PLOS ONE

RESEARCH ARTICLE

A generalized right truncated bivariate Poisson regression model with applications to health data

M. Ataharul Islam[1]*, **Rafiqul I. Chowdhury**[2]

1 Applied Statistics, East West University, Dhaka-1212, Dhaka, Bangladesh, **2** ISRT, University of Dhaka, Dhaka, Bangladesh

* mataharul@yahoo.com

OPEN ACCESS

Citation: Islam MA, Chowdhury RI (2017) A generalized right truncated bivariate Poisson regression model with applications to health data. PLoS ONE 12(6): e0178153. https://doi.org/10.1371/journal.pone.0178153

Editor: Antonio G. Pacheco, Fundacao Oswaldo Cruz, BRAZIL

Received: August 1, 2016

Accepted: May 9, 2017

Published: June 6, 2017

Data Availability Statement: All data files are freely available from the http://hrsonline.isr.umich.edu/. As registered user, we access the website. This data set is a public domain data and freely available from Health and Retirement Study website. After registration, anybody can have access to the dataset. The same variables and observations could be obtained from this site using the description in the Application section of the submitted paper.

Abstract

A generalized right truncated bivariate Poisson regression model is proposed in this paper. Estimation and tests for goodness of fit and over or under dispersion are illustrated for both untruncated and right truncated bivariate Poisson regression models using marginal-conditional approach. Estimation and test procedures are illustrated for bivariate Poisson regression models with applications to Health and Retirement Study data on number of health conditions and the number of health care services utilized. The proposed test statistics are easy to compute and it is evident from the results that the models fit the data very well. A comparison between the right truncated and untruncated bivariate Poisson regression models using the test for nonnested models clearly shows that the truncated model performs significantly better than the untruncated model.

Introduction

The bivariate Poisson models have emerged to address an increasingly important area of research with a wide range of applications in various fields where paired count data are correlated. Data on paired counts arise commonly in various areas of research such as traffic accidents (number of accidents and fatalities), medical research (number of visits to health professionals and utilization of health care services), epidemiology (number of episodes of depression and number of visits to doctors), marketing (number of sales of different products), sports (number of goals scored by opponent teams), econometrics (job switching voluntarily or involuntarily), etc. As mentioned, a typical example of such dependence arises in the traffic accidents where the extent of physical injuries may lead to fatalities. Leiter and Hamdan [1] suggested bivariate probability models applicable to traffic accidents and fatalities. A similar problem was addressed by Cacoullos and Papageorgiou [2]. Several other studies defined and studied the bivariate Poisson distribution [3–10].

The bivariate Poisson distributions have been developed following various assumptions. Among those, the most comprehensive one has been proposed by Kocherlakota and Kocherlakota [11]. The bivariate Poisson form is shown by using a trivariate reduction method [8] allowing for correlation between the variables, which is considered as a nuisance parameter.

Funding: The study is supported by the HEQEP sub-project 3293, University Grants Commission of Bangladesh and the World Bank.

Competing interests: The authors have declared that no competing interests exist.

This bivariate Poisson regression is used by Jung and Winkelmann [8], Vernic [12] and Karlis and Ntzoufras [9, 13] among others. However, there is no attempt for modelling right truncated bivariate Poisson regression model so far which can play an important role for analyzing bivariate count data arising from a wider range of problems emerging in various fields.

Leiter and Hamdan [1] suggested joint distributions for number of accidents and number of fatalities by Poisson-Bernoulli (or Binomial) and Poisson-Poisson distribution. An alternative Poisson-Binomial model was proposed by Cacoullos and Papageorgiou [2]. In this paper, we have used the Poisson-Poisson distribution and the model for covariate dependence based on extended generalized linear model proposed by Islam and Chowdhury [14]. The marginal and conditional models are employed which are expressed in terms of the family of exponential distributions. Then a generalized linear model for this bivariate form is developed and the link functions are identified and expressed as functions of covariates. In some cases, we need to use right truncated Poisson distribution for both the outcome variables due to restrictions in the definition of variables. A model is shown for taking account of right truncated bivariate Poisson and a generalized bivariate count regression model is proposed using marginal-conditional approach. As both the tests for goodness of fit and over(under)dispersion are of prime concern in dealing with these models, some test procedures are shown in this paper. Gurmu and Trivedi [15, 16] suggested overdispersion tests for truncated Poisson regression models for univariate case. In this study, Gurmu and Trivedi [15] test is extended for bivariate Poisson for both untruncated and right truncated Poisson regression. For testing the goodness of fit tests for right truncated bivariate data, the test proposed by Islam and Chowdhury [14] is extended for truncated data. The estimation and test procedures are applied to health data, namely, the Health and Retirement Study [17] on number of health conditions ever had and utilization of healthcare services. Both the models with or without truncation are fitted to the same data and goodness of fit tests are applied to compare the models.

The bivariate Poisson-Poisson model

Let Y_1 be the number of occurrences of the first event in a given interval follow a Poisson distribution with parameter λ_1 and the probability of the second event, Y_2, for given Y_1, where $Y_2 = Y_{21} + \ldots + Y_{2y_1}$, be Poisson with parameter, $\lambda_1 y_1$, $y_{21}, \ldots, y_{2y_1} = 0, 1, \ldots$. Then the marginal, conditional and joint distributions of Y_1 and Y_2 can be shown as follows

$$g(y_1) = \frac{e^{-\lambda_1}\lambda_1^{y_1}}{y_1!}, y_1 = 0, 1, \ldots$$

$$g(y_2|y_1) = \frac{e^{-\lambda_2 y_1}(\lambda_2 y_1)^{y_2}}{y_2!}, y_2 = 0, 1, \ldots$$

Multiplying the marginal and conditional distributions shown above, we obtain the following joint distribution

$$g(y_1, y_2) = g(y_2 \mid y_1)g(y_1) = \frac{e^{-\lambda_1}\lambda_1^{y_1}e^{-\lambda_2 y_1}(\lambda_2 y_1)^{y_2}}{(y_1! y_2!)}, y_1 = 0, 1, \ldots \quad (1)$$

The probability mass function for Y_2 is

$$g_2(y_2) = \frac{e^{-\lambda_1}\lambda_2^{y_2}}{y_2!}\sum_{y_1=0}^{\infty}\frac{(\lambda_1 e^{-\lambda_2})^{y_1} y_1^{y_2}}{y_1!}, \quad y_2 = 0, 1, \ldots \quad (2)$$

It is noteworthy that in the proposed model, the marginal models for Y_1 and Y_2 do not allow same joint probability mass function from $g(y_2)g(y_1 \mid y_2)$ and $g(y_1)g(y_2 \mid y_1)$ which is expected as we observe from corresponding marginal probabilities.

