Bivariate exponentiated-exponential geometric regression model

Felix Famoye
Central Michigan University, Department of Mathematics, Mt. Pleasant, MI 48859

Abstract
A bivariate exponentiated-exponential geometric regression (BEEGR) model that allows any type of correlation is defined and studied. The regression model is based on the univariate exponentiated-exponential geometric distribution and the marginal means of the bivariate model are functions of the explanatory variables. The parameters of the bivariate regression model are estimated by using the maximum likelihood method. Some test statistics including goodness-of-fit are discussed. One numerical data set is used to illustrate the applications of the regression model.

Key Words: Correlated count data; dispersion; estimation; goodness-of-fit

1. Introduction

The univariate regression models have been used to model count data where the sample mean and sample variance are about the same. When the sample mean and the sample variance are almost equal, we have an equi-dispersion situation. If the sample variance is greater (or smaller) than the sample mean, we have an over-dispersion (or under-dispersion) situation relative to the Poisson assumption. When the sample mean and sample variance are different, other univariate count data regression models have been developed and studied. Some of these univariate models have been extended to bivariate, and a few have been extended to multivariate count data regression models. See the books by Cameron and Trivedi (2013), Winkelmann (2008) and the references therein.

There are many ways to define a bivariate probability distribution (for example, see Kocherlakota and Kocherlakota, 1992; Johnson et al., 1997 and the references therein). One way is the trivariate reduction method. A disadvantage of the bivariate distribution from a trivariate reduction method is that the correlation between the variables is always positive. Other approaches that include using correlated random effects, conditional probabilities, or copula functions are mentioned by Famoye (2010b), Famoye (2015) and the references therein. Famoye (2010a) remarked that the bivariate distributions based on copula functions allow positive or negative correlation, but the bivariate distributions are very complicated in forms.

Famoye (2010a) defined and studied a new bivariate generalized Poisson distribution (BGPD) that allows for any type of correlation and any type of dispersion. The properties of Sarmanov (1966) bivariate distributions were discussed by Lee (1996) who gave the bivariate Poisson distribution (BPD) as an example. The BPD was later discussed by Lakshminarayana et al. (1999). Hofer and Leitner (2012) modified the BGPD in Famoye (2010a) and defined a bivariate Sarmanov regression model with generalized Poisson marginals. Famoye (2010b) defined and studied a bivariate Sarmanov regression model with negative binomial marginals and called it a new bivariate negative binomial regression (BNBR) model. The two negative binomial variates are characterized by any type of correlation. However, the variates allow for over-dispersion but not under-dispersion.

There are not many regression models that allow for both under-dispersion and over-dispersion. Winkelmann (2008, pp. 45-56) discussed three such models: the generalized

event count model, the double Poisson distribution, and the gamma count distribution. The drawbacks of these models include probability mass function in complicated forms and their means and variances are not in closed forms. Another model that allows for under-dispersion and over-dispersion is the generalized Poisson distribution with mean and variance in closed forms and the probability mass function is not complicated.

Alzaatreh et al. (2013) developed a general method for generating a univariate probability distributions and named it the *T-R* family. Let $f_T(t)$ be a probability density function (PDF) of a continuous random variable $T \in [a, b]$ for $-\infty \le a < b \le \infty$. Suppose $W(F_R(y))$ is a monotonic and absolutely continuous function of the cumulative distribution (CDF), $F_R(y)$, of any discrete or continuous random variable *R*. The CDF $F_Y(y)$ of a new random variable *Y* is given by

$$F_Y(y) = \int_a^{W(F_R(y))} f_T(t)dt = F_T\left\{W\left(F_R(y)\right)\right\}. \qquad (1)$$

Many families of continuous distributions have been defined by using (1). Alzaatreh et al. (2012) used (1) to define the *T*-geometric family, which consists of the discrete analogue to the distribution of any continuous non-negative random variable *T*. A member of this family is the exponentiated-exponential geometric distribution (EEGD) which was studied in details by Alzaatreh et al. (2012).

Famoye and Lee (2015) defined the exponentiated-exponential geometric regression (EEGR) model, which was based on the EEGD. The EEGR was fitted to three observed data sets and it was found that the model is very competitive or performed better than the generalized Poisson regression (GPR) model studied by Famoye (1993).

In this paper, a new Sarmanov bivariate regression model based on the exponentiated-exponential geometric marginal is defined and studied. Among the important characteristics of this new regression model includes (i) the model allows for any type of correlation (ii) the model allows for both under-dispersion and over-dispersion for each variate, and (iii) the model allows the correlation and the dispersion to be independently determined. Even though the probability mass function and CDF are in nice forms, the mean and the variance are not in closed forms. The bivariate exponentiated-exponential geometric regression (BEEGR) model is defined in section 2. Also in section 2, a zero-inflated BEEGR model is defined. In section 3, we discuss parameter estimation for the BEEGR. Some tests are provided in section 4. One numerical data set is used to illustrate the BEEGR model in section 5 and the results are compared with that of bivariate generalized Poisson regression model. In section 6, we provide some concluding remarks.

