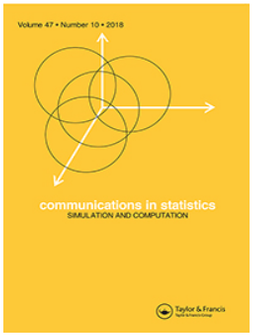

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Poisson-Modification of Quasi Lindley regression model for over-dispersed count responses

Ramajeyam Tharshan & Pushpakanthie Wijekoon

To cite this article: Ramajeyam Tharshan & Pushpakanthie Wijekoon (2022): Poisson-Modification of Quasi Lindley regression model for over-dispersed count responses, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2022.2044052

To link to this article: https://doi.org/10.1080/03610918.2022.2044052

Published online: 28 Feb 2022.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2022.2044052

Poisson-Modification of Quasi Lindley regression model for over-dispersed count responses

Ramajeyam Tharshan[a,b] and Pushpakanthie Wijekoon[c]

[a]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka; [b]Department of Mathematics and Statistics, University of Jaffna, Jaffna, Sri Lanka; [c]Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka

ABSTRACT

This paper introduces an alternative linear regression model for over-dispersed count responses with appropriate covariates. It is an extended work of univariate Poisson-Modification of the Quasi Lindley (PMQL) distribution via the generalized linear model approach. A re-parametrized PMQL distribution is considered to demonstrate the flexible properties of the distribution on its regression model. Further, the performance of its maximum likelihood estimation method is examined by a simulation study based on the asymptotic theory. The maximum likelihood estimator is used to estimate the parameters of the regression model. Finally, three simulated data sets and a real-world data set are taken to show the applicability of the PMQL regression model against the Poisson, Negative binomial (NB), Poisson-Quasi Lindley (PQL), and Generalized Poisson-Lindley (GPL) regression models. The results of applications show that the newly introduced model provides a better fit for over-dispersed count responses with covariates than the Poisson, NB, PQL, GPL regression models.

ARTICLE HISTORY
Received 16 March 2021
Accepted 12 February 2022

KEYWORDS
Generalized linear model; Mixed Poisson regression models; Over-dispersed count responses; Poisson distribution; Quasi Lindley distribution

1. Introduction

The Poisson regression model is the standard model for the count responses with appropriate covariates (Cameron and Trivedi 2013). It has a well-known property that its conditional mean counts equals to the conditional variance for a given set of covariates, and this property is commonly known as equidispersion (Johnson, Kotz, and Kemp 1992; Cameron and Trivedi 2013). In some real-world applications especially, actuarial, ecology, and genetics, the assumption of equidispersion is violated in such a way that the conditional variance exceeds the theoretical conditional variance. This phenomenon is explained as over-dispersion or variation inflation (Greenwood and Yule 1920).

To handle the apparent of over-dispersion in a count variable, several univariate mixed Poisson distributions have been widely introduced with their explicit forms of probability mass functions (pmf). Their pmfs are computationally flexible and they have successfully applied for various real-world applications. Some notables are, Poisson-gamma/Negative binomial (NB) distribution (Hilbe 2011), Poisson-Lindley distribution (Sankaran 1970), Generalized Poisson-Lindley (GPL) distribution (Mahmoudi and Zakerzadeh 2010), Poisson-Two parameter Lindley distribution (Bhati, Sastry, and Qadri 2015), Poisson-Quasi Lindley (PQL) distribution (Grine and Zeghdoudi 2017) and among others. They may have various flexible statistical properties to

CONTACT Ramajeyam Tharshan tharshan10684@gmail.com Department of Mathematics and Statistics, University of Jaffna, Jaffna 40000, Sri Lanka.

accommodate the several horizontal symmetry, right tail heaviness, and variance-to-mean ratios based on their mixing distributions (Lynch 1988; Tharshan and Wijekoon 2020a, 2020b).

Tharshan and Wijekoon (2021a) introduced a continuous distribution named Modification of the Quasi Lindley (MQL) distribution bounded to $(0, \infty)$. By amalgamating the Poisson distribution with the MQL distribution, Tharshan and Wijekoon (2022) obtained an univariate mixed Poisson distribution named Poisson-Modification of the Quasi Lindley (PMQL) distribution for the over-dispersed count data. Authors have shown that its pmf is an explicit form and the performance of the PMQL distribution is better than some of the existing competent mixed Poisson distributions. It has the flexibility to capture the various ranges of right-tail heaviness measured by excess kurtosis (*EK*), horizontal symmetry measured by skewness (*SK*), and the heterogeneity measured by index of dispersion (τ) for an over-dispersed count data.

To predict the over-dispersed count responses for given covariates, several mixed Poisson regression models have been introduced, in literature, by using the generalized linear model (GLM) approach. Examples of such models are NB regression model (Hilbe 2011), Poisson-Inverse Gaussian regression model (Shoukri et al. 2004), Poisson-Weighted exponential regression model (Zamani, Ismail, and Faroughi 2014), GPL regression model (Wongrin and Bodhisuwan 2016), and PQL regression model (Altun 2019).

The main contribution of this paper is to introduce the Poisson-Modification of the Quasi Lindley regression model for over-dispersed count responses with covariates. Then, we compare its performance with some existing competent predecessors. Here, we re-parametrize the PMQL distribution in terms of its mean and two over-dispersion parameters. Some of the important structural properties of the re-parametrized PMQL distribution are derived to show its applicability for an over-dispersed count response on its regression model. Then, we setup the re-parametrized PMQL distribution to approach the GLM.

The remaining part of the paper is structured as follows: In Sec. 2, we introduce the re-parametrized PMQL distribution in terms of PMQL distribution's mean and two over-dispersion parameters with some of its important structural properties, an algorithm to simulate its random variables, and the maximum likelihood estimator (MLE) to estimate the unknown parameters of the re-parametrized PMQL distribution. Section 3 discusses a Monte Carlo simulation study to examine the asymptotic property of the MLE and the performance of the MLE at different values of mean. In Sec. 4, we introduce the PMQL regression model via the GLM approach with its MLE part. Finally, three simulated data sets and a real-world example are taken to show the applicability of the newly introduced regression model by comparing it with some existing predecessors.

2. Re-parameterization of the PMQL distribution

In this section, we introduce the re-parametrized PMQL distribution with some important structural properties, an algorithm to simulate its random variables, and its unknown parameter estimation.

2.1. Probability mass and cumulative distribution functions (pmf and cdf)

Suppose the random variable Y be the total counts of a specific experiment with mean λ. Then, the traditional distribution to find the probabilities of such outcomes is the Poisson distribution, and its pmf is defined as:

$$f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}; y = 0, 1, 2, \ldots; \lambda > 0. \tag{1}$$

Here, $E(Y) = Var(Y) = \lambda$, and then the variance-to-mean ratio called index of dispersion, $\tau = 1$. Then, it is not a suitable distribution to accommodate the over-dispersion problem.

Tharshan and Wijekoon (2021a) proposed a new Lindley family of distributions, namely the Modification of Quasi Lindley distribution, MQL (θ, α, δ). Its probability density function (pdf) is defined as:

$$f_\Lambda(\lambda;\theta,\alpha,\delta)=\frac{\theta(\Gamma(\delta)\alpha^3+(\theta\lambda)^{\delta-1})e^{-\theta\lambda}}{(\alpha^3+1)\Gamma(\delta)};\ \ \lambda>0;\ \ \theta>0,\alpha^3>-1,\delta>0, \tag{2}$$

where α and δ are shape parameters, θ is the scale parameter, λ is the respective random variable, and $\Gamma(a)$ is the gamma function and defined as: $\Gamma(a)=\int_0^\infty s^{a-1}e^{-s}ds$.

