Austrian Journal of Statistics
July 2023, Volume 52, 96–113.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i3.1465

Using the Discrete Lindley Distribution to Deal with Over-dispersion in Count Data

Mien T.N. Nguyen
Mahidol University

Man V.M. Nguyen
Mahidol University

Ngoan T. Le
Duy Tan University,
International University
of Health & Welfare

Abstract

Count data in environmental epidemiology or ecology often display substantial over-dispersion, and failing to account for the over-dispersion could result in biased estimates and underestimated standard errors. This study develops a new generalized linear model family to model over-dispersed count data by assuming that the response variable follows the discrete Lindley distribution. The iterative weighted least square is developed to fit the model. Furthermore, asymptotic properties of estimators, the goodness of fit statistics are also derived. Lastly, some simulation studies and empirical data applications are carried out, and the generalized discrete Lindley linear model shows a better performance than the Poisson distribution model.

Keywords: count data, generalized linear model, discrete Lindley distribution, over-dispersion, distributed lag nonlinear model.

1. Introduction

The Poisson distribution is standard in modeling count data because it is easy to use and interpret. For instance, Almeida, Casimiro, and Calheiros (2010) fit the generalized additive model (GAM) to describe the effect of daily average temperature on Poisson daily mortality in Lisbon and Oporto cities of Portugal. Linares and Diaz (2010) investigate the relationship between daily $PM_{2.5}$ concentrations and hospital admissions in Madrid city (Spain) by the generalized linear model (GLM) with the Poisson distribution. However, a critical property of Poisson distribution is that the variance equals the expectation. For most of the observed count data, this property is often violated. Count data often display substantial over-dispersion, in which the variance of the outcome is greater than its mean. Over-dispersion is attributed to omitted-variable problems, the existence of outliers, heteroskedasticity (Rigby, Stasinopoulos, and Akantziliotou 2008). Crawley (2012) and Hilbe (2011) state that applying the Poisson model on over-dispersed data could result in biased estimates because of ignoring dispersion parameters in the fitted model. Another disadvantage is that the standard errors of the estimates are downward biased, so the explanatory variables tend to have significant impacts on the response variable, although they do not (Faddy and Smith 2011).

Various statistical procedures have been developed to overcome the over-dispersed problem in the last several years. A common way is to apply the Poisson quasi-likelihood (McCullagh and Nelder 2019) by specifying the parameters relating to the dependence of mean on explanatory variables and the variance written as a multiplicative constant of the mean. Besides, the negative binomial model is another valuable technique for over-dispersion. Also, Zuur, Ieno, Walker, Saveliev, Smith *et al.* (2009) introduces zero-inflated Poisson as a tool for an excess of zeroes in the data; Harrison (2014) suggests the Poisson-lognormal deal with the observation-level random-effect model.

Abebe and Shanker (2018) use a discretization method based on an infinite series to generate the discrete Lindley distribution from the continuous Lindley distribution (Lindley 1958). They discuss statistical properties of the discrete Lindley distribution consisting of moments, skewness, kurtosis, and parameter estimation. They conclude that its index of dispersion is greater than 1, and the discrete Lindley distribution is suitable for a model with an over-dispersed response variable.

In this study, we introduce a new method to solve the problem of over-dispersion in modeling count data by combining the discrete Lindley distribution with the *GLM*, named as the generalized discrete Lindley linear model (*GDLLM*). Section 2 is about estimating parameters and some statistics derived from the *GDLLM*. The simulation study is carried out to explore the performance of the discrete Lindley model compared to the Poisson model at different levels of over-dispersion in section 3. Furthermore, demonstrations of the *GDLLM* using a few realistic data sets are included in section 4. The first data set relates to the ecologic field extracting from the *faraway* package of software R Core Team (2013), and the discrete Lindley model is compared to Poisson and negative binomial models to select the best model. The second example is an empirical analysis to discover the impact of daily temperature on the all-cause mortality at Dien Chau district, Nghe An province, Vietnam.

2. Mathematical background

2.1. Discrete Lindley (DL) distribution

The single random variable Y follows the discrete Lindley distribution with a single parameter θ, denoted by $Y \sim DL(\theta)$, if its probability mass function can be written in the following form (Abebe and Shanker 2018)

$$f(y,\theta) = \left(1 - e^{-\theta}\right)^2 (1+y)\, e^{-\theta y}\,, \tag{1}$$

for $y \in \mathbb{N}$ and $\theta > 0$. The function in Equation (1) can be rewritten as

$$f(y,\theta,\phi) = \exp\left\{-\theta\, y + 2\ln(e^{\theta} - 1) - 2\theta + \ln(1+y)\right\},$$

showing that discrete Lindley distribution is in the exponential family. In which, $\gamma = -\theta$ is the canonical parameter, the cumulant function is $b(\gamma) = -2[\ln(e^{-\gamma} - 1) + \gamma]$, the dispersion parameter equals 1 and the normalizing function is $\ln(1+y)$. From the properties of distribution in the exponential family, we derive the expectation and the variance of the random variable Y following discrete Lindley distribution as follows

$$\mathbb{E}(Y) = b'(\gamma) = -2\left[\frac{-e^{-\gamma}}{e^{-\gamma} - 1} + 1\right] = \frac{2}{e^{-\gamma} - 1} = \frac{2}{e^{\theta} - 1}\,, \tag{2}$$

$$\mathbb{V}(Y) = b''(\gamma) = \frac{2e^{-\gamma}}{(e^{-\gamma} - 1)^2} = \frac{2e^{\theta}}{(e^{\theta} - 1)^2}\,. \tag{3}$$

The index of dispersion (IOD) is defined by

$$IOD = \frac{\mathbb{V}(Y)}{\mathbb{E}(Y)} = \frac{e^{\theta}}{e^{\theta} - 1}\,. \tag{4}$$

The index of dispersion is greater than 1 for all $\theta > 0$. Hence, the discrete Lindley distribution is used to describe the over-dispersed data set.
Besides, for further analysis, denoting $\mu = \mathbb{E}(Y)$ and rewriting $\mathbb{V}(Y)$ with respect to μ, we have

$$\mathbb{V}(Y) = \frac{(2+\mu)\mu}{2}. \tag{5}$$

2.2. Generalized discrete Lindley linear model (GDLLM)

Fitting the model

Suppose $Y_1, Y_2, \cdots, Y_n$ are independent random variables, each with discrete Lindley distribution and

$$\eta_i = \ln(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{6}$$

where $\mu_i = \mathbb{E}(Y_i)$, $\boldsymbol{X}_i$ is a i^{th} row of the design matrix containing data on the k explanatory variables, $\boldsymbol{\beta}$ is the $k-$vector of parameters.
Then, model (6) can be rewritten in an equivalent form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}, \tag{7}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)^T$.
The log-likelihood function for all $Y_i's$ is

$$\ell(\boldsymbol{\eta}, \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i = \sum_{i=1}^{n} \Big[-\theta_i y_i + 2\ln(e^{\theta_i} - 1) - 2\theta_i + \ln(1+y_i)\Big]. \tag{8}$$

