



Performance of Generalized Poisson Regression Model and Negative Binomial Regression Model in case of Over-dispersion Count Data

Farhana Sadia

Lecturer, Institute of Statistical Research and Training (ISRT)

University of Dhaka

Dhaka-1000, BANGLADESH

Abstract: This paper represents the comparison between Negative Binomial Regression model and Generalized Poisson Regression model for over-dispersion count data. For this comparison, we used BDHS 2007 data in where the response variable is the total children ever born which is a count data. When the response variable is count, then Poisson Regression Model as a Generalized Linear Model is widely and popularly used to analyze such type of response variable and Poisson Regression model gives better result than the usual regression model for analyzing count data. In this paper, the descriptive statistics of the total children ever born data exhibit the presence of over-dispersion in the data set. Since the total children ever born data used in this study exhibit over-dispersion, we can use Negative Binomial Regression Model and Generalized Poisson Regression Model. These two models have statistical advantages over standard Poisson regression model and are suitable for analysis of count data that exhibit either over-dispersion or under-dispersion.

Keywords: Negative Binomial Regression (NBR); Generalized Poisson Regression (GPR); Over-dispersion; Count data; Children ever born.

I. Introduction

Bangladesh is a small progressive country but its population is the 8th largest in the world. It is most densely populated country in the world. Population growth refers to change in the size of a population which can be either positive or negative over time depending on the balance of births and deaths. In Bangladesh, every year a large number of children born for various causes or factors. To find out the nature of the determinants of population growth or fertility, very little sophisticated and statistically sound methods have been performed. In recent years, Poisson type regression models have been used to model count response variable affected by one or more covariates. **King (1989) and Winkelmann and Zimmermann (1994)** developed the generalized even count models based on the Poisson, negative binomial, and the binomial distributions.

To understand the nature of socio-economic and demographic factors related to population growth, a generalized linear modeling approach has been used in this current study [5]. But there is more variability around the model's fitted values than is consistent with a Poisson formulation, that is, Over dispersion. To correct this problem, we have used Negative Binomial Regression and Generalized Poisson Regression which are also the special case of Generalized Linear Model. But the main concern of this study is to compare Negative Binomial Regression and Generalized Poisson Regression for explaining over dispersion. That means this study wants to explain that which model is better among Negative Binomial Regression and Generalized Poisson Regression for describing over dispersion of count data. **Femoye, Wulu and Singh (2001)** noted that the Poisson regression model is not appropriate when a data set exhibit over-dispersion, a condition where the variance is more than the mean [1].

Bangladesh demographic and health survey data (BDHS-2007) have been used for this study [7]. A number of candidate factors have been considered. The number of children ever born is taken as the response variable. The selected demographic and socio-economic differentials like mother's educational level, types of place of residence, age at first marriage, current working status, husband's occupation, division, having radio etc. are taken as predictor variables for this study.

II. Methods

A random variable Y is said to have a Poisson distribution with parameter θ if it takes integer values

$z = 0, 1, 2, \dots$ with probability [6]

$$\Pr\{Z = z\} = \frac{e^{-\theta} \theta^z}{z!}, \quad z = 0, 1, 2, \dots, \infty$$

The parameter θ on the **Poisson regression** model may be written as a log-linear model:

$$\text{Log} \theta_i = x_i \beta_i$$

The Poisson Regression is to be fitted to the mean number of children, θ can be expressed as below for our ten independent variables (considering X_1 = types of place of residence, X_2 = has radio, X_3 = has television, X_4 = religion, X_5 = educational attainment, X_6 = wealth index, X_7 = use of contraceptive method, X_8 = age at first marriage, X_9 = partner's educational level, X_{10} = current working status).

$$\text{Log}(\theta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}$$

The parameters β can be estimated by the maximum likelihood estimated method. Under the Poisson, the mean θ_i , is assumed to be constant or homogeneous within the classes. However, by defining a specific distribution for θ_i , heterogeneity within the classes is now allowed. For example, by assuming θ_i to be a gamma with mean $E(\theta_i) = \mu_i$ and variance $\text{Var}(\theta_i) = \mu_i^2 \nu_i^{-1}$ and $Z_i | \theta_i$ to be a Poisson with conditional mean $E(Z_i | \theta_i) = \theta_i$, it can be shown that the marginal distribution of Z_i follows a **Negative binomial distribution** with probability density function [3]

$$\Pr[Z_i = z_i] = \int \Pr(Z_i = z_i | \theta_i) f(\theta_i) d\theta_i = \frac{\Gamma(z_i + \nu_i)}{\Gamma(z_i + 1) \Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i} \left(\frac{\mu_i}{\nu_i + \mu_i}\right)^{z_i}$$