The bivariate Poisson-Poisson expression in Eq (1) can be expressed as bivariate exponential form as follows

$$g(y_1, y_2) = e^{\{y_1 \ln \lambda_1 + y_2 \ln \lambda_2 - \lambda_1 - \lambda_2 y_1 + y_2 \ln y_1 - \ln y_1! - \ln y_2!\}}. \tag{3}$$

The link functions are $\ln \lambda_1 = x'\beta_1$, $\ln \lambda_2 = x'\beta_2$, where,
$x' = (1, x_1, ..., x_p), \beta_1' = (\beta_{10}, \beta_{11}, ..., \beta_{1p}), \beta_2' = (\beta_{20}, \beta_{21}, ..., \beta_{2p})$.

Hence, we can show that $\lambda_1 = e^{x'\beta_1}$, and $\lambda_2 = e^{x'\beta_2}$. It is noteworthy that $E(Y_1) = \mu_1 = \lambda_1$ and $E(Y_2) = \mu_2 = \lambda_1\lambda_2$. Hence, $\mu_1(\beta_1) = e^{x'\beta_1}$ and $\mu_2(\beta_1, \beta_2) = e^{x'\beta_1 + x'\beta_2}$.

The log likelihood function is

$$\ln L = \sum_{i=1}^{n} \{y_{1i} x_i'\beta_1 + y_{2i} x_i'\beta_2 - e^{x_i'\beta_1} - e^{x_i'\beta_2} y_{1i} + y_{2i} \ln y_{1i} - \ln y_{1i}! - \ln y_{2i}!\} \tag{4}$$

Bivariate right truncated Poisson-Poisson model

Let Y_1 be the number of occurrences of the first event in a given interval which has a truncated Poisson distribution with mass function

$$g_1(y_1) = c_1 \frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!}, \quad y_1 = 0, 1, ..., k_1 \tag{5}$$

where

$$c_1 = \frac{1}{\sum_{y_1=0}^{k_1} \frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!}}.$$

Then the probability of the second event, Y_2, for given y_1, where $Y_2 = Y_{21} + \ldots + Y_{2y_1}$, the total number of second event cases among the truncated y_1 events in the specified time interval, is:

$$g(y_2 \mid y_1) = c_2 \frac{e^{-\lambda_2 y_1} (\lambda_2 y_1)^{y_2}}{y_2!}, \quad y_2 = 0, 1, ..., k_{y_1} \tag{6}$$

where

$$c_{2y_1} = \frac{1}{\sum_{y_2=0}^{k_{2y_1}} \frac{e^{-\lambda_2 y_1} (\lambda_2 y_1)^{y_1}}{y_2!}}.$$

Multiplying the conditional and marginal probability functions, the joint distribution of Y_1 and Y_2 can be shown as follows

$$g(y_1, y_2) = g(y_2 \mid y_1) g(y_1) = c_1 c_2 e^{-\lambda_1} \lambda_1^{y_1} e^{-\lambda_2 y_1} (\lambda_2 y_1)^{y_2} / (y_1! y_2!). \tag{7}$$

The truncated Poisson-Poisson expression can be expressed as bivariate exponential form as follows

$$g(y_1, y_2) = e^{\{y_1 \ln \lambda_1 + y_2 \ln \lambda_2 - \lambda_1 - \lambda_2 y_1 + y_2 \ln y_1 - \ln y_1! - \ln y_2! + \ln c_1 + \ln c_2\}}. \quad (8)$$

The expected value of Y_1 can be shown as

$$\mu_{Y_1} = E(Y_1) = \sum_{y_1=0}^{k1} \left[y_1 \left(\frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!} \right) \left(\frac{1}{\sum_{y_1=0}^{k_1} \frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!}} \right) \right] = \frac{\lambda_1 \Gamma(k1, \lambda_1) k1}{\Gamma(k1+1, \lambda_1)},$$

where incomplete gamma function approximations are obtained based on relationships between cumulative Poisson function and incomplete gamma functions shown in Abramowitz and Stegun Chapters 6, 7 and 26 [18].

Similarly the conditional expected value for Y_2 can be expressed as

$$\begin{aligned} \mu_{Y_2/y_1} &= E(Y_2|y_1) = \sum_{y_2=0}^{k2} \left[y_2 \left(\frac{e^{-\lambda_2 y_1} (\lambda_2 y_1)^{y_2}}{y_2!} \right) \left(\frac{1}{\sum_{y_2=0}^{k_2} \frac{e^{-\lambda_2 y_1} (\lambda_2 y_1)^{y_2}}{y_2!}} \right) \right] \\ &= \frac{(\lambda_2 y_1) \Gamma(k2, \lambda_2 y_1) k2}{\Gamma(k2+1, \lambda_2 y_1)}. \end{aligned}$$

The link functions in model Eq (7) are $\ln \lambda_1 = x'\beta_1$, $\ln \lambda_2 = x'\beta_2$, where $\beta' = (\beta_1', \beta_2')$, $x' = (1, x_1, \ldots, x_p)$, $\beta_1{}' = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})$, $\beta_2' = (\beta_{20}, \beta_{21}, \ldots, \beta_{2p})$ and the log likelihood and is shown below:

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \{ y_{1i} \ln \lambda_1 + y_{2i} \ln \lambda_2 - \lambda_1 - \lambda_2 y_{1i} + y_{2i} \ln y_{1i} - \ln y_{1i}! - \ln y_{2i}! \\ &\quad + lnc_1 + \ln c_{2_{y_1}} \\ &= \sum_{i=1}^{n} \{ y_{1i} x_{1i}' \beta_1 + y_{2i} x_{1i}' \beta_2 - e^{x_{1i}' \beta_1} - e^{x_{1i}' \beta_2} y_{1i} + y_{2i} \ln y_{1i} - \ln y_{1i}! - \ln y_{2i}! \\ &\quad + lnc_1 + \ln c_{2_{y_1}} \}. \end{aligned} \quad (9)$$

Test for goodness of fit and under(Over) dispersion

In Sections 2 and 3, the proposed generalized models for untruncated or right truncated data are based on marginal and conditional probabilities as functions of covariates, we need to use a test for goodness of fit suitable for such models. A test for goodness of fit is proposed for the

bivariate Poisson-Poisson model [14] using marginal and conditional means is:

$$T_1 = \sum_{y_1=0}^{k_1} T_1 y_1 = \sum_{y_1=0}^{k_1} \begin{pmatrix} \bar{y}_1 & - & \hat{\mu}_{u,y_1} \\ \bar{y}_2 \mid y_1 & - & \hat{\mu}_{u,y_2|y_1} \end{pmatrix}' \begin{pmatrix} \sum_{i=1}^{n_{y_1}} \hat{\lambda}_{1i}/n_{y_1} & 0 \\ 0 & \sum_{i=1}^{n_{y_1}} \hat{\lambda}_{2i} y_1/n_{y_1} \end{pmatrix}^{-1} \begin{pmatrix} \bar{y}_1 & - & \hat{\mu}_{u,y_1} \\ \bar{y}_2 \mid y_1 & - & \hat{\mu}_{u,y_2|y_1} \end{pmatrix} \quad (10)$$

where $T_1 y_1$ can be shown asymptotically as a χ^2 with 2 degrees of freedom. Hence T_1 is distributed asymptotically as χ^2 with $2(k_1 + 1)$ degrees of freedom as each $T_1 y_1$ is independent chi-square, $k_1 + 1$ is the number of groups of distinct y_1 values such as $Y_1 = 0$ with frequency n_0, $Y_1 = 1$ with frequency $n_1, \ldots, Y_1 = k_1$ with frequency n_{k_1}, $\hat{\mu}_{y_1} = \sum_{i=1}^{n_{y_1}} \hat{\lambda}_{1i}/n_{y_1}$, and $\hat{\mu}_{y_2|y_1} = \sum_{i=1}^{n_{y_1}} \hat{\lambda}_{2i} y_1/n_{y_1}$. Here $\hat{\lambda}_{1i} = e^{x_i'\hat{\beta}_1}$, and $\hat{\lambda}_2 = e^{x_i'\hat{\beta}_2}$.

Similarly, the goodness of fit test statistic as shown in Eq (10) can be modified for the bivariate truncated model too. In this case, the null and alternative hypotheses are: H_0: the data follow the proposed truncated bivariate Poisson model and, H_1: the data do not follow the proposed truncated bivariate Poisson model. The modified test statistic for testing the goodness of fit is:

$$\sum_{y_1=0}^{k_1} \begin{pmatrix} \bar{y}_1 & - & \hat{\mu}_{t,y_1} \\ \bar{y}_2 \mid y_1 & - & \hat{\mu}_{t,y_2|y_1} \end{pmatrix}' \begin{pmatrix} \sum_{i=1}^{n_{y_1}} \hat{V}_{1i}/n_{y_1} & 0 \\ 0 & \sum_{i=1}^{n_{y_1}} \hat{V}_{2i|1}/n_{y_1} \end{pmatrix}^{-1} \begin{pmatrix} \bar{y}_1 & - & \hat{\mu}_{t,y_1} \\ \bar{y}_2 \mid y_1 & - & \hat{\mu}_{t,y_2|y_1} \end{pmatrix} \quad (11)$$

where $T_1 y_1$ can be shown asymptotically as a χ^2 with 2 degrees of freedom. Hence T_1 is distributed asymptotically as χ^2 with $2(k_1 + 1)$ degrees of freedom as each $T_1 y_1$ is independent chi-square, $k_1 + 1$ is the number of groups of distinct y_1 values such as $Y_1 = 0$ with frequency n_0, $Y_1 = 1$ with frequency $n_1, \ldots, Y_1 = k_1$ with frequency n_{k_1}, $\hat{\mu}_{y_1} = \frac{\hat{\lambda}_1 \Gamma(k_1, \hat{\lambda}_1) k_1}{\Gamma(k_1+1, \hat{\lambda}_1)}$, $\hat{\mu}_{y_2|y_1} = \frac{(\hat{\lambda}_2 y_1)\Gamma(k_{2y_1}, \hat{\lambda}_2 y_1) k_{2y_1}}{\Gamma(k_{2y_1}+1, \hat{\lambda}_2 y_1)}$, V_1 and $V_{2|1}$ are the variances of Y_1 and Y_2 given Y_1 respectively (see Appendix). Using $\hat{\lambda}_{1i} = e^{x'_i\hat{\beta}^*_1}$, $\hat{\lambda}_{2i} = e^{x'_i\hat{\beta}^*_2}$, we obtain $\hat{V}_{1i}$ and $\hat{V}_{2i|1}$ from V_1 and $V_{2|1}$ for $i = 1, 2, \ldots, n_{y_1}$.

