Multivariate Generalized Poisson Distribution for Interference on Selected Non-Communicable Diseases in Lagos State, Nigeria

Adewara Johnson Ademola
University of Lagos, Akoka, Lagos, Nigeria

Mbata Ugochuckwu Ahamefula
University of Lagos, Akoka, Lagos, Nigeria

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Available at: http://digitalcommons.wayne.edu/jmasm/vol11/iss2/23

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.
Multivariate Generalized Poisson Distribution for Interference on Selected Non-Communicable Diseases in Lagos State, Nigeria

Cover Page Footnote
The authors express their profound gratitude to the entire staff and management of General Hospital, Lagos Island, Lagos, Nigeria, for their wonderful support during data collection.
Multivariate Generalized Poisson Distribution for Inference on Selected Non-Communicable Diseases in Lagos State, Nigeria

Adewara Johnson Ademola  Mbata Ugochukwu Ahamefula
University of Lagos,
Akoka, Lagos, Nigeria

Multivariate Generalized Poisson Distribution (MGPD) models are applied to make inferences regarding non-communicable diseases, diabetes, hypertension, stroke and ulcer in Lagos State, Nigeria. The generalized Poisson distribution is employed due to its usefulness in modeling count data in the presence of either over- or under- dispersion. Results show that the correlation between ulcer and stroke is not significant. Other pairwise comparisons of diseases are significant, thus implying that a patient who suffers from diabetes or stroke has a high propensity to also be hypertensive.

Key words: Multivariate generalized Poisson distribution, non-communicable disease, correlation, pairwise comparison.

Introduction

A multivariate generalized Poisson distribution (MGPD) is a discrete probability distribution that has the capacity to estimate the marginal mean and variance, which are univariate generalized Poisson distributions (GPD). MGPD allows for any form of correlation and can be used to describe count data with any type of dispersion. Several studies and applications have been conducted in count data modeling but fewer numbers of studies use MGPD models. Ordinary Poisson models have little ability to manage dispersion; in such cases, generalized Poisson distributions are used. According to Famoye, et al. (2011), a Poisson distribution satisfies the equi-dispersion property because its mean equals the variance, however, this property does not hold when the data is over-dispersed (when the variance exceeds the mean) or under-dispersed (when the variance is less than the mean). When data is over- or under-dispersed, it is necessary to employ a probability model.

Among the alternatives to the Poisson distribution are the negative binomial distribution (NBD), the generalized Poisson distribution (GPD) and the Poisson log-normal distribution (PLD). The generalized Poisson distribution is relevant because it has the capacity to measure either over- or under-dispersion. Notable works in this area include Consul (1989), Vernic (1997, 1999, 2000), Tsiamyrtzis and Karlis (2004) and Famoye (2010). This article estimates the mean prevalence of selected non-communicable diseases using correlation. A monthly retrospective investigation and routine check of hospital records of patients was used to count the number of patients living with any of these health problems at the General Hospital, Lagos, Nigeria from January 2005 to March 2012.

Non-Communicable Diseases

Onwasigwe (2010) defined non-communicable diseases (NCDs) as diseases which are not contagious, that is, they cannot be transmitted from one person to another. Non-communicable diseases are chronic conditions that do not result from an acute infectious process but cause death, dysfunction or
impairment in quality of life. NCDs typically develop over a relatively long time without causing symptoms but after the disease manifests there may be a period of prolonged impaired health.

The World Health Organization (WHO) in (2011a) released the first Global Status Report on Non-Communicable Diseases which outlines the statistics, evidence and experiences needed for a more forceful response to the growing threat posed by NCDs. WHO (2011b) further provided an overview of each country’s profile, the number, rates and causes of deaths from NCDs, the prevalence of selected risk factors, trends in metabolic risk factors in each country and information describing current prevention and control of NCDs.

Non-communicable diseases constitute a leading cause of functionary impairment and death globally. Nigeria loses about 400 million dollars yearly in national income from premature deaths from diseases such as diabetes mellitus, hypertension, cancer, renal failure and stroke and the economic cost from premature deaths due to NCDs is expected to rise to 8 billion dollars (Ogbebo, 2011). The WHO (2011b) report of year 2011 put mortality from NCDs at 59.8% of the global mortality and NCDs were estimated to account for 27% of all deaths in Nigeria. Due to the growing problem with these diseases in Nigeria among both young and old people, this study examines the mean prevalence and correlation of diabetes, hypertension, stroke and ulcer.