Tharshan and Wijekoon (2022) have obtained a mixed Poisson distribution by amalgamating the Poisson distribution with the MQL distribution, namely Poisson-MQL distribution, PMQL (θ, α, δ). Its pmf is defined as:

$$f_Y(y;\theta,\alpha,\delta)=\frac{\theta(\Gamma(\delta)\Gamma(y+1)\alpha^3(1+\theta)^{\delta-1}+\theta^{\delta-1}\Gamma(y+\delta))}{y!(\alpha^3+1)(1+\theta)^{y+\delta}\Gamma(\delta)};y=0,1,2,\ldots;\theta>0,\alpha^3>-1,\delta>0, \tag{3}$$

and the corresponding cdf of equation (3) is given by:

$$\begin{aligned}F_Y(y)&=\sum_{t=0}^{y}f(t)\\&=\frac{\delta(1+\theta)^{\delta-1}\Gamma(\delta)\alpha^3\Gamma(y+1)((1+\theta)^{y+1}-1)+\theta^\delta\Gamma(y+\delta+1)\,{}_2F_1\left(1,y+\delta+1;\delta+1;\frac{\theta}{1+\theta}\right)}{(\alpha^3+1)\Gamma(\delta)y!\delta(1+\theta)^{y+\delta+1}}\end{aligned} \tag{4}$$

where ${}_2F_1(c,d;r;w)$ is the Gaussian hypergeometric function defined as: ${}_2F_1(c,d;r;w)=\sum_{i=0}^{\infty}\frac{(c)_i(d)_iw^i}{(r)_ii!}$, which is a special case of the generalized hypergeometric function given by the expression: ${}_aF_b(p_1,p_2,\ldots,p_a;q_1,q_2,\ldots,q_b;w)=\sum_{i=0}^{\infty}\frac{(p_1)_i\ldots(p_a)_iw^i}{(q_1)_i\ldots(q_b)_ii!}$, and $(p)_i=\frac{\Gamma(p+i)}{\Gamma(p)}=p(p+1)\ldots(p+i+1)$ is the Pochhammer symbol (Slater 1966).

From its statistical properties, the mean and variance of the PMQL (θ, α, δ) are given in equations 5 and 6, respectively:

$$E(Y)=\mu=\frac{\alpha^3+\delta}{(\alpha^3+1)\theta} \tag{5}$$

$$Var(Y)=\frac{\alpha^6(1+\theta)+\alpha^3(2+\theta+\delta(\theta+\delta-1))+\delta(\theta+1)}{(\alpha^3+1)^2\theta^2} \tag{6}$$

To examine the PMQL distribution's flexible properties in terms of its mean (μ) and two over-dispersion parameters (α, δ) for the over-dispersed count responses on its regression model, the PMQL (θ, α, δ) is re-parametrized.

Proposition 1: By substituting $\theta=\frac{(\alpha^3+\delta)}{(\alpha^3+1)\mu}$ in equation (3), the pmf of the re-parametrized PMQL (μ, α, δ) can be written as:

$$f_Y(y;\mu,\alpha,\delta)=\frac{((\alpha^3+1)\mu)^y(\alpha^3+\delta)(\Gamma(\delta)\Gamma(y+1)\alpha^3A^{\delta-1}+(\alpha^3+\delta)^{\delta-1}\Gamma(y+\delta))}{y!(\alpha^3+1)A^{y+\delta}\Gamma(\delta)}; \tag{7}$$

$y=0,1,2,\ldots;\mu>0,\alpha^3>-1,\delta>0,$ where $A=(\alpha^3+1)\mu+\alpha^3+\delta$.

Proposition 2: By substituting $\theta=\frac{(\alpha^3+\delta)}{(\alpha^3+1)\mu}$ in equation (4), the cdf of the re-parametrized PMQL (μ, α, δ) can be written as:

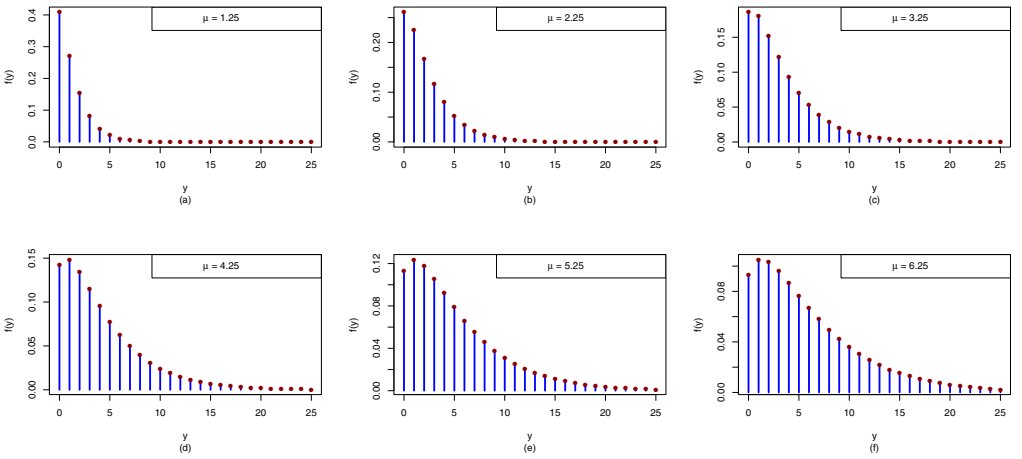

Figure 1. The probability mass function of the re-parametrized PMQL distribution at $\alpha = 0.50, \delta = 1.50,$ and different values of μ.

$$F_Y(y;\mu,\alpha,\delta) = \frac{\gamma_1(y) + \gamma_2(y)_2F_1\left(1, y+\delta+1; \delta+1; \frac{\alpha^3+\delta}{A}\right)}{\gamma_3(y)}, \tag{8}$$

where A is defined as in proposition 1, $\gamma_1(y) = (\alpha^3+1)\mu\delta A^{\delta-1}\Gamma(\delta)\alpha^3\Gamma(y+1)(A^{y+1} - ((\alpha^3+1)\mu)^{y+1}), \gamma_2(y) = ((\alpha^3+1)\mu)^{y+1}(\alpha^3+\delta)^{\delta}\Gamma(y+\delta+1)$, and $\gamma_3(y) = (\alpha^3+1)\Gamma(\delta)y!\delta A^{y+\delta+1}$.

The variance of the re-parametrized PMQL (μ,α,δ) can be obtained easily by substituting $\theta = \frac{(\alpha^3+\delta)}{(\alpha^3+1)\mu}$ in equation (6) as:

$$Var(Y) = \mu + \mu^2\left(\frac{\alpha^3(\alpha^3+2+\delta(\delta-1))+\delta}{(\alpha^3+\delta)^2}\right),$$

and then its $Var(Y)/E(Y) = \tau$ is obtained as:

$\tau = 1 + \mu(\frac{\alpha^3(\alpha^3+2+\delta(\delta-1))+\delta}{(\alpha^3+\delta)^2}) =$ term I$+\mu$ term II, say, respectively.

We can clearly see that $\tau > 1$ for the positive value of term II. Then, the re-parametrized PMQL (μ,α,δ) is an over-dispersed distribution.

Shape properties of the re-paramerized PMQL (μ,α,δ) can be easily examined by re-parametrizing the shape properties of PMQL (θ,α,δ) as discussed in Tharshan and Wijekoon (2022). Here, we display the right-tail behavior of the re-parametrized PMQL (μ,α,δ) in Figure 1 for different parameter values of μ while the parameters α and δ are fixed. We may note that, for fixed parameter values α and δ, the distribution captures more right-tail when the parameter μ increases.

Figure 2 depicts the surface plots of the index of dispersion, τ at different parameter values of $\mu,\alpha,$ and δ. It is clear that when the parameter μ increases, τ also increases at various rates based on the values of parameters α and δ. Further, they indicate that the re-parametrized PMQL (μ,α,δ) has the capability to accommodate various ranges of the index of dispersion for an over-dispersed count data.

2.2. Simulation of random variables

Here, we provide an algorithm by using the inverse transform method to simulate the random variables $y_1, y_2, ..., y_n$ from the re-parametrized PMQL (μ,α,δ) of size n.

Algorithm

i. Simulate random variables, $U_i \sim$ Uniform$(0,1); i = 1,2,...,n$.

