Modeling Under-dispersed Count Data Using Generalized Poisson Regression Approach

Md. Maidul Husain¹, Md. Shaddam Hossain Bagmar²*

Abstract
This paper models household fertility decisions by using a generalized Poisson regression model. Since the fertility data used in the paper exhibit under-dispersion, the generalized Poisson regression model has statistical advantages over both standard Poisson regression and negative binomial regression models, and is suitable for analysis of count data that exhibit either over-dispersion or under-dispersion. The model is estimated by maximum likelihood procedure. Approximate tests for the dispersion and goodness-of-fit measures for comparing alternative models are discussed. Based on observations from the Bangladesh Demographic Health Survey 2011, the empirical results support the generalized Poisson regression to model the under dispersed data.

Key words: Fertility, Count data, Under-dispersion, Generalized Poisson regression, Goodness of fit.

Introduction
Bangladesh is the world's most densely populated country with about 1,100 people per square kilo-meter (World Bank, 2015). We know that we live in limited resources, for that to best utilize of our resources we need to control our population. Bangladesh has made remarkable achievements in reducing the average number of children per woman of reproductive age (total fertility rate – TFR). Total fertility rate in Bangladesh is now merely 2.2 which is the lowest TFR in South Asia which also discussed previously by Gubhaju (2007). Individual household fertility decisions have been modeled in various ways in literature. Barmby and Cigno (1990) estimated fertility patterns using a sequential probability model. Sobel and Arminger (1992) used a non linear simultaneous probit model. In recent years, the modeling of household fertility decisions has utilized Poisson type models. Along the lines of King (1989), Winkelmann and Zimmermann (1994) developed the generalized event count model which subsumes the Poisson, the negative binomial and the binomial models. Caudill and Mixon Jr (1995) developed censored regression models for fertility data. Gschlobl and Czado (2008) consider regression models for count data allowing for overdispersion in a Bayesian framework. Kokonendji (2014) discussed some count statistical models, which are tied to the phenomenon of over, equi or underdispersion. Famoye (2015) studied the generalized Poisson regression model for count data on multivariate context. Rodríguez-Avi and Olmo-Jiménez (2015) developed a regression model for overdispersed count data and compared it with the traditional models i.e., generalized Poisson regression model, the negative binomial regression model and the zero inflated Poisson regression model.

In many empirical studies of fertility (Wang and Famoye, 1997), the number of children in a household is modeled as a function of social and economical variables such as wife's education level and family income. The commonly used models are standard Poisson and negative binomial. But these models are restricted in some situations. Under the Poisson regression model, the conditional mean of the dependent variable is equal to variance (equidispersion) for each observation. In practice this assumptions is often false since

---

¹ Institute of Statistical Research and Training, University of Dhaka, Bangladesh, mhusain@isrt.ac.bd
² Institute of Statistical Research and Training, University of Dhaka, Bangladesh, shbagmar@isrt.ac.bd
*Corresponding author
variance is either smaller (under-dispersion) or larger (over-dispersion) than mean. If there exists over or under dispersion in count data then estimates from Poisson regression is consistent but standard error is different compare to generalized Poisson regression (GPR). Therefore, inference based on standard error is no longer valid. In demographic contexts, fertility refers to the actual production of offspring, rather than the physical capability to produce which is termed fecundity. In this study we took total number of living children as a dependent variable to determine the socio economic factor that would be effect on the fertility decision. This is also suggest by (Wang and Famojye, 1997) in their paper to modeling household fertility. As noted by Winkelmann and Zimmermann (1994), the number of children in household often exhibits under-dispersion when mode is two. Therefore, the standard Poisson model which assumes equidispersion and negative binomial which assumes over-dispersion are not appropriate. Winkelmann and Zimmermann (1995) found under-dispersion in German fertility data by using the generalized event count model proposed by King (1989).

The paper proceeds in the following way. Section 2 presents the data used in the paper, outlines the GPR model of household fertility decisions. We also consider goodness-of-fit statistics for the model as well as test statistics for assessing the significance of the dispersion parameter. Section 3 provides and discusses the resulting estimates and makes the comparisons between standard Poisson and GPR models. The paper concludes with some remarks in Section 4.

