A Multivariate Negative Binomial Regression Model

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

Abstract
A multivariate negative binomial regression model based on the multivariate negative binomial distribution is defined and studied. The regression model can be used to describe a count data with over-dispersion. The model allows for both positive and negative correlation between any pair of the response variables. The parameters of the regression model are estimated by using the maximum likelihood method. Some test statistics are discussed and a numerical data set is used to illustrate the applications of the multivariate count data regression model.

Key Words: count data; correlated data; over-dispersion; estimation.

1. Introduction

Many univariate count data regression models have been defined and studied. Some of these models have been extended to bivariate and few have been extended to multivariate count data models. See the books by Cameron and Trivedi (2013) and Winkelmann (2008) and the references there-in. The univariate Poisson regression model has been extended to the multivariate Poisson regression (MPR) model, see for example Winkelmann (2008, p. 205). The MPR model assumes that the conditional mean and the conditional variance of the response variable are equal. Thus, the MPR model does not allow for over- or under-dispersion. Over-dispersion (or under-dispersion) relative to the Poisson distribution is a condition in which the conditional variance is more (or less) than the conditional mean.

According to Cameron and Trivedi (2013, p. 311), there is no unique multivariate Poisson distribution. There are many ways to derive a multivariate model with Poisson marginals. A multivariate Poisson distribution can be obtained by the method of mixtures and convolutions. For example, the bivariate Poisson distribution was obtained through the method of trivariate reduction (Kocherlakota and Kocherlakota (1992, p.88)). Thus, an m-variate Poisson distribution can be obtained by the method of $(m+1)$-variate reduction. This m-variate Poisson distribution only allows positive correlation between the count response variables.

According to Alfò and Trovato (2004), the univariate approach is insufficient and it has to be extended when the primary focus of analysis of multivariate count data is to describe association among the counts. Three of the five multivariate count regression models discussed by Winkelmann (2008, Chapter 7) are the MPR model, the multivariate negative binomial regression (MNBR) model and the multivariate Poisson-gamma mixture (MPGM) model that allow non-negative correlations. These multivariate count regression models are obtained through the method of variate reduction. The MNBR and MPGM models allow for over-dispersion. A potential disadvantage of the MPGM is that its covariances cannot be determined independently of its dispersion. However, the covariances and the dispersion for MNBR can be independently determined. A disadvantage of the MNBR and MPGM models is that each is characterized by one dispersion parameter. In view of this, one cannot determine if a response variable is over-dispersed or under-dispersed.

Attention has shifted to multivariate mixing models, such as Poisson-lognormal regression model (Chib and Winkelmann, 2001) in response to the drawbacks of only positive correlation and equi-dispersion. Estimation of multivariate Poisson-lognormal regression (MPLR) model involves numerical integration and this can be extremely time-consuming, especially for large number of response variables. Several authors (e.g. Ma et al., 2008) suggested using the Bayesian method to estimate the parameters of MPLR. Karlis and Meligkotsidou (2005) proposed a MPR model which allows for different covariances between the pairs of the count variables, which is an improvement on the multivariate distribution defined by Karlis (2003), where all the pairs have the same covariance. The model by Karlis and Meligkotsidou (2005) does not allow for negative correlation or over-dispersion. Another approach to extend the univariate count data regression to the multivariate model is to model dependence among count variables through copula functions (see for examples van Ophem (1999), Lee (1999) and Cameron et al. (2004)).

The properties of the Sarmanov (1966) bivariate distributions were discussed by Lee (1996), who gave as an example, the bivariate Poisson distribution which was later discussed by Lakshminarayana et al. (1999). Suppose $h_i(y_i)$, i = 1, 2, are univariate probability mass functions. If $\varphi_i(t)$, is a bounded non-constant function such that $\Sigma_t \varphi_i(t) h_i(t) = E[\varphi_i(T)] = 0$, i = 1, 2, then the function $h(y_1, y_2) = h_1(y_1) h_2(y_2) \{1 + \omega\varphi_1(y_1)\varphi_2(y_2)\}$ defined by Sarmanov (1966) is a bivariate joint probability mass function, where ω is a real number satisfying the condition 1 + $\omega\varphi_1(y_1)\varphi_2(y_2) \geq 0$ for all y_1 and y_2. One general method given by Lee (1996) for finding φ_i is by using the Laplace transform of h_i. By definition, the Laplace transform of h_i is given by $L_i(t) = E(e^{-tY_i})$. Thus, one can define $\varphi_i(y_i) = e^{-y_i} - L_i(1) = e^{-y_i} - E(e^{-Y_i})$.

