Mathematical Statistician and Engineering Applications
ISSN: 2326-9865

# Count Data Modeling Under Over Dispersion Issue: A Comparative Study

**Mohammed Khalid Mohammed Nory**

Department of Statistics and Informatics, College of Computer science & Mathematics, University of Mosul, Mosul, Iraq

E-mail: mm98989844@gmail.com

**Tahadhaher Abd.**

Department of Statistics and Informatics, College of Computer science & Mathematics, University of Mosul, Mosul, Iraq

E-mail: tahadabd@gmail.com

**Zakariya Yahya Algamal**

Department of Statistics and Informatics, College of Computer science & Mathematics, University of Mosul, Mosul, Iraq

E-mail: zakariya.algamal@uomosul.edu.iq

***Article Info***

***Page Number: 1398 - 1406***

***Publication Issue:***

***Vol 71 No. 3 (2022)***

***Article History***

***Article Received****: 12 January 2022*

***Revised****: 25 February 2022*

***Accepted****: 20 April 2022*

***Publication****: 09 June 2022*

***Abstract***

Statistical modelling of count data has been of extreme interest to researchers. However, in practice, the count data is often identified with overdispersion or underdispersion. The Conway-Maxwell-Poisson regression model (CMPR) has been proven powerful in modelling count data with a wide range of dispersion. In this study, the performance of CMPR is tested under different value of dispersions. Our Monte Carlo simulation results suggest that the CMPR can bring significant improvement relative to Poisson regression model, in terms of AIC, BIC, and Deviance.

**Keywords:**Overdispersion; Conway-Maxwell-Poisson regression model; Poisson regression model; Monte Carlo simulation.

## 1. Introduction

The count response variable is widely included in modeling several real data problems, such as social, automobile insurance claims, healthcare economics, physical sciences, and medical science [1-8]. Specifically, count data regression model is used when the response variable under the study is discrete distributions representing counts and proportions [9, 10].

Consequently, the Poisson regression model is one of the most models that used in modeling count data. However, it assumes that the equidispersion property in which the variance which is a measure of dispersion is equal to the mean for Poisson distribution. This property is often not hold in real data resulting the incapability of fitting the Poisson regression model [11-13].

The Conway–Maxwell–Poisson (CMP) distribution which is introduced by Conway and Maxwell in 1962 [14] is a great tool to overcome the equidispersion issue. This is because CMP can model a wide range of dispersion. In addition, CMP belongs to an exponential family [15].

## 2. Poisson regression model

Most popular distribution when analyzing count data is Poisson regression, where this type of data used in economic, social and medicine. We know that the form of Poisson distribution is:

$$f(y_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}, \quad y_i = 0,1,2,...,i = 1,2,...,n \tag{1}$$

Where $y_i$ the response variable, the expected value of poisson regression is equal to the exponential distribution, such as:

$$\mu_i = e^{x_i \beta} \tag{2}$$

Here $x_i$ is the i-th row of independent variables $X$ which is $n \times p$ with p variables and $\beta$ is a vector of $1 \times p$ of coefficient. By using maximum likelihood method to estimate the coefficient of poisson regression model which is considered non-linear model as follows:

$$\prod_{i=1}^{n} f\ (y_i) = \prod_{i=1}^{n} \left( e^{-\mu_i}\ \frac{{\mu_i}^{y_i}}{y_i\ !} \right) \tag{3}$$

The log-likelihood function of (3) is:

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{n} \left( y_i x_i^T \beta - \exp(x_i^T \beta) - \ln(y_i\ !) \right) \\ &= \sum y_i x_i^T \beta - \sum \exp(x_i^T \beta) - \sum \ln(y_i\ !) \end{aligned} \tag{4}$$

By using the maximum likelihood method to solve the following equation:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \exp(x_i^T \beta)) x_i \tag{5}$$

Since equation (5) is non-linear in $\beta$, then by using weighted least square algorithm, we have:

$$\hat{\beta}_{PR} = (X\ \hat{W} X\ )^{-1} X\ \hat{W} \hat{\mathbf{s}} \tag{6}$$

Where $\hat{W}\ = diag\ (\hat{\mu}_i)$ and $\hat{s}$ is a vector where the diagonal elements are equal to $\log(\hat{\mu}_i) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$.