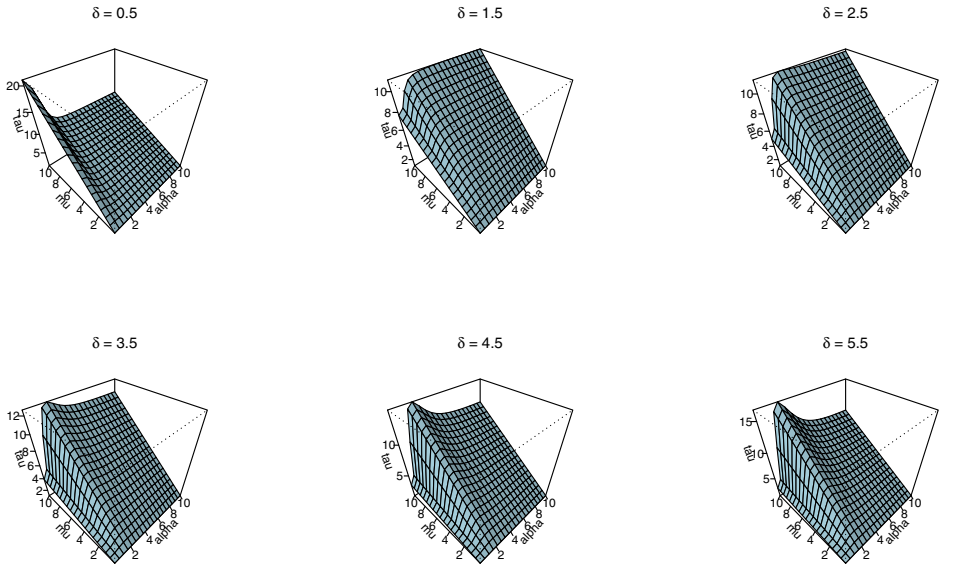

Figure 2. Surface plots for the index of dispersion, τ at different values of $\mu, \alpha,$ and δ.

ii. Solve the non-linear equation for $[y_{u_i}]$;
$\gamma_1(y_{u_i}) + \gamma_2(y_{u_i})_2F_1(1, y_{u_i} + \delta + 1; \delta + 1; \frac{\alpha^3+1}{A}) - u_i\gamma_3\ (y_{u_i}) = 0,$
where A is defined as in proposition 1, $\gamma_1(y_{u_i}), \gamma_2(y_{u_i})$ and $\gamma_3(y_{u_i})$ are defined as in proposition 2. [.] denotes the integer part. Since it is a non-linear in y_{u_i}, the numerical method called the Newton-Raphson method can be employed to find the roots.

2.3. Parameter estimation of the re-parametrized PMQL distribution

This subsection introduces the method of the unknown parameter estimation of the re-parametrized PMQL (μ, α, δ) based on the maximum likelihood estimation method. Let $y_1, y_2, ..., y_n$ be a random sample of size n from the re-parametrized PMQL (μ, α, δ). Then, the log-likelihood function of its pmf is given as:

$$\ell(\mu, \alpha, \delta|y) = n \ln\left(\frac{\alpha^3 + \delta}{(\alpha^3+1)\delta}\right) + \sum_{i=1}^{n} \ln\left(\Gamma(\delta)\Gamma(y_i+1)\alpha^3 A^{\delta-1} + \alpha^3 + \delta)^{\delta-1}\Gamma(y_i - 1)\right) - \sum_{i=1}^{n} \ln(y_i!) - \sum_{i=1}^{n}(y_i + \delta)\ln(A) + \sum_{i=1}^{n} y_i \ln((\alpha^3+1)\mu). \tag{9}$$

The score functions are:

$$\frac{\partial \ell(\mu, \alpha, \delta|y)}{\partial \mu} = \sum_{i=1}^{n}\frac{y_i}{\mu} - \sum_{i=1}^{n}\frac{(y_i+\delta)(\alpha^3+1)}{A} + \sum_{i=1}^{n}\frac{\Gamma(\delta)\alpha^3\Gamma(y_i+1)(\delta-1)A^{\delta-2}(\alpha^3+1)}{\Gamma(\delta)\Gamma(y_i+1)\alpha^3A^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)},$$

$$\frac{\partial \ell(\mu, \alpha, \delta|y)}{\partial \alpha} = 3n\alpha^2\left(\frac{1-\delta}{(\alpha^3+\delta)(\alpha^3+1)}\right) + \sum_{i=1}^{n}\frac{3y_i\alpha^2}{\alpha^3+1} - \sum_{i=1}^{n}\frac{3\alpha^2(y_i+\delta)(\mu+1)}{A} + \sum_{i=1}^{n}\frac{\Gamma(\delta)\Gamma(y_i+1)(3\alpha^2A^{\delta-2}(A+\alpha^3(\delta-1)(\mu+1)) + 3\alpha^2\Gamma(y_i+\delta)(\delta-1)(\alpha^3+\delta)^{\delta-2}}{\Gamma(\delta)\Gamma(y_i+1)\alpha^3A^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)},$$

$$\frac{\partial \ell(\mu, \alpha, \delta|y)}{\partial \delta} = n\left(\frac{1}{\alpha^3+\delta} - \psi(\delta) - \ln(A)\right) - \sum_{i=1}^{n}\frac{y_i+\delta}{A} + \sum_{i=1}^{n}\frac{\alpha^3\Gamma(y_i+1)B + \Gamma(y_i+\delta)C + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)\psi(y_i+\delta)}{\Gamma(\delta)\Gamma(y_i+1)\alpha^3A^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)},$$

where A is defined as in proposition 1, and

Table 1. Performance of the MLE for the re-parametrized PMQL ($\mu, \alpha = 0.25, \delta = 0.20$).

	$n = 60$		$n = 100$		$n = 200$		$n = 300$	
	MLE	Bias(MSE)	MLE	Bias(MSE)	MLE	Bias(MSE)	MLE	Bias(MSE)
$\mu = 1.25$								
μ	0.9330	−0.3170(0.5778)	1.1145	−0.1355(0.46l5)	1.1462	−0.1038(0.2903)	1.1784	−0.0716(0.1658)
α	0.6388	0.3888(0.2170)	0.5917	0.3417(0.1582)	0.5644	0.3144(0.1183)	0.5331	0.2831(0.1054)
δ	0.1120	−0.0880(0.0105)	0.1214	−0.0786(0.0081)	0.1315	−0.0685(0.0074)	0.1408	−0.0592(0.0055)
$\mu = 2.25$								
μ	1.7428	−0.5072(0.6601)	2.0712	−0.1788(0.5638)	2.1091	−0.1409(0.4402)	2.1434	−0.1066(0.2056)
α	0.6891	0.4391(0.2358)	0.6357	0.3857(0.1713)	0.5972	0.3472(0.1396)	0.5706	0.3206(0.1285)
δ	0.0946	−0.1054(0.0145)	0.1008	−0.0992(0.0121)	0.1211	−0.0789(0.0100)	0.1251	−0.0749(0.0062)
$\mu = 3.25$								
μ	2.6270	−0.6230(1.0487)	2.8963	−0.3537(0.6180)	2.9958	−0.2542(0.5134)	3.0123	−0.2377(0.2940)
α	0.7221	0.4721(0.2428)	0.6442	0.3942(0.1829)	0.6244	0.3744(0.1532)	0.5902	0.3402(0.1430)
δ	0.0855	−0.1145(0.0159)	0.0951	−0.1049(0.0133)	0.1023	−0.0977(0.0110)	0.1118	−0.0882(0.0093)
$\mu = 4.25$								
μ	3.5202	−0.7298(1.5230)	3.7907	−0.4593(1.3024)	3.8423	−0.4077(0.6714)	3.9140	−0.3360(0.4263)
α	0.7320	0.4820(0.2674)	0.6608	0.4108(0.2180)	0.6465	0.3965(0.1684)	0.6280	0.3780(0.1514)
δ	0.0761	−0.1239(0.0172)	0.0847	−0.1153(0.0156)	0.0918	−0.1082(0.0127)	0.1005	−0.0995(0.0110)
$\mu = 5.25$								
μ	4.3720	−0.8780(2.1880)	4.5752	−0.6748(1.5614)	4.6259	−0.6241(1.1164)	4.8347	−0.4153(0.6177)
α	0.7674	0.5174(0.2818)	0.7207	0.4707(0.2396)	0.6720	0.4220(0.1911)	0.6351	0.3851(0.1611)
δ	0.0644	−0.1356(0.0210)	0.0779	−0.1220(0.0170)	0.0869	−0.1131(0.0136)	0.0920	−0.1080(0.0127)
$\mu = 6.25$								
μ	5.3238	−0.9262(3.2433)	5.5291	−0.7209(2.4180)	5.5652	−0.6848(1.1522)	5.7286	−0.5214(0.9079)
α	0.7993	0.5493(0.3181)	0.7425	0.4925(0.2597)	0.6937	0.4437(0.2140)	0.6631	0.4131(0.1871)
δ	0.0525	−0.1475(0.0232)	0.0679	−0.1321(0.0199)	0.0768	−0.1232(0.0165)	0.0881	−0.1119(0.0140)

$B = \Gamma(\delta)A^{\delta-2}(\delta - 1 + A\ \ln(A) + A\ \psi(\delta))$, and $C = (\alpha^3 + \delta)^{\delta-2}(\delta - 1 + (\alpha^3 + \delta)\ln(\alpha^3 + \delta))$. Furher, $\psi(a)$ is the digamma function and defined as: $\psi(a) = \frac{\partial}{\partial a}\ln(\Gamma(a)) = \frac{\Gamma'(a)}{\Gamma(a)}$.

