Modeling the Number of Children Ever Born in a Household in Bangladesh Using Generalized Poisson Regression

Mariam Begum Ratna, Hossain Ahmed Khan, Md Anower Hossain

Abstract— In this paper, an attempt has been made to model the total number of children ever born in a household in Bangladesh by using a generalized Poisson regression model. The generalized Poisson regression model has statistical advantages over standard Poisson regression model and is suitable for analysis of count data that exhibit either over-dispersion or under-dispersion. The maximum likelihood method is used to estimate the model. Approximate tests are performed for the dispersion and goodness-of-fit measures for comparing alternative models.

Keywords— Generalized Poisson regression model, dispersion, goodness-of-fit.

1 Introduction

WHEN the response or dependent variable is a count generated by processes in which the number of incidences is due to a rare or chance event, and that rare or chance event follows the principle of randomness. In such cases, Poisson regression model is applied to fit this type of data. In theory, data of the Poisson distribution should have its mean equal to its variance. But in practice, data arising from groups or individuals may be statistically dependent, so the observed variance of the data may be larger or smaller than the corresponding mean.

There are number of approaches to dealing with count data, or data arising from accumulated or aggregated binomial (or multinomial) trials. The more familiar is the Poisson regression (PR) model. But the generalized Poisson regression (GPR) model has shown statistical advantages over standard Poisson regression, negative binomial regression, generalized negative binomial regression and generalized linear models in the event of fitting count data that may be over-dispersed or under-dispersed or equi-dispersed. The GPR provides a versatile approach for analyzing count random variables and their relationships to other variables or covariates.

Consul [1] presented pioneering work on a generalization of Poisson distribution. Singh and Femoye [2] used and suggested the GPR model instead of the PR model in their analysis of life table and follow-up data. They suggested that the PR model was not appropriate to analyze an extra-Poisson variation survival data set. A number of works have suggested various models to deal with extra-Poisson variation in data. (See, for example, Cox [3]; Breslow [4]; Lawless [5]).

In many empirical studies of fertility, the number of children ever born in a household in Bangladesh is modeled as a function of socio-economic variables. The commonly used model is the standard Poisson. This model is considered because the number of children ever born in a family is non-negative. However, this model has some restrictions in some situations. In standard Poisson regression model, the conditional mean and variance of the dependent variable is constrained to be equal (equidispersion) for each observation. In practice, this assumption is often violated since the variance can either be larger or smaller than the mean. That is, both over-dispersion and under-dispersion can exist in the count data. If the equidispersion assumption is violated, the estimates in Poisson regression model are still consistent but inefficient. As a result, inference based on the estimated standard errors is no longer valid. As noted in Winkelmann and Zimmermann [6], the number of children ever born in a household often does not follow equal-dispersion assumption when mode is 2. Therefore, the standard Poisson regression model which assumes equal-dispersion is not appropriate to model data about household fertility decision.

The paper proceeds in the following way. Section 2 describes the data and variables used in this paper. Section 3 outlines the generalized Poisson regression model, goodness of fit and comparison measures and test of dispersion. Section 4 presents and discusses the estimated results. The paper concludes in section 5.

- *Mariam Begum Ratna is with School of Business, University of Liberal Arts Bangladesh. Email: mariam.begum@ulab.edu.bd*
- *Hossain Ahmed Khan is with BRAC. E-mail: hossain.ahmed@yahoo.com*
- *Md Anower Hossain is with the Institute of Statistical Research and Training (ISRT), University of Dhaka. E-mail: anower@isrt.ac.bd*

Manuscript received on20 July 2012 and accepted for publication on 30 September 2012.

2. DATA AND VARIABLES

The data from Bangladesh Demographic and Health Survey (BDHS) 2007 have been used in this study. The 2007 BDHS employs a nationally representative sample that covers the entire population residing the private dwelling units in Bangladesh. The survey used the sampling frame provided by the list of census enumeration areas (EAs) with population and household information from 2001 Population Census. Bangladesh is divided into six administrative divisions: Barisal, Chittagong, Dhaka, Khulna, Rajshahi and Sylthet. In turn, each division is divided into zillas, and each zilla into upazillas. Rural areas in an upazila are divided into union parishads (UPs), and UPs are further divided into mouzas. Urban areas in an upazila are divided into wards, and wards are subdivided into mahallas. EAs from the census were used as the Primary Sampling Units (PSUs) for the survey. The survey was based on a two-stage stratified sample of households. At the first stage of sampling, 361 PSUs were selected where 227 were rural PSUs and 134 urban PSUs. The survey was designed to obtain 11,485 completed interviews with ever-married women age 10-49. According to the sample design, 4360 interviews were allocated to urban areas and 7125 rural areas.

A household fertility decision may depend on different factors. Following is the list of dependent and independent variables used in this study. Table 1 shows the variable definition and descriptive statistics of each variable.