Lee (1996) extended the bivariate Sarmanov distribution to multivariate case with several parameters measuring covariances of order 2, 3, …, *m* for *m*-variate distribution. Miravete (2009) presented two multivariate count data regression based on the Sarmanov family of distributions. The two models have the double Poisson and the gamma count distributions as their marginals. The mean and variance of the models are either not exact or they are not in closed forms (see Winkelmann (2008), pp. 49 and 59).

Famoye (2010a) defined a new bivariate negative binomial regression model as a product of negative binomial marginals with a multiplicative factor. This is a bivariate Sarmanov regression model with negative binomial marginals. The correlation between the two negative binomial variates can be positive, zero or negative. However, the regression model only allows for over-dispersion.

In this paper, we define a multivariate negative binomial regression (MNBR) model based on Sarmanov multivariate negative binomial distribution. Among the importance of the multivariate count data regression model are (i) it allows for any type of correlation between any two variates, (ii) it allows correlations and dispersions to be determined independently and (iii) the means and the variances of the variates are in closed forms. The MNBR model is defined in section 2. We discuss parameter estimation for the MNBR in section 3. We present some tests in section 4 and in section 5; one numerical dataset is used to illustrate the MNBR model. Finally in section 6, some concluding remarks are given.

2. The Multivariate Negative Binomial Regression Model

The probability mass function of the negative binomial distribution (NBD) is given by

$$P(Y = y) = \binom{m^{-1} + y - 1}{y} \theta^{y} (1-\theta)^{m^{-1}}, \; y = 0, 1, 2, 3, \ldots \qquad (1)$$

for $0 < \theta < 1$ and $m > 0$. The model reduces to the Poisson distribution when the dispersion parameter $m \to 0$. The NBD in (1) is over-dispersed when $m > 0$.

Famoye (2010a) defined a bivariate negative binomial regression (BNBR) model as a product of negative binomial marginals. The regression model is given as

$$P(y_1, y_2) = \prod_{t=1}^{2} \binom{m_t^{-1} + y_t - 1}{y_t} \left(\frac{\mu_t}{m_t^{-1} + \mu_t} \right)^{y_t} \left(\frac{m_t^{-1}}{m_t^{-1} + \mu_t} \right)^{m_t^{-1}} \left[1 + \lambda (e^{-y_1} - c_1)(e^{-y_2} - c_2) \right], \qquad (2)$$

where $c_t = E(e^{-Y_t}) = [(1-\theta_t)/(1-\theta_t e^{-1})]^{m_t^{-1}}$, $m_t > 0$, μ_t is the conditional mean of y_t given the predictor variables, and $\theta_t = \mu_t / (m_t^{-1} + \mu_t)$, ($t$ = 1, 2). By using a similar approach, a new multivariate negative binomial regression (MNBR) model, which is an extension of the model in (2) can be defined as

$$P(\underline{y}) = \prod_{t=1}^{d} \binom{m_t^{-1} + y_t - 1}{y_t} \left(\frac{\mu_t}{m_t^{-1} + \mu_t} \right)^{y_t} \left(\frac{m_t^{-1}}{m_t^{-1} + \mu_t} \right)^{m_t^{-1}} \left[1 + \sum_{t<v}^{d} \lambda_{tv} (e^{-y_t} - c_t)(e^{-y_v} - c_v) \right]. \qquad (3)$$

When λ_{tv} = 0 the pair Y_t and Y_v are independent. When λ_{tv} > 0, the model in (3) allows positive correlation and when λ_{tv} < 0, the model allows negative correlation. The parameter m_t measures the dispersion. If $m_t \to 0$ for all *t*, the multivariate negative binomial regression model reduces to the multivariate Poisson regression model. The variable Y_t is over-dispersed when m_t > 0.