Now, we can find the MLEs of $\mu, \alpha,$ and δ, by setting the score functions equal to zero and solving these systems of non-linear equations simultaneously. Here, we can use the *optim* function in the R package *stats* (R Core Team 2020) in order to solve these non-linear equations by applying the Newton-Raphson method. For the interval estimations of the corresponding parameters based on the asymptotic theory, we provide the observed information matrix:

Proposition 3: The observed information matrix of the maximum likelihood estimators of $\mu, \alpha,$ and δ is given by:

$$I(\mu,\alpha,\delta) = \begin{pmatrix} -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\mu^2} & -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\mu\partial\alpha} & -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\mu\partial\delta} \\ -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\alpha\partial\mu} & -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\alpha^2} & -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\alpha\partial\delta} \\ -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\delta\partial\mu} & -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\delta\partial\alpha} & -\frac{\partial^2\ell(\mu,\alpha,\delta|x)}{\partial\delta^2} \end{pmatrix}$$

at $\mu = \hat{\mu}$, $\alpha = \hat{\alpha}$, and $\delta = \hat{\delta}$. The elements of the observed information matrix are given in Appendix 1. The asymptotic confidence intervals for the parameters $\mu, \alpha,$ and δ can be found by the asymptotic theory as $n \to \infty$.

Note:

According to the asymptotic theory, the distribution of $\sqrt{n}(\hat{\mu} - \mu, \hat{\alpha} - \alpha, \hat{\delta} - \delta)$ as $n \to \infty$ is a three-variate normal with zero means and variance covariance matrix $I^{-1}(\hat{\mu}, \hat{\alpha}, \hat{\delta})$.

Therefore, the asymptotic $(1 - a)100\%$ confidence interval for the parameters $\mu, \alpha,$ and δ are given, respectively:

$$\hat{\mu} \pm z_{a/2}\sqrt{Var(\hat{\mu})},\ \hat{\alpha} \pm z_{a/2}\sqrt{Var(\hat{\alpha})},\ \hat{\delta} \pm z_{a/2}\sqrt{Var(\hat{\delta})},$$

where the $Var(\hat{\mu}), Var(\hat{\alpha})$, and $Var(\hat{\delta})$ are the variance of $\hat{\mu}, \hat{\alpha}$, and $\hat{\delta}$, respectively. They are diagonal elements of $I^{-1}(\hat{\mu}, \hat{\alpha}, \hat{\delta})$ and $z_{a/2}$ is the $a/2$ quantile of the standard normal distribution.

3. Simulation study

Here, we do a simulation study to evaluate the performance of the MLEs, $\hat{\mu}, \hat{\alpha}$, and $\hat{\delta}$ for sample size n. The algorithm given in subsection 2.2, is used to simulate the random variables from the re-parametrized PMQL (μ, α, δ). The simulation study was repeated 1000 times. The following steps are considered:

i. Simulate 1000 samples of size n.
ii. Compute MLEs for the 1000 samples, say $(\hat{\mu}_i, \hat{\alpha}_i, \hat{\delta}_i)$, $i = 1, 2, ..., 1000$.
iii. Compute the average MLEs, biases, and mean square errors (MSEs) by using the following equations:
$\hat{s}(n) = \frac{1}{1000}\sum_{i=1}^{1000} \hat{s}_i$, $bias_s(n) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{s}_i - s)$, and $MSE_s(n) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{s}_i - s)^2$, for $s = \mu, \alpha, \delta$, and the sample size n.

We have repeated these steps for $n = 60, 100, 200$, and 300 with $\mu = 1.25, 2.25, 3.25, 4.25, 5.25$, and 6.25; $\alpha = 0.25$; and $\delta = 0.20$.

Table 1 summarizes the average MLEs, biases, and MSEs (in parentheses) of μ, α, and δ for different values of μ which are $1.25, 2.25, 3.25, 4.25, 5.25$, and 6.25 by fixing $\alpha = 0.25$ and $\delta = 0.20$. We may note that the biases and MSEs decrease as n increases for all parameters. It reveals that the maximum likelihood estimation method verifies the asymptotic property for all parameter estimates. Further, for all given μ values; μ are under-estimated and estimation of μ is good for small value of μ and considerably higher sample size based on average biases and MSEs.

4. PMQL regression model

This section discusses the regression model of the PMQL distribution and its parameter estimation.

4.1. Probability mass function

The Poisson regression model is the traditional model to model the count responses with appropriate covariates. When the response variable is over-dispersed the NB regression model is commonly used to model the over-dispersed count responses.

Here, we introduce the PMQL regression model as an alternative regression model for the over-dispersed count response variable, i.e., we present the generalized linear model (GLM) approach for the PMQL distribution say PMQL_{GLM}.

Let $y_1, y_2, ..., y_n$ be the random sample of n observations from the PMQL distribution. The log of the mean is taken as the link function of the PMQL_{GLM} to confirm positive estimated values of responses, i.e., the link between p-dimensional covariates and the mean responses, y is given as:

$$\eta_i = g(\mu_i) = \log(\mu_i) = \sum_{j=0}^{p} \beta_j x_{ij} = x_i'\beta, \quad i = 1, 2, ..., n, \tag{10}$$

where $x_i' = (1, x_{i1}, x_{i2}, ..., x_{ip})$ is the vector of covariates supplemented with a 1 in front for the intercept, and $\beta' = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of unknown regression coefficients. Then, the mean and variance of the regression model are given, respectively:

$$E(Y_i|x_i') = \mu_i = \exp(x_i'\beta) \tag{11}$$

$$Var(Y_i|x_i') = \mu_i + \mu_i^2\left(\frac{\alpha^3(\alpha^3+2+\delta(\delta-1))+\delta}{(\alpha^3+\delta)^2}\right)$$
$$= \exp(x_i'\beta) + (\exp(x_i'\beta))^2\left(\frac{\alpha^3(\alpha^3+2+\delta(\delta-1))+\delta}{(\alpha^3+\delta)^2}\right) \quad (12)$$

Proposition 4: Suppose $Y_i|x_i'$ be a response variable for a given set of covariates, x_i'; then by substituting $\mu = \mu_i = \exp(x_i'\beta)$ in equation (7), we can write the pmf of $Y_i|x_i' \sim \text{PMQL}_{GLM}(\beta, \alpha, \delta)$ as:

$$f(y_i|x_i') = \frac{((\alpha^3+1)\exp(x_i'\beta))^{y_i}(\alpha^3+\delta)(\Gamma(\delta)\Gamma(y_i+1)\alpha^3 A_i^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta))}{y_i!(\alpha^3+1)A_i^{y_i+\delta}\Gamma(\delta)}, \quad (13)$$

where $A_i = ((\alpha^3+1)\exp(x_i'\beta) + (\alpha^3+\delta)), i = 1, 2, ..., n$.

4.2. Parameter estimation of the PMQL$_{GLM}$

This subsection introduces the parameter estimation of the $\text{PMQL}_{GLM}(\beta, \alpha, \delta)$. The maximum likelihood estimation method is used to estimate the parameters. The log-likelihood function of its pmf is given by:

$$\ell(\beta, \alpha, \delta|y, x) = \sum_{i=1}^{n} y_i \ln((\alpha^3+1)\exp(x_i'\beta)) + n\ln(\alpha^3+\delta) - \sum_{i=1}^{n}\ln(y_i!) - n\ln(\alpha^3+1) - n\ln(\Gamma(\delta))$$
$$+ \sum_{i=1}^{n}(\Gamma(\delta)\Gamma(y_i+1)\alpha^3 A_i^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)) - \sum_{i=1}^{n}(y_i+\delta)\ln(A_i). \quad (14)$$

Then, the score functions of the regression coefficients $\beta_0, \beta_1, ..., \beta_p$ are:

$$\frac{\partial\ell(\beta, \alpha, \delta|y, x)}{\partial\beta_r} = \sum_{i=1}^{n} y_i x_{ir} - \sum_{i=1}^{n}\frac{(y_i+\delta)(\alpha^3+1)\exp(x_i'\beta)x_{ir}}{A_i}$$
$$+ \sum_{i=1}^{n}\frac{\Gamma(\delta)\Gamma(y_i+1)\alpha^3(\delta-1)A_i^{\delta-2}(\alpha^3+1)\exp(x_i'\beta)x_{ir}}{\Gamma(\delta)\Gamma(y_i+1)\alpha^3 A_i^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)}, r = 0, 1, ..., p,$$

and the score functions of the parameters α, and δ can be easily derived by substituting $\mu = \mu_i = \exp(x_i'\beta)$ in the respective score functions of α and δ defined in subsection 2.3.