In general, the $n-$vector score and $n \times n$ information matrix are computed by

$$\boldsymbol{u} = \frac{\partial \ell}{\partial \boldsymbol{\eta}},$$

$$\boldsymbol{J} = \mathbb{E}\Big(-\frac{\partial^2 \ell}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^T}\Big).$$

Applying to the discrete Lindley distribution, the elements of vector $\boldsymbol{u}$ and matrix $\boldsymbol{J}$ are

$$u_i = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \frac{2(y_i - \mu_i)}{2+\mu_i}, \tag{9}$$

$$J_{il} = \begin{cases} \dfrac{2\mu_i}{2+\mu_i} & \text{if } i = l, \\ 0 & \text{if } i \neq l. \end{cases} \tag{10}$$

Also, we easily have that

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \boldsymbol{X}. \tag{11}$$

To obtain the maximum likelihood estimate for the parameters, we need to find the differentiation of (8) with respect to $\beta's$ and set them equal to zero. It turns out that the estimates $\widehat{\boldsymbol{\beta}}$ are the solutions of

$$\boldsymbol{X}^T\boldsymbol{u} = 0. \tag{12}$$

Applying the Newton-Raphson, the updated $\widehat{\boldsymbol{\beta}}^{(m+1)}$ at the $(m+1)^{th}$ iteration is

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \big(\boldsymbol{X}^T\boldsymbol{J}^{(m)}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{J}^{(m)}\boldsymbol{w}^{(m)}, \tag{13}$$

where $\boldsymbol{w}$ is the vector of ”pseudo data” and calculated by

$$\boldsymbol{w}^{(m)} = \boldsymbol{\eta}^{(m)} + \big(\boldsymbol{J}^{(m)}\big)^{-1}\boldsymbol{u}^{(m)}.$$

The iterative weighted least square is implemented to update $\boldsymbol{\eta}, \boldsymbol{\mu}$, and $\boldsymbol{w}$ each iteration. Developed from the general algorithm by Nelder and Wedderburn (1972), our specific algorithm for *GDLLM* is presented as follows

Algorithm 1

1. **Step 1** Set $\mu_i^{(0)} = \bar{y}$ and $D_0 = 0$. Calculate $\eta_i^{(0)} = \ln(\mu_i^{(0)})$.

2. **Step 2** Update

   - Construct "pseudo data" $w_i^{(0)} = \eta_i^{(0)} + \dfrac{y_i - \mu_i^{(0)}}{\mu_i^{(0)}}$,

   and weight $J_{ii}^{(0)} = \dfrac{2\mu_i^{(0)}}{2 + \mu_i^{(0)}}$.

   - Calculate $\widehat{\boldsymbol{\beta}}^{(1)}$ from (13).
   - Obtain the fitted value $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\mu}^{(1)}$.
   - Compute adequate convergence

   $\Delta = \left|\dfrac{D_1 - D_0}{D_0}\right|$ where $D_1 = D(y, \eta_1)$: the deviance of the model.

3. **Step 3** Repeat Step 2, replacing $\mu_i^{(0)}$ by $\mu_i^{(1)}$ and D_0 by D_1 until Δ below some small threshold.

Each step of the iterative weighted least square adjusts the pseudo dependent variable $\boldsymbol{w}$. We define an asymptotic approximation $\boldsymbol{w_0}$ of $\boldsymbol{w}$. We get

$$\boldsymbol{w} \approx \boldsymbol{w}_0 = \boldsymbol{\eta}_0 + (\boldsymbol{J}_0)^{-1}\boldsymbol{u}_0 \,.$$

Then, $\boldsymbol{w}$ has asymptotic mean $\boldsymbol{\eta}_0$ and asymptotic variance $(\boldsymbol{J}_0)^{-1}$. We, therefore, have the variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$

$$\begin{aligned}
\operatorname{cov}(\widehat{\boldsymbol{\beta}}) &= \operatorname{cov}\Big[\big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{w}_0\Big] \\
&= \Big[\big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{J}_0\Big]\operatorname{cov}(\boldsymbol{w}_0)\Big[\big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{J}_0\Big]^T \\
&= \big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{J}_0(\boldsymbol{J}_0)^{-1}\boldsymbol{J}_0\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1} \\
&= \big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}. \qquad (14)
\end{aligned}$$

When all the regularity conditions are met, we can extend to show that $\widehat{\boldsymbol{\beta}}$ is asymptotically normal distributed

$$N\Big(\boldsymbol{\beta}, \big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}\Big), \qquad (15)$$

or in the form of Wald statistic

$$\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)^T\big(\boldsymbol{X}^T\boldsymbol{J}_0\boldsymbol{X}\big)^{-1}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) \sim \chi^2(k). \qquad (16)$$

Deviance

The definition of deviance for a fitted model is

$$D = 2[\ell(\widehat{\boldsymbol{\mu}}_{max}, \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\mu}}, \boldsymbol{y})]\,, \qquad (17)$$

where $\ell(\widehat{\boldsymbol{\mu}}_{max}, \boldsymbol{y})$ indicates the maximized log-likelihood of the full model: the model with the maximum number of parameters that can be estimated, while $\ell(\widehat{\boldsymbol{\mu}}, \boldsymbol{y})$ denotes the maximum value of the log-likelihood function for the model of interest.

If the response variables $Y_1, Y_2, \cdots, Y_n$ are independent and each $Y_i \sim DL(\theta_i)$, the likelihood function is mentioned in (8). The maximum likelihood estimator for each θ_i is

$$\widehat{\theta}_i = \ln\Big(\frac{2+y_i}{y_i}\Big). \tag{18}$$

Then, the maximum log-likelihood for the full model

$$\ell(\widehat{\boldsymbol{\mu}}_{max}, \boldsymbol{y}) = -\sum_{i=1}^{n}(y_i+2)\ln\Big(\frac{2+y_i}{y_i}\Big) + 2\sum_{i=1}^{n}\ln\Big(\frac{2}{y_i}\Big) + \sum_{i=1}^{n}\ln(1+y_i). \tag{19}$$

Because $\mu_i = 2/(e^{\theta_i}-1)$, the log-likelihood as a function with respect to $\boldsymbol{\mu}$

$$-\sum_{i=1}^{n}(y_i+2)\ln\Big(\frac{2+\mu_i}{\mu_i}\Big) + 2\sum_{i=1}^{n}\ln\Big(\frac{2}{\mu_i}\Big) + \sum_{i=1}^{n}\ln(1+y_i). \tag{20}$$

Substitute the estimate $\widehat{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}$, $\ell(\widehat{\boldsymbol{\mu}}, \boldsymbol{y})$ is derived. Therefore, the deviance is

$$D = 2\sum_{i=1}^{n}\Big[y_i\ln\Big(\frac{y_i}{\widehat{\mu}_i}\Big) - (y_i+2)\ln\Big(\frac{y_i+2}{\widehat{\mu}_i+2}\Big)\Big]. \tag{21}$$

Under the hypothesis that the model is correct, the deviance D approximately has the Chi-square distribution

$$D \sim \chi^2(n-k)\,,$$

where k is the number of parameters in the model.