Dependent variable:

- Number of children ever born in a family

Independent variables:

- Age of respondent
- Has electricity (1 = yes, 0 = no)
- Has Television (1 = yes, 0 = no)
- Age at marriage
- Partner's education level (1 = HSC or more, 0 = otherwise)
- Type of place of residence (1 = Urban, 0 = else)
- Literacy of the respondent (1 = SSC or more, 0 = else)
- Religion of the respondent (1 = Islam, 0 = otherwise)
- Contraceptive use (1 = yes, 0 = no).

3. THE GENERALIZED POISSION REGRESSION MODEL

Suppose a count response variable follows a generalized Poisson distribution. To model number of children ever born, we define as the number of children ever born per household. Following Singh and Famoye [2], the probability of mass function is given by

$$f(y_i;\mu_i,\alpha)=\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\times \exp\left(-\frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right) \quad (1)$$

$$y_i=0,1,2,.... \qquad \mu_i=\mu_i(x_i)=\exp(x_i\beta)$$

where x_i is a $(k-1)$ dimensional vector of explanatory variables including personal characteristics of both husband and wife in a family as well as some demographic attributes of the family, and β is a k dimensional vector of regression parameters. The mean and variance of Y_i are given by $E(Y_i|x_i)=\mu_i$ and $V(Y_i|x_i)=\mu_i(1+\alpha\mu_i)^2$, respectively.

TABLE 1
VARIABLE DEFINITION AND DESCRIPTIVE STATISTICS (SAMPLE SIZE = 10058)

Variable	Proportion of 1's	Mean	Std. Dev. (SD)
Number of children ever born in a family		2.88	2.07
Has electricity	0.526		
Has Television	0.374		
Age at marriage		15.39	2.86
Partner's education level	0.397		
Place of residence	0.379		
Literacy of the respondent	0.667		
Religion of the respondent	0.902		
Contraceptive use	0.521		

The generalized Poisson regression model (1) is a generalization of the standard Poisson regression (PR) model. When $\alpha=0$ the probability mass function in (1) reduces to the PR model and then

$$E(Y_i|x_i)=V(Y_i|x_i),$$

which means equidispersion.

In practical applications, this assumption is often not true since the variance can either be larger or smaller than the mean. If the variance is not equal to the mean, the estimates in PR model are still consistent but not efficient, which lead to the invalidation of inference based on the estimated standard errors.

For $\alpha>0$, $V(Y_i|x_i)>E(Y_i|x_i)$ and the generalized Poisson regression (GPR) model in (1) represents overdispersed count data. For $\alpha<0$, $V(Y_i|x_i)<E(Y_i|x_i)$

and the GPR model in (1) represents under-dispersed count data. In (1), α is called the dispersion parameter and can be estimated simultaneously with the coefficients in the GPR model (1).

To estimate (β, α) in the GPR model (1), we need the log-likelihood function of the GPR model, that is,

$$\ell(\alpha,\beta) = \ln L(\alpha,\beta; y_i) = \sum_{i=1}^{n} y_i \log\left(\frac{\mu_i}{1+\alpha\mu_i}\right) + (y_i - 1)\log(1+\alpha y_i) - \frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i} - \log(y_i!)$$

The maximum likelihood equations for estimating α and β are obtained by taking the partial derivatives and equating to zero. Thus we get

$$\frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n}\left\{\frac{-y_i u_i}{1+\alpha\mu_i} + \frac{y_i(y_i-1)}{1+\alpha y_i} - \frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2}\right\} = 0 \quad (2)$$

and

$$\frac{\partial \ell(\alpha,\beta)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{y_i-\mu_i}{\mu_i(1+\alpha\mu_i)^2}\frac{\partial\mu_i}{\partial\beta_r} = 0\ ,$$

$r = 1, 2, 3, ..., k$

Substituting $\mu_i = \exp(x_i\beta)$, Eq. (2) becomes

$$\frac{\partial \ell(\alpha,\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \frac{y_i-\mu_i}{(1+\alpha\mu_i)^2} = 0\ , \quad (3)$$

and

$$\frac{\partial \ell(\alpha,\beta)}{\partial \beta_r} = \sum \frac{(y_i-\mu_i)x_i}{(1+\alpha\mu_i)^2} = 0\ ,\ r = 2,3,....,k \quad (4)$$

By using an iterative algorithm equations (2), (3) and (4) are solved simultaneously. The final estimate of β from fitting a Poisson regression model to the data is used as initial estimate of β for the iteration process. The initial estimate of α can be taken as zero or it may be obtained by equating the chi-square statistic to its degrees of freedom. This is given by

$$\sum \frac{(y_i-\mu_i)^2}{V(Y_i|x_i)} = n-k$$

When $\alpha < 0$ (the case of under dispersion), the value of α is such that $1+\alpha\mu_i > 0$ and $1+\alpha y_i > 0$, i.e., $\alpha > \min(-1/\max(\mu_i), -1/\max(y_i))$, as required in equation in (1). An R-program is used to solve Eqs. (2), (3), and (4) simultaneously.