2. Bivariate exponentiated-exponential geometric regression model

The exponentiated-exponential geometric distribution (EEGD) was defined and studied by Alzaatreh et al. (2012). The probability mass function is given as

$$P(Y = y) = \left(1-\theta^{y+1}\right)^b - \left(1-\theta^y\right)^b, \; y = 0,1,2,\ldots, \qquad (2)$$

where $0 < \theta < 1$ and $b > 0$. The EEGD is unimodal and skewed to the right. Alzaatreh et al. (2012) showed that the EEGD is always over-dispersed when $0 < b \le 2$ and when $b > 2$, the EEGD can be equi- or under- or over-dispersed.

Famoye and Lee (2015) defined the EEGR model as

$$P(Y = y_i \mid x_i) = \left(1-[\theta(x_i)]^{y_i+1}\right)^b - \left(1-[\theta(x_i)]^{y_i}\right)^b,\ y_i = 0,1,2,\ldots, \qquad (3)$$

where $\theta_i = \theta(x_i) = f(x_i, \beta) = 1/[1+\exp(-x_i'\beta)]$, $x_i = (x_{i0}=1, x_{i1}, x_{i2}, \ldots, x_{i,k-1})'$ is a (k – 1)-dimensional vector of predictor variables, and $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1})'$ is k-dimensional vector of regression parameters. In the EEGR model (3), the shape parameter b is taken to be a nuisance parameter and this parameter is also the dispersion parameter. When b = 1, the EEGR in (3) reduces to the geometric regression model given by

$$P(Y = y_i \mid x_i) = [\theta(x_i)]^{y_i}\left(1-\theta(x_i)\right),\ y_i = 0,1,2,\ldots$$

By using the model in (2), a bivariate exponentiated-exponential geometric distribution (BEEGD) can be defined using the system of bivariate Sarmanov (1966) distributions. The probability mass function of BEEGD is given by

$$P(y_1, y_2) = \prod_{t=1}^{2}\left[(1-\theta_t^{y_t+1})^{b_t} - (1-\theta_t^{y_t})^{b_t}\right] \times [1+\lambda(e^{-y_1}-c_1)(e^{-y_2}-c_2)], \qquad (4)$$

where $c_t = E(e^{-Y_t})$ for t = 1, 2. In order to determine c_t, we need to find the moment generating function of the EEGD in (2). The moment generating function of the EEGD is given by

$$\begin{aligned} E(e^{sY}) &= \sum_{y=0}^{\infty} e^{sy}\left[(1-\theta^{y+1})^b - (1-\theta^y)^b\right] \\ &= \sum_{y=0}^{\infty} e^{sy}\left[\sum_{r=0}^{\infty}\frac{b(b-1)\cdots(b-r+1)(-\theta)^{r(y+1)}}{r!} - \sum_{r=0}^{\infty}\frac{b(b-1)\cdots(b-r+1)(-\theta)^{ry}}{r!}\right] \\ &= \sum_{r=0}^{\infty}\frac{b(b-1)\cdots(b-r+1)(-1)^r}{r!}\frac{\theta^r-1}{1-\theta^r e^s}. \end{aligned}$$

Hence,

$$c_t = E(e^{-Y_t}) = \sum_{r=0}^{\infty}\frac{b(b-1)\cdots(b-r+1)(-1)^r}{r!}\frac{\theta^r-1}{1-\theta^r e^{-1}}. \qquad (5)$$

If b is an integer, the result in (5) reduces to

$$c_t = E(e^{-Y_t}) = \sum_{r=1}^{b}\binom{b}{r}(-1)^r\frac{\theta^r-1}{1-\theta^r e^{-1}}.$$

Let Y_{it} (t = 1, 2; i = 1, 2, …, n; where n is the sample size) be a count response variable, and let $x_{it}' = (x_{it0}=1, x_{it1}, x_{it2}, \ldots x_{itk})$ be a vector of predictors. For a bivariate exponentiated-exponential geometric regression model, the joint probability distribution of (Y_{i1}, Y_{i2}) for any given (x_{i1}, x_{i2}) is that of BEEGD given in Equation (4). Suppose the parameter θ_t in (4) is a function of x_{it} given by $\theta_t(x_{it}) = f(x_{it}, \beta_t)$, where $0 < f(x_{it}, \beta_t) < 1$ is differentiable with respect to the vector parameter β_t. It is, in general, difficult to know which covariates affect each of the response variables Y_{i1} and Y_{i2}. To simplify the analysis, we assume that the same covariates affect each count response variable Y_{it}. Under this assumption, $x_{i1} = x_{i2} = \cdots = x_{id} = x_i$, however, the vector

parameters β_1 and β_2 are not assumed to be equal. We take $f(x_{it}, \beta_t)$ to be the logit function

$$\theta_t(x_i) = \theta_t = f(x_{it}, \beta_t) = \exp(x_i'\beta_t)/[1+\exp(x_i'\beta_t)] = 1/[1+\exp(-x_i'\beta_t)]. \quad (6)$$