**Methodology**

In this study we consider the modeling approach for count data namely, standard Poisson and GPR models. The dependent variable, number of living children which exhibits under-dispersion. That’s why we don’t consider the negative binomial model, which is appropriate only for over-dispersed count data.

**Data and variables**

In our study, we use the Bangladesh Demographic Health Survey (BDHS) 2011 data. Table 1 shows the different variables and their descriptive statistics. In total there were 8753 living children with a mean of 2.31, who were considered throughout the analysis. As the fertility decisions directly related to the number of living children, it is considered as dependent variable. For independent variables, employment, education and age are related to mother, status, residence, religion and media are related to socio-economic and household characteristics. The only one continuous variable considered here is age difference of the women with her husband has mean of 8.74 years (8.57 years for Urban and 8.81 years for Rural). Briefly, most of the women's are currently unemployed may be worked as housewife, has secondary education and belongs to age group 15-24. Under household characteristics, highest percent of households are poor, belongs to rural, religiously Muslim and exposed to media.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>Mean/Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children*</td>
<td>Counts</td>
<td>2.31</td>
</tr>
</tbody>
</table>

**Mother characteristics**

<table>
<thead>
<tr>
<th>Employment</th>
<th>Yes = 1</th>
<th>9.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>No = 0</td>
<td>90.1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>No education = 0</th>
<th>19.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary = 1</td>
<td>30.7</td>
<td></td>
</tr>
<tr>
<td>Secondary = 2</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>Higher = 3</td>
<td>7.9</td>
<td></td>
</tr>
</tbody>
</table>

| Age | 15-24 = 0 | 49.5 |
Standard Poisson regression model

Let $Y_i$ be the random variable takes non negative values, $i = 1, 2, \ldots, n$ where $n$ is the number of observations. If $Y_i$ is the number of counts for $i$th occasion and follows a Poisson distributions with the probability function:

$$f(y_i) = \Pr(Y_i = y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$  (1)

With mean and the variance are equal

$$E(Y_i) = \text{Var}(Y_i) = \lambda_i,$$  (2)

where $\lambda_i = \exp(X_i^T\beta)$, is the mean number of counts, $X_i$ is the $i$th row of covariate matrix, and $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$ are unknown $k$-dimensional vector of regression parameters. The mean of $Y_i$ is given by $E(Y_i | X_i)$ and the variances of $Y_i$ is given by $\text{Var}(Y_i | X_i)$. The parameters $\beta$ can be estimated by maximum likelihood approach,

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$  (3)

The log-likelihood function is given by

$$l(\beta) = \sum_{i=1}^{n} \left[-\lambda_i + y_i\ln\lambda_i - \ln(y_i!)ight] = \sum_{i=1}^{n} \left[-\exp(X_i^T\beta) + y_i(X_i^T\beta - y_i!)ight]$$  (4)

By differentiating Equation 4 with respect to $\beta$ and equating to zero we will get

$$\frac{d\ln(l)}{d\beta_j} = \sum_{i=1}^{n} \left[-\exp(X_i^T\beta)X_{ij} + y_iX_{ij}ight],$$  (5)

yields $k$ nonlinear equations and solve these equations by Newton-Raphson method or by iteratively weighted least square procedure the parameters are estimated.

Generalized Poisson regression model
Suppose \( Y_i \) is a count response variable that follows a generalized Poisson distribution. The probability mass function of \( Y_i, i = 1, 2, \ldots, n \) is given by Famoye (1993), Wang and Famoye (1997):

\[
f(y) = \Pr(Y_i = y) = \left( \frac{\lambda_i}{1 + \alpha \lambda_i} \right)^{y_i-1} \exp\left[ - \frac{\lambda_i (1 + \alpha y_i)}{1 + \alpha \lambda_i} \right], \quad y_i = 0, 1, 2, \ldots
\]  

(6)

The mean and variance of \( Y_i \) are given by:

\[
E(Y_i / X_i) = \lambda_i, \quad \text{Var}(Y_i / X_i) = \lambda_i (1 + \alpha \lambda_i),
\]