3.1 Goodness-of-fit and model comparison

When more than one regression models are available for a given data set, one can compare performance of alternative models based on some measures of goodness-of-fit. Several measures of goodness-of-fit have been proposed in the literature. One commonly used measure is the Akaike information criterion AIC, which is defined as

$$AIC = -\ell + K$$

where ℓ is the log-likelihood value of the estimated model and K is the number of estimated parameters. The smaller is the AIC, the better is the model.

Merkle and Zimmermann also suggested several Pseudo-R^2 measures. One of these statistics is defined as

$$R_G^2 = \frac{l(\hat{\alpha},\hat{\mu}_i) - l(\hat{\alpha},\bar{y})}{l(\hat{\alpha},y_i) - l(\hat{\alpha},\bar{y})} \quad (5)$$

where

$$l(\hat{\alpha},\hat{\mu}_i) = \sum_{i=1}^{n}\left\{y_i\log\left(\frac{\hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}\right) + (y_i-1)\log(1+\hat{\alpha}y_i) - \frac{\hat{\mu}_i(1+\hat{\alpha}y_i)}{1+\hat{\alpha}\hat{\mu}_i} - \log(y_i!)\right\},$$

$$l(\hat{\alpha},\bar{y}) = \sum_{i=1}^{n}\left\{y_i\log\left(\frac{\bar{y}}{1+\hat{\alpha}\bar{y}}\right) + (y_i-1)\log(1+\hat{\alpha}y_i) - \frac{\bar{y}(1+\hat{\alpha}y_i)}{1+\hat{\alpha}\bar{y}} - \log(y_i!)\right\},$$

$$l(\hat{\alpha}, y_i) = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{1+\hat{\alpha} y_i} \right) + (y_i - 1)\log(1+\hat{\alpha} y_i) - \frac{y_i(1+\hat{\alpha} y_i)}{1+\hat{\alpha} y_i} - \log(y_i!) \right\}$$

$$= \sum_{i=1}^{n} \{ y_i \log(y_i) - \log(1+\hat{\alpha} y_i) - y_i - \log(y_i!) \}$$

R_G^2 measures the explained maximum possible increase in the log-likelihood.

TABLE 2

DETERMINANTS OF HOUSEHOLD FERTILITY: COMPARISON BETWEEN POISSON AND GENERALIZED POISSON REGRESSION MODELS.

1 VARIABLE	POISSON REGRESSION (PR)				GENERALIZED POISSON REGRESSION (GPR)		
	ESTIMATES	STANDARD ERROR (SE)	$t-$ VALUE		ESTIMATES	STANDARD ERROR (SE)	$t-$ VALUE
I…NTERCEPT	0.3364	0.0707	4.76		0.3071	0.0451	6.81
HAS_ELEC	-0.0254	0.0154	-1.65		-0.0391	0.0201	-1.95
HAS_TV	-0.0872	0.0163	-5.35		-0.095	0.0256	-3.71
AGE_MAR	-0.0472	0.0024	-19.67		-0.0534	0.0105	-5.09
PAT_EDU	-0.1167	0.0144	-8.1		-0.1199	0.0195	-6.15
RESI	-0.0655	0.0138	-4.75		-0.0639	0.0215	-2.97
EDU	-0.0723	0.01378	-5.25		-0.0845	0.0159	-5.31
RELIGION	0.1135	0.0213	5.33		0.1096	0.0302	3.63
CON_USE	-0.1466	0.0119	-12.32		-0.1843	0.0185	-9.96
α				0.0627		0.0021	29.85

3.2 Test for Dispersion

The generalized Poisson regression model reduces to the Poisson regression model when the dispersion parameter α equals to zero. To assess justification of using GPR model over the PR model, we test the hypothesis

$$H_0: \alpha = 0 \quad \text{against } H_1: \alpha \neq 0 \tag{6}$$

The test of H_0 in (6) is for the significance of the dispersion parameter. Whenever H_0 is rejected, it is recommended to use the GPR model in place of the PR model. To carry out the test in (6), one can use the asymptotically normal Wald type "t" statistic defined as the ratio of the estimate of α to its standard error. Another way to test the null hypothesis of α equals to zero is to use the likelihood ratio statistic, which is approximately chi-square distribution with one degree of freedom when the null hypothesis is true. Both the likelihood ratio test and the Wald type "t" test are asymptotically equivalent.

4. RESULTS AND DISCUSSION

Both Poisson regression (PR) and generalized Poisson regression (GPR) models are estimated using sample data. Table 2 represents the parameter estimates, their standard errors, and t-value. Table 3 presents several measures of goodness-of-fit including Pearson's chi-square, deviance, AIC and R_G^2.