Goodness of fit

Testing the goodness of fit of the generalized linear model is not often based on the raw residuals $\widehat{\epsilon}_i = y_i - \widehat{y}_i$ because it does not imply the mean and variance relationship. Common residual types used in the generalized linear model are the Pearson residuals and the deviance residuals.

For the discrete Lindley *GLM*, the concrete formula of the Pearson residual is

$$\widehat{\epsilon}_i^{pear} = \frac{(y_i - \widehat{\mu}_i)}{\sqrt{(1+\widehat{\mu}_i/2)\widehat{\mu}_i}}\,. \tag{22}$$

Besides, if we write the deviance as the summation of n components d_i computed on the i^{th} datum

$$d_i = 2\Big[y_i\ln\Big(\frac{y_i}{\widehat{\mu}_i}\Big) - (y_i+2)\ln\Big(\frac{y_i+2}{\widehat{\mu}_i+2}\Big)\Big],$$

the deviance residuals can be defined

$$\widehat{\epsilon}_i^{dev} = \operatorname{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}\,. \tag{23}$$

These residuals should follow a distribution being approximately standardized Normal. Furthermore, the residuals should not show any apparent pattern in trend when plotted against the fitted value or any explanatory variable.

Akaike Information Criteria (AIC) is used to select different models. Also, for a model with over-dispersed outcome, the choice of model is based on the modified Akaike Information Criteria *QAIC* (Hastie and Tibshirani 1990), given by

$$QAIC = -2\big[\ell(\boldsymbol{\eta}, \boldsymbol{y}) - k\phi\big],$$

in which ϕ is the over-dispersed parameter, and it can be estimated by

$$\widehat{\phi} = \frac{\sum_{i=1}^{n}(\widehat{\epsilon}_i^{pear})^2}{n-k}. \tag{24}$$

3. Simulation study

To generate the over-dispersed data, we apply the result derived by Cameron and Trivedi (1998). Let Y_i have a Poisson distribution with parameter θ_i where

$$\begin{aligned}\theta_i &= \mu_i \varepsilon_i,\\ \mu_i &= \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}),\\ \epsilon_i &\sim Gamma(\alpha, \alpha),\end{aligned}$$

where $\boldsymbol{x}_i$ is a $k \times 1$ vector of explanatory variables. Note that $\mathbb{E}(\epsilon_i) = 1$ and $\mathbb{V}(\epsilon_i) = \dfrac{1}{\alpha}$. The unconditional distribution of y will be negative binomial $NB\Big(r = \alpha, p = \dfrac{\mu}{\mu+\alpha}\Big)$. Then, the mean and the variance will be $\mathbb{E}(Y) = \mu\mathbb{E}(\varepsilon_i) = \mu$ and $\mathbb{V}(Y) = \mu\Big(1 + \dfrac{\mu}{\alpha}\Big)$, presenting for over-dispersion.

We simulate 5000 samples, with sample size $n \in \{50, 100, 200\}$, through the following steps:

1. Two independent variables x_1 and x_2 are generated: x_1 is randomly chosen from the Normal distribution $N(1, 0.5)$, and x_2 is drawn from the Bernoulli distribution with success probability $p = 0.5$.

2. Using the independent variables, the mean of the dependent variable (μ) is generated as
$$\ln(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are assigned values $(1, 2, -0.5)$.

3. Generate $\varepsilon_i \sim Gamma(\alpha, \alpha)$.
Four values of α, including $4, 2, 0.5, 0.1$, are selected to present different levels of over-dispersion.

4. Compute $\lambda_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})\varepsilon_i$ and generate $Y_i \sim Poi(\lambda_i)$

Once the data are generated, the generalized linear model, which assumes the response variable following the discrete Lindley distribution, is employed. The model of generalized linear with Poisson response variable is applied, too. We are then going to compare two methods for each condition by using the common bias (CB), the standard deviation (SD), the mean squared error (MSE) of each estimated parameter based on the 5000 replicates. We have

$$\begin{aligned}\text{CB}(\beta_j) &= \frac{1}{5000}\sum_{r=1}^{5000}(\widehat{\beta}_{jr} - \beta_j),\\ \text{SD}(\beta_j) &= \frac{1}{5000}\sum_{r=1}^{5000}(\widehat{\beta}_{jr} - \overline{\widehat{\beta}_j})^2,\\ \text{MSE}(\beta_j) &= \frac{1}{5000}\sum_{r=1}^{5000}(\widehat{\beta}_{jr} - \beta_j)^2,\end{aligned}$$

where $\widehat{\beta}_{jr}$ is the estimated value of parameter β_j $(j = 0, 1, 2)$ at r^{th} simulation experiment and $\overline{\widehat{\beta}_j} = \dfrac{1}{5000}\displaystyle\sum_{r=1}^{5000}\widehat{\beta}_{jr}$.

Table 1: Table of common bias (CB), standard deviation (SD), mean square error (MSE) of parameters for varying levels of over-dispersion in case $n = 50$

Overdispersion	Parameter	CB		SD		MSE	
		DL	Poi	DL	Poi	DL	Poi
0.1	β_1	-0.1026	-0.2870	1.0655	1.2070	1.1457	1.5391
	β_2	0.0046	0.0374	1.0606	1.3110	1.1248	1.7201
0.25	β_1	-0.0455	-0.1395	0.6242	0.8506	0.3917	0.7430
	β_2	-0.0095	-0.0041	0.6380	0.8368	0.4071	0.7002
2	β_1	-0.0025	-0.0258	0.2288	0.3459	0.0524	0.1203
	β_2	-0.0016	-0.0025	0.2318	0.3087	0.0537	0.0953
4	β_1	0.0026	-0.0116	0.1737	0.2509	0.0302	0.0631
	β_2	-0.0049	-0.0025	0.1698	0.2215	0.0289	0.0490

Table 2: Table of common bias (CB), standard deviation (SD), mean square error (MSE) of parameters for varying levels of over-dispersion in case $n = 100$