We can modify the test for goodness of fit as shown in Eqs (10) and (11) to develop test statistics for over(under)dispersion. As overdispersion and underdispersion may influence the fit of the proposed untruncated Poisson regression models, we have used the method of moments estimator [14, 19–21] to estimate dispersion parameter for untruncated model, $\varphi_{u,r}$ where $\hat{\varphi}_{u,r} = \frac{1}{n-p} \sum_{i=1}^{n} \left[\frac{(y_{ri} - \hat{\mu}_{u,ri})^2}{V(\hat{\mu}_{u,ri})} \right] = \frac{\chi^2_{r,n-p}}{n-p}$, $r = 1, 2$, where $V(\hat{\mu}_{u,ri}) = \hat{\mu}_{u,ri}$.

Using the mean, variance and correction factor [15] for right truncated Poisson marginal and conditional models, we can define $\hat{\mu}_{t,ri} = \hat{\lambda}_{ri} + \hat{\delta}_{ri}$ and $\hat{V}(\hat{\mu}_{t,ri}) = \hat{\lambda}_{ri} - \hat{\delta}_{ri}(\hat{\mu}_{ri} - k_r - 1)$ where $\hat{\delta}_{ri} = \hat{\mu}_{ri} - \hat{\lambda}_{ri} = -\hat{\lambda}_{ri}.\hat{\alpha}(k_r, \hat{\lambda}_{ri})$, $\hat{\alpha}(k_r, \lambda_{ri}) = \frac{h(k_r, \lambda_{ri})}{1 - H(k_r, \lambda_{ri})}$, $h(k_r, \lambda_{ri}) = P(Yri = k_r)$, $H(k_r, \lambda_{ri}) = P(Yri \leq k_r)$ and then using these values we can estimate φ_r for the truncated model.

A test for over(under)dispersion, T2, is proposed assuming adjusted variances for the marginal mean of Y1 and conditional means of Y2 given Y1 = y1. If there is over(under)dispersion,

then the test results in acceptance of the null hypothesis that the expected values are equal to the adjusted variances.

Then T_2 for untruncated bivariate Poisson regression model is:

$$T_2 = \sum_{y_1=0}^{k_1} T_2 y_1 = \begin{pmatrix} \bar{y}_1 - \hat{\mu}^*_{u,y_1} \\ \bar{y}_2|y_1 - \hat{\mu}^*_{u,y_2|y_1} \end{pmatrix}' \begin{pmatrix} \dfrac{\hat{\phi}_{u,1}\sum_{i=1}^{n_{y_1}} \hat{\lambda}_{1i}}{n_{y_1}} & 0 \\ 0 & \dfrac{\hat{\phi}_{u,2}\sum_{i=1}^{n_{y_1}} \hat{\lambda}_{2i} y_1}{n_{y_1}} \end{pmatrix}^{-1} \begin{pmatrix} \bar{y}_1 - \hat{\mu}^*_{u,y_1} \\ \bar{y}_2|y_1 - \hat{\mu}^*_{u,y_2|y_1} \end{pmatrix} \tag{12}$$

where $T_2 y_1$ can be shown asymptotically as a χ^2 with 2 degrees of freedom. Hence T_2 is distributed asymptotically as χ^2 with $2(k_1 + 1)$ degrees of freedom as each $T_2 y_1$ is independent chi-square, $\hat{\mu}^*_{u,y_1} = \hat{\phi}_{u,1}\hat{\mu}_{u,y_1}$, and $\hat{\mu}^*_{u,y_2|y_1} = \hat{\phi}_{u,2}\hat{\mu}_{u,y_2|y_1}$, u denotes untruncated model, $\hat{\lambda}_{1i} = e^{x_i'\hat{\beta}_1}$, $\hat{\lambda}_2 = e^{x_i'\hat{\beta}_2}$ and $k_1 + 1$ is the number of distinct counts observed for Y_1. Similarly, for right truncated model T_2 can be defined as follows:

$$T_2 = \sum_{y_1=0}^{k_1} T_2 y_1 = \sum_{y_1=0}^{k_1} \begin{pmatrix} \bar{y}_1 - \hat{\mu}^*_{t,y_1} \\ \bar{y}_2|y_1 - \hat{\mu}^*_{t,y_2|y_1} \end{pmatrix}' \begin{pmatrix} \dfrac{\hat{\phi}_{t,1}\sum_{i=1}^{n_{y_1}} \hat{V}_{1i}}{n_{y_1}} & 0 \\ 0 & \dfrac{\hat{\phi}_{t,2}\sum_{i=1}^{n_{y_1}} V_{2i|1}}{n_{y_1}} \end{pmatrix}^{-1} \begin{pmatrix} \bar{y}_1 - \hat{\mu}^*_{t,y_1} \\ \bar{y}_2|y_1 - \hat{\mu}^*_{t,y_2|y_1} \end{pmatrix} \tag{13}$$

where $T_2 y_1$ can be shown asymptotically as a χ^2 with 2 degrees of freedom. Hence T_2 is distributed asymptotically as χ^2 with $2(k_1 + 1)$ degrees of freedom as each $T_2 y_1$ is independent chi-square, $\hat{\mu}^*_{t,y_1} = \hat{\varphi}_{t,1}\hat{\mu}_{t,y_1}$ *and* $\hat{\mu}^*_{t,y_2|y_1} = \hat{\varphi}_{t,2}\hat{\mu}_{t,y_2|y_1}$, $\hat{\lambda}_{1i} = e^{x_i'\hat{\beta}^*_1}$ and $\hat{\lambda}_{2i} = e^{x_i'\hat{\beta}^*_2}$, t denotes truncated model.

In the above tests for the goodness of fit, specific models (proposed models) are tested. Hence, the test for goodness of fit may be termed equivalently as the test for specification of the proposed models.