In the MNBR model in (3), Y_{it} (t = 1, 2, …, d; i = 1, 2, …, n; and n is the sample size) is a count response variable, and $x_{it}' = (x_{it0} = 1, x_{it1}, x_{it2}, \ldots, x_{itk})$ is a vector of covariates. Thus, the joint distribution of $(Y_{i1}, Y_{i2}, \ldots, Y_{id})$ for any given $(x_{i1}, x_{i2}, \ldots, x_{id})$ is that of MNBR with mean

$$E(Y_{i1} \mid x_{i1}) = \mu_{i1}(x_{i1}) = \gamma_{i1} f(x_{i1}, \beta_1), \; \ldots, \; E(Y_{id} \mid x_{id}) = \mu_{id}(x_{id}) = \gamma_{id} f(x_{id}, \beta_d), \qquad (4)$$

where $f(x_{it}, \beta_t) > 0$ (t = 1, 2, …, d) is a known function of x_{it} and a vector $\beta_t^T = (\beta_{t0}, \beta_{t1}, \beta_{t2}, \ldots, \beta_{tk})$ of regression parameters, and γ_{it} is a measure of exposure. The function $f(x_{it}, \beta_t)$ is differentiable with respect to β_t. It may be difficult to know which covariates affect each or a combination of Y_1, Y_2, …, Y_d. To simplify our analysis in this paper, we will assume that the same covariates affect every count response variable Y_t. Under this assumption, $x_{i1} = x_{i2} = \cdots = x_{id} = x_i$, however, the parameter vectors β_1, β_2, …, β_d are not assumed to be equal.

The marginal distribution of Y_t (t = 1, 2, …, d) in (3) is negative binomial with mean $\mu_t = m_t^{-1}\theta_t / (1-\theta_t)$ and variance $\sigma_t^2 = m_t^{-1}\theta_t / (1-\theta_t)^2$. To find the covariance between Y_t and Y_ν, we need $E(Y_t Y_\nu)$. The covariance between any pairs Y_t and Y_ν is $\lambda_{t\nu} c_t c_\nu A_t A_\nu$, where $A_t = m_t^{-1}\theta_t e^{-1} / (1-\theta_t e^{-1}) - m_t^{-1}\theta_t / (1-\theta_t)$ (t = 1, 2, …, d). The $d(d-1)/2$ correlation coefficients between any two variates are given by $\rho_{t\nu} = \lambda_{t\nu} c_t c_\nu A_t A_\nu / (\sigma_t \sigma_\nu)$, ($t$, ν, = 1, 2, …, d and $t < \nu$). Thus, the parameter $\lambda_{t\nu}$ can be written in terms of the correlation coefficient $\rho_{t\nu}$. For the marginal bivariate distribution of Y_t and Y_ν, $1+\lambda_{t\nu}(1-c_t)(1-c_\nu) \geq 0$ since $P(y_t, y_\nu) \geq 0$. Therefore, $\lambda_{t\nu}$ satisfies $|\lambda_{t\nu}| \leq 1/[(1-c_t)(1-c_\nu)]$, which allows $\rho_{t\nu}$ to take on negative, zero or positive values. By using this result, the correlation coefficient satisfies the condition $|\rho_{t\nu}| \leq \min(1,\ c_t c_\nu A_t A_\nu / [\sigma_t \sigma_\nu (1-c_t)(1-c_\nu)])$. Note that the second quantity in the minimum function may exceed 1, especially when c_t and c_ν are very close to 1.

The covariances and the dispersion parameters for the MNBR in (3) can be determined independently. This is an advantage over the multivariate count regression models derived through variate reduction in Winkelmann (2008).

The model in (3) belongs to the multivariate Sarmanov family with parameters measuring covariances of order 2. It is straightforward to extend the model in (3) to include parameters measuring covariances of order 3, 4, …, d as proposed by Lee (1996). In order to keep the number of covariance parameters to a minimum, we will consider in this paper the MNBR in (3).