Overdispersion	Parameter	CB		SD		MSE	
		DL	Poi	DL	Poi	DL	Poi
0.1	β_1	-0.0732	-0.1999	0.7327	0.9313	0.5422	0.9072
	β_2	-0.0252	-0.0534	0.6986	0.9074	0.4887	0.8262
0.25	β_1	-0.0257	-0.1041	0.4472	0.6351	0.2007	0.4142
	β_2	-0.0158	-0.0177	0.4186	0.5752	0.1754	0.3311
2	β_1	0.0001	-0.0115	0.1618	0.2494	0.0262	0.0623
	β_2	-0.0009	-0.0007	0.1550	0.2226	0.0240	0.0495
4	β_1	0.0031	-0.0063	0.1216	0.1792	0.0148	0.0322
	β_2	-0.0025	-0.0013	0.1150	0.1570	0.0132	0.0247

Table 3: Table of common bias (CB), standard deviation (SD), mean square error (MSE) of parameters for varying levels of over-dispersion in case $n = 200$

Overdispersion	Parameter	CB		SD		MSE	
		DL	Poi	DL	Poi	DL	Poi
0.1	β_1	-0.0496	0.1192	0.4821	0.6858	0.2349	1.799
	β_2	-0.0019	0.006	0.4743	0.6704	0.2250	0.4495
0.25	β_1	-0.0178	-0.0609	0.2968	0.4579	0.0884	1.2877
	β_2	0.0008	0.0011	0.2971	0.4313	0.0882	0.1861
2	β_1	-0.0039	-0.0121	0.1095	0.1682	0.0120	1.0240
	β_2	-0.0026	-0.0020	0.1069	0.1516	0.0114	0.0230
4	β_1	0.0007	-0.0022	0.0843	0.1210	0.0071	1.0241
	β_2	-0.0020	-0.0023	0.0821	0.1130	0.0068	0.0128

Table 1 reveals the CB, SD, and MSE for the estimated coefficients of the continuous covariate x_1 and the dichotomous variable x_2 in case sample size $n = 50$. Similarly, the measures of bias, standard deviation and MSE for cases $n = 100$ and $n = 200$ are summarized in Table 2 and 3.

The results in Table (1) - (3) present that, in all levels of over-dispersion and sample sizes, estimators based on the assumption of discrete Lindley distributions perform better than the case of Poisson distribution in all terms of common bias, standard error, and mean squared error. Besides, the common bias, standard error, and MSE of estimators in both models increase as the level of over-dispersion increases (α decreases). A larger sample size leads to more precise estimates of parameters in the discrete Lindley model, and the same conclusion is obtained for the Poisson model. In summary, the *GDLLM* appears to be a good recommendation for over-dispersion.

4. Empirical data analysis

4.1. Example 1: Species diversity on the Galapagos Islands

The *gala* data set is available in the *faraway* R package (Faraway 2004). In order to demonstrate the performance of the *GDLLM* to handle over-dispersion, various popular *GLM* on count data, including Poisson regression and negative binomial regression, are applied to the data set and compared to the *GDLLM* through AIC and QAIC criteria. The model with the smallest values of these criteria is considered the best-fitting model for the data set.

The response variable is the number of endemic species found in 30 Galapagos islands. The predictors used in this study include four variables:

- Area (km^2): the area of the island
- Elevation (m): the highest elevation of the island
- Nearest (km): the distance from the nearest island
- Adjacent (km^2): the area of the adjacent island

The sample mean of the response variable is 26.1, and the sample variance is 746.9897. Hence, there exists an over-dispersed problem in the data.

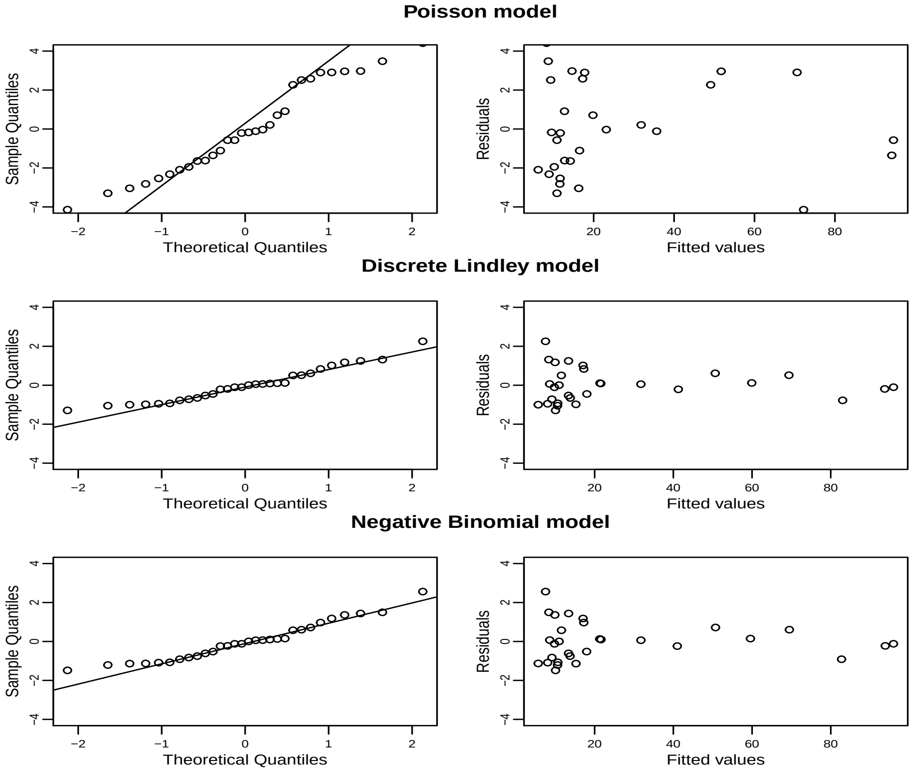
Table (4) shows the results of the fitting models based on Poisson, discrete Lindley, and negative binomial distributions. The results show that both discrete Lindley and negative binomial models provide better fits to the data than the Poisson regression model, according to the AIC and QAIC. There is a non-significant difference between the discrete Lindley and negative binomial models.

The residual plots are presented in Figure (1). The left plots are the normal Q-Q plot. The normal probability plots of discrete Lindley and negative binomial models are straighter than the Poisson model. The right panel plots the Pearson residuals against the fitted values $\widehat{\mu}_i$. The residual ranges in the discrete Lindley and negative binomial are narrower than the Poisson. Besides, the residual fluctuation in the Poisson model shows a more evident trend than the others. Hence, the Poisson model is not fit the data well. No difference between the discrete Lindley and negative binomial model indicates the reasonableness of the discrete Lindley distribution.