Test for model selection

Voung [22] developed a general test for nonnested models. This statistic tests the hypothesis that two nonnested parametric models are equally distant in the Kullback-Leibler sense from the true data distribution. This test statistic is used to select a parsimonuous model. Consider following exponential density form for generalized linear models:

$$f(y, \theta) = e^{[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)]},$$

where θ is the natural link function which is a function of expected value of Y, $b(\theta)$ is a function of θ, $a(\phi)$ is dispersion parameter and $c(y, \phi)$ is a function of y and ϕ. We can obtain expected

value and variance of Y from $E(Y) = b'(\theta)$ and $Var(Y) = a(\phi)b''(\theta)$ where $b''(\theta)$ is the variance function, $Var[E(Y)]$. From the probability function we can show that $\theta = \ln \mu_1$, $E(Y_1) = \mu_1$, and $Var(Y_1) = \mu_1/\phi_1$. Similarly, using the conditional probability function for Y_2 given $Y_1 = y_1$, we find $\theta = \ln \mu_2 y_1$, $E(Y_2|y_1) = \mu_2 y_1$ and $Var(Y_2|y_1) = \mu_2 y_1/\phi_2$.

The Voung test is a t or standard normal test for large sample where

$$V = \frac{\bar{m}\sqrt{n}}{s_m}, \quad \bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}, \quad m_i = \ln[f(y_{1i}, y_{2i}; \theta)/g(y_{1i}, y_{2i}; \theta')],$$

where $f(y_{1i}, y_{2i};\theta)$ is the probability function for model with parameter vector θ and $g(y_{1i}, y_{2i};\theta')$ the probability function for model with parameter vector θ' and

$$s_m^2 = \frac{1}{n}\left[\sum_{i=1}^{n}\left\{\ln\left(\frac{f(y_{1i}, y_{2i}; \theta)}{g(y_{1i}, y_{2i}; \theta')}\right)\right\}^2 - \frac{1}{n}\left[\sum_{i=1}^{n}\left\{\ln\left(\frac{f(y_{1i}, y_{2i}; \theta)}{g(y_{1i}, y_{2i}; \theta')}\right)\right\}\right]^2\right]$$

Adjusted Vuong test is

$$V = \frac{\bar{m}'\sqrt{n}}{s_{m'}}, \quad \bar{m}' = \frac{\sum_{i=1}^{n} m_i'}{n}, \quad \ln\left[\frac{f(y_{1i}, y_{2i}; \theta)}{g(y_{1i}, y_{2i}; \theta')}\right] - \frac{(p-q) - \ln n}{2n},$$

where p = number of parameters in numerator in model $f(y_{1i}, y_{2i};\theta)$ and q = number of parameters in denominator in model $g(y_{1i}, y_{2i};\theta')$. If $V > 1.96$ then model in the numerator is favored and if $V < -1.96$ then model in the denominator is favored.

Application

The models proposed in this paper are applied to the tenth wave of the Health and Retirement Study [17]. The outcome variables are number of conditions ever had (Y_1) as mentioned by doctors and utilization of healthcare services (Y_2) where utilization of healthcare services include services from hospital, nursing home, doctor, and home care. The explanatory variables are: gender (1 male, 0 female), age (in years), race (1 Hispanic, 0 others) and veteran status (1 yes, 0 no). The sample size is 5568.

Table 1 displays the average number of conditions (Y_1) and utilization of health care services (Y_2) and sample variances. The average number of conditions is 2.63 and the variance (2.05) appears to be lower than the mean. The average number of utilization of health care services is 0.77 and the corresponding variance is slightly higher (0.79). The measure of correlation between number of conditions and utilization of health care services can be estimated by $r = [\hat{\lambda}_2/(\hat{\lambda}_2 + 1)]^{\frac{1}{2}}$ where $\hat{\lambda}_2 = \frac{\bar{y}_2}{\bar{y}_1}$ (see Leiter and Hamdan, [1]). The estimated correlation is 0.48

Table 1. Descriptive measures of outcome variables.

Measures	Health and Retirement Study (HRS, 2010)
$\bar{y}_1$	2.642716
$\widehat{Var}(Y_1)$	2.117924
$\bar{y}_2$	0.769176
$\widehat{Var}(Y_2)$	0.792021

https://doi.org/10.1371/journal.pone.0178153.t001

and the corresponding standard error is very small indicating significant association between the bivariate counts. The estimated standard error of r is $\widehat{se(r)} = [\bar{y}_1^2/\{4n(\bar{y}_1+\bar{y}_2)\}^3]^{\frac{1}{2}}$.

The dispersion parameters for number of conditions (Y_1) and utilization of health care services (Y_2) are 0.80 and 1.05, respectively. A test for over(under)dispersion using T_2 indicates that there might be statistically significant over(under) dispersion in the bivariate count data. This indicates slight overdispersion for utilization of health care services and underdispersion for number of conditions. Using the mean, variance and correction factor [15] for right truncated Poisson marginal and conditional models, we can estimate the dispersion parameters.

Table 2 displays estimates of the parameters of the marginal and conditional models for the untruncated model as well as tests for significance of the parameters with and without considering adjustment for over or under dispersion. The selected explanatory variables in the models are gender, age, race and veteran status. All the variables except race show statistically significant association with number of conditions. From the marginal model for number of conditions it appears that the number of conditions is significantly lower for males compared to females but age and veteran status are positively associated with number of conditions. This shows that the number of conditions in health increases steadily with age, as expected, and the veterans have higher number of conditions compared to non-veterans. There is no significant difference in number of conditions by race. As there is evidence of under dispersion in the number of conditions, an adjustment is made using the dispersion parameter, After adjustment, the standard errors are smaller which are displayed in Table 2 and p-values appear to be smaller. We observe that the conditional model for utilization of healthcare services for given number of conditions indicate significant positive association with gender (male = 1, female = 0) and veteran status (yes = 1, no = 0) while age and race (race = 1 for Hispanic, race = 0, otherwise) show negative association. Although number of conditions is lower for males, utilization of healthcare services appears to be higher among males compared to females. Another important finding reveals that the utilization of healthcare services reduces significantly with age for given number of indications which implies that the with ncreased age elderly people do not take healthcare services from different sources. Race is not associated significantly with number of conditions but utilization of healthcare services show a negative association with race indicating lower utilization for Hispanics as compared to other races.