Multivariate Generalized Poisson Distribution

Famoye, et al. (2011) defined a new multivariate (d-variate) generalized Poisson distribution (MGPD) as a product of generalized Poisson marginals with a multiplicative factor. The probability function of the MGPD is

\[
P(y_1, ..., y_d) = \prod_{t=1}^{d} \frac{\theta_t^{y_t} (1+\alpha_t y_t)^{\kappa_t}}{y_t!} \exp\left[-\theta_t (1+\alpha_t y_t)\right] \\
\times \left[1+ \sum_{t<v} \lambda_{tv} \left(e^{-\gamma_t} - c_t\right)\left(e^{-\gamma_v} - c_v\right)\right] \\
\]

(1)

where \( c_t = E\left(e^{-Y_t}\right) = \exp\left[\theta_t (s_t - 1)\right] \) with \( \ln s_t - \alpha_t \theta_t (s_t - 1) + 1 = 0 \);

\[
c_{tv} = E\left(Y_t e^{-Y_v}\right) = \theta_t (1-\alpha_t \theta_t s_t)^{\kappa_t} \exp\left[\theta_t (1-\alpha_t \theta_t (s_t - 1) - 1\right] \\
\text{with} \quad \ln s_t - \alpha_t \theta_t (s_t - 1) + 1 = 0 \quad \text{and} \quad y_1, ..., y_d = 0, 1, 2, ..., 
\]

The variate \( y_t \) exhibits some dispersion when \( \alpha_t \neq 0 \). It is under-dispersed when \( \alpha_t < 0 \) and is over-dispersed when \( \alpha_t > 0 \). The variates \( y_t \) and \( y_v \) are correlated when \( \lambda_{tv} \neq 0 \). The pair is negatively (or positively) correlated \( \lambda_{tv} < 0 \) as (or \( \lambda_{tv} > 0 \)). According to Famoye (2011), the \( d \) marginal means and \( d \) marginal variances of the MGPD are:

\[
\mu_t = \theta_t (1-\alpha_t \theta_t)^{-1}, \quad t = 1, 2, ..., d \\
\mu_t = \theta_t (1-\alpha_t \theta_t)^{-3}, \quad t = 1, 2, ..., d \\
\]

(2) 

(3)

respectively. The \( d \) \((d - 1)/2\) covariances between any two variates are

\[
\sigma_{tv} = \lambda_{tv} \left(c_t - c_t \mu_t\right)\left(c_v - c_v \mu_v\right), \quad t, v = 1, 2, ..., d, \text{ and} \quad t < v \\
\]

(4)

Using the covariance \( \sigma_{tv} \) between the variables \( Y_t \) and \( Y_v \) in equation (4), the correlation coefficient between \( Y_t \) and \( Y_v \) is

\[
\rho_{tv} = \frac{\sigma_{tv}}{\sqrt{\sigma_{tv}} \sqrt{\sigma_{tv}}} = \lambda_{tv} \left(c_t - c_t \mu_t\right)\left(c_v - c_v \mu_v\right) / \left(\sigma_t \sigma_v\right). \\
\]

(5)

Thus, the parameter \( \lambda_{tv} \) can be written in terms of the correlation coefficient \( \rho_{tv} \). The correlation coefficient can be positive, zero or negative depending on the value of \( \lambda_{tv} \). The parameter
\( \lambda_n \) satisfies \( |\lambda_n| \leq 1 \left[ \left( 1 - c_t \right) \left( 1 - c_v \right) \right] \). Using this result, the correlation coefficient satisfies the condition:

\[
|\rho_v| \leq \frac{(c_v - c_t \mu_t)(c_{tv} - c_t \mu_y)}{\sigma_t \sigma_v \left( 1 - c_t \right) \left( 1 - c_v \right)}.
\]

Parameter Estimation

Assuming \( n \) independent vectors \( (y_{it}, y_{iv}, \ldots, y_{in}) \), where the \( i \)th vector has the MGPD in (1). The sample mean and sample variance are denoted by \( \bar{y}_t = \frac{\sum_{i=1}^n y_{it}}{n} \) and \( s_i^2 = \frac{\sum_{i=1}^n (y_{it} - \bar{y}_t)^2}{(n-1)} \), where \( t = 1, 2, \ldots, d \) respectively. The sample covariance between the variables \( Y_t \) and \( Y_v \) is denoted by

\[
s_{tv} = \frac{\sum_{i=1}^n (y_{it} - \bar{y}_t)(y_{iv} - \bar{y}_v)}{(n-1)}.
\]