Table 4: Results from the Poisson, discrete Lindley and negative binomial regression models

Predictors	Poisson	Discrete Lindley	Negative Binomial
Intercept	2.12200***	1.99825***	2.00139***
	(0.00810)	(0.22827)	(0.19845)
Area	-0.00049***	-0.00052*	-0.00052**
	(0.00005)	(0.00027)	(0.00023)
Elevation	0.00299***	0.00314***	0.00314***
	(0.00015)	(0.00065)	(0.00055)
Nearest	-0.00212	0.00302	0.00285
	(0.00260)	(0.01008)	(0.00867)
Adjacent	-0.00058***	-0.00057***	-0.00057***
	(0.00005)	(0.00021)	(0.00018)
AIC	319.1	237.89	238.88
QAIC	373.056	236.3093	237.9763
$\widehat{\phi}$	6.39572	0.84229	1.10953

Note: *p<0.1; **p<0.05; ***p<0.01. Standard deviations are in parentheses.


Poisson model

Discrete Lindley model

Negative Binomial model

Figure 1: Residual plots for models fitted to the *gala* data

4.2. Example 2: Effect of temperature on all-cause mortality in Nghe An, Vietnam

Many previous studies confirm the nonlinear and delayed effects of environmental exposures on heath or mortality (Paravantis, Santamouris, Cartalis, Efthymiou, and Kontoulis 2017;

Guo, Barnett, Pan, Yu, and Tong 2011). The distributed lag nonlinear model (*DLNM*) proposed by Gasparrini, Armstrong, and Kenward (2010) has become a powerful technique in the epidemiological field because it controls the bi-dimensional effect of the exposure and lags on the outcome. The main idea of *DLNM* is to generate cross-basis variables and include them in a design matrix of a generalized linear model to estimate the parameters.

In this example, we apply the *DLNM* to investigate the effect of temperature on all-cause mortality with the assumption that the response variable follows the discrete Lindley distribution. The procedures for generating the cross-basis variables and applying the *GDLLM* are discussed clearly.

Data description

The empirical analysis is based on a time-series data set conducted at Dien Chau, a coastal plain district in Nghe An Province, Vietnam. Geographically, Dien Chau is located in the central part of Vietnam. The climate of Dien Chau is influenced by strong monsoon with hot, dry summer and cold, cloudy, drizzly winter.

For the study, the daily number of death from all causes in 2014 at Dien Chau is obtained from the A6 death register, which records every death event occurring across the country by medical workers at commune health stations. The meteorologic indicators are daily observations measured at Nghe An Observatory and provided by the Vietnam Meteorological and Hydrological Administration, including temperature (^0C) and total precipitation (mm).

Table 5: Distribution of daily mortality cases, average temperature, and total precipitation

Variables	Mean ± sd	Min	Q1	Median	Q3	Max
All-cause mortality	4.39 ± 2.98	0.00	2.00	4.00	6.00	15.00
Temperature (^oC)	24.97 ± 5.58	11.80	20.20	25.90	29.70	34.30
Total precipitation (mm)	4.01 ± 17.10	0.00	0.00	0.00	0.76	195.07

The descriptive statistics of these variables are presented in Table 5. Besides, the histogram for all mortality cases is shown in Figure 2, with the observed and expected frequencies of the number of death fitted by Poisson and discrete Lindley distributions. The expected frequencies from discrete Lindley seem closer to the observed frequencies. The variance of mortality cases is about two times the mean, so it is available to apply the *GLM* with discrete Lindley distribution.

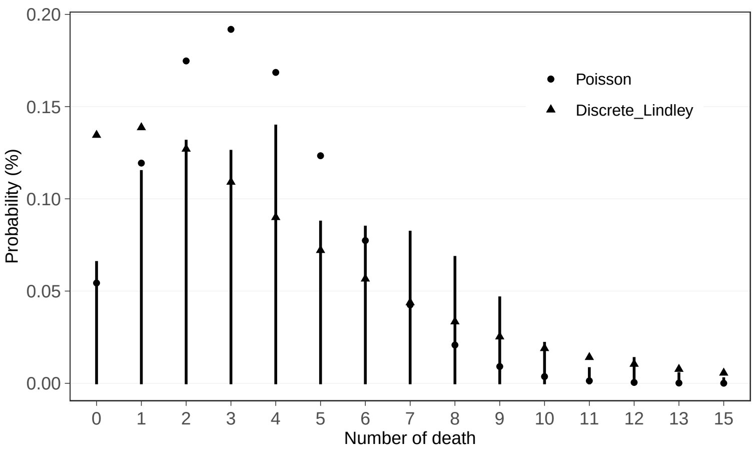


Figure 2: Histogram for all-cause mortality

The model

In this example, we focus on the relationship between daily temperature and the number of all-cause death at Dien Chau, Nghe An, Vietnam. The suggested model is

$$\log(\mu_t) = \alpha_0 + \sum_{\ell=0}^{L} f \cdot g(Temp_{t-\ell}, \ell) + s(Pre) + s(trend) + \sum_{j=1}^{6} \theta_j DOW_{jt}, \tag{25}$$

where Y is the daily all-cause mortality following the discrete Lindley distribution, $\mu_t = \mathbb{E}(Y_t)$, $Temp$ is the daily temperature, $f \cdot g$ is the bi-dimensional function exploring relationship along with the temperature over their lags with the outcome. Also, $s(trend)$ and $s(Pre)$ denote for smooth functions of the long-time trend and total precipitation (Pre). DOW is a system including six dummy variables for days of the week, with Monday being the base category to which no dummy is assigned.

$f \cdot g$ combines the exposure-outcome and lag-outcome functions, in which f and g can be parameterized as a linear combination of basis functions. Denote $\boldsymbol{R}$ to be a $n \times (L+1)v_x$ matrix containing the evaluations of lagged occurrence of each basis function at all observations of variable *Temp*, where v_x is the number of basis functions and n is the sample size. In the same vein, $\boldsymbol{C}$ is a $(L+1) \times v_\ell$ matrix ($v_\ell < L$) representing the basis functions applied to the lag vector.

Gasparrini *et al.* (2010) construct a cross-basis matrix $\boldsymbol{W}$ encompassing values of the bi-dimensional function by re-arranging and summing along with the lag of the following matrix

$$\boldsymbol{A} = \left(\boldsymbol{R} \otimes \boldsymbol{1}_{v_\ell}^T\right) \odot \left(\boldsymbol{1}_{v_x}^T \otimes \left(\mathrm{vec}(\boldsymbol{C}^T)\right)^T \otimes \boldsymbol{1}_n\right), \tag{26}$$

where $\boldsymbol{1}$ refers to an all-one vector with length denoted by a subscript, $\otimes$ and $\odot$ denote the Kronecker product and Hadamard product, respectively.

The representation of $f \cdot g$ is

$$f \cdot g = \boldsymbol{W}\boldsymbol{\gamma}\,, \tag{27}$$

with $\boldsymbol{\gamma}$ being a vector of new parameters corresponding to the new design matrix $\boldsymbol{W}$.