Table 2. Parameter estimates for the bivariate poisson models using the data on number of conditions and utilization of healthcare services (HRS, 2010).

Variables	Estimate	S.E.	z-value	p-value	$\sqrt{\varphi_r \boldsymbol{V}(\hat{\beta})}$	p-value
Marginal model for (Y_1)						
β_{10}	−0.055	0.1953	−0.281	0.779	0.1720	0.750
Gender	−0.055	0.0214	−2.573	0.010	0.0189	0.003
Age (years)	0.014	0.0026	5.315	0.000	0.0023	0.000
Race	0.007	0.0288	0.234	0.815	0.0254	0.790
Veteran status	0.048	0.025	1.937	0.053	0.0220	0.028
Conditional model for (Y_2)						
β_{20}	0.265	0.3627	0.730	0.466	0.3713	0.476
Gender	0.345	0.0385	8.953	0.000	0.0395	0.000
Age (years)	−0.023	0.0049	−4.608	0.000	0.0050	0.000
Race	−0.174	0.0582	−2.997	0.003	0.0590	0.003
Veteran status	0.093	0.0423	2.208	0.027	0.0432	0.031

Gender (male = 1, female = 0); Race (Hispanic = 1, others = 0); Veteran status (yes = 1, no = 0).

https://doi.org/10.1371/journal.pone.0178153.t002

Table 3. Parameter estimates for the truncated bivariate poisson models using the data on number of conditions and utilization of healthcare services (HRS, 2010).

Variables	Estimate	S.E.	z-value	p-value	$\sqrt{\varphi_r \boldsymbol{V}(\hat{\beta})}$	p-value
Marginal model for (Y_1)						
β_{10}	−0.180	0.2107	−0.854	0.393	0.1657	0.278
Gender	−0.063	0.0229	−2.765	0.006	0.0182	0.001
Age (years)	0.016	0.0029	5.719	0.000	0.0022	0.000
Race	0.008	0.0310	0.250	0.802	0.0244	0.751
Veteran status	0.056	0.0268	2.072	0.038	0.0213	0.009
Conditional model for (Y_2)						
β_{10}	0.888	0.3592	2.472	0.013	0.3126	0.005
Gender	0.453	0.0381	11.883	0.000	0.0337	0.000
Age (years)	−0.030	0.0049	−6.144	0.000	0.0043	0.000
Race	−0.249	0.0577	−4.310	0.000	0.0517	0.000
Veteran status	0.219	0.0418	5.252	0.000	0.0354	0.000

Gender (male = 1, female = 0); Race (Hispanic = 1, others = 0); Veteran status (yes = 1, no = 0).

https://doi.org/10.1371/journal.pone.0178153.t003

There is positive association between utilization of healthcare services and veteran status, similar to the association found with number of conditions observed from the marginal model.

The estimates for the right truncated bivariate Poisson model are displayed in Table 3. The results are generally similar to the fit of the untruncated model displayed in Table 2 but estimates of the parameters for the right truncated model show substantial difference from that of the untruncated model and it is more evident for the conditional part of the joint model. Let us denote $\beta_1* = (\beta_{11}, \ldots, \beta_{1p})'$, $\beta_2* = (\beta_{21}, \ldots, \beta_{2p})'$, $\beta* = (\beta_1{}', \beta_2{}')$ then for the overall test for the null hypothesis H_0: $\beta* = 0$, the following likelihood ratio test is used $\chi^2 = -2[\ln L(\hat{\beta}_{10}, \hat{\beta}_{20}) - \ln L(\hat{\beta})]$ where $\beta = (\beta_1{}', \beta_2{}')$. Here the likelihood ratio test statistic is asymptotically chi square with 2p degrees of freedom. Both the models for untruncated and truncated appear statistically significant. The estimated regression coefficients for gender for the untruncated model is 0.345 compared to 0.453 in the truncated model. Similarly, the estimated regression coefficient for veteran status is 0.093 and 0.219 in the untruncated and truncated models respectively. However, standard errors have not shown much change. The likelihood ratio test statistics show p-values less than 0.001 for both the untruncated and truncated models. It may be noted here that both the untruncated and truncated models are fitted to examine the changes in the estimation of parameters and tests attributable to truncation. Test for over (under) dispersion for both the models show that the fitted models are not significantly over or under dispersed.

The estimated dispersion parameters for untruncated models, $\hat{\varphi}_{u,1}$ and $\hat{\varphi}_{u,2}$, are 0.80 and 1.05 respectively and from the truncated models $\hat{\varphi}_{t,1}$ and $\hat{\varphi}_{t,2}$ are 0.90 and 0.83 respectively (see Table 4). Adjusting for both untruncted models and truncated models, we observe that the test results for significance of the parameters remain similar.

Table 4 summarizes the results for both untruncated and truncated bivariate Poisson regression models. We observe from goodness of fit test statistic, T_1, that both untruncated and truncated bivariate Poisson regression models fit the data very well. In other words, the null hypotheses may be accepted in favor of the proposed models. To make a more precise decision about the choice of the better model, the Voung test is employed. We have used the Voung test for comparing non-nested models because non-nested models can not be

Table 4. Model statistics for untruncated and truncated bivariate poisson models for Health and Retirement Study data (HRS) Data.