(7)

(8)

where \( \alpha \) is called the dispersion parameter. The generalized Poisson distribution is a natural extension of the Poisson distribution when \( \alpha = 0 \), the probability mass function in Equation (6) reduces to the probability mass function in Equation (1) so that the mean is equal to the variance and this a case of equi-dispersion. In practical application, this assumption is often untrue since the variance can either longer or smaller than the mean. If the variance is not equal to the mean, the estimates in Poisson regression model are still consistent but are underestimated i.e., presence of negative bias, which leads to the invalidation of inference based on the estimated standard errors (Famoye et al., 2004). When \( \alpha > 0 \), then the variance is larger than the mean, and for this situation, the GPR model represents count data with over–dispersion, and when \( \alpha < 0 \), the variance is smaller than the mean, and for this situation, the GPR model represents count date with under–dispersion. The estimates of \( \alpha \) and \( \beta \) in the GPR model are obtained using the method of maximum likelihood.

The log–likelihood function of the GPR model is given by:

\[
\ell(\alpha, \beta) = \sum_{i=1}^{n} \left[ - \ln \left( \frac{\lambda_i}{1 + \alpha \lambda_i} \right) + (y_i - 1) \ln (1 + \alpha y_i) - \frac{\lambda_i (y_i - 1) + \lambda_i (y_i - \lambda_i)}{(1 + \alpha \lambda_i)^2} \right]
\]  

(9)

By finding the partial derivatives with respect to \( \alpha \) and \( \beta \) the likelihood equations are given by:

\[
\frac{d\ell}{d\alpha} = \sum_{i=1}^{n} \left[ - \frac{\lambda_i y_i}{1 + \alpha \lambda_i} + \frac{y_i (y_i - 1) + \lambda_i (y_i - 1)}{(1 + \alpha \lambda_i)^2} \right] = 0
\]  

(10)

\[
\frac{d\ell}{d\beta_j} = \sum_{i=1}^{n} \left[ \frac{y_i - \lambda_i}{\lambda_i (1 + \alpha \lambda_i)^2} \right] \frac{d\lambda_i}{d\beta_j} = 0, \quad j = 1, 2, \ldots, k
\]  

(11)

As we know from generalized linear model the link function for Poisson regression is \( \log(\lambda_i) = \beta_j^T \beta \). Substituting \( \lambda_i = \exp(\beta_j^T \beta) \) on Equation (10) and (11) can be rewritten as

\[
\frac{d\ell}{d\alpha} = \sum_{i=1}^{n} \left[ \frac{y_i - \lambda_i}{\lambda_i (1 + \alpha \lambda_i)^2} \right] = 0
\]  

(12)

\[
\frac{d\ell}{d\beta_j} = \sum_{i=1}^{n} \left[ \frac{y_i - \lambda_i \beta_j}{\lambda_i (1 + \alpha \lambda_i)^2} \right] = 0, \quad j = 1, 2, \ldots, k
\]  

(13)

The parameters \( \alpha \) and \( \beta \) are estimated by the Newton–Raphson method. When \( \alpha < 0 \) (underdispersion) the value of \( \alpha \) is such that \( (1 + \alpha \lambda_i) > 0 \) and \( (1 + \alpha y_i) > 0 \). For the condition \( (1 + \alpha y_i) > 0 \), the program checks if \( (1 + \alpha \max(y_i)) > 0 \) on each iteration. Whenever this condition is not true, a new initial estimate for \( \alpha \) is set to \( -1 / \max(y_i+1) \). A similar check is implemented for the condition \( (1 + \alpha \lambda_i) > 0 \) (Wang and Famoye, 1997).

Also we can estimate \( \alpha \) by using method of moments, under this method \( \alpha \) may be estimated by equating the Pearson chi-squarer with \( (n - k) \) degree of freedom (Breshow and Lin, 1995), is given by:

\[
\sum_{i=1}^{n} \left[ \frac{(y_i - \lambda_i) x_j}{\lambda_i (1 + \alpha \lambda_i)^2} \right] = n-k,
\]  

(14)
where \( n \) denotes the number of values and \( k \) the number of regression parameters.