The type of smooth functions for the temperature and lag are chosen independently. The natural cubic spline is chosen to describe the relationship in each dimension because of its flexibility at two boundary points where some degree of non-linearity is expected (Goldberg, Gasparrini, Armstrong, and Valois 2011). The knots are evenly spread over the temperature range and placed at equal intervals over the logarithm of lags. These choices are motivated by substantive papers on epidemiological literature (Gasparrini *et al.* 2010; Guo *et al.* 2011). Besides, the natural cubic spline with knots placed at all points (degree of freedom equals 1) is chosen to control the long-time trend and total precipitation. The model (25) now becomes

$$\log(\mu_t) = \boldsymbol{W}\boldsymbol{\gamma} + \alpha B_{trend} + \beta B_{Pre} + \sum_{j=1}^{6} \theta_j DOW_{jt}\,, \tag{28}$$

where B_{trend} and B_{Pre} refer to the basis relating to the long-time trend and the total precipitation.

With the given basis functions of the exposures and the lags, we can generate a cross-basis function by using the command *crossbasis* of the package **dlnm** in R-software (Gasparrini, Armstrong, and Scheipl 2011). The model (28) with the new system of predictors is treated as the *GDLLM*. Then, the estimates and the inferences of the parameters can be carried out through the iterative weighted least square discussed in section 2.2. Detailed pieces of code to generate cross-basis variables and estimate the *GDLLM* in R are available in the Appendix.

For the mortality analysis, we are often interested in the relative risk of an exposure x with the given reference value x_0

$$RR_i = \frac{\mathbb{E}(Y|X = x_i)}{\mathbb{E}(Y|X = x_0)} = e^{\beta(x_i - x_0)}\,.$$

Results and discussion

The $QAIC$ is used to select the best degree of freedom (df) for the smooth functions of the predictor *Temp* and the lag. The best degree of freedom for the exposure dimension is two, while four for the lag dimension. The maximum lag is set to 20 days. The relative risks and their 90% confidence intervals are presented.

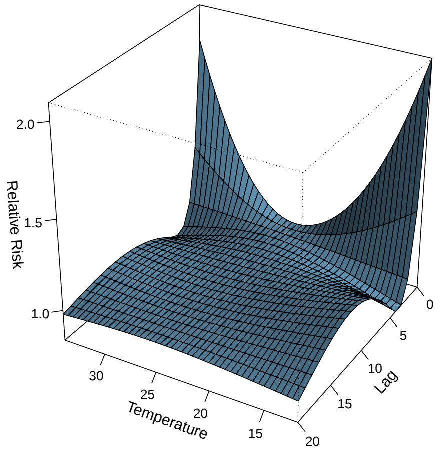


Figure 3: Relative risk along with temperature and lags

We choose 25^0C (the median temperature) as the reference to the data analysis. Figure (3) presents the three-dimensional plot of the relative risk along with the temperature and lags. It is found that the effects of heat and cold happen strongly and immediately but last for only two days. Furthermore, the shape of the temperature and mortality relationship changes along with lags.

Figure (4) shows the effects of the extreme cold and extreme hot temperature, corresponding to 13^0C (the 1st percentile) and 33^0C (the 99th percentile), on the all-cause mortality risk throughout 20-day lag. The strong effects appear in the first two days and rapidly decline afterward. There is no significant effect of the heat after lag 1. However, extremely cold temperatures also suggest delayed effects on all-cause mortality with significant relative risks 11 to 15 days after exposure.

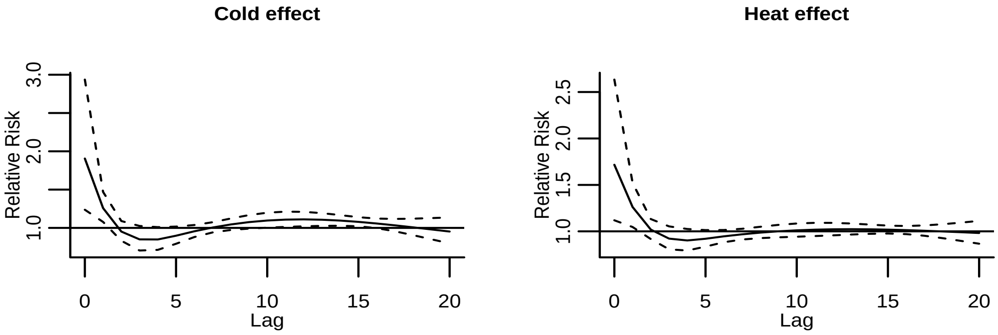


Figure 4: Relative risk by the temperature at different lags

Figure (5) reveals the relative risk by the temperature at specific lags. We start from lag 0 upwards and show the plots when there exists any change in the pattern of effects. Generally, the temperature effects on mortality manifest a nonlinear and U-shaped curve. At lag 0 and 1 days, the minimum mortality temperature is 24^0C, and the temperature below 18^0C or above 29^0C causes a significant increase in mortality risk. Though the causes of death are not examined in this study, we suppose that sudden death can be attributed to heart attacks, cardiac arrests, and strokes. Besides, the mortality risks relating to cold temperatures (less than 18^0C) are assessed over the 11 to 15 lag period. This result tallies with Carder, McNamee, Beverland, Elton, Cohen, Boyd, and Agius (2005), and Wang, Shi, Zanobetti, and Schwartz (2016) that mortality risk can persist within 2 or 3 weeks after exposure to the cold temperature.

