Hurdle model is consistent and easy to interpret; easy to implement (Potts and Elith, 2006) when the model has zero-inflation along with over dispersion. The main feature isthat the probability of hurdle increases with the increase of covariates and quickly decrease as covariates are decreased.

Dawit Sekata (2015) found that Hurdle Poisson (HP) regression model is the best model among all count data models. Both Zero-inflated negative binomial (ZINB) and Hurdle Negative Binomial (HNB) have an equal preference and have similar model fits. But it is noted that HNB provides nicer interpretation (Zeileis, Kleiber and

Jackman, 2008) and data characterized by excess zeros along with high variability in the non-zero outcomes.

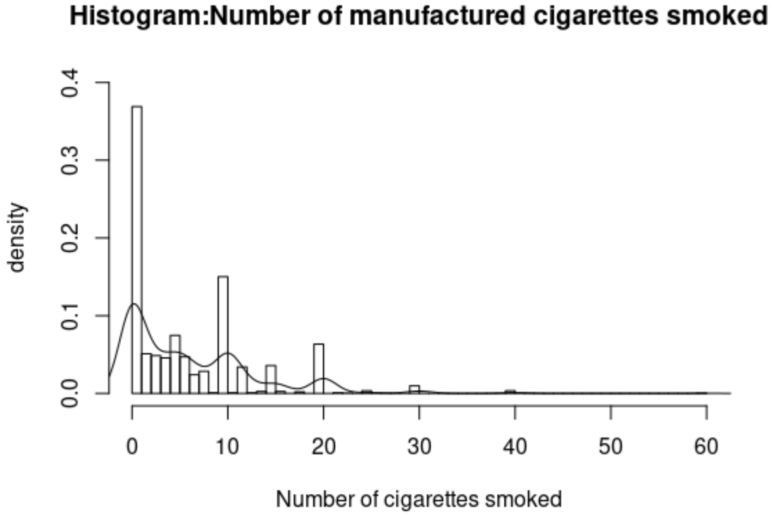
The paper is organized as follows. The data source is discussed in section 2. Variable selection is provided in section 3. Models comparison and parameter estimate of HNB is demonstrated in section 4. Finally, a result interpretation and conclusion elucidate in section 5.

2. Source of Data

"GATS (2009) "is conducted by the Bangladesh Bureau of Statistics (BBS) and data is interviewed from adults aged 15 years and over. The response rate is 93.6%.11200 households were selected, 10751 were screened and finally 9626 were interviewed. We draw a subset of size 2038 in such a way that all are take tobacco especially manufactured cigarettes by using a standard global protocol.

3. Variables

According to my research selected response variable "On average, how many of the following products do you currently smoke each day: A. Manufactured cigarettes?" in such a way that "0" represent "Don't take Manufactured cigarettes but Take others Tobacco" and 1,2,3…60 represent number of smoking "Manufactured cigarettes" per day. We take a subset of 2038 such respondent. 8 predictor variables have been included for these analyses which are respondent's age, sex, type of place of residence, respondent's education, wealth index, news media, health warning and addictive. The selected variable has a significant relation at the 5% level with the response variable.



From the Figure, it is observed that the number of cigarettes smoked at the point "0" is inflated.

4. Methodology

In order to modeling and analysis “Number of Manufactured cigarettes currently smokes each day” data in Bangladesh, R statistical software version R i386 3.2.3 was used. The Generalized Linear Model (GLM) procedure with PD specified using the log link function. The Hurdle Poisson regression and Hurdle Negative Binomial regression were also used to overcome the excess number of zeros and over-dispersion in the data as well as to study the positive count part and zero part separately. We use packages: MASS, PSCL, datasets and GGPLOT2 for our analysis. Vuong test is also used to compare the models.