The covariance matrix of $\hat{\beta}_{ML}$ is equal to the second derivatives to equation (6) by using ML method as follows:

$$Cov\ (\hat{\beta}_{ML}) = (X\ \hat{W} X\ )^{-1} \tag{7}$$

## 3. Conway-Maxwell-Poisson regression model

In real application, count data have often been shown to exhibit overdispersion, meaning that the variance is greater than the mean, and have sometimes shown characteristics of underdispersion, meaning that the variance is less than the mean. The Conway–Maxwell–Poisson distribution (CMPD) offers a simple way to accommodate the overdispersion and underdispersion[16, 17]. The CMPD is an extension of the Poisson distribution with two parameters $\lambda$ (centering parameter related to the observations mean) and $\theta$ (the shape parameter) [18]. Suppose $y \in \{0,1,2,.....\}$ is a random variable that follows a CMPD, then the probability mass function is defined as

$$P\left(Y = y;\lambda,\theta\right) = \frac{\lambda^{y}}{\left(y!\right)^{\theta} Z\left(\lambda,\theta\right)}, \quad \lambda > 1,\ \theta \geq 0, \tag{8}$$

where $Z\left(\lambda,\theta\right) = \sum_{s=0}^{\infty}\left(\lambda^{s}/(s!)^{\theta}\right)$ is a normalizing constant. The CMPD can model both underdispersed ($\theta > 1$) and overdispersed ($\theta < 1$) data.

According to Eq. (8), there is no closed form representation available for the mean. This is because that the normalizing constant, $Z\left(\lambda,\theta\right)$, is an infinite series with no closed form representation [19]. Shmueli, Minka [20] used the asymptotic expression for $Z\left(\lambda,\theta\right)$ in Eq. (8) to express the mean and variance of the CMPD as

$$\begin{aligned} E(Y) &\approx \lambda^{\frac{1}{\theta}} - \frac{\theta - 1}{2\theta}, \\ Var(Y) &\approx \frac{1}{\theta}\lambda^{\frac{1}{\theta}} \end{aligned} \tag{9}$$

For regression modeling in which the count responses may change depending on a set of explanatory variables, it is more convenient and interpretable to model the mean of the CMPD directly. By setting $\mu = \lambda^{\frac{1}{\theta}}$ [21], a re-parameterization of Eq. (8) to provide a clear centering parameter is can be defined as

$$P\left(Y = y;\mu,\theta\right) = \left(\frac{\mu^{y}}{y!}\right)^{\theta} \frac{1}{S\left(\mu,\theta\right)}, \tag{10}$$

where $S\left(\mu,\theta\right) = \sum_{n=0}^{\infty}\left(\mu^{n}/n!\right)^{\theta}$.

Depending on Eq. (10) and in terms of generalized linear model framework, the Conway–Maxwell–Poisson regression model (CMPR) can be formulated as

$$\ln(\mu) = \beta_0 + \sum_{j=1}^{p}\beta_j \mathbf{x}_j, \tag{11}$$

$$\ln(\theta) = \gamma_0 + \sum_{k=1}^{q} \gamma_k \mathbf{m}_j . \quad (12)$$

In Eqs. (13) and (14), $\mathbf{x}_j$ and $\mathbf{m}_j$ are expandatory variables, and there are assumed to be $p$ covariates used in the centering link function and $q$ covariates used in the shape link function. Assuming $\theta$ as a dispersion parameter and using single link function, Eq. (13), with $\eta = \ln(\mu) = \boldsymbol{\beta}\mathbf{x}$ as a linear predictor with log link, where $\boldsymbol{\beta}$ is the vector of regression coefficients including intercept, the log likelihood function can be written a [11]

$$\ell(\beta) = \theta \sum_{i=1}^{n} y_i (\boldsymbol{\beta}\mathbf{x}_i) - \theta \sum_{i=1}^{n} \ln(y_i !) - \sum_{i=1}^{n} \ln\left[S(\boldsymbol{\beta}\mathbf{x}_i, \theta)\right]. \quad (13)$$