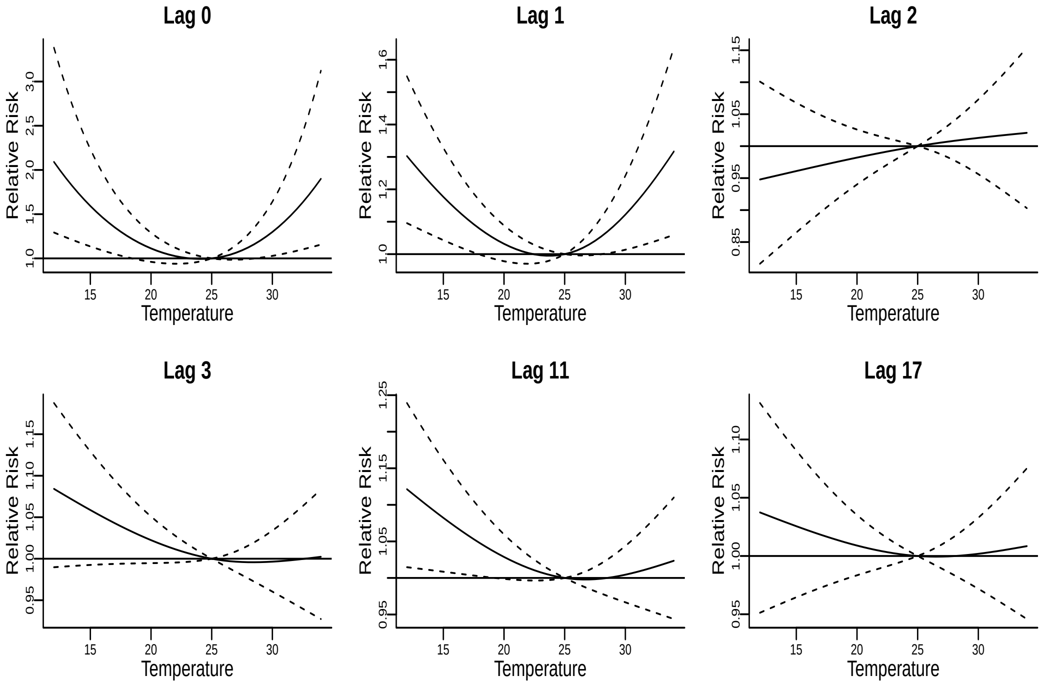


Figure 5: Relative risk by temperature at different lags

5. Conclusion

It becomes essential to construct specialized models that adapt well to the over-dispersed dependent count data. This study focuses on the generalized linear model, assuming that the response variable follows the discrete Lindley distribution, and shows how this model can fit the response variable with the over-dispersed problem. The iterative weighted least square is employed to estimate the model's parameters.

The performance of the discrete Lindley distribution model is evaluated through a simulation study at different levels of over-dispersion. The discrete Lindley model gets better behavior than the Poisson model under some criteria, including common bias, standard deviation, and mean squared error (MSE). Data applications are also carried out to demonstrate the modeling and assessing issues. The first example is cross-sectional data in ecological fields. The discrete Lindley model is better than the Poisson while performing equally well with the negative binomial model in AIC, QAIC, and residual plots. Hence, the discrete Lindley model can be recommended for the over-dispersed data.

Finally, the empirical analysis in the second example offers evidence for the usefulness of the proposed model in environmental epidemiology. Because of the nonlinear and delayed associations in the temperature and all-cause mortality relationship, the distributed lag non-linear model is carried out to generate cross-basis variables included in the regression model. The final model has a representation of the generalized linear model. Applying the estimating technique for the *GDLLM* model, we obtain the estimates of the parameters relating to these cross-basis variables, which describe the bi-dimensional effect of the predictor and the lag. This example provides a flexible method that can be used to investigate the complex association in exposure-health studies.

References

Abebe B, Shanker RA (2018). "Discrete Lindley Distribution with Applications in Biological Sciences." *Biometrics & Biostatistics International Journal*, **7**(1), 48–52. doi:10.15406/bbij.2018.07.00189.

Almeida SP, Casimiro E, Calheiros J (2010). "Effects of Apparent Temperature on Daily Mortality in Lisbon and Oporto, Portugal." *Environmental Health*, **9**(1), 1–7. doi:10.1186/1476-069X-9-12.

Cameron AC, Trivedi PK (1998). *Regression Analysis of Count Data*, volume 53. Cambridge university press.

Carder M, McNamee R, Beverland I, Elton R, Cohen GR, Boyd J, Agius RM (2005). "The Lagged Effect of Cold Temperature and Wind Chill on Cardiorespiratory Mortality in Scotland." *Occupational and Environmental Medicine*, **62**(10), 702–710. doi:10.1136/oem.2004.016394.

Crawley MJ (2012). *The R Book*. John Wiley & Sons.

Faddy MJ, Smith DM (2011). "Analysis of Count Data with Covariate Dependence in Both Mean and Variance." *Journal of Applied Statistics*, **38**(12), 2683–2694. doi:10.1080/02664763.2011.567250.

Faraway JJ (2004). *Linear Models with R*. Chapman and Hall/CRC.

Gasparrini A, Armstrong B, Kenward MG (2010). "Distributed Lag Non-linear Models." *Statistics in Medicine*, **29**(21), 2224–2234. doi:10.1002/sim.3940.

Gasparrini A, Armstrong B, Scheipl F (2011). "Distributed Lag Linear and Non-linear Models in R: The Package dlnm." *R package version 2.4.7*. URL https://cran.r-project.org/web/packages/dlnm/index.html.

Goldberg MS, Gasparrini A, Armstrong B, Valois MF (2011). "The Short-term Influence of Temperature on Daily Mortality in the Temperate Climate of Montreal, Canada." *Environmental Research*, **111**(6), 853–860. doi:10.1016/j.envres.2011.05.022.

Guo Y, Barnett AG, Pan X, Yu W, Tong S (2011). "The Impact of Temperature on Mortality in Tianjin, China: A Case-crossover Design with a Distributed Lag Nonlinear Model." *Environmental Health Perspectives*, **119**(12), 1719–1725. doi:10.1289/ehp.1103598.

Harrison XA (2014). "Using Observation-level Random Effects to Model Overdispersion in Count Data in Ecology and Evolution." *PeerJ*, **2**, e616. doi:10.7717/peerj.616.

Hastie T, Tibshirani R (1990). *Generalized Additive Model*. Edmundsbury Press.

Hilbe JM (2011). *Negative Binomial Regression*. Cambridge University Press.

Linares C, Diaz J (2010). "Short-term Effect of PM2. 5 on Daily Hospital Admissions in Madrid (2003–2005)." *International Journal of Environmental Health Research*, **20**(2), 129–140. doi:10.1080/09603120903456810.

Lindley DV (1958). "Fiducial Distributions and Bayes' Theorem." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 102–107. doi:10.1111/j.2517-6161.1958.tb00278.x.

McCullagh P, Nelder JA (2019). *Generalized Linear Models*. Routledge.

Nelder JA, Wedderburn RWM (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384. doi:10.2307/2344614.

Paravantis J, Santamouris M, Cartalis C, Efthymiou C, Kontoulis N (2017). "Mortality Associated with High Ambient Temperatures, Heatwaves, and the Urban Heat Island in Athens, Greece." *Sustainability*, **9**(4), 606. doi:10.3390/su9040606.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, Vienna, Austria. URL http://www.R-project.org/.

Rigby RA, Stasinopoulos DM, Akantziliotou C (2008). "A Framework for Modelling Overdispersed Count Data, Including the Poisson-shifted Generalized Inverse Gaussian Distribution." *Computational Statistics & Data Analysis*, **53**(2), 381–393. doi:10.1016/j.csda.2008.07.043.

Wang Y, Shi L, Zanobetti A, Schwartz JD (2016). "Estimating and Projecting the Effect of Cold Waves on Mortality in 209 US Cities." *Environment International*, **94**, 141–149. doi:10.1016/j.envint.2016.05.008.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM, *et al.* (2009). *Mixed Effects Models and Extensions in Ecology with R*, volume 574. Springer.