Statistics	Reduced Model		Full Model	
	Untruncated	Right truncated	Untruncated	Right truncated
Log Likelihood	−16707.66	−16596.21	−16586.07	−16500.71
AIC	33419.33	33196.42	33192.13	33021.41
BIC	33432.58	33209.67	33258.38	33087.66
T_1 (D.F., p-value)	9.34 (12, 0.622)	11.89 (12, 0.455)	9.73 (12, 0.640)	11.80 (12, 0.462)
T_2 (D.F., p-value)	12.26 (12, 0.425)	13.49 (12, 0.335)	12.03 (12, 0.444)	13.43 (12, 0.339)
LRT: Reduced vs. full model (D.F., p-value)	–	–	243.18 (8, 0.000)	191.01 (8, 0.000)
φ_1	0.78	0.90	0.80	0.90
φ_2	1.03	0.83	1.05	0.83
V: Voung test (p-value)		12.67 (0.000)		6.87 (0.000)

https://doi.org/10.1371/journal.pone.0178153.t004

compared using the classical likelihood ratio based tests. Table 4 shows comparison between untruncated and right truncated models and it is found that the bivariate right truncated Poisson model appears to be significantly better than the untruncated model ($p<0.001$).

Conclusion

The use of bivariate count regression models can provide very useful insights for analyzing repeated measures data emerging from various fields including health. In this paper, a comprehensive methodology has been displayed to deal with different types of problems associated with bivariate count data. In this paper, a new model is proposed using the Poisson-Poisson marginal-conditional approach for both untruncated and truncated data. In reality, there may be influence of right truncation for analyzing such data and the application displayed in this paper clearly demonstrates substantial change in the estimates of regression parameters. A generalized linear model for the bivariate Poisson-Poisson regression model is developed for the right truncated data. The proposed model is applied to the Health and Retirement Study data on number of conditions and utilization of healthcare services. The problem of under or overdispersion in the count data is also examined and appropriate procedures for adjusting the standard errors are also shown. Tests for both goodness of fit and overdispersion have been used and it is observed that the models fit well to the bivariate count data and the extent of under or overdispersion are not found to be significant in the application to the data on number of conditions and utilization of healthcare services from the Health and Retirement Study. In order to select a better model, an extended test for non-nested models for bivariate count data is shown in this paper and it is found that the right truncated bivariate Poisson model is a better choice for the health data used in this study.

Appendix

$$E(Y_1) = \sum_{y_1=0}^{k_1} \left[y_1 \left(\frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!} \right) \left(\frac{1}{\sum_{y_1=0}^{k_1} \frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!}} \right) \right] = \frac{\lambda_1 \Gamma(k_1, \lambda_1) k_1}{\Gamma(k_1 + 1, \lambda_1)}.$$

$$E(Y_1^2) = \sum_{y_1=0}^{k1}\left[y_1^2\left(\frac{e^{-\lambda_1}\lambda_1^{y_1}}{y_1!}\right)\left(\frac{1}{\sum_{y_1=0}^{k_1}\frac{e^{-\lambda_1}\lambda_1^{y_1}}{y_1!}}\right)\right] = \frac{\Gamma(k1+1)}{\Gamma(k1+1,\lambda_1)}$$
$$\left(\lambda_1(\lambda_1+1) - \frac{(k1+1)^2\lambda_1^{k1+1}hypergeometric([1,k1+2],[k1+1,k1+1],\lambda_1)}{(k1+1)!e_1^{\lambda}}\right).$$

$$Var(Y_1) = V_1 = E(Y_1^2) - [E(Y_1)]^2 = -\frac{(\lambda_1^{k1+1})^2(\Gamma(k1,\lambda_1))^2k1^2}{(\Gamma(k1+1,\lambda_1))^2(\lambda_1^{k1})^2} + \frac{\Gamma(k1+1)}{\Gamma(k1+1,\lambda_1)}$$
$$\left(\lambda_1(\lambda_1+1) - \frac{(k1+1)^2\lambda_1^{k1+1}hypergeometric([1,k1+2],[k1+1,k1+1],\lambda_1)}{(k1+1)!e^{\lambda_1}}\right).$$

Similarly,

$$E(Y_2) = \sum_{y_2=0}^{k2}\left[y_2\left(\frac{e^{-\lambda_2}\lambda_2^{y_2}}{y_2!}\right)\left(\frac{1}{\sum_{y_2=0}^{k_2}\frac{e^{-\lambda_2}\lambda_2^{y_2}}{y_2!}}\right)\right] = \frac{\lambda_2\Gamma(k2,\lambda_2)k2}{\Gamma(k2+1,\lambda_2)}.$$

$$Var(Y_2) = E(Y_2^2) - [E(Y_2)]^2 = -\frac{(\lambda_2^{k2+1})^2(\Gamma(k2,\lambda_2))^2k2^2}{(\Gamma(k2+1,\lambda_2))^2(\lambda_2^{k2})^2} + \frac{\Gamma(k2+1)}{\Gamma(k2+1,\lambda_2)}$$
$$\left(\lambda_2(\lambda_2+1) - \frac{(k2+1)^2\lambda_2^{k1+1}hypergeometric([1,k2+2],[k2+1,k2+1],\lambda_2)}{(k2+1)!e^{\lambda_2}}\right).$$

$$E(Y_2 \mid y_1) = \sum_{y_2=0}^{k2}\left[y_2\left(\frac{e^{-\lambda_2 y_1}(\lambda_2 y_1)^{y_2}}{y_2!}\right)\left(\frac{1}{\sum_{y_2=0}^{k2}\frac{e^{-\lambda_2 y_1}(\lambda_2 y_1)^{y_2}}{y_2!}}\right)\right]$$
$$= \frac{(\lambda_2 y_1)\Gamma(k2,\lambda_2 y_1)k2}{\Gamma(k2+1,\lambda_2 y_1)}.$$