By setting the score functions equal to zero and solving the system of non-linear equations, the MLEs of β_r, α, and δ, abbreviated as $\hat{\beta}_{r(MLE)}, \hat{\alpha}_{(MLE)}$, and $\hat{\delta}_{(MLE)}$ can be derived.

The asymptotic confidence intervals for the regression coefficients, $\beta_0, \beta_1, ..., \beta_p$ can be found by the asymptotic theory as $n \to \infty$.

Note:

Based on the asymptotic theory, $\sqrt{n}(\hat{\beta} - \beta) \to N_{p+1}(0, I^{-1}(\hat{\beta}))$ as $n \to \infty$, where $I(\hat{\beta})$ is the observed information matrix of $\hat{\beta}$ and its elements are given in Appendix 2.

Iteratively weighted least square (IWLS) algorithm

As an alternative to this, the iteratively weighted least square (IWLS) algorithm can be also utilized to estimate the β by maximizing equation (14) with respect to β (Dutang 2017).

Let us define $\beta^{(s-1)}$ be the estimated value of β with $(s-1)$ iterations. Then, the Fisher scoring method may be given as:

$$\beta^{(s)} = \beta^{(s-1)} + I^{-1}(\beta^{(s-1)})S(\beta^{(s-1)}), \quad (15)$$

where $S(\beta^{(s-1)})$ be the score function of the regression coefficients evaluated at $\beta^{(s-1)}$ and it is given as:

$$S(\beta^{(s-1)}) = \frac{\partial l(\beta^{(s-1)}, \alpha, \delta|y, x)}{\partial \beta^{(s-1)}} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{(y_i + \delta)(\alpha^3 + 1)\exp(x_i'\beta^{(s-1)})x_i}{A_i}$$
$$+ \sum_{i=1}^{n} \frac{\Gamma(\delta)\Gamma(y_i + 1)\alpha^3(\delta - 1)A_i^{\delta-2}(\alpha^3 + 1)\exp(x_i'\beta^{(s-1)})x_i}{\Gamma(\delta)\Gamma(y_i + 1)\alpha^3 A_i^{\delta-1} + (\alpha^3 + \delta)^{\delta-1}\Gamma(y_i + \delta)}$$
$$= X'W(\beta^{(s-1)})U(\beta^{(s-1)}),$$

where

$$X = \begin{pmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & . & x_{2p} \\ & & & . & & \\ & & & . & & \\ & & & . & & \\ x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix},$$

W be the weight matrix evaluated at $\beta^{(s-1)}$ and it can be obtained by using equations (10) and (12) as:

$$W = diag\left(\frac{1}{(g'(\mu_i(\beta^{(s-1)})))^2 Var(\mu i)(\beta^{(s-1)})}\right)$$
$$= diag\left(\frac{\mu_i(\hat{\beta^{(s\bar{1})}})(\alpha^3 + \hat{\delta})^2}{(\alpha^3 + \hat{\delta})^2 + \hat{\mu}_i(\beta^{(s-1)})(\alpha^3(\alpha^3 + 2 + \delta(\delta - 1)) + \delta)}\right),$$

$U(\beta^{(s-1)})$ be a vector and its i^{th} element is given as:

$$(y_i - \mu_i(\beta^{(s-1)}))g'(\mu_i(\beta^{(s-1)}) = \frac{(y_i - \mu_i(\beta^{(s-1)}))}{\mu_i(\beta^{(s-1)})},$$

and

$$I(\beta^{(s-1)}) = -E\left(\frac{\partial^2 l(\beta^{(s-1)}|y, x)}{\partial \beta^{(s-1)} \partial(\beta^{(s-1)})'}\right) = X'W(\beta^{(s-1)})X.$$

Then, the final part of equation (15) may be written as:

$$X'W(\beta^{(s-1)})X\beta^{(s)} = X'W(\beta^{(s-1)})Xz(\beta^{(s-1)}),$$

by defining $z(\beta^{(s-1)})$ be a vector and its i^{th} element is given as:

$$g(\mu_i(\beta^{(s-1)})) + (y_i - \mu_i(\beta^{(s-1)}))g'(\mu_i(\beta^{(s-1)})) = \log(\mu_i(\beta^{(s-1)})) + \frac{(y_i - \mu_i(\beta^{(s-1)}))}{\mu_i(\beta^{(s-1)})}.$$

Hence,

$$\beta^{(s)} = (X'W(\beta^{(s-1)})X)^{-1}X'W(\beta^{(s-1)})z(\beta^{(s-1)}). \tag{16}$$

The root of equation (16) need to be found iteratively since both $W(\beta^{(s-1)})$ and $z(\beta^{(s-1)})$ contain $\beta^{(s-1)}$. Then, this method is commonly known as IWLS. In the final step of the IWLS algorithm, the $\hat{\beta}_{MLE}$ is obtained as:

$$\hat{\beta}_{MLE} = (X'\hat{W}X)^{-1}X'\hat{W}\hat{z}, \tag{17}$$

where $\hat{W}$ and $\hat{z}$ are the values of the $W(\beta^{(s-1)})$ and $z(\beta^{(s-1)})$ at the final step, respectively.The covariance matrix of this estimator is given as:

$$Cov(\hat{\beta}_{MLE}) = (X'\hat{W}X)^{-1}. \tag{18}$$

Table 2. Statistical measures for data sets 1 to 3.

Data	Min	Max	τ	*SK*	*EK*	VIFs of covariates			
						1	2	3	4
Data 1 ($\alpha = 0.95, \delta = 19.50, \rho = 0.005$)	8	88	13.00	2.00	3.50	1.004	1.003	1.003	1.004
Data 2 ($\alpha = 0.95, \delta = 10.50, \rho = 0.005$)	5	104	20.00	2.50	6.25	1.008	1.003	1.004	1.011
Data 3 ($\alpha = 0.45, \delta = 15.50, \rho = 0.005$)	0	80	28.00	5.50	33.00	1.015	1.004	1.094	1.087

5. Applications

In this section, we use three simulated data sets and a real-world data set to study the applicability of the PMQL regression model over the Poisson regression model (P_{GLM}), NB regression model (NB_{GLM}), and Poisson-Quasi Lindley regression model (PQL_{GLM}), and Generalized Poisson-Lindley regression model (GPL_{GLM}). The best-fitted regression model will be selected based on the negative log-likelihood ($-2logL$) and Akaike Information Criterion (AIC) values. The unknown parameter of regression coefficients and over-dispersed parameters are estimated by using the maximum likelihood estimation method. Further, The system of non-linear equations in the regression coefficients and over-dispersion parameters mentioned in subsection 4.2 were solved by using the *optim* function in the R package *stats*.

5.1. Simulated data applications

Here, we simulate three over-dispersed data sets with a number of covariates (p) is equal to 4 and the sample size (n) is equal to 500. To design the simulation for a data set, the following steps are considered:

i. Simulate the covariates by using the following formula as proposed by McDonald and Galarneau (1975)

$$x_{ij} = \sqrt{(1-\rho^2)}m_{ij} + \rho m_{i,5}, i = 1, 2..., 500, j = 1, 2, 3, 4, \tag{19}$$

where ρ is the correlation among the covariates and m_{ij}'s are independent standard normal pseudo-random numbers.

ii. Based on the simulated covariates in step (i), simulate the response variable $y_i (i = 1, 2, ..., 500)$ from the re-parametrized PMQL $(\mu_i = \exp(x_i'\beta), \alpha, \delta), i = 1, 2, ..., 500$ by using the method discussed in subsection 2.2, where the starting values of the slope parameters are selected such that $\sum_{j=1}^{4} \beta_j^2 = 1$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4$.

Further, to avoid the multicollinearity issue, we set the ρ value such that the variation inflation factor (VIF) of four covariates are less than 5.

Some important statistical measures for simulated data sets 1 to 3 are summarized in Table 2. We may observe that all simulated data sets have a higher sample index of dispersion (τ), skewness (*SK*), and excess kurtosis (*EK*) values. Then, it reveals that the data sets' response variables are extremely over-dispersed and positively skewed distributions with very long right-tails. Further, the VIFs of four covariates in each data set are less than 5. Then, we can say there is no multicollinearity issue in all simulated data sets. Figure 3 shows the distributions of the simulated response variables for data sets 1 to 3.

The estimated parameters with corresponding standard errors (SEs) reported in parenthesis, $-2logL$, and AIC values for different regression models are summarized in Tables 3–5. We may note that the $PMQL_{GLM}$ is the best model based on the minimum $-2logL$, and AIC values.