This leads to the bivariate exponentiated-exponential geometric regression (BEEGR) model given by

$$P(y_{i1}, y_{i2} \mid x_i) = \prod_{t=1}^{2}\left[(1-[\theta_t(x_i)]^{y_{it}+1})^{b_t} - (1-[\theta_t(x_i)]^{y_{it}})^{b_t}\right] \times \left[1+\lambda(e^{-y_{i1}}-c_1)(e^{-y_{i2}}-c_2)\right], \quad (7)$$

where θ_t is given by (6) and $c_t = E(e^{-Y_t})$ is given by (5) for $t = 1, 2$. When both parameters $b_1 = b_2 = 1$, then the BEEGR model in (7) reduces to the bivariate geometric regression model given by

$$P(y_{i1}, y_{i2} \mid x_i) = \prod_{t=1}^{2}\left[[\theta_t(x_i)]^{y_{it}}\left(1-\theta_t(x_i)\right)\right] \times \left[1+\lambda(e^{-y_{i1}}-c_1)(e^{-y_{i2}}-c_2)\right],$$

where $c_t = (1-\theta_t(x_i))/(1-\theta_t(x_i)e^{-1})$.

The result in (7) can be extended to the multivariate exponentiated-exponential geometric regression (MEEGR) model and this is given by

$$P(y_1, y_2, \ldots, y_d) = \prod_{t=1}^{d}\left[(1-\theta_t^{y_t+1})^{b_t} - (1-\theta_t^{y_t})^{b_t}\right]\left[1+\sum_{t<v}^{d}\lambda_{tv}(e^{-y_t}-c_t)(e^{-y_v}-c_v)\right].$$

A count data may be truncated in such a way that there are no zeros. It is also possible that the proportion of zeros may be inflated. To address these types of situations, one can define the zero-truncated or zero-inflated model. We now define a zero-inflated BEEGR model for which the proportion of (0, 0) cell is too high. A zero-inflated BEEGR model has the probability mass function given by

$$f(y_{i1}, y_{i2} \mid x_i, z_i) = \begin{cases} \varphi_i + (1-\varphi_i)P(y_{i1}, y_{i2} \mid x_i), & y_{i1} = y_{i2} = 0 \\ (1-\varphi_i)P(y_{i1}, y_{i2} \mid x_i), & y_{i1} \text{ and } y_{i2} \text{ are not both zeros,} \end{cases} \quad (8)$$

where the probability φ_i is taken to be a function of covariates $z_i = (z_{i0} = 1, z_{i1}, \ldots, z_{i,m-1})'$ and it is defined by the logit function $\varphi_i = 1/[1+\exp(-z_i'\delta)]$, where δ is an m-dimensional vector $\delta = (\delta_0, \delta_1, \delta_2, \ldots, \delta_{m-1})'$ of parameters. The covariates z_i may be a subset of the x_i or may be completely different from the x_i. It is possible to assume that φ_i is a nuisance parameter instead of taking it to be a function of covariates z_i. The probability $P(y_{i1}, y_{i2} \mid x_i)$ in (8) is the BEEGR model given by (7) with

$$P(y_{i1} = 0, y_{i2} = 0 \mid x_i) = [1-\theta_1(x_i)]^{b_1}[1-\theta_2(x_i)]^{b_2}[1+\lambda(1-c_1)(1-c_2)].$$

By using a similar method that leads to equation (8), one can define a zero-truncated BEEGR model. A simple zero-truncated BEEGR model is a situation when both y_1 and y_2 are not allowed to be zeros, and it is given by

$$f(y_{i1}, y_{i2} \mid x_i, z_i) = P(y_{i1}, y_{i2} \mid x_i)/[1-P(y_{i1} = 0, y_{i2} = 0 \mid x_i)],$$

where y_{i1} and y_{i2} are not both zeros.