TABLE 3:

GOODNESS-OF-FIT TEST MEASURE

GOODNESS-OF-FIT MEASURES	PR	GPR
PEARSON'S CHI-SQUARE	7486.13	7571.00
DEVIANCE	72935.10	13493.69
AIC	25988.93	17929.04
R_G^2	0.3079	0.4608

We note from table 2 that the estimate of dispersion parameter using GPR model is positive indicating over-dispersion. The asymptotic t-statistics for testing the null hypothesis $H_0 : \alpha = 0$ is significant ($t-$value = 29.85). The dispersion parameter α is significantly different from zero. So the PR model is not appropriate for this data since we reject the null hypothesis $H_0 : \alpha = 0$. From table 3, the generalized Poisson regression model is preferred to the Poisson regression model based on all four goodness-of-fit measures: Pearson's chi-square, deviance, AIC and R_G^2. For example, the generalized Poisson regression model has a smaller deviance value (13493.69) than the deviance value (72935.10) of the standard Poisson regression model. The value of Pearson's chi-square is 7571.00 for generalized Poisson regression model, whereas it is 7486.13 for the Poisson model, which indicates that modeling over-dispersion data using the GPR is more appropriate than the PR model.

The parameter estimates are almost similar for both Poisson regression and GPR models. This is expected since estimates from both models are consistent. The results from table 2 show that the standard errors of estimates from PR model are under estimated because the PR model does not consider the over-dispersion exhibited by the data. In this case the standard errors of the estimates from GPR are more accurate since it considers the over-dispersion showed by the data. Therefore, the t-statistic for testing the significance of the parameter estimates is upward biased for Poisson regression model.

From the results in table 2, the coefficient of partner's education and respondent education are negative and significant. These imply that households with educated parents have fewer children. Also the explanatory variables have electricity and TVs are significant and are inversely related to the family size. This is expected because households are aware about the problem of more children through the different TV programs about population problem.

The effect of place of residence (1 = Urban, 0 = else) on family size is negative and significant. The urban people prefer less number of children in the family. The variable Religion (1 = Islam, 0 = otherwise) has positive effect on family fertility decision and statistically significant. Contraceptive use (1 = yes, 0 = no) has negative effect on number of children in a family and us significant. The variable age at first marriage has negative coefficient in the fitted model which implies that the number of children in a family decreases as the value of age at first marriage increases.

5. SUMMARY

In this paper, we have described nonlinear regression techniques (namely, generalized Poisson regression and Standard Poisson regression) appropriate for the analysis of number of children in a household of Bangladesh. It has been shown that when over-dispersion exists in the data generalized Poisson regression model gives better fits than standard Poisson regression model. Several goodness-of-fit techniques and asymptotic t-test for over-dispersion imply that the generalized Poisson regression model is more appropriate for the data about the number of children in a household for Bangladesh.

REFERENCES

[1] P. C. Consul, Generalized Poisson distribution: Properties and applications, Dekker, Inc., New York, 1989.

[2] Singh and F. Famoye Restricted generalized Poisson regression model. *Communications in* regression, *Canadian Journal of Statistics*, 1993, vol 15, pp 209-225.

[3] D. R. Cox, Some remarks on over-dispersion. Biometrica, 1983, vol 70, pp 269-272.

[4] N. Breslow, Tests of hypothesis in over-dispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association,* 1990, vol 85, pp 565-571.

[5] J. F. Lawless, Negative binomial and mixed Poisson regression, Canadian Journal of Statistics, 1987, vol 15, pp 209-225.

[6] R. Winkelmann and K. F. Zimmermann, Count data models for demographic data. *Mathematical population Studies*, 1994, vol 4, pp 205-221.

Mariam Begum Ratna was born in Noakhali, Bangladesh in 1981. She completed her B. Sc. (Hons.) and M. S. in Applied Statistics from Institute of Statistical Research and Training (ISRT), University of Dhaka, in 2002 and 2003, respectively. Currently she is working as a Lecturer of Statistics in ULAB. Before that she worked as a Lecturer of Statistics in American International University and Stamford University Bangladesh.

Hossain Ahmed Khan completed his B. Sc. (Hons.) and M. S. in Applied Statistics from Institute of Statistical Research and Training (ISRT), University of Dhaka, in 2006 and 2007, respectively. Currently he is working as a researcher in BRAC.

Md Anower Hossain was born in Comilla, Bangladesh in 1980. He did his B. Sc. (Hons.) and M. S. in Applied Statistics from Institute of Statistical Research and Training (ISRT), University of Dhaka, in 2002 and 2003, respectively. He joined as a Lecturer of Applied Statistics at ISRT in 2007 and currently he is an Assistant Professor of this institute.