3. Parameter Estimation

Suppose we have n independent vectors ($y_{i1}, y_{i2}, \ldots, y_{id}$), where the i-th vector has the MNBR model in (3). The log-likelihood function, $\log L(\mu; \underline{y}) = \log L(\mu, m, \lambda; \underline{y})$, for the MNBR model is

$$\log L(\mu; \underline{y}) = \sum_{i=1}^{n} \left\{ \sum_{t=1}^{d} \left[y_{it} \log\left(\frac{m_t \mu_{it}}{1+m_t \mu_{it}} \right) - m_t^{-1} \log(1+m_t \mu_{it}) + \log \Gamma(m_t^{-1} + y_{it}) \right] \right.$$
$$\left. - \sum_{t=1}^{d} \left[\log \Gamma(m_t^{-1}) + \log(y_{it}!) \right] + \log\left[1 + \sum_{t<\nu}^{d} \lambda_{t\nu} (e^{-y_{it}} - c_t)(e^{-y_{i\nu}} - c_\nu) \right] \right\}. \qquad (5)$$

The log-likelihood in (5) is maximized over the parameters β_t, m_t (t = 1, 2, …, d), and $\lambda_{t\nu}$ (t, ν, = 1, 2, …, d and $t < \nu$). It is not difficult to obtain the first and second partial derivatives with respect to the parameters. On taking the expectations of the second partial derivatives and multiplying them by –1, we obtain the Fisher information matrix. For estimation purposes, it is better to express the parameter $\lambda_{t\nu}$ in terms of the correlation coefficient $\rho_{t\nu}$ since the correlation coefficient lies between –1 and +1. One can use the Newton-Raphson (or any other optimization routine) iterative technique to obtain the maximum likelihood estimates of the MNBR parameters. If the optimization routine requires an initial estimates for the model parameters β_t, m_t (t = 1, 2, …, d), and $\rho_{t\nu}$ (t, ν, = 1, 2, …, d and $t < \nu$), maximum likelihood estimates from the univariate NBR models can be used for β_t and m_t. The initial estimate of $\rho_{t\nu}$ may be taken as the sample correlation coefficient or zero. The asymptotic Wald statistic for testing the significance of each model parameter can be computed. In the numerical example provided in section 5, we expressed $\lambda_{t\nu}$ in terms of $\rho_{t\nu}$.

We used the PROC NLMIXED in SAS to maximize the log-likelihood function. The Hessian matrix is obtained by taking the second partial derivatives of (5), multiplying by –1 and finally evaluating at the maximum likelihood estimates. In PROC NLMIXED, the inverse of this Hessian matrix is the estimated covariance matrix of the parameter estimates. In addition to the goodness of fit statistics, PROC NLMIXED gives the parameter estimates with their standard errors, which are the square roots of the diagonal entries in the estimated covariance matrix.

A measure of goodness of fit for the MNBR may be based on the log-likelihood statistic given in (5). In order to account for the number of estimated parameters in the MNBR model, one can use the Akaike Information Criterion (AIC) which has the form $\text{AIC} = -2\log L(\mu; \underline{y}) + 2p$, where p is the number of estimated parameters in the model. A model with a smaller AIC is generally preferred.

4. Some Tests

In this section, we are interested in some hypotheses on the MNBR model. We will test for the independence of the two random variables Y_t and Y_ν (t, ν, = 1, 2, …, d and $t \neq \nu$). We will also compare the MPR model with the MNBR model to determine if MNBR model is more suitable. The last test will be used to check if the dispersion parameters are the same.

4.1 Test for Independence

A pair of random variables Y_t and Y_ν are independent when the parameter $\lambda_{t\nu}$ (or $\rho_{t\nu}$) is equal to zero. For Y_t and Y_ν to be independent, we test the null hypothesis

$$H_0 : \lambda_{t\nu} = 0 \text{ against } H_a : \lambda_{t\nu} \neq 0. \qquad (6)$$

Let L_{ind} be the likelihood function under H_0 and let L_a be the likelihood function under H_a. The test statistic is given by $\chi_{ind} = -2\log(L_{ind} / L_a)$, which has an approximate chi-square distribution with 1 degree of freedom. The hypotheses in (6) can be extended to test for pairwise independence. For this situation, the null hypothesis becomes $H_0 : \lambda_{t\nu} = 0$ for all pairs (t, ν), where t, ν = 1, 2, …, d. Suppose $L_{p.ind}$ is the likelihood function under H_0, the test statistic is given by $\chi_{p.ind} = -2\log(L_{p.ind} / L_a)$, which has an approximate chi-square distribution with $d(d$ – 1)/2 degrees of freedom. An alternative to using the likelihood ratio statistic to test the null hypothesis in (6) is to use a score statistic. The reader is referred to Famoye (2010b) for the score test for the BGPD.