Where the mean is $E(Z_i) = \mu_i$ and the variance is $\text{Var}(Z_i) = \mu_i + \mu_i^2 \nu_i^{-1}$. The likelihood for **Negative Binomial regression** model may be written as:

$$\ell(\beta, \alpha) = \sum_i \left\{ \sum_{r=1}^{z_i-1} \log(1 + \alpha^r) \right\} - z_i \log(\alpha) - \log(z_i!) + z_i \log(\alpha \mu_i) - (z_i + \alpha^{-1}) \log(1 + \alpha \mu_i)$$

Therefore, the maximum likelihood estimates, $(\hat{\beta}, \hat{\alpha})$, may be obtained by maximizing $\ell(\beta, \alpha)$ with respect to β and α .

Generalized poisson regression model uses response variable Z_i which is count. Here, Z_i is defined as the total number of children ever born. The probability function of Z_i is given by [1]

$$f_i(z_i, \theta_i, \alpha) = \left(\frac{\theta_i}{1 + \alpha \theta_i}\right)^{z_i} \frac{(1 + \alpha z_i)^{z_i-1}}{z_i!} \exp\left(-\frac{\theta_i(1 + \alpha z_i)}{1 + \alpha \theta_i}\right), \quad z_i = 0, 1, \dots, \infty$$

$$\text{And } \theta_i = \theta_i(x_i) = \exp(x_i \beta)$$

Where, x_i is a (k-1) dimensional vector of explanatory variables and β is a k-dimensional vector of regression parameters. The mean and variance of z_i are given by [1]

$$E(Z_i | X_i) = \theta_i \text{ and } V(Z_i | X_i) = \theta_i(1 + \alpha \theta_i)^2$$

To estimate (β, α) in the GPR model, we take the partial derivatives of the log-likelihood function of the GPR model [2].

$$\text{Log}L(\alpha, \beta; z_i) = \sum_{i=1} \left\{ z_i \log\left(\frac{\theta_i}{1 + \alpha \theta_i}\right) + (z_i - 1) \log(1 + \alpha z_i) - \frac{\theta_i(1 + \alpha z_i)}{1 + \alpha \theta_i} - \log(z_i!) \right\}$$

III. Results

Table-1: Poisson regression verses Negative binomial regression

Variable	Poisson Regression			Negative Binomial Regression		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Intercept	2.6197544	0.0671172	0.0000	2.221103	0.048956	0.0000
Residence	-0.0422345	0.0138268	0.0011	-0.042171	0.015368	0.00607
Has Radio	0.0122917	0.0142422	0.1941	0.011397	0.015826	0.47143
Has Television	-0.0165784	0.0176640	0.1740	-0.016441	0.019603	0.40164
Religion	0.0673319	0.0212768	0.0008	0.065269	0.023463	0.00541
Wealth Index	-0.0015129	0.0180022	0.4665	-0.002454	0.020029	0.90249
Age At First Marriage	-0.0643101	0.0025339	0.0000	-0.064814	0.002773	0.0000
Partner's Education	-0.0447968	0.0145081	0.0010	-0.045759	0.016058	0.00438
Working Status	-0.0868602	0.0129356	0.0000	-0.084298	0.014411	0.0000
Educational level	-0.3870177	0.0131581	0.0000	-0.387970	0.014724	0.0000
Current Contraceptive Method	0.0943425	0.0118879	0.0000	0.101512	0.013250	0.0000

Table-2: Goodness-of-fit tests between Poisson regression and Negative binomial regression

Goodness-of-fit Test Measures	Poisson Regression	Negative Binomial Regression
Deviance	73880.54	13119
AIC	26212.62	39351