3. Maximum likelihood estimation of BEEGR model parameters

Suppose a random sample of size *n* is taken from the BEEGR model in (7). We now discuss the estimation of the BEEGR model parameters by the method of maximum likelihood. The log-likelihood function for the BEEGR model in (7) is given by

$$\ell(b_t, \beta_t, \lambda) = \ell = \sum_{i=1}^{n} \log P(y_{i1}, y_{i2} \mid x_i)$$

$$= \sum_{i=1}^{n} \left\{ \sum_{t=1}^{2} \log\left[(1-[\theta_t(x_i)]^{y_{it}+1})^{b_t} - (1-[\theta_t(x_i)]^{y_{it}})^{b_t} \right] + \log\left[1+\lambda(e^{-y_{i1}}-c_1)(e^{-y_{i2}}-c_2) \right] \right\}, \tag{9}$$

On taking the first partial derivatives of (9) with respect to the $(2k+3)$ parameters, we obtain the maximum likelihood equations. The second partial derivatives can be used to compute the Hessian matrix which is used to obtain the standard errors of the parameter estimates. In the application section, these maximum likelihood estimates are computed in SAS by using the NLMIXED procedure. This procedure also gives the AIC and BIC as well as the standard errors of the parameter estimates. The initial estimates can be obtained by first fitting the univariate EEGR model to each of the dependent variables (See Famoye and Lee, 2015). To fit the BEEGR, these initial estimates from the EEGR model can be combined with taking parameter λ to be 1 as the starting solutions for the BEEGR model.

Similar to taking the log-likelihood function of BEEGR model in (7), one can take the log-likelihood function for the zero-inflated BEEGR (ZIBEEGR) model in (8) to obtain

$$\ell_{zi}(b_t, \beta_t, \delta, \lambda) = \sum_{y_{i1}=y_{i2}=0} \log\left[\varphi_i + (1-\varphi_i) P(y_{i1}=0, y_{i2}=0 \mid x_i) \right]$$
$$+ \sum_{\Omega} \left[\log(1-\varphi_i) + \log P(y_{i1}, y_{i1} \mid x_i) \right],$$

where Ω is the set for which y_{i1} and y_{i2} are not both zeros. Note that the φ_i may be a function of covariates z_i.

4. Tests and goodness-of-fit statistics

In this section, we test for independence of the two count response variables y_1 and y_2. We will compare the BEEGR with bivariate geometric regression (BGR) model to determine whether BEEGR is more suitable. This test is equivalent to checking whether the count data exhibit any form of dispersion. For the BEEGR, b_1 and b_2 are both dispersion parameters. We will test if these parameters are equal to an unknown nuisance parameter. This test is equivalent to constant dispersion parameter for both count response variables. We will test for zero-inflation and briefly mention some goodness-of-fit statistics.

4.1 Test for Independence

The count response variables y_1 and y_2 are independent when the parameter λ is zero. For independence, we test the null hypothesis

$$H_0 : \lambda = 0 \text{ against } H_a : \lambda \neq 0. \tag{10}$$

Suppose L_{ind} is the likelihood function when the null hypothesis is true and L_a is the likelihood function when the null hypothesis is false. The test statistic for testing the hypothesis in (10) is given by $\chi^2_{ind} = -2\log(L_{ind} / L_a)$, which is approximately chi-squared with one degree of freedom. An alternative to using the chi-square test is to use the Wald asymptotic *t*-test, which is given by $t_{ind} = \hat{\lambda} / se(\hat{\lambda})$, where $\hat{\lambda}$ is the MLE of λ and $se(\hat{\lambda})$ is the standard error of $\hat{\lambda}$. The test statistic is asymptotically normal and H_0 is rejected when $|t_{ind}| \geq z_{\alpha/2}$.

4.2 Test for dispersion or test of BEEGR model against BGR model

The BEEGR model reduces to the bivariate geometric regression (BGR) model when the parameters $b_t = 1$ (t = 1, 2). To test if the BEEGR model should be used in place of BGR model, we test the hypothesis that b_t is 1. This is equivalent to a situation in which there is no dispersion. To test for no dispersion, we test the null hypothesis

$$H_0 : b_1 = b_2 = 1 \text{ against } H_a : H_0 \text{ is not true.} \quad (11)$$

Let L_{dis} be the likelihood function when the null hypothesis is true and L_a be the likelihood function when the null hypothesis is false. The test statistic for testing the hypothesis in (11) is given by $\chi^2_{dis} = -2\log(L_{dis} / L_a)$, which is approximately chi-squared with two degrees of freedom.

4.3 Test for constant dispersion parameter

The two dispersion parameters for the BEEGR model are b_1 and b_2. To test for a constant dispersion parameter, we test the null hypothesis

$$H_0 : b_1 = b_2 = b \text{ against } H_a : H_0 \text{ is not true.} \quad (12)$$

Suppose L_{con} is the likelihood function when the null hypothesis is true and L_a is the likelihood function when the null hypothesis is false. The test statistic for testing the hypothesis in (12) is given by $\chi^2_{con} = -2\log(L_{con} / L_a)$, which is approximately chi-squared with one degree of freedom.