4.2 Test of MPR Model Against MNBR Model

The MNBR model in (3) reduces to the MPR model when the parameters $m_t \to 0$ (t = 1, 2, …, d). To assess the usefulness of the MNBR model over the MPR model, one can test the null hypothesis that all m_t is zero against the alternative that at least one m_t > 0. When all $m_t \to 0$, this corresponds to a situation in which there is no dispersion. To test for no dispersion in any of the d-variates, we test the null hypothesis

$$H_0 : m_t = 0 \ (t = 1, 2, \ldots, d) \text{ against } H_a : H_0 \text{ is not true.} \qquad (7)$$

If L_{dis} is the likelihood function under H_0 and L_a is the likelihood function under H_a, the test statistic $\chi_{dis} = -2\log(L_{dis} / L_a)$ cannot be approximated by chi-square distribution with d degrees of freedom because m_t = 0 are on the boundary of the parameter space. By using the results of Chernoff (1954) [see also Self and Liang (1987)], the test statistic χ_{dis} is asymptotically distributed as a random variable which has a probability mass of 2^{-d} at the point 0, and a mixture of chi-square distributions above zero. The mixing probability for the χ_k^2 component is ${}^dC_k 2^{-d}$, where k = 1, 2, …, d. If d = 4, we have a mixture of $\frac{1}{4}\chi_1^2$, $\frac{3}{8}\chi_2^2$, $\frac{1}{4}\chi_3^2$, and $\frac{1}{16}\chi_4^2$ for probability above zero.

4.3 Test for Homogeneity of Dispersion Parameters

In the formulation of the MNBR, we have the dispersion parameters $m_t > 0$ (t = 1, 2, …, d). To test for constant dispersion or homogeneity of dispersion parameters, we test the hypothesis

$$H_0 : m_1 = m_2 = \cdots = m_d = m \text{ against } H_a : H_0 \text{ is not true.}$$

Let L_{con} be the likelihood function under H_0 and let L_a be the likelihood function under H_a. The test statistic is given by $\chi_{con} = -2\log(L_{con} / L_a)$, which has an approximate chi-square distribution with d – 1 degrees of freedom.

5. Applications

In this section, the MNBR defined in this paper will be illustrated with one numerical multivariate count dataset. Cameron et al. (1988) analyzed various measures of health-care utilization by using a sample of 5190 single-person households from the 1977-78 Australian Health Survey. The data are obtained from the Journal of Applied Econometrics 1997 Data Archive. Various authors, including Mullahy (1997) and Cameron and Johansson (1997) used the data to illustrate univariate regression models. Gurmu and Elder (2000) used the data to illustrate a generalized bivariate negative binomial regression model. Famoye (2010a) used the data to illustrate the Sarmanov bivariate negative binomial regression model. Famoye (2013) used the data to illustrate the multivariate generalized Poisson regression model.

In this paper, we model four health care utilization response variables by the MNBR model. The four response variables (y_1, y_2, y_3, y_4) are the number of consultations with a doctor during the past two weeks to the survey (y_1 = doctor), the number of consultations with non-doctor health professional during the past four weeks to the survey (y_2 = non-doctor), the total number of prescribed medications used in the past two days (y_3 = prescribe) and the total number of non-prescribed medications used in the past two days (y_4 = non-prescribe). The complete data has six response variables. The four response variables used for illustration are chosen from the six response variables because of the following: y_1 and y_2 have been analyzed by various authors to illustrate some bivariate count models. Among the other response variables, the only one that has a negative correlation with y_1 or y_2 is y_4. Furthermore, the SAS NLMIXED procedure used to fit MPR model to all six variables could not converge after 16 hours of execution. Thus, we use the variables y_1, y_2, y_3, and y_4. The descriptive statistics and the correlations between the pairs of the variables are given in Table 1.