**Goodness of fit measures**

When many regression models are available for a given data set, one can compare the performance of alternative models based on some measures of goodness-of-fit. Several measures of goodness-of-fit have been proposed in literature.

*Deviance* is a way of assessing the adequacy of the model of interest with a saturated model with maximum number of parameters that can be estimated. Let \( \hat{b}_{\text{max}} \) and \( \hat{b} \) denote the estimated parameter vector for the saturated model and the model of interest, corresponding likelihood functions are \( L(\hat{b}_{\text{max}}; y) \) and \( L(\hat{b}; y) \). Then the likelihood ratio

\[
\Lambda = \frac{L(\hat{b}_{\text{max}}; y)}{L(\hat{b}; y)}
\]

\[
\ln \Lambda = l(\hat{b}_{\text{max}}; y) - l(\hat{b}; y),
\]

provides a way of assessing the goodness of fit for the model. As the values of \( \log \lambda \) much closer to 1, suggests the model of interest is describe the data well as saturated model (Nelder and Baker, 1972).

*The Akaike Information Criteria (AIC)* is defined by

\[
AIC = -\ln L + K,
\]

where \( \ln L \) is the log-likelihood value of the estimated model and \( K \) is the number of estimated parameters. Smaller the value of \( AIC \) indicates better the model is.

*Pearson chi-squared statistic* includes the test for independence in two way contingency tables. It has been extended in generalized linear model theory to a test for the adequacy of the current fitted model. Given a generalized linear model with responses \( y_i \), weights \( w_i \), fitted means \( \mu_i \), variance function \( \text{Var}(\mu_i) \) and dispersion \( \alpha = 1 \), the Pearson goodness of fit statistic is

\[
\chi^2 = \sum w_i \frac{(y_i - \mu_i)^2}{\text{Var}(\mu_i)}
\]

If the fitted model is correct and the observations \( y_i \) are approximately normal, then the \( \chi^2 \) statistic in Equation (17) is approximately distributed as \( \chi^2 \) on the residual degrees of freedom for the model.

*Test of dispersion parameter* is used to assess the adequacy of the GPR model over the Poisson regression model. The GPR reduces to the Poisson regression model when the dispersion parameter \( \alpha \) equals zero. One may test the hypothesis

\[
H_0: \alpha = 0 \quad \text{vs} \quad H_1: \alpha \neq 0
\]

We can use log likelihood ratio test (LRT) as

\[
\text{LRT} = -2[\ln(\text{Poisson}) - \ln(\text{GPR})]
\]

which follows a \( \chi^2 \) distribution with 1 df. We could reject the null hypothesis \( H_0 \), if the calculated value from Equation (19) is larger than the tabulated value from \( \chi^2 \) table at chosen level of significance. The inclusion of the dispersion parameter \( \alpha \) in the regression model is justified when the \( H_0 \) is rejected.

**Results and discussions**
At the beginning of the analysis, we showed some preliminary descriptive statistics on dependent variable. After that we estimate both of the models Poisson and generalized Poisson. Compare these models considering parameter estimates and goodness of fit measures. Lastly we justify the modeling by testing dispersion parameter. Figure 1 shows the histogram along with normal curve of the dependent variable number of living children. From the histogram we easily identify that the mode of the data is 2, which indicates that the data may exhibits under-dispersed (Winkelmann and Zimmermann, 1994). In the figure upper part shows that the variable number of living children has mean 2.31 and standard error 1.38 (variance is 1.90), i.e., the conditional variance is less than the conditional mean, which also indicates that the data exhibits under-dispersion (Cameron and Trivedi, 1999).