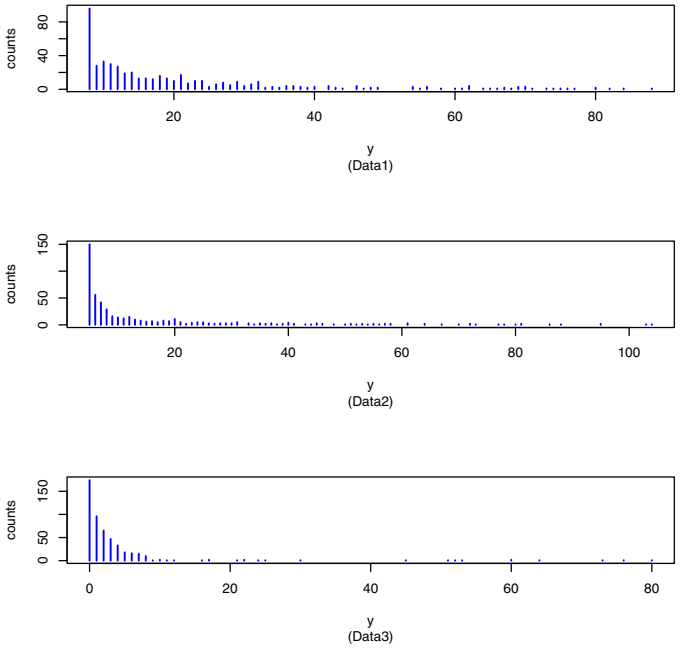

Figure 3. The distribution of response variables for data sets 1 to 3.

Table 3. Inference of the parameters of the regression models for Data 1.

	P_{GLM}		NB_{GLM}		PQL_{GLM}		GPL_{GLM}		$PMQL_{GLM}$	
	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value
β_0	2.702 (0.011)	<0.001	2.716 (0.036)	<0.001	2.811 (0.081)	<0.001	2.713 (0.037)	<0.001	4.056 (0.659)	<0.001
β_1	−0.230 (0.012)	<0.001	−0.144 (0.036)	<0.001	−0.213 (0.022)	<0.001	−0.194 (0.035)	<0.001	−0.133 (0.022)	<0.001
β_2	−0.201 (0.011)	<0.001	−0.144 (0.037)	<0.001	−0.264 (0.024)	<0.001	−0.137 (0.034)	<0.001	−0.152 (0.023)	<0.001
β_3	−0.152 (0.010)	<0.001	−0.059 (0.032)	0.065	−0.165 (0.020)	<0.001	−0.121 (0.033)	<0.001	−0.081 (0.020)	<0.001
β_4	−0.217 (0.011)	<0.001	−0.114 (0.032)	<0.001	−0.048 (0.020)	0.016	−0.172 (0.034)	<0.001	−0.100 (0.010)	<0.001
DP 1	–	–	1.650 (0.106)	<0.001	−0.724 (0.032)	<0.001	0.118 (0.008)	<0.001	−0.987 (0.007)	<0.001
DP 2	–	–	–	–	–	–	–	–	1.512 (0.164)	<0.001
$-2logL$	7863.5		3715.9		3512.9		3715.9		**2169.1**	
AIC	7873.5		3727.9		3500.9		3727.9		**2183.1**	

Note: DP, dispersion parameter.

5.2. Real-world application

The data set was obtained from 1495 Arizona Cardiovascular patients medicare files in 1991 and it is limited to only diagnosis-related group 112 (Hilbe 2011). It represents length of stay (in days) of patients in the hospital (LOS); patient identifies themselves as primarily Caucasian with levels 1-White, 0-Nonwhite (W); patient died with levels 1-Died, 0-Alive (D); admission types with levels 1-Urgent, 2-Emergency, 3-Elective. Wherein, the LOS (response variable) can be explained by the rest of the variables (explanatory variables). The Elective admission type is considered as a referent type for the admission types and Urgent and Emergency admission types are coded U and E, respectively.

Table 4. Inference of the parameters of the regression models for Data 2.

	P_{GLM}		NB_{GLM}		PQL_{GLM}		GPL_{GLM}		$PMQL_{GLM}$	
	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value
β_0	0.284 (0.039)	<0.001	−0.117 (0.063)	0.063	−0.160 (0.072)	0.026	0.516 (0.058)	<0.001	0.632 (0.038)	<0.001
β_1	0.595 (0.023)	<0.001	0.802 (0.040)	<0.001	0.692 (0.039)	<0.001	0.599 (0.034)	<0.001	0.809 (0.033)	<0.001
β_2	0.781 (0.027)	<0.001	0.979 (0.040)	<0.001	1.068 (0.047)	<0.001	0.641 (0.040)	<0.001	0.859 (0.036)	<0.001
β_3	0.776 (0.021)	<0.001	0.931 (0.037)	<0.001	1.112 (0.040)	<0.001	0.644 (0.032)	<0.001	0.734 (0.029)	<0.001
β_4	0.504 (0.020)	<0.001	0.766 (0.041)	<0.001	0.828 (0.034)	<0.001	0.503 (0.030)	<0.001	0.824 (0.030)	<0.001
DP 1	–	–	10.681 (2.834)	<0.001	−0.165 (0.038)	<0.001	0.636 (0.058)	<0.001	−0.937 (0.019)	<0.001
DP 2	–	–	–	–	–	–	–	–	2.543 (0.492)	<0.001
$-2logL$	2130.2		1447.1		1210.8		1694.3		**976.9**	
AIC	2140.2		1459.1		1222.8		1706.3		**990.9**	

Note: DP, dispersion parameter.

Table 5. Inference of the parameters of the regression models for Data 3.

	P_{GLM}		NB_{GLM}		PQL_{GLM}		GPL_{GLM}		$PMQL_{GLM}$	
	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value
β_0	3.000 (0.010)	<0.001	2.932 (0.098)	<0.001	2.852 (0.033)	<0.001	2.760 (0.042)	<0.001	3.216 (0.028)	<0.001
β_1	0.098 (0.010)	<0.001	0.468 (0.068)	<0.001	−0.361 (0.024)	<0.001	0.455 (0.031)	<0.001	−0.083 (0.029)	0.004
β_2	0.015 (0.010)	0.128	0.826 (0.075)	<0.001	−0.361 (0.020)	<0.001	0.246 (0.034)	<0.001	−0.227 (0.025)	<0.001
β_3	0.085 (0.009)	<0.001	−0.043 (0.049)	0.380	−0.019 (0.015)	0.205	0.306 (0.028)	<0.001	0.264 (0.027)	<0.001
β_4	0.094 (0.009)	<0.001	0.877 (0.071)	<0.001	−0.201 (0.015)	<0.001	0.332 (0.028)	<0.001	0.261 (0.026)	<0.001
DP 1	–	–	0.331 (0.035)	<0.001	−0.563 (0.142)	<0.001	0.100 (0.025)	<0.001	−0.839 (0.006)	<0.001
DP 2	–	–	–	–	–	–	–	–	1.468 (0.061)	<0.001
$-2logL$	7129.5		5109.6		4814.2		4029.5		**3909.4**	
AIC	7139.5		5121.6		4826.2		4041.5		**3923.4**	

Note: DP, dispersion parameter.

The index dispersion (τ) of the LOS is 7.917 which is greater than one. The skewness and excess kurtosis of this response variable are 3.648 and 26.163, respectively. These statistics indicate that the distribution of the responses is extremely over-dispersed, highly positive skewed, and having a very long right-tail. Further, the variation inflation factor (VIF) of variables W, D, U, and E are 1.012, 1.013, 1.049, and 1.047, respectively. All VIF < 5 and indicate that there is no multicollinearity for this data set. Figure 4 illustrates the distribution of the response variable.

Table 6 summarizes the estimated regression coefficients and over-dispersed parameters with corresponding standard errors (SEs) reported in parenthesis, $-2logL$, and AIC values for the selected regression models. We can note that $PMQL_{GLM}(\beta, \alpha, \delta)$ provides minimum $-2logL$ and AIC values to this over-dispersed real-data set by comparing with other models. Further, the asymptotic confidence interval for the regression coefficients and over-dispersed parameters are: $-0.2848 < \beta_1 < -0.0461$, $-0.45272 < \beta_2 < -0.3077, 0.0266 < \beta_3 < 0.2042, 0.2635 < \beta_4 < 0.4775$, $-0.8416 < \alpha < -0.8268$, and $1.3648 < \delta < 1.4494$, respectively. It is clear that all intervals do not contain the value zero, which means all parameters are statistically significant. Since

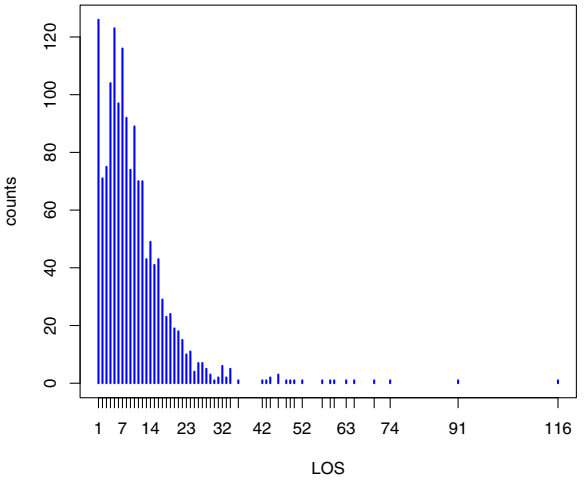

Figure 4. The distribution of length of stay of patients.