5. Modeling and Analysis

Using the “Number of Manufactured cigarettes currently smoke each day” as dependent variable our proposed model is

$$\text{Log}(\mu_i) = \beta_0 + \beta_{1\text{j}}X_{1\text{j}} + \beta_{2\text{j}}X_{2\text{j}} + \beta_{3\text{j}}X_{3\text{j}} + \beta_{4\text{j}}X_{4\text{j}} + \beta_{5\text{j}}X_{5\text{j}} + \beta_{6\text{j}}X_{6\text{j}} + \beta_{7\text{j}}X_{7\text{j}} + \beta_{8\text{j}}X_{8\text{j}}$$

Where, the variables used in the model are defined as: X1 = Respondent’s age, X_2 = Sex, X_3 = Residence, X_4 = Respondent’s education, X_5 = Wealth index, X_6 = News media, X_7 = Health warnings, X_8 = Addictive and suffix j indicates the category of the variables.

Using the Poisson regression model for the Number of Manufactured cigarettes currently smoke per day. The AIC of the above-fitted model is 18039, residual deviance 12942 on 2021 degree of freedom (df) the following chi-square with 1 df. Dispersion parameter 6.404 indicates that the model is over dispersed with p-value 0.000 and model is significance at 5% level .Using Negative Binomial (NB) regression AIC of above model is 11015,residual deviance 2290.4 on 2021 df following Chi-square with 1 df and dispersion parameter is 1.23. NB reduce over-dispersion problem but standard error of estimated parameters are being suddenly increased. It is noted that all variables are remain fixed in applying all through the models.

SL	Models Name	LogLikelihood	df	AIC	BIC
1	Poisson	-9002.6	17	18039.2	18134.7
2	NB	-5490.5	17	11015	11110.6
3	ZIP	-6150.9	34	12369.8	12560.9
4	ZINB	-5000.7	34	10069.4	10260.4
5	HP	-6150.9	34	12369.7	12560.8
6	HNB	-4997.9	34	10063.8	10254.9

Table 01: Comparison of different models in count data with AIC and BIC

On the basis of data set for n=2038 we have seen that the magnitudes of the AIC and BIC values obtained from the data are in the following order of models (Table 01): HNB< ZINB< NB< HP< ZIP< Poisson.

5.1Comparison the models using Vuong test

Our all competing models are misspecified and non-nested too. The comparison among the Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB), Hurdle Poisson (HP) and Hurdle Negative Binomial (HNB) regression models using Vuongtest and(Desmarais, Bruce A. and Jeffrey J. Harden (2013)) comparisonis given below:

Model	Vuong	Hypothesis	p-value	Comment
	z-statistic			
Poisson Model Vs ZIP Model	-21.523	H_0: model1 = model2 H_A: model2 > model1	0.000	ZIP Model is better
ZIP Model Vs HP model	-0.7214	H_0: model1 = model2 H_A: model2 > model1	0.000	HP model is better
HP model Vs NB model	-5.657	H_0: model1 = model2 H_A: model2 > model1	0.000	NB model is better
NB model Vs ZINB model	-19.769	H_0: model1 = model2 H_A: model2 > model1	0.000	ZINB model is better
ZINB model Vs HNB model	-2.874	H_0: model1 = model2 H_A: model2 > model1	0.002	HNB model is appropriate