4.4 Test for zero-inflation

If $\varphi_i = \varphi$ is a nuisance parameter, then it is not a function of the covariates. For this case, we test the null hypothesis

$$H_0 : \varphi = 0 \text{ against } H_a : \varphi \neq 0. \quad (13)$$

Suppose L_{01} is the likelihood function when the null hypothesis is true and L_a is the likelihood function when the null hypothesis is false. The test statistic for testing the null hypothesis in (13) is given by $\chi^2_{01} = -2\log(L_{01} / L_a)$, which is approximately chi-squared with one degree of freedom. One can also use the Wald asymptotic *t*-test given by $t_* = \hat{\varphi} / se(\hat{\varphi})$, where $\hat{\varphi}$ is the MLE of φ and $se(\hat{\varphi})$ is the standard error of $\hat{\varphi}$.

If φ_i is a function of the covariates, then we have $\varphi_i = 1/[1+\exp(-z_i'\delta)]$, where δ is an *m*-dimensional vector of parameters. For this case, we test the null hypothesis

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \cdots = \delta_{m-1} = 0 \text{ against } H_a : H_0 \text{ is not true.} \quad (14)$$

Suppose L_{02} is the likelihood function when the null hypothesis is true and L_a is the likelihood function when the null hypothesis is false. The test statistic for testing the null hypothesis in (14) is given by $\chi^2_{02} = -2\log(L_{02} / L_a)$, which is approximately chi-squared with *m* degrees of freedom.

4.5 Goodness-of-fit statistics

A goodness-of-fit statistic for the BEEGR is the log-likelihood statistic in (9). In addition to the log-likelihood, alternative measures of goodness-of-fit are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These statistics, which are based on the log-likelihood statistic, are defined as follows: The AIC is defined as $\text{AIC} = -2\ell + 2p$ while the BIC is defined as $\text{BIC} = -2\ell + p\log(n)$, where *n* is the sample size, *p* is the number of estimated parameters in the model, and ℓ is the log-likelihood statistic in (9). Both the AIC and BIC take into consideration the number of parameters in the regression model to control over-parameterization. These measures are provided by SAS NLMIXED procedure.

A goodness-of-fit statistic for the BEEGR can be based on the Pearson's chi-squared statistic, which is defined as $\chi^2 = \sum_{i,j}(O_{ij} - E_{ij})^2 / E_{ij}$, where O_{ij} is the observed frequency in cell (*i*, *j*) and E_{ij} is the expected frequency in cell (*i*, *j*). The expected frequency is calculated by summing all the probabilities $P(Y_1 = i, Y_2 = j)$ for all observations in the data set. Note that the probabilities are not the same for two observations, except if the two observations have exactly the same values for all predictor variables. If the expected values are too small, one can combine some of the cells. But there is no unique way to combine some of the cells.

5. Application

In this section we apply a domestic violence data to illustrate the usefulness of the BEEGR model and compare the results with that of BGPR model, which is a special case of the multivariate generalized Poisson regression model defined by Famoye (2015). The BGPR model is given by

$$P(y_{i1}, y_{i2} \mid x_i) = \prod_{t=1}^{2}\left(\frac{\mu_{it}}{1+b_t\mu_{it}}\right)^{y_{it}} \frac{(1+b_t y_{it})^{y_{it}-1}}{y_{it}!}\exp\left[\frac{-\mu_{it}(1+b_t y_{it})}{1+b_t\mu_{it}}\right] \times\left[1+\lambda(e^{-y_{i1}}-c_1)(e^{-y_{i2}}-c_2)\right], \quad (15)$$

where mean $E(Y_{it} \mid x_i) = \mu_{it}(x_i) = \exp(x_i'\beta_t)$ for *t* = 1, 2, b_t is a dispersion parameter, and c_t is $c_t = \exp[\mu_{it}(s_t - 1)/(1+b_t\mu_{it})]$, with $\ln s_t - b_t\mu_{it}(s_t - 1)/(1+b_t\mu_{it}) + 1 = 0$.

The BGPR model when $y_1 = y_2 = 0$ is given by

$$P(y_{i1} = 0, y_{i2} = 0 \mid x_i) = \exp\left[-\frac{\mu_{i1}}{1+b_1\mu_{i1}} - \frac{\mu_{i2}}{1+b_2\mu_{i2}}\right]\times\left[1+\lambda(1-c_1)(1-c_2)\right].$$

Similar to the zero-inflated BEEGR model, one can write down the probability mass function for the zero-inflated BGPR by using the above results. We will first describe the data used for the analysis and then provide the results of the data analysis.