Table 6. Inference of the parameters of the regression models for the real-world data.

	P_{GLM}		NB_{GLM}		PQL_{GLM}		GPL_{GLM}		$PMQL_{GLM}$	
	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value	MLE (SE)	p-value
β_0	2.380 (0.027)	<0.001	2.289 (0.065)	<0.001	2.438 (0.065)	<0.001	2.401 (0.058)	<0.001	2.509 (0.061)	<0.001
β_1	−0.141 (0.027)	<0.001	−0.032 (0.065)	0.311	−0.190 (0.064)	0.002	−0.097 (0.059)	0.100	−0.166 (0.061)	0.003
β_2	−0.245 (0.018)	<0.001	−0.251 (0.039)	<0.001	−0.251 (0.037)	<0.001	−0.310 (0.038)	<0.001	−0.380 (0.037)	<0.001
β_3	0.244 (0.021)	<0.001	0.302 (0.050)	<0.001	0.263 (0.047)	<0.001	0.181 (0.044)	<0.001	0.115 (0.045)	0.010
β_4	0.755 (0.026)	<0.001	0.762 (0.073)	<0.001	0.637 (0.070)	<0.001	0.486 (0.062)	<0.001	0.371 (0.055)	<0.001
DP 1	–	–	2.466 (0.115)	<0.001	−0.209 (0.019)	<0.001	0.255 (0.011)	<0.001	−0.834 (0.004)	<0.001
DP 2	–	–	–	–	–	–	–	–	1.407 (0.022)	<0.001
$-2logL$	13679		9569.8		9513.9		9623.1		**9441.2**	
AIC	13689		9581.8		9525.9		9635.1		**9455.2**	

Note: DP, dispersion parameter.

regression coefficients of White and Died variables are negative, if they are white as well as have died, the length of stay of patients in hospital will decrease. Further, LOS for the urgent and emergency admitted patients are 0.1154 and 0.3705 times higher than the elective admitted patients in hospital. Then, we may fit this data set by the following regression model:

$$\mu_i = \exp(2.509 - 0.166W_i - 0.380D_i + 0.115U_i + 0.371E_i).$$

6. Conclusion

In this paper we have introduced and studied a re-parametrized Poisson-Modification of the Quasi Lindley distribution and the Poisson-Modification of the Quasi Lindley (PMQL) regression model. Structural properties of the re-parametrized PMQL distribution in terms of PMQL distribution's mean and two over-dispersion parameters show various flexible properties to cover the over-dispersion on its regression model. The PMQL regression model provides a statistical tool for the modeling of the over-dispersed count responses with appropriate covariates. The

maximum likelihood estimation method is utilized to estimate the unknown parameters of the regression model, and the observed information matrix has also derived. A results of the three simulated data sets and a real-world data set show that the PMQL regression model exhibits a better fit than the Poisson, Negative binomial, and Poisson-Quasi Lindley, Generalized Poisson-Lindley regression models.

1. Appendix 1: The elements of the observed information matrix, $I(\mu, \alpha, \delta)$ defined in proposition 3

Let us define:

$$T_1 = \Gamma(\delta)\alpha^3\Gamma(y_i+1)(\delta-1)A^{\delta-2}(\alpha^3+1),$$
$$T_2 = \Gamma(\delta)\Gamma(y_i+1)\alpha^3A^{\delta-1} + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta),$$
$$T_3 = \Gamma(\delta)\Gamma(y_i+1)(3\alpha^2A^{\delta-2}(A+\alpha^3(\delta-1)(\mu+1)) + 3\alpha^2\Gamma(y_i+\delta)(\delta-1)(\alpha^3+\delta)^{\delta-2},$$

and

$$T_4 = \alpha^3\Gamma(y_i+1)B + \Gamma(y_i+\delta)C + (\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)\psi(y_i+\delta),$$

where A, B, and C are defined as in Sec. 2. Then, the second order partial derivatives are:

$$\frac{\partial^2\ell(\mu,\alpha,\delta|y)}{\partial\mu^2} = \sum_{i=1}^{n}\frac{(y_i+\delta)(\alpha^3+1)^2}{A^2} + \sum_{i=1}^{n}\frac{T_2\frac{\partial T_1}{\partial\mu} - T_1^2}{T_2^2} - \sum_{i=1}^{n}\frac{y_i}{\mu^2},$$

$$\frac{\partial^2\ell(\mu,\alpha,\delta|y)}{\partial\mu\partial\alpha} = \sum_{i=1}^{n}\frac{T_2\frac{\partial T_1}{\partial\alpha} - T_1T_3}{T_2^2} - \sum_{i=1}^{n}\frac{(y_i+\delta)(3\alpha^2(\alpha^3+1)(\mu+1) - 3\alpha^2A)}{A^2},$$

$$\frac{\partial^2\ell(\mu,\alpha,\delta|y)}{\partial\mu\partial\delta} = \sum_{i=1}^{n}\frac{T_2\frac{\partial T_1}{\partial\delta} - T_1T_4}{T_2^2} + \sum_{i=1}^{n}\frac{(\alpha^3+1)(y_i+\delta-A)}{A^2},$$

$$\frac{\partial^2\ell(\mu,\alpha,\delta|y)}{\partial\alpha^2} = \frac{3n\alpha(2\delta-\alpha^3)}{(\alpha^3+\delta)^2} - \frac{3n\alpha(2-\alpha^3)}{(\alpha^3+1)^2} + \sum_{i=1}^{n}\frac{3y_i\alpha(2-\alpha^3)}{(\alpha^3+1)^2} + \sum_{i=1}^{n}\frac{T_2\frac{\partial T_3}{\partial\alpha} - T_3^2}{T_2^2},$$

$$\frac{\partial^2\ell(\mu,\alpha,\delta|y)}{\partial\alpha\partial\delta} = \sum_{i=1}^{n}\frac{T_2\frac{\partial T_3}{\partial\delta} - T_3T_4}{T_2^2} - \frac{3n\alpha^2}{(\alpha^3+\delta)^2} - \sum_{i=1}^{n}\frac{3\alpha^2(\mu+1)(A-(y_i+\delta))}{A^2},$$

$$\frac{\partial^2\ell(\mu,\alpha,\delta|y)}{\partial\delta^2} = \sum_{i=1}^{n}\frac{T_2\frac{\partial T_4}{\partial\delta} - T_4^2}{T_2^2} - \frac{n}{(\alpha^3+\delta)^2} - n\psi_1(\delta) - \sum_{i=1}^{n}\frac{A-(y_i+\delta)}{A^2} - \frac{n}{A},$$

where $\psi_1(s)$ is the trigamma function and defined as

$$\psi_1(s) = \frac{d^2}{ds^2}\ln(\Gamma(s)) = \sum_{i=1}^{n}\frac{1}{(s+k)^2},$$

$$\frac{\partial T_1}{\partial\mu} = (\alpha^3+1)^2\Gamma(\delta)\Gamma(y_i+1)\alpha^3(\delta-1)(\delta-2)A^{\delta-3},$$

$$\frac{\partial T_1}{\partial\alpha} = 3\alpha^2\Gamma(\delta)\Gamma(y_i+1)(\delta-1)A^{\delta-3}(\alpha^3(\alpha^3+1)(\delta-2)(\mu+1) + (2\alpha^3+1)A),$$

$$\frac{\partial T_1}{\partial\delta} = (\alpha^3+1)\Gamma(y_i+1)\alpha^3(\Gamma(\delta)(\delta-1)A^{\delta-3}((\delta-2)+A\ \ln(A)) + A^{\delta-2}\Gamma(\delta)(1+(\delta-1)\psi(\delta)),$$