Table 1: Mean, variance and correlation for health-care utilization data

	Correlation			Mean	Variance
	y_2	y_3	y_4		
y_1	0.1481*	0.3078*	–0.0149	0.3017	0.6370
y_2		0.1634*	0.0089	0.2146	0.9318
y_3			–0.0435*	0.8626	2.0032
y_4				0.3557	0.5075

* significant at 5% level

The predictor variables are made up of four socio-economic variables and eight insurance and health status variables. The socio-economic variables are dummy variable for whether or not the patient is female (gender), age in years divided by 100 (age), the square of “age in years divided by 100” (agesq), and annual income in ten-thousands of dollars (income). The insurance and health status variables are indicator variable for private insurance coverage (levyplus), free government insurance coverage due to low income (freepoor), free government insurance coverage due to old age, disability or veteran status (freerepa), number of illnesses in the past two weeks (illness), number of days of reduced activity in the past two weeks due to illness or injury (actdays), general health questionnaire score using Goldberg’s method with high score indicating bad health (hscore), indicator variable for chronic condition not limiting activity (chcon1), and indicator variable for chronic condition limiting activity (chcon2). The summary statistics for these predictor variables are contained in Cameron et al. (1988).

The marginal mean of Y_{it} is assumed to have a log-linear relationship with the covariates x_i through

$$\log[E(Y_{it})] = x_{it0}\beta_{t0} + x_{it1}\beta_{t1} + x_{it2}\beta_{t2} + \ldots + x_{it12}\beta_{t12}\,, \qquad (8)$$

for t = 1, 2, 3, 4 and i = 1, 2, …, 5190. The regression function (8) relates the logarithm of the marginal means to the explanatory variables. The correlation between Y_{it} and Y_{iv} is specified in terms of parameter λ_{tv}, which can be written in terms of ρ_{tv}. The predictors x_{it1} through x_{it12} are listed in Table 2. We fitted the MPR and MNBR models to the data and the results are presented in Table 2.

To check the adequacy of MNBR model over the MPR model, we observe that all the dispersion parameters are significant. This shows that the data exhibit substantial over-dispersion. Furthermore, the log-likelihood values show that the MNBR model provides a better fit than the MPR model. The multivariate generalized Poisson regression (MGPR) model yielded a log-likelihood value of –14584.06. This value is very close to the log-likelihood value for the MNBR model given in Table 2. The AIC statistics for the MPR, MNBR and MGPR are respectively 31803.12, 29293.34, and 29292.12. Under the MPR model, all pairs of correlations are significant. However, under the MGPR and MNBR models, the pairs of response variables (y_1, y_2) and (y_2, y_4) are not significant.

The significant parameters under the MNBR and the MGPR models are the same. The MNBR model estimates show that only illness and hscore are important determinants of all the four response variables. The MNBR model shows that the predictors gender, freepoor, actdays and chronic condition 1 are significant for doctor's visits (y_1). However, the predictors freerepa, actdays, chronic conditions 1 and 2 are significant at 5% for non-doctor's visits (y_2). The number of prescribed medications (y_3) are responsive to gender, age, levyplus, freerepa, actdays, chronic conditions 1 and 2 while the predictors gender, age, agesq, freerepa, and chronic condition 1 are important determinants of the number of non-prescribed medications (y_4).

Gurmu and Elder (2000) stated that there is enough evidence that doctor and non-doctor are dependent counts and therefore they should be jointly estimated. Famoye (2010a) observed that the response variables y_1 and y_2 appeared not to be dependent from using the bivariate negative binomial regression model. A similar result is obtained in the MNBR model estimates in Table 2. The response variables y_1 and y_2 appear to be independent since ρ_{12} (*p*-value = 0.1835) is not significant.