Table-3: Poisson regression verses Generalized poisson regression

Variable	Poisson Regression			Generalized Poisson Regression		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Intercept	2.6197544	0.0671172	0.0000	2.6208475	0.0200179	0.0000
Residence	-0.0422345	0.0138268	0.0011	-0.0421906	0.0041446	0.0000
Has Radio	0.0122917	0.0142422	0.1941	0.0118439	0.0042669	0.0027
Has Television	-0.0165784	0.0176640	0.1740	-0.0165244	0.0052983	0.0009
Religion	0.0673319	0.0212768	0.0008	0.0664167	0.0063337	0.0000
Wealth Index	-0.0015129	0.0180022	0.4665	-0.0019201	0.0054040	0.3599
Age At First Marriage	-0.0643101	0.0025339	0.0000	-0.0645548	0.0007451	0.0000
Partner's Education	-0.0447968	0.0145081	0.0010	-0.0452272	0.0043242	0.0000
Working Status	-0.0868602	0.0129356	0.0000	-0.0856665	0.0038892	0.0000
Educational level	-0.3870177	0.0131581	0.0000	-0.3874412	0.0039705	0.0000
Current Contraceptive Method	0.0943425	0.0118879	0.0000	0.0976496	0.0035808	0.0000

Table-4: Goodness-of-fit tests between Poisson regression and Generalized poisson regression

Goodness-of-fit Test Measures	Poisson Regression	Generalized Poisson Regression
α		0.0148217
Deviance	73880.54	16411.70
AIC	26212.62	19712.25

Table-5: Negative binomial regression Vs. Generalized poisson regression

Variable	Negative Binomial Regression			Generalized Poisson Regression		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Intercept	2.221103	0.048956	0.0000	2.6208475	0.0200179	0.0000
Residence	-0.042171	0.015368	0.00607	-0.0421906	0.0041446	0.0000
Has Radio	0.011397	0.015826	0.47143	0.0118439	0.0042669	0.0027
Has Television	-0.016441	0.019603	0.40164	-0.0165244	0.0052983	0.0009
Religion	0.065269	0.023463	0.00541	0.0664167	0.0063337	0.0000
Wealth Index	-0.002454	0.020029	0.90249	-0.0019201	0.0054040	0.3599
Age At First Marriage	-0.064814	0.002773	0.0000	-0.0645548	0.0007451	0.0000
Partner's Education	-0.045759	0.016058	0.00438	-0.0452272	0.0043242	0.0000
Working Status	-0.084298	0.014411	0.0000	-0.0856665	0.0038892	0.0000
Educational level	-0.387970	0.014724	0.0000	-0.3874412	0.0039705	0.0000
Current Contraceptive Method	0.101512	0.013250	0.0000	0.0976496	0.0035808	0.0000

Table-6: Goodness-of-fit tests between Negative binomial regression and Generalized poisson regression

Goodness-of-fit Test Measures	Negative Binomial Regression	Generalized Poisson Regression
α		0.0148217
Deviance	13119	16411.70
AIC	39351	19712.25

IV. Discussion

The comparison between Negative Binomial Regression and the Poisson Regression models have been presented in **Table 1**. The regression parameters estimates for these two models are somewhat similar. This is expected, as estimates from these models are consistent. Of the explanatory variables, the coefficient for the educational level of mother is negative and significant which indicates that educated women have fewer children than uneducated women. As expected, the coefficient for the dummy variable residence is negative indicating that urban families have fewer children than rural families. However, this effect is statistically significant. In summary, the estimation results support the neoclassical theory of fertility. The standard errors of Negative Binomial Regression model are slightly larger than the standard errors of the Poisson model.

The deviance for the Poisson regression model (**Table 2**) is relatively larger than the deviance of the Negative Binomial Regression model and thus indicating possible existence for over-dispersion. To test for over-dispersion, AIC of Poisson against Negative Binomial Regression model is implemented. Based on the AIC, the Negative Binomial Regression model is not a better model for explaining over-dispersion in our data set. Now we will compare Generalized Poisson Regression model against Poisson model which is the alternative model of the Negative Binomial Regression model for explaining over-dispersion.