$$E(Y_2^2|y_1) = \sum_{y_2=0}^{k2}\left[y_2^2\left(\frac{e^{-\lambda_2 y_1}(\lambda_2 y_1)^{y_2}}{y_2!}\right)\left(\frac{1}{\sum_{y_2=0}^{k2}\frac{e^{-\lambda_2 y_1}(\lambda_2 y_1)^{y_2}}{y_2!}}\right)\right] = \frac{\Gamma(k2+1)}{\Gamma(k2+1,\lambda_2 y_1)}$$
$$\left(\lambda_2^2y_1^2 + \lambda_2 y_1 - \frac{(k2+1)^2(\lambda_2 y_1)^{k2+1}hypergeometric([1,k2+2],[k2+1,k2+1],\lambda_2 y_1)}{e^{\lambda_2 y_1}(k2+1)!}\right).$$

$$Var(Y_2 \mid y_1) = V_{2|1} = E(Y_2^2|y_1) - [E(Y_2 \mid y_1)]^2 =$$
$$-\frac{((\lambda_2 y_1)^{k2+1})^2(\Gamma(k2, \lambda_2 y_1))^2 k2^2}{(\Gamma(k2+1, \lambda_2 y_1))^2((\lambda_2 y_1)^{k2})^2} + \frac{\Gamma(k2+1)}{\Gamma(k2+1, \lambda_2 y_1)}$$
$$\left(\lambda_2^2 y_1^2 + \lambda_2 y_1 - \frac{(k2+1)^2(\lambda_2 y_1)^{k2+1} hypergeometric([1, k2+2], [k2+1, k2+1], \lambda_2 y_1)}{e^{\lambda_2 y_1}(k2+1)!}\right).$$

Acknowledgments

We acknowledge gratefully that the study is supported by the HEQEP sub-project 3293, University Grants Commission of Bangladesh and the World Bank. The authors also acknowledge gratefully to the HRS (Health and Retirement Study) conducted by the University of Michigan for making the data publicly available.

Author Contributions

Conceptualization: MAI RIC.

Formal analysis: RIC.

Funding acquisition: MAI.

Investigation: MAI RIC.

Methodology: MAI RIC.

Project administration: MAI.

Resources: MAI RIC.

Software: RIC.

Supervision: MAI RIC.

Validation: MAI RIC.

Visualization: MAI RIC.

Writing – original draft: MAI RIC.

Writing – review & editing: MAI RIC.

References

1. Leiter RE, Hamdan MA. Some Bivariate Probability Models Applicable to Traffic Accidents and Fatalities. International Statistical Review. 1973 41: 87–100. https://doi.org/10.2307/1402790
2. Cacoullos T, Papageorgiou H. On Some Bivariate Probability Models Applicable to Traffic Accidents and Fatalities. International Statistical Review. 1980; 48: 345–356. https://doi.org/10.2307/1402946
3. Consul PC, Jain GC. A Generalization of the Poisson Distribution. Technometrics. 1973 15: 791–799. https://doi.org/10.1080/00401706.1973.10489112
4. Consul PC. Some Bivariate Families of Lagrangian Probability Distributions. Communications in Statistics—Theory and Methods. 1994 23: 2895–2906. https://doi.org/10.1080/03610929408831423
5. Consul PC. Generalized Poisson Distributions: Properties and Applications. Marcel Dekker. 1989.

6. Consul PC, Shoukri MM. The generalized Poisson distribution when the sample mean is larger than the sample variance. Communications in Statistics—Simulation and Computation. 1985 14: 1533–1547.
7. Holgate P. Estimation for the Bivariate Poisson Distribution. Biometrika. 1964 51: 241–245. https://doi.org/10.1093/biomet/51.1-2.241
8. Jung R, Winkelmann R. Two Aspects of Labor Mobility: A Bivariate Poisson Regression Approach. Empirical economics. 1993 18: 543–556. https://doi.org/10.1007/BF01176203
9. Karlis D Ntzoufras I. Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R. Journal of Statistical Software. 2005 14: 1–36.
10. Hofer V, Letiner JA. bivariate Sarmanov regression model for count data with generalised Poisson marginals. Journal of Applied Statistics. 2012 39, 2599–2617. https://doi.org/10.1080/02664763.2012.724661
11. Kocherlakota S, Kocherlakota K. Bivariate Discrete Distributions. Marcel Dekker. 1992.
12. Vernic R. On The Bivariate Generalized Poisson Distribution. ASTIN Bulletin: The Journal of the International Actuarial Association. 1997 27: 23–32. https://doi.org/10.2143/AST.27.1.542065
13. Karlis D, Ntzoufras I. Analysis of Sports Data by Using Bivariate Poisson Models. Journal of the Royal Statistical Society Series D (The Statistician). 2003 52: 381–393. https://doi.org/10.1111/1467-9884.00366
14. Islam MA, Chowdhury RI. A Bivariate Poisson Models with Covariate Dependence. Bulletin of Calcutta Mathematical Society. 2015 107: 11–20.
15. Gurmu S, Trivedi PK. Overdispersion Tests for Truncated Poisson Regression Models. Journal of Econometrics. 1992 54: 347–370. https://doi.org/10.1016/0304-4076(92)90113-6
16. Cameron A, Trivedi P. Regression Analysis of Count Data. Oxford University Press. 1998.
17. Health and Retirement Study (HRS). Wave 10- Public Use Dataset. Produced and Distributed by the University of Michigan with Funding from the National Institute on Aging (Grant Number NIA U01AG09740). Ann Arbor, MI. 2010.
18. Abramowitz M, Stegun I.A. Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables. National Bureau of Standards, United States Department of Commerce. 1964.
19. McCullagh P. The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data, Journal of the American Statistical Association. 1986 81: 104–107. https://doi.org/10.1080/01621459.1986.10478244
20. McCullagh P, Nelder JA. Generalized Linear Models, Second Edition. Chapman and Hall/CRC. 1989.
21. Long JS. Regression Models for Categorical and Limited Dependent Variables. SAGE Publications, Thousand Oaks. 1997.
22. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989 57: 307–333. https://doi.org/10.2307/1912557