5.1 Description of the Data

In 1995-1996, the National Violence Against Women (NVAW) Survey was conducted and a public-use data set was obtained. Completed interviews were obtained from men and women, but the data used in this paper is a subset of the 8000 interviews from women who were 18 years of age or older residing in United States households. Respondents to the survey were asked questions on various topics including (a) their general fear of violence and how they managed their fears, (b) emotional abuse they had experienced by their partners and (c) physical assault they had experienced as adults by any type of perpetrator. For more details, the reader is referred to Tjaden and Thoennes (1999), ICPSR 2566.

The count response variables used in the data analysis are control (this will be denoted by y_1) and physical assault or violence (this will be denoted by y_2). Control is the total number of controlling behaviors the current partner and/or all former partners exerted on the woman. A controlling behavior is made up of thirteen actions and this variable ranges from 0 to 12. Violence is the number of incidents of physical assault. This is the total number of twelve possible violent physical actions directed toward a woman by her current and/or former partners. This variable also ranges from 0 to 12. A high score on any of these variables indicates the woman experienced severe control or violence. See Cheng and Lo (2015) who used some of the variables to examine racial disparities in women's experience of intimate partner violence.

The eight explanatory variables used in the data analysis are as follows: age in years; level of education is one of the seven school levels (0 = no schooling to 6 = postgraduate); race (1 = white, 0 = others); number of children under 18 years of age (nchild); respondent's income level is one of 10 levels (1 = below \$5,000 to 10 = over \$100000); being stalked (stalk) is a binary variable with 1 = yes and 0 = no; health level is one of 5 levels (0 = poor to 4 = excellent), and drug is a binary variable that indicates illicit drug use with 1 = yes and 0 = no. The variable drug indicates if a woman had used marijuana, cocaine, heroin, angel dust, etc. in the past month. After excluding the cases having missing information, we have 4171 observations. The descriptive statistics for the variables used in the analysis are given in Table 1. A simple correlation between control and violence is computed and it is 0.244 with a p-value of less than 0.0001. Thus, both variables are significantly correlated. By using the sample means and sample variances of y_1 and y_2, both response variables appear to be over-dispersed.

Table 1: Descriptive statistics for the variables

Variable	Description	Mean ± SD	Proportion of 1's
age	Age in years	41.92 ± 13.32	
educ	Education level	3.85 ± 1.14	
race	Race		0.856
nchild	Number under 18 yrs.	1.09 ± 1.23	
income	1995 income level	4.07 ± 2.53	
stalk	Ever been stalked		0.119
health	Health condition	2.82 ± 1.05	
drug	Illicit drug use		0.013
control (y_1)	Dependent variable	0.88 ± 1.60	
violence (y_2)	Dependent variable	1.09 ± 2.20	

SD = standard deviation

5.2 Data Analysis and Results

The BEEGR model in (7) and the BGPR model in (15) are applied to fit the data with 'control' and 'violence' as the two response variables. We also applied their zero-inflated regression models. We computed the expected frequencies for BEEGR, ZIBEEGR and ZIBGPR and compared them with the observed frequencies. Even though the ZIBGPR provided the best expected frequency for the (0, 0) cell, but overall, its expected frequencies for many cells are furthest from the corresponding observed frequencies. The response variables y_1 and y_2 range from 0 to 12. The observed frequencies for possible combinations of y_1 and y_2 values range from 0 to 1960. Because of the small expected frequencies, we combine all classes for $y_1 \geq 7$ and all classes for $y_2 \geq 9$ and this lead to an 8 by 10 contingency table. The chi-square statistics based on this contingency table is computed. The chi-square values for ZIBGPR, ZIBEEGR and BEEGR models are respectively 403.97, 254.60, 252.97. The ZIBEEGR and BEEGR models provide a much closer expected values to the observed frequencies than the ZIBGPR model.

In Table 2, we report part of the 8 by 10 contingence table. Table 2 shows the observed and expected frequencies for the data where majority (over 80%) of the observed frequencies are distributed. From Table 2, the model with the worst expected frequency is ZIBGPR. Even though, the definition of BEEGR model does not include a special consideration for the zero inflation, it does well in taking care of the zero-inflation when the data has such a characteristic.