Table 02: Comparison of different models using Vuong test

From above model selection criteria AIC, BIC and Vuong test HNB is more appropriate than all other models. So we use the HNB model as a best in our consideration. Our findings demonstrate that 1.4% respondent below 20 years, 19.8% respondents in age interval 21-30 years and 30.7% respondents in age interval 31-40 years and 48.1% respondents above 40 years. Among the respondents 96.8% are male and rest of are female. Among male 33.2% don't take manufactured cigarettes and largely 66.8% take manufactured cigarettes. Respondent lives in urban 84.1% take manufactured cigarettes and who lives in rural 47.9% take manufactured cigarettes. Overall 47.3% are living in urban and rests are in rural. In the group of illiterate 47.0%, primary educated 27.3%, secondary educated 19.8% and above secondary 5.9%.Among illiterate 50.2% take manufactured cigarettes. Among primary education level 69.4% take manufactured cigarettes, it also found that the percentage of respondent who take manufactured cigarettes is increased with increasing of wealth index where poor 49.5%, middle 18.4% and high 32.1%.Having high wealth index respondent take more manufactured cigarettes then others. Among high wealth index 88.4% take manufactured cigarettes. It is clear that 9.9% respondent notice information in news media about effect of manufactured cigarettes, 32.0% don't notice and 58.0% respondent don't concern about newspaper's discouragement. Surprisingly who notice information in news media about the dangers of use 84.2% are takes manufactured cigarettes. Among respondent 78.4% notice any health warnings on cigarette packages, 17.9% don't notice any health warnings and 3.7% don't concern about health. Miraculously among whom notice health warnings 74.6% take manufactured cigarettes. It is found that 91.6% strongly believe cigarettes are addictive, 7.4% don't believe cigarettes are addictive and 1.0% respondent don't know that cigarettes are addictive or not. It is most wondering that among who believe cigarettes are addictive 65.5% of them takes cigarettes.

5.2 Parameter Estimate of HNB Regression Model

Independent variable		Estimate	Std. Error	z value	Pr (>|z|)	Lower Limit	Upper Limit	Odds ratio
Intercept		0.725	0.425	1.704	0.088 .	-0.232	1.476	2.065
Respon-dent's age	BELOW 20	…	…	…	…	…	…	1.00
	21-30	0.156	0.054	2.849	0.004 **	0.033	0.653	1.169
	31-40	0.002	0.053	0.035	0.972	0.192	0.807	1.002
	Above 40	-0.34	0.158	-2.17	0.030 *	0.039	0.651	0.709
Sex	Female	…	…	…	…	…	…	1.00
	Male	0.571	0.292	1.953	0.050 .	-0.002	1.144	1.771
Residence	Rural	…	…	…	…	…	…	1.00
	Urban	0.169	0.041	4.069	0.00 4**	0.088	0.252	1.185
Level of Education	Illiterate	…	…	…	…	…	…	1.00
	Primary	0.262	0.089	2.931	0.003 **	-0.23	-0.023	1.299
	Secondary	0.135	0.084	1.613	0.107	-0.26	-0.028	1.145
	Above	0.114	0.08	1.425	0.154	-0.43	-0.087	1.121
Wealth Index	Low	…	…	…	…	…	…	1.00
	Middle	-0.02	0.052	-0.421	0.674	-0.13	0.088	0.978

	High	-0.04	0.058	-0.794	0.427	-0.08	0.124	0.955
News media	Don't Concern	…	…	…	…	…	…	1.00
	No	0.077	0.047	1.659	0.097 .	-0.01	0.17	1.081
	Yes	0.093	0.069	1.349	0.177	-0.04	0.229	1.098
Health warnings	Don't Concern	…	…	…	…	…	…	1.00
	No	0.067	0.166	0.408	0.683	-0.25	0.394	1.070
	Yes	0.126	0.154	0.817	0.414	-0.17	0.429	1.135
Addictive	Don't Know	…	…	…	…	…	…	1.00
	No	0.418	0.288	1.448	0.148	-0.14	0.984	1.519
	Yes	0.431	0.279	1.543	0.123	-0.11	0.979	1.539
	Log(Theta)	0.945	0.058	16.214	.000***	0.83	1.06	