Table 2: Parameter estimates (standard errors in parentheses) for health-care data

	y_1		y_2	
Variable	MPR model	MNBR model	MPR model	MNBR model
Constant (x_0)	–2.221 (0.181)*	–2.306 (.227)*	–2.475 (0.235)*	–2.837 (.434)*
Gender (x_1)	0.131 (0.055)*	0.175 (0.068)*	0.323 (0.068)*	0.217 (0.124)
Age (x_2)	1.144 (0.943)	0.794 (1.231)	–2.893 (1.206)*	–2.184 (2.429)
Agesq (x_3)	–0.863 (0.997)	–0.558 (1.342)	3.975 (1.282)*	3.322 (2.625)
Income (x_4)	–0.237 (0.087)*	–0.187 (0.107)	–0.056 (0.109)	–0.079 (0.191)
Levyplus(x_5)	0.136 (0.071)	0.126 (0.085)	0.319 (0.095)*	0.294 (0.157)
Freepoor (x_6)	–0.467 (0.185)*	–0.535 (.214)*	0.043 (0.201)	–0.147 (0.346)
Freerepa (x_7)	0.120 (0.088)	0.190 (0.113)	0.460 (0.114)*	0.573 (0.218)*
Illness (x_8)	0.175 (0.017)*	0.198 (0.023)*	0.064 (0.021)*	0.137 (0.047)*
Actdays (x_9)	0.123 (0.005)*	0.140 (0.008)*	0.100 (0.006)*	0.136 (0.017)*
Hscore (x_{10})	0.034 (0.009)*	0.035 (0.014)*	0.044 (0.011)*	0.075 (0.028)*
Chcon1 (x_{11})	0.169 (0.065)*	0.141 (0.078)	0.504 (0.085)*	0.414 (0.142)*
Chcon2 (x_{12})	0.239 (0.080)*	0.247 (0.100)*	1.063 (0.096)*	1.120 (0.182)*
$\hat{m}_t$		1.094 (0.104)*		8.912 (0.676)*

	y_3		y_4	
	MPR model	MNBR model	MPR model	MNBR model
Constant (x_0)	–2.651 (0.125)*	–2.668 (.145)*	–2.283 (0.168)*	–2.315 (.193)*
Gender (x_1)	0.482 (0.036)*	0.548 (0.042)*	0.270 (0.051)*	0.282 (0.058)*
Age (x_2)	2.232 (0.579)*	1.806 (0.705)*	4.438 (0.943)*	4.556 (1.079)*
Agesq (x_3)	–0.426 (0.597)	0.039 (0.745)	–5.700 (1.067)*	–5.81 (1.216)*
Income (x_4)	0.006 (0.055)	0.045 (0.064)	0.130 (0.075)	0.120 (0.085)
Levyplus (x_5)	0.278 (0.051)*	0.257 (0.057)*	–0.028 (0.057)	–0.037 (0.066)
Freepoor (x_6)	–0.019 (0.124)	–0.024 (0.136)	–0.027 (0.124)	–0.037 (0.141)
Freerepa (x_7)	0.280 (0.058)*	0.271 (0.067)*	–0.269 (0.092)*	–0.269 (.104)*
Illness (x_8)	0.198 (0.010)*	0.210 (0.013)*	0.201 (0.018)*	0.210 (0.021)*
Actdays (x_9)	0.036 (0.004)*	0.035 (0.005)*	0.005 (0.008)	0.005 (0.009)
Hscore (x_{10})	0.025 (0.006)*	0.024 (0.008)*	0.028 (0.010)*	0.029 (0.012)*
Chcon1 (x_{11})	0.774 (0.046)*	0.768 (0.050)*	0.157 (0.056)*	0.153 (0.063)*
Chcon2 (x_{12})	0.986 (0.053)*	0.991 (0.061)*	0.018 (0.082)	–0.003 (0.094)
$\hat{m}_t$		0.307 (0.031)*		0.742 (0.083)*
$\hat{\rho}_{tv}$ for MPR	$\hat{\rho}_{12}$ = .0453 (.0123)*; $\hat{\rho}_{13}$ = 0.1473 (.0086)*; $\hat{\rho}_{14}$ = –.0316 (.0116)* $\hat{\rho}_{23}$ = .0788 (.0122)*; $\hat{\rho}_{24}$ = 0.0362 (.0140)*; $\hat{\rho}_{34}$ = –.0712 (.0104)*			
$\hat{\rho}_{tv}$ for MNBR	$\hat{\rho}_{12}$ = .0141 (.0106); $\hat{\rho}_{13}$ = 0.1587 (.0090)*; $\hat{\rho}_{14}$ = –0.0421 (.0107)* $\hat{\rho}_{23}$ = .0412 (.0111)*; $\hat{\rho}_{24}$ = 0.0171 (.0117); $\hat{\rho}_{34}$ = –0.0818 (.0105)*			
Log-likelihood	For MPR: –15843.56		For MNBR: –14584.67	
AIC	For MPR: 31803.12		For MNBR: 29293.34	