**Figure 1:** Histogram of the number of living children.

![Histogram of the number of living children](image)

**Table 2:** Comparing the results of Poisson and generalized Poisson regression model.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Poisson</th>
<th>Generalized Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>.487</td>
<td>.0376</td>
</tr>
<tr>
<td>Employment[Yes]</td>
<td>-.045</td>
<td>.0252</td>
</tr>
<tr>
<td>Education[Higher]</td>
<td>-.544</td>
<td>.0361</td>
</tr>
<tr>
<td>Education[Secondary]</td>
<td>-.269</td>
<td>.0215</td>
</tr>
<tr>
<td>Education[Primary]</td>
<td>-.097</td>
<td>.0192</td>
</tr>
<tr>
<td>Mother's age[35-49]</td>
<td>.913</td>
<td>.0227</td>
</tr>
<tr>
<td>Mother's age[25-34]</td>
<td>.557</td>
<td>.0163</td>
</tr>
<tr>
<td>Status[Rich]</td>
<td>-.029</td>
<td>.0208</td>
</tr>
<tr>
<td>Status[Middle]</td>
<td>-.015</td>
<td>.0206</td>
</tr>
<tr>
<td>Residence[Rural]</td>
<td>.037</td>
<td>.0180</td>
</tr>
</tbody>
</table>
Table 2 shows that the parameter estimates from both the Poisson regression and GPR models are quite similar (same estimates). This is not unexpected since estimates from the Poisson and generalized both are consistent. In the table we present only the interested categories (not the reference category) and the significant (at 5% level of significance) values are shown in bold type. The standard errors of the parameter estimates for Poisson regression are smaller than that of GPR.

From the list of explanatory variables, the characteristics of the wife stand out as significant in analyzing the fertility decisions. Based on GPR results the mother education levels are statistically significant and are inversely related to the number of the children in household. The social status is not statistically significant effect on the determination of the fertility decision of household. But we see that as family status increases households prefer to have less but higher quality children. The coefficients of dummy variable residence are positive and significant for standard Poisson model, not in GPR. It indicates that rural household has more children than urban household. The coefficient of religion is positive and significant. Which indicates that muslim household have more children than any other religion, which supports our socio-economic context. Age difference between wife and husband has significant effect on the fertility.

**Table 3:** Goodness of fit measures of Poisson and generalized Poisson regression model.

<table>
<thead>
<tr>
<th></th>
<th>Poisson</th>
<th>Generalized Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>3187.489</td>
<td>1946.718</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>3174.836</td>
<td>1938.991</td>
</tr>
<tr>
<td>DF</td>
<td>8856</td>
<td>8856</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-12728.585</td>
<td>-7773.821</td>
</tr>
<tr>
<td>AIC</td>
<td>25483.169</td>
<td>15573.642</td>
</tr>
</tbody>
</table>

Table 3 shows the values of measures of goodness-of-fit for two model, we found that deviance, Pearson chi-square and AIC for GPR are less than standard Poisson. Again the log-likelihood of GPR is greater than log-likelihood of standard Poisson. So we can say that the GPR is better described the data than Poisson and we would select GPR as an appropriate model. Considering the standard error estimates Poisson would be the better choice as the standard error estimates of standard Poisson is smaller compare to GPR (Table 2). So there the dilemma in choosing a model between Poisson and GPR. Finally we investigate the presence of dispersion in the data. Using the Equation (19) we obtain a value of the statistic $LRT$ equals 4954.746, comparing it with tabulated value from $\chi^2$ table with 1 df (3.84) we can reject the $H_0$ of zero dispersion ($\alpha = 0$) as defined in Equation (18). That implies there is presence of dispersion. The Poisson regression which overlook the dispersion in the data, underestimate the standard errors of the
parameter. Consequently Wald statistic and their significance are generally upward biased for the Poisson regression.

Conclusions
The aim of our analysis is to obtain the simplest model that reasonably explains the variation, i.e., under-dispersion in the data. Poisson regression model provides the consistent and more precise parameter estimates (smaller standard error) compare to generalized Poisson regression. For comparing Poisson regression model and generalized Poisson regression model, we apply several measures of goodness-of-fit such as deviance, AIC and log likelihood for choosing better models. These values indicate that the generalized Poisson regression model is more appropriate for the living children and leads to more reliable parameter estimates. To overcome this ambiguity in choosing the appropriate model, we test the existence of dispersion in the data. The test and the descriptive measures gives the evidence of underdispersion in the data. As the (under) dispersion in the data were overlooked by Poisson regression, it underestimates the parameter. So generalized Poisson regression provides us with reliable estimates of the parameter for the count data modeling with under dispersion.

References