Table 2: Observed and expected frequencies for ZIBGPR, ZIBEEGR and BEEGR

$y_1 \backslash y_2 \rightarrow$		0	1	2	3
0	Observed	1960	170	109	82
	ZIBGPR	1962.06	251.51	121.51	68.87
	ZIBEEGR	1975.88	219.59	115.79	71.26
	BEEGR	1941.96	255.55	121.48	70.68
1	Observed	560	69	40	42
	ZIBGPR	392.50	144.70	71.94	41.37
	ZIBEEGR	428.28	110.36	64.34	41.28
	BEEGR	477.52	106.19	61.78	39.06
2	Observed	189	32	23	20
	ZIBGPR	200.95	76.14	38.42	22.34
	ZIBEEGR	210.08	59.70	36.03	23.53
	BEEGR	208.97	56.26	34.51	22.38
3	Observed	85	23	19	15
	ZIBGPR	105.52	40.59	20.72	12.19
	ZIBEEGR	110.52	32.72	20.10	13.29
	BEEGR	104.67	30.47	19.17	12.66

The fit by the BEEGR, ZIBEEGR and ZIBGPR are reported in Table 3. The log-likelihood for the BGPR model (which is not reported) is -10414.065. This value is worse than any of the values reported in Table 3 for BEEGR, ZIBEEGR and ZIBGPR. The fit by BGPR is very poor, hence the ZIBGPR is applied and it provides a much better fit by using the AIC. From Table 3, the model with the best fit is the ZIBEEGR by using the AIC. The BEEGR model performs better than the ZIBGPR model.

In all the regression models, including BGPR model, only the ZIBGPR model shows an insignificant correlation parameter (see Table 2). This is quite a surprise and we have no explanation for it. Both the BEEGR and ZIBEEGR in Table 2 has significant correlation parameter. All the tests proposed in sections 4.1 to 4.4 are significant when tested for both the ZIBGPR and ZIBEEGR models. The test with the smallest statistic is the zero-inflation for ZIBEEGR model. The observed value of the test statistic is 70.97, and the statistic has a chi-square distribution with 9 degrees of freedom. For this test, the p-value is less than 0.0001.

Table 2: Parameter estimates (standard errors in parentheses) for BEEGR and BGPR

Variable	ZIBGPR model	ZIBEEGR model	BEEGR model
constant (x_{10})	1.295 (0.170)*	1.349 (0.171)*	1.750 (0.160)*
age (x_{11})	-0.003 (0.002)	-0.003 (0.002)	-0.007 (0.002)*
educ (x_{12})	-0.128 (0.028)*	-0.141 (0.029)*	-0.156 (0.025)*
race (x_{13})	-0.296 (0.073)*	-0.294 (0.073)*	-0.373 (0.069)*
nchild (x_{14})	0.065 (0.025)*	0.067 (0.025)*	0.056 (0.023)*
income (x_{15})	-0.007 (0.013)	-0.008 (0.013)	0.0002 (0.011)
stalk (x_{16})	0.133 (0.074)	0.205 (0.073)*	0.365 (0.074)*
health (x_{17})	-0.120 (0.028)*	-0.125 (0.028)*	-0.169 (0.026)*
drug (x_{18})	0.484 (0.196)*	0.518 (0.193)*	0.535 (0.194)*
constant (x_{20})	1.525 (0.263)*	2.029 (0.227)*	2.463 (0.218)*
age (x_{21})	-0.014 (0.004)*	-0.012 (0.003)*	-0.017 (0.003)*
educ (x_{22})	-0.100 (0.039)*	-0.104 (0.035)*	-0.118 (0.033)*
race (x_{23})	-0.050 (0.101)	-0.033 (0.091)	-0.125 (0.090)
nchild (x_{24})	0.0002 (0.034)	0.013 (0.030)	0.002 (0.029)
income (x_{25})	-0.001 (0.017)	-0.002 (0.015)	0.004 (0.014)
stalk (x_{26})	0.868 (0.107)*	0.896 (0.089)*	1.107 (0.092)*
health (x_{27})	-0.111 (0.036)*	-0.111 (0.033)*	-0.156 (0.032)*
drug (x_{28})	0.500 (0.311)	0.527 (0.251)*	0.584 (0.262)*
constant (z_0)	-2.880 (0.423)*	-3.921 (0.755)*	
age (z_1)	0.017 (0.005)*	0.024 (0.008)*	
educ (z_2)	0.121 (0.057)*	0.102 (0.088)	
race (z_3)	0.422 (0.177)*	0.623 (0.329)	
nchild (z_4)	0.020 (0.051)	0.053 (0.075)	
income (z_5)	-0.023 (0.025)	-0.032 (0.036)	
stalk (z_6)	-2.552 (0.909)*	-17.0 (1284.9)	
health (z_7)	0.239 (0.060)*	0.297 (0.094)*	
drug (z_8)	-0.809 (0.661)	-0.846 (1.176)	

Dispersion ($\hat{b}_1$)	0.413 (0.033)*	0.863 (0.053)*	0.627 (0.025)*
Dispersion ($\hat{b}_2$)	0.856 (0.056)*	0.392 (0.022)*	0.298 (0.012)*
Correlation ($\hat{\lambda}$)	0.141 (0.186)	0.625 (0.150)*	1.310 (0.101)*
Log-likelihood	-10331.9153	-10235.4025	-10270.8863
AIC	20724.0	20531.0	20584.0

*Significant at 5% level

The two dispersion parameters are significant indicating that the bivariate Poisson regression model and the bivariate geometric regression model will not perform well in fitting this data set. The dispersion parameter estimates lie between 0 and 2, which indicates over-dispersion. The data is over-dispersed and hence, the bivariate negative binomial regression can be used as an alternate model. However, it cannot handle situations with under-dispersion. This is one main advantage possessed by the BGPR and BEEGR models.