Table 03: Count Model coefficients (truncated negbin with log link) with 95% CI

Independent variable		Estimates	Std. Error	Z value	Pr (>|z|)	Lower limit	Upper limit	Odds ratio
Intercept		-1.164	0.857	-1.358	0.174	-4.295	-0.764	0.312
Respon-dent's age	Below 20	…	…	…	…	…	…	1.00
	21-30	-0.377	0.170	-2.218	0.026*	-1.787	0.537	0.686
	31-40	-0.782	0.160	-4.898	.0009***	-2.153	0.150	0.458
	Above 40	0.625	0.593	1.054	0.292	-2.553	-0.261	1.868
Sex	Female	…	…	…		…	…	1.00
	Male	1.348	0.420	3.211	.00132**	0.525	2.171	3.851
Residence	Rural	…	…	…	…	…	…	1.00
	Urban	1.556	0.124	12.547	.0002***	1.313	1.800	4.742
Level of Education	Illiterate	…	…	…	…	…	…	1.00
	Primary	-0.744	0.442	-1.682	0.092	-0.156	0.391	0.475
	Secondary	-0.626	0.437	-1.431	0.152	-0.003	0.770	0.535
	Above	-0.360	0.441	-0.816	0.414	-0.123	1.610	0.698
Wealth	Low	…	…	…	…	…	…	1.00

Index	Middle	-1.246	0.167	-7.447	.0009***	0.009	0.586	0.288
	High	-0.949	0.188	-5.057	.0004***	0.918	1.574	0.387
News media	Don't Concern	…	…	…	…	…	…	1.00
	No	0.581	0.140	4.151	.000 ***	0.307	0.856	1.788
	Yes	0.484	0.245	1.979	0.0477*	0.005	0.964	1.623
Health warnings	Don't Concern	…	…	…	…	…	…	1.00
	No	-0.069	0.322	-0.215	0.830	-0.700	0.562	0.933
	Yes	1.32	0.308	4.306	0.000***	0.722	1.930	3.766
Addictive	Don't Know	…	…	…	…	…	…	1.00
	No	0.439	0.571	0.769	0.442	-0.681	1.559	1.552
	Yes	0.68	0.536	1.282	0.200	-0.364	1.737	1.988

Table 04: Zero hurdle model coefficients (binomial with logit link) with 95% CI

Significance codes: 0 %('***'), 0.1 %('**'), 1 %('*'), 5 %('.')

The AIC of the above-fitted model is 10063.83 and theta = 2.5753. The first part of the Table 01 contains truncated Negative Binomial regression coefficient for each of the variables. A second part corresponds to the inflation model which includes logit coefficients for predicting excess zeros.

6. Result Interpretation

Respondent's age: This is the estimated HNB Regression coefficient's odds ratio comparing the age group below 20, considering other variables are held constant in the model. People who take manufactured cigarettes 1.169 times more for age group 21-30, 1.008 times higher for the age group 31-40 and 0.709 times higher for the age above 40 compared to the age group below 20.

But people who take tobacco but do not take manufactured cigarettes 0.686 times higher for age group 21-30, 0.458 times higher for the age group 31-40 and 1.868 times higher for the age above 40 compared to age group below 20.

Sex: People who take manufactured cigarettes 1.7703 times more for males compared to females. However, people who take tobacco but do not take manufactured cigarettes 3.851 times higher for males compared to females.

Residence: People who take manufactured cigarettes 1.185 times higher comparing urban to rural. People who take others tobacco are 4.471 times higher for urban compared to rural, remaining all other variables constant in the model.

Level of Education: This is the estimated coefficient's odds ratio comparing illiterate. People who take manufactured cigarettes 1.2997 times higher for primary education level 1.145 times higher for secondary education level, 1.121 times for above secondary education level compared to illiterate. On the contrary, People who take others tobacco are 0.4754 times higher for primary education level 0.5348 times higher for secondary education level, 0.6976 times for above secondary education level compared to illiterate

Wealth Index: People who take manufactured cigarette 0.979 times higher owing middle wealth index, 0.955 times higher owing high wealth index compared to low wealth index. But people who take tobacco but do not take manufactured cigarettes 0.288 times higher owing middle wealth index, 0.3872 times higher owing high wealth index compared to low wealth index.

News media: This is the estimated HNB Regression coefficient comparing the group who don't concern notice information in news media about the effect of manufactured cigarettes. The difference in the logs of expected counts would be expected to increase by 1.081 times for who don't notice compared to the group who don't concern. And1.097 times more for who do notice compared to the group who don't concern. On the other side people who take tobacco but do not take manufactured cigarettes 1.788 times higher for who don't notice and 1.623 times higher for who do notice compared to the group who don't concern.