Solving Eq. (13), the estimation of the regression parameters, $\boldsymbol{\beta}$, and the estimation of the dispersion parameter, $\theta$, can be obtained as, respectively,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i \theta - \frac{\partial}{\partial \eta_i} \ln[S(\eta_i, \theta)]) x_{ij} \quad (14)$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \theta} = \sum_{i=1}^{n} (-\ln(y_i !) - \frac{\partial}{\partial \theta} \log[S(\eta_i, \theta)]) \quad (15)$$

Iterative reweighted least square (IRLS) is used to solve both Eq. (14) and Eq. (15). By fixing $\theta$, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_{CMPR} = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{\mathbf{u}}, \quad (16)$$

where $\hat{\mathbf{u}} = \ln(\hat{\mu}) + \frac{(y - \hat{\mu})}{\hat{\mu}^2}$ is a vector of the adjusted response variable, and $\hat{\mathbf{W}}$ is a matrix of weights [19].

## 4. Simulation study

In this section, an extensive Monte Carlo simulation study is conducted to evaluate the performance of Poisson regression model, PRM, and Conway–Maxwell–Poisson regression model, CMPR, under different conditions. The response variable of $n \in \{30, 50, 150\}$ observations from CMP regression

model is generated as $y_i \sim CMP(\mu_i, \theta)$, where $\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ with $\sum_{j=1}^{p} \beta_j^2 = 1$ and $\beta_1 = \beta_2 = ... = \beta_p$ [22]. Three different values of the dispersion parameter, $\theta$, are considered to capture overdispersion $(\theta = 1.5)$and $(\theta = 10)$. The explanatory variables $\mathbf{x}_i^T = (x_{i1}, x_{i2}, ..., x_{in})$ have been generated from the following formula

$$x_{ij} = (1-\rho^2)^{1/2} w_{ij} + \rho w_{ip}, \; i = 1, 2, ..., n, \quad j = 1, 2, ..., p, \tag{27}$$

where $\rho = 0.5$ represents the correlation between the explanatory variables and $w_{ij}$ 's are independent standard normal pseudo-random numbers. In addition, the number of the explanatory variables is considered as $p = 5$ and $p = 10$. Three evaluating criteria were used: Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), and Deviance.

The averaged AIC, BIC, and Deviance for all the combination of $n, \theta$, and $p$, are respectively summarized in Tables 1 – 3. The best value is highlighted in bold. As Tables 1, 2, and 3 show, CMPR, achieved smaller averaged AIC, BIC, and Deviance than PRM. In general, this finding specifies that the CMPR is significantly decreasing the bias in estimating the parameter. In terms of Deviance, it is evident from Tables 1, 2, and 3 that CMPR are is quite better than the PRM.

Regarding the number of explanatory variables, it is easily seen that there is a negative impact on Deviance, where there are increasing in their values when the $p$ increasing from 5 variables to 10 variables. In Addition, in terms of the sample size $n$, the Deviance decrease when $n$ increases, regardless the value of $\theta$ and $p$.

Table 1: Averaged AIC, BIC, and Deviance when $n = 30$

| | | $\theta = 1.5$ | | $\theta = 10$ | |
|---|---|---|---|---|---|
| | | PRM | CMPR | PRM | CMPR |
| P=5 | AIC | 137.6614 | **133.66144** | 207.7209 | **203.72086** |
| | BIC | 127.2543 | **125.45426** | 197.3137 | **195.31368** |
| | Deviance | 23.3434 | **22.1555** | 27.14671 | **26.14671** |
| P=10 | AIC | 248.5789 | **248.57892** | 217.3627 | **214.76274** |

| | | | | | |
|---|---|---|---|---|---|
| | BIC | 227.5467 | **227.54667** | 200.9496 | **198.94957** |
| | Deviance | 48.89717 | **47.89717** | 23.90295 | **23.40295** |

Table 2: Averaged AIC, BIC, and Deviance when $n = 50$

| | | $\theta = 1.5$ | | $\theta = 10$ | |
|---|---|---|---|---|---|
| | | PRM | CMPR | PRM | CMPR |
| P=5 | AIC | 244.9697 | **240.96971** | 313.1853 | **308.18527** |
| | BIC | 231.4976 | **229.49758** | 298.7131 | **296.41313** |
| | Deviance | 58.81776 | **57.81776** | 25.90338 | **24.20338** |
| P=10 | AIC | 252.5789 | **243.57892** | 359.6199 | **355.61986** |
| | BIC | 229.5467 | **227.54667** | 336.8876 | **334.5876** |
| | Deviance | 49.89717 | **47.49717** | 56.81274 | **55.41274** |