Equating these sample moments to the corresponding population moments, the moment estimates for the MGPD are given by

\[
\hat{\theta}_t = \sqrt{\frac{\bar{y}_t^2}{s_i^2}}, \quad t = 1, 2, \ldots, d
\]

\[
\hat{\alpha}_t = \hat{\theta}_t^{-1} - \bar{y}_t^{-1}, \quad t = 1, 2, \ldots, d
\]

\[
\hat{\lambda}_{tv} = s_{tv} \left( \hat{c}_t - \hat{c}_t \bar{y}_t \right) \left( \hat{c}_v - \hat{c}_v \bar{y}_v \right)^{-1}
\]

\[
\frac{\hat{\alpha}_t}{\hat{\theta}_t} \times 100 = \text{Coefficient of Dispersion},
\]

where \( \hat{c}_t \) and \( \hat{c}_v \) are the estimated values of \( c_t \) and \( c_v \) (where \( t, v = 1, 2, \ldots, d \) and \( t < v \)). In general, for \( d \)-variates MGPD, equations (6)-(8) provide a total of \( 2d + d(d-1)/2 \) equations, which are solved simultaneously to obtain the moment estimates. For the log-likelihood function and estimation of the parameters, see Famoye (2010) and Famoye, et al. (2011).

The log-likelihood function, \( \log L = \log L(\theta, \alpha, \lambda; y) \), for the MGPD is:

\[
\log L = \sum_{i=1}^n \sum_{t=1}^d \left[ \frac{y_{it} \log \theta_t + (y_{it} - 1) \log (1 + \alpha_t y_{it})}{\sigma_t \sigma_v \left( 1 - c_t \right) \left( 1 - c_v \right)} \right]
\]

The log-likelihood in (10) is maximized over the parameters \( \theta_t, \alpha_t (t = 1, 2, \ldots, d) \), and \( \lambda_{tv} (t, v = 1, 2, \ldots, d \text{ and } t < v) \). Famoye, et al. (2011) concluded that a measure of goodness of fit for the MGPD may be based on the log-likelihood statistic given in (10).

Methodology

The data for this research was collected from Island Hospital, Lagos. The numbers of patients suffering from diabetes, hypertension, stroke and/ulcer from January 2005 to March 2012 were counted. The total number of patients observed suffering from diabetes, hypertension, stroke and ulcer were: 7,898, 10,055, 8,565 and 5,604 respectively. These figures sum to 32,122 out of 61,786 patients observed, constituting 51.99%. The parameters \( \alpha_t \) and \( \alpha_v \) were estimated using Excel and R statistical packages; descriptive statistics and correlations \( \rho_v \) between the pairs of the diseases were also estimated.

Test for Constant Dispersion Parameter

A test of dispersion was conducted of the parameters for MGPD. The test was given as \( \alpha = 0 (t = 1, 2, \ldots, d) \), with a test hypothesis for constant dispersion:

\[
H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_d
\]

\[
\text{vs}
\]

\[
H_1: \alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_d
\]

526
Let $L_{\text{con}}$ be the likelihood function when $H_0$ is true and let $L_a$ be the likelihood function when $H_0$ is false. The test statistic is given by $\chi^2 = -2\log\left(\frac{L_{\text{con}}}{L_a}\right)$, which is approximately a Chi-square with $d - 1$ degrees of freedom (Famoye, et al., 2011).

Results
Data was analyzed using the MGPD model to obtain descriptive statistics and to determine the parameters $\tilde{\theta}_t$, $\tilde{a}_t$, $\tilde{\mu}_t$, and $\tilde{\sigma}_t^2$. The correlation between the variables under consideration (see Table 1) shows that ulcer and stroke are not significant, however, the remaining paired variables are significant. This indicates that a patient suffering from any of these diseases also has a high degree of risk for the other diseases. For example, a diabetic patient may also be expected to be hypertensive. However, inferences regarding the selected non-communicable diseases were based on the fitted MGPD model (see Table 2).