The comparison between Poisson model and Generalized Poisson Regression is shown in **Table 3**. The parameter estimates from the both models (PR and GPR) are somewhat similar. This is expected, as estimated from these models are consistent. Poisson regression model founds three insignificant factors whereas the Generalized poisson regression model founds only one insignificant factor indicating better explanation of fertility. Tabular results indicate that not accounting for the over-dispersion; the standard errors from the Poisson Regression model are under estimated. Consequently, the t-statistic for testing the significance of each regression parameter is generally upward biased for the PR model.

Table 4 gives the estimates of dispersion parameter using GPR model is positive indicating over-dispersion. From **Table 4**, the GPR model is preferred to the PR model based on two goodness-of-fit measures: deviance and AIC. For example, the GPR model has a smaller deviance than the PR model. The estimated AIC for GPR model is 19712.25 whereas it is 26212.62 for the PR model indicating better fits using GPR models.

Table 5 presents the comparison between Negative Binomial Regression models and the Generalized Poisson Regression models. The parameter estimates for these two models are similar. The standard errors of the GPR model are smaller than the standard errors of the Negative Binomial Regression models.

Table 6 represents that the GPR model is preferred to the Negative Binomial regression model based on two goodness-of-fit measures: deviance and AIC. In this table, the GPR model has a smaller deviance than the PR model. The estimated AIC for GPR model is 19712.25 whereas it is 26212.62 for the PR model indicating better fits using GPR models.

V. Conclusion

This study demonstrates the performance of better model among Negative binomial regression and the Generalized poisson regression for explaining over dispersion count data. The empirical results presented in this study support the neoclassical theory of fertility. The fitted models suggest that the educational level of mother is a determinant of more children. The educated women are more likely to bear fewer children than the women who have no education. Residence is found to be the key factor of more child bearing in Bangladesh. In rural areas, women desire more children than women in urban areas. As a result, the family size increases in rural areas than urban areas. Religion has an obliging effect on childbearing. The populations of Bangladesh are almost pious. But some of them are extremely conservative. They think that the children are the gift of God. They don't care for the bad effects of over number of children. For this reason, family size increases. Early age at first marriage is also one of the main reasons of more childbearing to the Bangladeshi women. Partner's occupation, having radio, televisions, working status, current contraceptive methods etc. are also appeared to be important as determinants of childbearing.

The aim of our analysis is to obtain the simplest model that reasonably explains the variation, i.e., over-dispersion in the data. For comparing Negative Binomial Regression model and Generalized Poisson Regression model, we use two measures of goodness-of-fit like deviance and AIC for over-dispersion. These values indicate that the Generalized Poisson Regression model is more appropriate for the present born data and leads to more efficient parameter estimates. Generalized Poisson Regression model founds only one insignificant factor but Negative binomial regression model and the Poisson regression model found three insignificant factors. That means, Generalized Poisson Regression model exhibits better explanation of child bearing which is a count data. Generalized Poisson Regression model shows that the population is growing at a very first rate in Bangladesh because of the effects of some typical demographic and socio-economic factors. Therefore, government should take appropriate policy and programs to reduce the level of childbearing in consideration to the existing higher rate of childbearing among the women in Bangladesh.

References

- [1] F. Famoye , J.T. Wulu and K.P. Singh, "On the Generalized Poisson Regression Model with an application to Accident Data", *Data Science* 2, 2001, pp. 287-295.
- [2] W. Wang and F. Famoye, " Modeling Household Fertility Decisions with Generalized Poisson Regression", *Popul Econ* 10, 1997, pp. 273-283.
- [3] N. Ismail and A.A. Jemain, "Handling Over-dispersion with Negative Binomial and Generalized Poisson Regression Models", *Casualty Actuarial Society Forum*, 2007, pp. 103-158.

- [4] Y. Cui, D. Y. Kim and J. Zhu J, "On the Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci With Count Data", Genetics Society of America 174, 2006, pp. 2159-2172.
- [5] M. Asaduzzaman and M. H. R. Khan, "Factors Related to Childbearing in Bangladesh: A Generalized Linear Modelling approach", Brac University Journal, vol. V, no. 2, 2008, pp. 15-21.
- [6] A.J. Dobson, "Introduction to Generalized Linear Models", 2001, Second Edition, London: Chapman and Hall/CRC.
- [7] S. N. Mitra et al, Bangladesh Demographic and Health Survey 2007, NIPORT, Dhaka, Bangladesh. (2007).