6. Summary and Conclusion

A new bivariate count data regression model, the BEEGR, is defined and studied. The model can be applied to fit data with over-dispersion or under-dispersion relative to the Poisson assumption. The parameter measuring the association between the two response variables can be positive or negative. Thus, the model allows for positive or negative correlation.

It is interesting to note that the BGPR model did not perform as well as the ZIGPR model for the domestic violence data. We notice that the likelihood ratio test for zero-inflation in both ZIBGPR and ZIBEEGR models show significant results. In examining the predicted zero proportion, the ZIBGPR model provided the best prediction while the ZIBEEGR over-estimated the zero proportion. The ZIBGPR model has six of the nine parameters measuring the zero-inflation to be significant at 5%. On the other hand, the ZIBEEGR model shows that three of the nine parameters measuring zero-inflation are significant. This may not be unconnected with why the BEEGR model provided good expected frequencies to the data when compared with the zero-inflated models. It is conjectured that BEEGR model seems to perform well for cases where the data appears to show zero-inflation. Future work to investigate this conjecture will be undertaken.

A limitation of the Sarmanov bivariate/multivariate regression model is that the correlation coefficient could be restricted to a subset of the interval [-1, 1] depending on the parameters of the marginal distribution (Lee, 1996). A disadvantage of the BEEGR model is that its mean and variance are not in closed forms. The advantages of the BEEGR model includes a likelihood function that is in closed form. The parameter estimation is less time-consuming. The example provided in the paper is on a bivariate data with positive correlation. Will the BEEGR model perform well for negative correlation? This is a problem which will be explored in future work.

References

Alzaatreh, A., Lee, C. and Famoye, F. (2013) A new method for generating families of continuous distributions. *Metron*, 71(1), 63-79.

Alzaatreh, A, Lee, C. and Famoye, F. (2012) On the discrete analogues of continuous distributions. *Statistical Methodology*, 9, 589-603.

Cameron, A.C. and Trivedi, P.K. (2013) *Regression Analysis of Count Data, 2nd Edition*. Cambridge University Press, New York, NY.

Cheng, T.C. and Lo, C.C. (2015) Racial disparities in intimate partner violence and seeking help with mental health, *Journal of Interpersonal Violence*, 30(18), 3283-3307.

Famoye, F. (2015) A multivariate generalized Poisson regression model. *Communications in Statistics - Theory & Methods*, 44, 497-511.

Famoye, F. (2010a) A new bivariate generalized Poisson distribution. *Statistica Neerlandica*, 64(1), 112-124.

Famoye, F. (2010b) On the bivariate negative binomial regression model. *Journal of Applied Statistics*, 37(6), 969-981.

Famoye, F. (1993) Restricted generalized Poisson regression model. *Communications in Statistics - Theory & Methods*, 22(5), 1335-1354.

Famoye, F. and Lee, C. (2015) Exponentiated exponential-geometric regression model. In *Proceedings of 60th ISI World Statistics Congress, 2015*, Rio de Janeiro, Brazil.

Hofer, V. and Leitner, J. (2012) A bivariate Sarmanov regression model for count data with generalized Poisson marginal. *Journal of Applied Statistics*, 39(12), 2599-2617.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*. John Wiley and Sons, Inc., New York, NY

Kocherlakota, S. and Kocherlakota, K. (1992) *Bivariate Discrete Distributions*. Marcel Dekker, Inc., New York, NY.

Lakshminarayana, J., Pandit, S.N.N. and Rao, K.S. (1999) On bivariate Poisson distribution. *Communications in Statistics - Theory & Methods*, 28, 267-276.

Lee, M.-L.T. (1996) Properties and applications of the Sarmanov family of bivariate distributions. *Communications in Statistics - Theory and Methods*, 25, 1207-1222.

Sarmanov, O.V. (1966) Generalized normal correlation and two-dimensional Fréchet classes. *Soviet Mathematics Doklady*, 168, 596-599.

Tjaden, P. and Thoennes, N. (1999) Violence and threats of violence against women and men in the United States, 1994-1996 [Computer file]. ICPSR version. Denver, CO: Center for Policy Research [producer], 1998. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1999.

Winkelmann, R. (2008) *Econometric Analysis of Count Data* (5th ed.). Berlin: Springer Verlag.