As shown in Table 2, the estimates for the dispersion parameter $\tilde{a}_t$, which is greater than zero are 0.176, 0.162, 0.219 and 0.112 indicating an over-dispersion in the data collected for each disease. Moreover, the result of the coefficient of dispersion (CD) in percentage also supports a constant dispersion. Testing the hypothesis on constant dispersion of parameters; a Chi-square value of 7.34 on 3 degrees for freedom is obtained with a $p$-value of 0.3713. Result show that dispersion exists and that the dispersion parameters are constant. The parameter estimates from MGPD with their corresponding standard errors are presented in Table 2. It can be observed that the standard deviation from the MGPD and standard deviations from the sample information are approximately equal. Furthermore, it may be inferred that the mean prevalence of diabetic, hypertension, stroke and ulcer patients per month are approximately 91, 115, 99 and 64, respectively. In addition, the estimates are significant using the confidence interval provided by the estimate and standard error.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Diabetic</th>
<th>Hypertension</th>
<th>Stroke</th>
<th>Ulcer</th>
<th>$\bar{y}_t$</th>
<th>$s_t^2$</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>90.782</td>
<td>26265.661</td>
<td>162.067</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.772*</td>
<td>1</td>
<td></td>
<td></td>
<td>115.575</td>
<td>45120.922</td>
<td>212.417</td>
</tr>
<tr>
<td>Stroke</td>
<td>0.639*</td>
<td>0.612*</td>
<td>1</td>
<td></td>
<td>98.448</td>
<td>50604.320</td>
<td>224.954</td>
</tr>
<tr>
<td>Ulcer</td>
<td>0.286*</td>
<td>0.388*</td>
<td>0.046</td>
<td>1</td>
<td>64.414</td>
<td>4378.222</td>
<td>66.168</td>
</tr>
</tbody>
</table>

*Correlation is significant at $\alpha = 0.01$ level
Conclusion

Count data was modeled using MGPD, which allows a determination regarding whether data is over- or under-dispersed. This is an advantage of the MGPD model over other discrete probability models such as negative binomial distribution (NBD), Poisson distribution (PD), Poisson log-normal distribution (PLD), and multivariate Poisson distribution (MPD). Out of 61,786 patients observed, 32,122 suffered from a non-communicable disease, constituting about 51.99%. This figure somewhat supports the statistical evidence highlighted by WHO (2011a, 2011b) regarding non-communicable diseases. Results from this investigation show that there is a high correlation between the pairs of the diseases diabetes, hypertension and stroke. Thus, it may be stated that a patient who suffers from diabetes or stroke is also likely to be hypertensive. Hence, continued study is highly recommended to investigate the threats posed by non-communicable diseases globally.

Acknowledgements

The authors express their profound gratitude to the entire staff and management of General Hospital, Lagos Island, Lagos, Nigeria, for their wonderful support during data collection.

References


Table 2: Parameter Estimates for Non Communicable Diseases Studied

<table>
<thead>
<tr>
<th>Disease</th>
<th>$\tilde{\theta}_i$</th>
<th>$\tilde{a}_i$</th>
<th>$\tilde{\mu}_i$</th>
<th>$\tilde{\sigma}_i^2$</th>
<th>SD</th>
<th>Estimate ± Standard Error</th>
<th>CD(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic</td>
<td>5.337</td>
<td>0.176</td>
<td>90.762</td>
<td>26249.523</td>
<td>162.017</td>
<td>90.762 ± 17.370 *</td>
<td>3.298</td>
</tr>
<tr>
<td>Hypertension</td>
<td>5.849</td>
<td>0.162</td>
<td>115.451</td>
<td>44980.595</td>
<td>212.086</td>
<td>115.451 ± 22.738 *</td>
<td>2.770</td>
</tr>
<tr>
<td>Stroke</td>
<td>4.360</td>
<td>0.219</td>
<td>98.678</td>
<td>50546.837</td>
<td>224.826</td>
<td>98.678 ± 24.104 *</td>
<td>5.023</td>
</tr>
<tr>
<td>Ulcer</td>
<td>7.813</td>
<td>0.112</td>
<td>64.412</td>
<td>4377.858</td>
<td>66.165</td>
<td>64.412 ± 7.094 *</td>
<td>1.434</td>
</tr>
</tbody>
</table>

*Significant at $\alpha = 0.05$ level; CD, coefficient of dispersion