Health warnings: This is the estimated HNB Regression coefficient comparing the group who don't concern notice any health warnings on manufactured cigarettes package. People who take manufactured cigarettes 1.0702 times more for who don't notice, 1.1345 times more for who notice health warnings compared to the group who don't concern. Surprisingly, the most interesting fact is that people who take other tobacco rather than manufactured cigarettes are 0.933 times more for who don't notice, 3.766 times more for who notice health warnings compared to the group who don't concern.

Addictive: This is the estimated HNB Regression coefficient comparing the group who don't know cigarettes are addictive. People who take manufactured cigarettes 1.519 times higher for who don't believe cigarettes are addictive, 1.539 times for who believe cigarettes are addictive compared to the group who don't know. And other tobacco users who do not take cigarettes 1.552 times higher for who don't believe cigarettes are addictive, 1.988 times for who believe cigarettes are addictive compared to the group who don't know.

7. Conclusion

Based on the cross-tabulation with chi-square most significant variable were used in modeling. The Poison regression model is used in regression count variable as the 'Number of Manufactured cigarettes currently smoke each day' on significant predictor variables. This model showed over-dispersion problem which violates the equality assumption of mean and variance of the Poisson regression model. To overcome this problem we used the Negative Binomial regression model. But in this model, we also found the over-dispersion problem as well as excess zero in the count data. It is clear that the real data i.e. Number of Manufactured cigarettes currently smoke each day is over-dispersed as well as zero-inflated. To handle the over-dispersion with inflation of zeros, then we used ZIP regression model and ZINB Regression model. Another advanced model for over-dispersed and zero-inflated count data is Hurdle regression model. We applied Hurdle Negative Binomial (HNB) regression model to handle the over-dispersion problem with inflation of zeros as well as to study positive count part and zero count part of the count data separately.

Finally, it is assessed that the people who had aged 21-30 are 1.169 times higher to habituate to take manufactured cigarettes; but, among other tobacco takers aged above 40 is 1.868 times higher than others. Males who used to take manufactured cigarettes 1.77 fold more than females. Paradoxically, the male also 3.85 fold more to take other tobacco rather than a manufactured cigarette. Urban people are 4.742 times higher to lean on other tobacco. Ironically, people who cognized that, other tobacco might have more health haphazard, were 3.766 times higher than those who did not know. Surprisingly, tobacco users who knew that, tobacco is addictive in nature; manufactured cigarettes users 1.539 times and other tobacco users 1.988 times higher in taking than those who did not know.

Acknowledgement

We would like to thanks the Editor and referees whose constructive criticism led us to develop the presentation and maintain the quality of the paper. We also thank the Department of Statistics, University of Chittagong to give their almost all opportunity for conducting the research.

References

1. Aguirre-Torres, V. and Gallant, A. R. (1983). The null and non-null asymptotic distribution of the Cox test for multivariate nonlinear regression: alternatives and a new distribution-free Cox test, Journal of Econometrics, 21, p. 5-33.
2. Arndt, C. (2004). Information Measures: Information and its Description in Science and Engineering, Springer, ISBN 978-3-540-40855-0.
3. Baetschmann, G. and Winkelmann, R. (2014). A Dynamic Hurdle Model for Zero- Inflated Count Data: With an Application to Health Care Utilization, University of Zurich, Working Paper Series, Paper No. 151, JEL Classification: C25, I10.
4. Bandopadhyaya, A. (1994). an estimation of the hazard rate of firms under Chapter 11 protection,ReviewofEconomicsandStatistics, 76(2),p. 346-350.
5. Bohara, A.K. and Krieg R.G. (1996). A Poisson hurdle model of migration frequency, The Journal of Regional Analysis and Policy,26(1), p. 37-45.
6. Burnham, K.P. and Anderson, D.R. (1998). Model Selection and Inference, Springer – Verlag, New York.
7. Cameron, A. C. and Trivedi, P. K. (1998). Regression Analysis of Count Data, Cambridge: Cambridge University Press.
8. Cameron, A.C. and Windmeijer, F. A.G. (1996). R-squared measures for count data regression models with applications to health care utilization, Journal of Business and Economic Statistics, 14(2), p. 209-220.
9. Cameron, A.C. and P.K. Trivedi (1986).Econometric models based on count data: comparisons and applications of some estimators and tests, Journal of Applied Econometrics, 1, p. 29-54.
10. Cameron, A.C. and Trivedi, P.K. (1990). Regression based tests for over-dispersion in the Poisson model, Journal of Econometrics, 46(3), p. 347-364.