Appendix

R-code Used to Obtain the Results in Examples

R-code for estimating *GDLLM* (Algorithm 1)

```
glm_lin <- function(y,x){
        # Step 1: Setting initial values
        i = 1
        muy0 <- matrix(mean(y),nrow=NROW(y),1)
        x <- as.matrix(cbind(x))
        dev1 = 0
        # Step 2: Updating
        while (i < 100){
                eta0 <- log(muy0)
                J = diag(c(2*muy0/(2+muy0)),NROW(x),NROW(x))
                w <- eta0 + (y-muy0)/muy0
                beta <- solve(t(x)%*%J%*%x)%*%(t(x)%*%J%*%w)
                eta <- x%*%beta
                muy <- exp(eta)
                ll <-  sum(-(y+2)*log((2+muy)/muy) + 2*log(2/muy)
                     + log(1+y))
                old.dev1 <- dev1
                sig1 <- sum(2*(y-muy)^2/((2+muy)*muy))/(NROW(x)-NCOL(x))
                cov_beta  <-  solve(t(x)%*%J%*%x)
                cov_beta_ad  <-  sig1*solve(t(x)%*%J%*%x)
                dev1 <- 2*sum(ifelse(y==0,-2*log(2)+2*log(muy+2),
                        y*log(y/muy) - (y+2)*log((y+2)/(muy+2))))
                aic <- -2*ll+2*NROW(beta)
                i = i + 1
                if (abs(dev1-old.dev1) < 10^-5*dev1) break
                muy0 <- muy
        }
        # Extracting the results
        mylist <- list("coeff"= beta,"varcov"=cov_beta,
                        "adj_varcov" = cov_beta_ad, "deviance"=dev1,
                        "Loglikelihood"=ll, "AIC"=aic,
                        "fitted"=muy, "disperson"=sig1)
        return (mylist)
}
```

R-code used in Example 1:

```
# Obtaining the gala dataset
library(faraway)

# Fitting the GDLLM
X <- cbind(1,gala$Area, gala$Elevation, gala$Nearest, gala$Adjacent)
Y <- gala$Endemics
md1 <- glm_lin(Y,X)

# Fitting the Poisson model
md2 <- glm(Endemics~Area+Elevation+Nearest+Adjacent,
           family=poisson,data=gala)

# Fitting the Negative Binomial model
library(MASS)
md3 <- glm.nb(Endemics~Area+Elevation+Nearest+Adjacent,data=gala)
```

```
# Computing the fitted values and Pearson residuals
y_lin <- exp(X%*%md1[[1]])
res_lin <- (Y-y_lin)/sqrt((1+y_lin/2)*y_lin)
y_poi <-md2$fitted.values
res_poi <- residuals(md2,"pearson")
y_nb <-md3$fitted.values
res_pois <- residuals(md3,"pearson")

# Residual plots for model fitting
pdf("gala1.pdf",width=6,height=6)
par(mfrow=c(3,2))
qqnorm(res_poi,ylim=c(-4,4),main=NULL)
qqline(res_poi)
plot(y_poi,res_poi,xlab = "Fitted␣values",ylab="Residuals",ylim=c(-4,4))

qqnorm(res_lin,main=NULL,ylim=c(-4,4))
qqline(res_lin)
plot(y_lin,res_lin,xlab = "Fitted␣values",ylab="Residuals",ylim=c(-4,4))

qqnorm(res_nb,main=NULL,ylim=c(-4,4))
qqline(res_nb)
plot(y_nb,res_nb,xlab = "Fitted␣values",ylab="Residuals",ylim=c(-4,4))

dev.off()
```

R-code used in Example 2

```
# Loading the packages
library(dlnm);library(splines)

# Generating the cross-basis matrix for temperature along lags
lagnots <- logknots(20,2)
b <- crossbasis(dt$temp_c,lag=20, argvar=list(fun="ns",df=2),
arglag=list(fun="ns",knots=lagnots))

# Generating basis functions for variables "trend" and "precipitation"
 by natural cubic spline
x1 <- ns(dt$trend)
x2 <- ns(dt$pre)

# Generating the new dataset, removing omitted observation
dt1 <- na.omit(cbind(y=dt$mort,1,b,x1,x2,factor(dt$dayofweek)))

# Fitting the model
md1 <- glm_lin(dt1[,1],dt1[,-1])

# Plotting
bb <- md1$coef[2:9]
BB <- md1$CVar[2:9,2:9]
tem.pred <- crosspred(b,coef = bb, vcov=BB, model.link="log",
                      cen=25, ci.level=0.90, cumul=FALSE)
pdf("3d-tem-lin.pdf",width=6,height=6)
plot(tem.pred ,xlab="Temperature",zlab="Relative␣Risk",ylab="Lag")
dev.off()

pdf("lagrisk.pdf",width=5,height=2)
par(mfrow=c(1,2))
plot(tem.pred,"slices",xlab="Lag",ylab="Relative␣Risk",ci="lines",
var=13,cex.main = 0.6,main="Cold␣effect",ci.level=0.90)
plot(tem.pred ,"slices",xlab="Lag",ylab="Relative␣Risk",ci="lines",
var=33,cex.main = 0.6,main="Heat␣effect",ci.level=0.90)
dev.off()

pdf("temrisk.pdf",width=8,height=4)
par(mfrow=c(2,3))
plot(tem.pred ,"slices",xlab="Temperature",ylab="Relative␣Risk",
     ci="lines", lag=0, main = "Lag␣0",ci.level=0.90)
plot(tem.pred ,"slices",xlab="Temperature",ylab="Relative␣Risk",
     ci="lines", lag=1, main = "Lag␣1",ci.level=0.9)
plot(tem.pred ,"slices",xlab="Temperature",ylab="Relative␣Risk",
     ci="lines", lag=2, main = "Lag␣2",ci.level=0.9)
plot(tem.pred ,"slices",xlab="Temperature",ylab="Relative␣Risk",
     ci="lines", lag=9, main = "Lag␣3",ci.level=0.9)
plot(tem.pred ,"slices",xlab="Temperature",ylab="Relative␣Risk",
     ci="lines", lag=11,main = "Lag␣11",ci.level=0.9)
plot(tem.pred ,"slices",xlab="Temperature",ylab="Relative␣Risk",
     ci="lines", lag=17, main = "Lag␣17",ci.level=0.9)
dev.off()
```

Affiliation:

Mien T.N. Nguyen
Department of Mathematics, Faculty of Science, Mahidol University
Bangkok, Thailand
E-mail: mienthingoc.ngu@student.mahidol.edu

Man V.M. Nguyen
Department of Mathematics, Faculty of Science, Mahidol University
Centre of Excellence in Mathematics, CHE
Bangkok, Thailand
E-mail: man.ngu@mahidol.edu

Ngoan T. Le
Institute of Research and Development, Duy Tan University,
Da Nang City, Vietnam
School of Medicine, International University of Health & Welfare,
Chiba, Japan
E-mail: letngoan@hmu.edu.vn

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 52 *Submitted:* 2022-01-31
July 2023 *Accepted:* 2022-07-26