* significant at 5% level

6. Concluding Remarks

The univariate negative binomial regression model has been used to model over-dispersed count data. Famoye (2010a) defined and studied a bivariate negative binomial regression model. This regression model is extended to give the MNBR model in (3). The MNBR model can be used to model over-dispersed count data. The multivariate Poisson lognormal regression (MPLR) model is characterized by unrestricted correlation. However, the MPLR model can also be used to model over-dispersed count data. Another disadvantage of the MPLR model is that the model and the likelihood function have complicated forms. Famoye (2013) defined the MGPR model which can be used to model a count data with any type of dispersion. The MNBR, MPLR and MGPR models are competitors. However, only the MGPR model can be used to model under-dispersed response variables.

The MNBR model overcomes several drawbacks of other multivariate count data regression models mentioned in section 1. The model accounts for over-dispersion in the response variables. It allows for correlations of any sign among counts independently of the dispersion parameters. This adds flexibility to the MNBR model by separating the effect of dispersion and correlation among counts. The estimation of MNBR model is not time consuming compared to the MPLR model because the likelihood function for the MNBR model can be written in closed form.

Future research work will include the comparison of likelihood ratio test and the score test for the hypotheses in section 4, especially the hypotheses in (7) for testing the MPR against the MNBR. Another area of research is to include parameters measuring covariances of order 3, 4, …, *d* for *d*-variate distribution.

Acknowledgement

The author acknowledges the financial support of the Central Michigan University Faculty Research and Creative Endeavors committee to fund the presentation of the paper.

References

Alfò, M. and Trovato, G. (2004) Semiparametric mixture models for multivariate count data, with application. *Econometric Journal*, 7, 426-454.

Cameron, A.C. and Johansson, P. (1997) Count data regression using series expansion: With applications. *Journal of Applied Econometrics*, 12, 203-223.

Cameron, A.C., Li, T., Trivedi, P.K. and Zimmer, D.M. (2004) Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *Econometrics Journal*, 7, 566-584.

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

Cameron, A.C., Trivedi, P.K., Milne, F. and Piggott, J. (1988) A microeconomic model of the demand for health care and health insurance in Australia. *Review of Econometric Studies*, LV, 85-106.

Chernoff, H. (1954) On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25(3), 573-578.

Chib, S. and Winkelmann, R. (2001) Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4), 428-435.

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

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

Famoye, F. (2013) Multivariate generalized Poisson regression model. Accepted by *Communications in Statistics-Theory and Methods*.

Famoye, F., Okafor, R. and Adamu, M.O. (2011) A multivariate generalized Poisson distribution. *Journal of Statistical Theory and Applications*, 10(3), 519-531.

Gurmu, S. and Elder, J. (2000) Generalized bivariate count data regression models. *Economics Letters*, 68, 31-36.

Karlis, D. (2003) An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30, 63-67.

Karlis, D. and Meligkotsidou, L. (2005) Multivariate Poisson regression with covariance structure. *Statistics and Computing*, 15, 255-265.

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 a bivariate Poisson distribution. *Communications in Statistics-Theory and Methods*, 28(2), 267-276.

Lee, A. (1999) Modelling rugby league data via bivariate negative binomial regression. *Austral. & New Zealand Journal of Statistics*, 41(2), 141-152.

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

Ma, J., Kockelman, K.M. and Damien, P. (2008) A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, 40, 964-975.

Miravete, E.J. (2009) Multivariate Sarmanov count data models. CEPR Discussion Paper No. DP7463, University of Texas at Austin; Centre for Economic Policy Research (CEPR).

Mullahy, J. (1997) Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12, 337-350.

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

Self, S.G. and Liang, K. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605-610.

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

van Ophem, H. (1999) A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228-237.