11. Cameron, A.C. and Trivedi, P.K. (1993). Tests of independence in parametric models with applications and illustrations, Journal of Business and Economic Statistics, 11, p. 29-43.
12. Cameron, A.C., Trivedi, P.K., Milne, F. and Piggott, J. (1988). A micro econometric model of the demand for health insurance and health care in Australia', Review of Economic Studies, 55(1), p. 85-106.
13. Chandhiok, N., Dhillon, B. S., Kambo, I. and Saxena N.C. (2006). Determinants of antenatal care utilization in rural areas of India: A cross-sectional study from 28 districts (An ICMR task force study), J. Obstet. Gynecol. India, 56(1), p. 47-52.
14. Chao, A. and Shen, Tsung-Jen (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample, Environ. Ecol. Stat., 10, p. 429–443.
15. Chen, X.., Hong, H. and Shum, M. (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models, Journal of Econometrics, 141, p. 109-140.
16. Chernoff, H. (1954). On the distribution of the likelihood ratio, Annals of Mathematical Statistics, 25, p. 573-578.
17. Chipeta, M.G., Ngwira, B.M., Simoonga, C. and Kazembe L. N. (2014). Zero adjusted models with applications to analyzing helminths count data, BMC Research Notes, 7(1), p. 856.
18. Cover, T. M. and Thomas, J. A. (2006). Elements of Information Theory, 2nd Edition,Wiley Interscience, ISBN 0-471-24195-4.
19. Dawit Sekata (2015). Modeling the Number of Antenatal Care Service Visits Among Pregnant Women in Rural Ethiopia: Zero Inflated and Hurdle Model Specifications, International Journal of Healthcare Sciences, 3(1), p. 332-355.
20. Desmarais, Bruce A. and Jeffrey J. Harden (2013). Testing for Zero-Inflation in Count Models: Bias Correction for the Vuong Test, The Stata Journal, 13(4), p. 810–835.
21. Horvitz, D.G., and D. J. Thompson(1952). A generalization of sampling without replacement from a finite universe, J. Am. Stat. Assoc. 47, p. 663-685.
22. Linden, A. and Mantyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data, Ecology, 92(7), p. 1414–1421.
23. Liu,W.S. and Cela, J. (2008). Count Data Models in SAS®, Statistics and Data Analysis, SAS Global Forum, Statistics and Data Analysis, paper 371.
24. MacDonald, R. B. (2014, March 24). Overdispersion and Poisson regression, J Quant Criminol,p. 269-284.
25. Makalic, D. F. (November 22,2008). Model selection Tutorial#1:Akaike's Information Criterion, In Model selection with AIC, Melbourne.
26. Mullahy, J. (1986). Specification and testing of some modified count data models, Journal of Econometrics, 33, p. 341-365.
27. Simkhada, B., Teijlingen E. R., Porter M. and Simkhada P. (2008). Factor affecting the utilization of antenatal care in developing countries: systematic review of the literature, Journal of Advanced Nursing, 61(3), p. 244-260.
28. Winkelmann, R. (1994). Count Data Models: Econometric Theory and an Application to Labor Mobility, Springer-Verlag, Berlin.
29. Winkelmann, R. and Zimmermann (1991). A new approach for modeling economic count data, Economics Letters, 37, p.139-143.