$$\frac{\partial T_3}{\partial\alpha} = \Gamma(\delta)\Gamma(y_i+1)(3(\delta-1)(\mu+1)\alpha^4A^{\delta-3}(3\alpha^3(\delta-2)(\mu+1)+5A) + 3\alpha A^{\delta-2}(3\alpha^3(\delta-1)(\mu+1)+2A)) + 3\Gamma(y_i+\delta)(\delta-1)\alpha(5\alpha^3+2\delta),$$

$$\frac{\partial T_3}{\partial\delta} = \Gamma(y_i+1)(\Gamma(\delta)(3\alpha^5(\mu+1)A^{\delta-3}((\delta-1)((\delta-2)+A\ \ln(A))+A) 3\alpha^2A^{\delta-2}((\delta-1)+A\ \ln(A))) + 3\alpha^2A^{\delta-2}\Gamma(\delta)\psi(\delta)(\alpha^3(\delta-1)(\mu+1)+A)) + 3\alpha^2\Gamma(y_i+\delta)((\delta-1)(\alpha^3+\delta)\psi(y_i+\delta)+(2\delta+\alpha^3-1)),$$

and

$$\begin{aligned}\frac{\partial T_4}{\partial \delta} &= \alpha^3\Gamma(y_i+1)(\Gamma(\delta)((\delta-1)A^{\delta-3}((\delta-2)+A\ \ln(A))+2A^{\delta-2}+\ln(A)A^{\delta-2}((\delta-1)+A\ \ln(A)))\\ &+((\delta-1)+A\ \ln(A))A^{\delta-2}\Gamma(\delta)\psi(\delta)+A^{\delta-1}\Gamma(\delta)(\psi_1(\delta)+(\psi(\delta))^2)\\ &+\Gamma(\delta)\psi(\delta)A^{\delta-2}((\delta-1)+A\ \ln(A)))\\ &+\Gamma(y_i+\delta)(((\delta-1)(\alpha^3+\delta)^{\delta-3}((\delta-2)+(\alpha^3+\delta)\ln(\alpha^3+\delta))\\ &+2(\alpha^3+\delta)^{\delta-2}+\ln(\alpha^3+\delta)(\alpha^3+\delta)^{\delta-2}((\delta-1)+(\alpha^3+\delta)\ln(\alpha^3+\delta)))\\ &+(\alpha^3+\delta)^{\delta-2}((\delta-1)+(\alpha^3+\delta)\ln(\alpha^3+\delta))\Gamma(y_i+\delta)\psi(y_i+\delta)\\ &+(\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta)(\psi_1(y_i+\delta)+(\psi(y_i+\delta))^2)\\ &+\Gamma(y_i+\delta)\psi(y_i+\delta)(\alpha^3+\delta)^{\delta-2}((\delta-1)+(\alpha^3+\delta)\ln(\alpha^3+\delta)).\end{aligned}$$

2. Appendix 2: The elements of the observed information matrix, $I(\beta)$ defined in subsection 4.2

Let us define:

$$T_5 = \Gamma(\delta)\Gamma(y_i+1)\alpha^3(\delta-1)A_i^{\delta-2}(\alpha^3+1)\exp(x_i'\beta)x_{ir},$$
$$T_6 = \Gamma(\delta)\Gamma(y_i+1)\alpha^3A_i^{\delta-1}+(\alpha^3+\delta)^{\delta-1}\Gamma(y_i+\delta).$$

Then,

$$\frac{\partial^2\ell(\beta,\alpha,\delta|y)}{\partial\beta_r\partial\beta_s}=\sum_{i=1}^{n}\frac{A_i(y_i+\delta)(\alpha^3+\delta)\exp(x_i'\beta)x_{ir}x_{is}-(y_i+\delta)(\alpha^3+1)^2(\exp(x_i'\beta))^2x_{ir}x_{is}}{A_i^2}$$
$$+\sum_{i=1}^{n}\frac{T_6\dfrac{\partial T_5}{\partial\beta_s}-T_5\dfrac{\partial T_6}{\partial\beta_s}}{T_6^2},\ r,s=0,1,\ldots,p,$$

where A_i is defined as proposition 4,

$$\frac{\partial T_5}{\partial\beta_s}=\Gamma(\delta)\Gamma(y_i+1)\alpha^3(\delta-1)(\alpha^3+1)\exp(x_i'\beta)A_i^{\delta-3}x_{is}x_{ir}(A_i+(\delta-2)(\alpha^3+1)\exp(x_i'\beta)),$$

and

$$\frac{\partial T_6}{\partial\beta_s}=\Gamma(\delta)\Gamma(y_i+1)\alpha^3(\delta-1)A_i^{\delta-2}(\alpha^3+1)\exp(x_i'\beta)x_{is}.$$

Acknowledgement

We are greateful for the editor and the referee for their important comments, which led to significantly improve the paper.

Disclosure statement

The authors declare no conflicts of interest regarding the publication of this paper.

ORCID

Ramajeyam Tharshan http://orcid.org/0000-0002-6112-2517
Pushpakanthie Wijekoon http://orcid.org/0000-0003-4242-1017

References

Altun, E. 2019. A new model for over-dispersed count data: Poisson quasi-Lindley regression model. *Mathematical Sciences* 13 (3):241–7. doi:10.1007/s40096-019-0293-5.

Bhati, D., D. V. S. Sastry, and P. Z. Qadri. 2015. A new generalized Poisson Lindley distribution, applications and properties. *Austrian Journal of Statistics* 44 (4):35–51. doi:10.17713/ajs.v44i4.54.

Cameron, A. C., and P. K. Trivedi. 2013. *Regression analysis of count data*. 2nd ed. New York, USA: Cambridge University Press.

Dutang, C. 2017. Some explanations about the IWLS algorithm to fit generalized linear models. hal-01577698.

Greenwood, M., and G. U. Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* 83 (2):255–79. doi:10.2307/2341080.

Grine, R., and H. Zeghdoudi. 2017. On Poisson quasi-Lindley distribution and its applications. *Journal of Modern Applied Statistical Methods* 16 (2):403–17. doi:10.22237/jmasm/1509495660.

Hilbe, J. M. 2011. *Negative binomial regression*. 2nd ed. New York, USA: Cambridge University Press.

Johnson, N. L., S. Kotz, and A. W. Kemp. 1992. *Univariate discrete distributions*. 3rd ed. New Jersey, USA: John Wiley and Sons.

Lynch, J. 1988. Mixtures, generalized convexity and balayages. *Scandinavian Journal of Statistics* 15:203–10.

Mahmoudi, E., and H. Zakerzadeh. 2010. Generalized Poisson – Lindley distribution. *Communications in Statistics - Theory and Methods* 39 (10):1785–98. doi:10.1080/03610920902898514.

McDonald, G. C., and D. I. Galarneau. 1975. A monte carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* 70 (350):407–16. doi:10.2307/2285832.

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sankaran, M. 1970. The discrete Poisson-Lindley distribution. *Biometrics* 26 (1):145–9. doi:10.2307/2529053.

Shoukri, M. M., M. H. Asyali, R. VanDorp, and D. Kelton. 2004. The Poisson inverse Gaussian regression model in the analysis of clustered counts data. *Journal of Data Science* 2 (1):17–32. doi:10.6339/JDS.2004.02(1).135.

Slater, L. J. 1966. *Generalized hypergeometric functions*. Cambridge, UK: Cambridge University Press.

Tharshan, R., and P. Wijekoon. 2020a. Location based generalized Akash distribution: Properties and applications. *Open Journal of Statistics* 10 (02):163–87. doi:10.4236/ojs.2020.102013.

Tharshan, R., and P. Wijekoon. 2020b. A comparison study on a new five-parametergeneralized Lindley distribution with its sub-models. *Statistics in Transition New Series* 21 (2):89–117. doi:10.21307/stattrans-2020-015.

Tharshan, R., and P. Wijekoon. 2021a. A modification of the Quasi Lindley distribution. *Open Journal of Statistics* 11 (03):369–92. doi:10.4236/ojs.2021.113022.

Tharshan, R., and P. Wijekoon. 2022. A New mixed Poisson distribution for over-dispersed count data: Theory and Applications. *Reliability: Theory & Application*. (in press). arXiv:2110.13004. https://arxiv.org/abs/2110.13004.

Wongrin, W., and W. Bodhisuwan. 2016. Generalized Poisson-Lindley linear model for count data. *Journal of Applied Statistics* 44 (15):2659–71. doi:10.1080/02664763.2016.1260095.

Zamani, H., N. Ismail, and P. Faroughi. 2014. Poisson-weighted exponential univariate version and regression model with applications. *Journal of Mathematics and Statistics* 10:148–54. doi:10.3844/jmssp.2014.148.154.