Table 3: Averaged AIC, BIC, and Deviance when $n = 150$

| | | $\theta = 1.5$ | | $\theta = 10$ | |
|---|---|---|---|---|---|
| | | PRM | CMPR | PRM | CMPR |
| P=5 | AIC | 636.8294 | **631.8294** | 945.8044 | **940.8044** |
| | BIC | 620.7656 | **607.7656** | 924.7406 | **922.7406** |
| | Deviance | 130.9503 | **130.3503** | 152.8789 | **151.8789** |
| P=10 | AIC | 656.0188 | **651.0188** | 946.825 | **942.825** |
| | BIC | 619.9018 | **617.1018** | 910.908 | **908.708** |
| | Deviance | 131.8299 | **131.3299** | 116.9988 | **116.1988** |

## 5. Conclusions

The Conway–Maxwell–Poisson regression model is very popular statistical model to analyze data whose response variable are counts. This paper addresses issue of overdispersion. According to Monte Carlo simulation studies, it has been seen that the CMPR can bring significant improvement relative to Poisson regression model, in terms of AIC, BIC, and Deviance.

REFERENCES

1. Alkhateeb, A. and Z. Algamal, *Jackknifed Liu-type Estimator in Poisson Regression Model.* Journal of the Iranian Statistical Society, 2020. **19**(1): p. 21-37.
2. Rashad, N.K. and Z.Y. Algamal, *A New Ridge Estimator for the Poisson Regression Model.* Iranian Journal of Science and Technology, Transactions A: Science, 2019. **43**(6): p. 2921-2928.
3. Algamal, Z.Y. and M.H. Lee, *Adjusted adaptive lasso in high-dimensional Poisson regression model.* Modern Applied Science, 2015. **9**(4): p. 170-176.
4. Al-Taweel, Y. and Z. Algamal, *Almost unbiased ridge estimator in the zero-inated Poisson regression model.* TWMS Journal Of Applied And Engineering Mathematics, 2022. **12**(1): p. 235-246.
5. Algamal, Z.Y., *Diagnostic in Poisson regression models.* Electronic Journal of Applied Statistical Analysis, 2012. **5**(2): p. 178-186.
6. Rasheed, H.A., et al., *Jackknifed Liu-type estimator in the Conway-Maxwell Poisson regression model.* 2022. **13**(1): p. 3153-3168.
7. Algamal, Z.Y. and M.H. Lee, *Penalized Poisson regression model using adaptive modified elastic net penalty.* Electronic Journal of Applied Statistical Analysis, 2015. **8**(2): p. 236-245.
8. Algamal, Z.Y. and M.M.J.E.J.o.A.S.A. Alanaz, *Proposed methods in estimating the ridge regression parameter in Poisson regression model.* 2018. **11**(2): p. 506-515.
9. Vanegas, L.H., L.M.J.J.o.S.C. Rondon, and Simulation, *A data transformation to deal with constant under/over-dispersion in Poisson and binomial regression models.* 2020. **90**(10): p. 1811-1833.
10. Forthmann, B. and P. Doebler, *Reliability of researcher capacity estimates and count data dispersion: a comparison of Poisson, negative binomial, and Conway-Maxwell-Poisson models.* Scientometrics, 2021.
11. Francis, R.A., et al., *Characterizing the performance of the conway-maxwell poisson generalized linear model.* 2012. **32**(1): p. 167-183.
12. Abdella, G.M., et al., *Penalized Conway-Maxwell-Poisson regression for modelling dispersed discrete data: The case study of motor vehicle crash frequency.* Safety Science, 2019. **120**: p. 157-163.

13. Huang, A.J.S.M., *Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts.* 2017. **17**(6): p. 359-380.
14. Conway, R.W. and W.L.J.J.o.I.E. Maxwell, *A queuing model with state dependent service rates.* 1962. **12**(2): p. 132-136.
15. Choo-Wosoba, H., S.M. Levy, and S.J.B. Datta, *Marginal regression models for clustered count data based on zero-inflated Conway–Maxwell–Poisson distribution with applications.* 2016. **72**(2): p. 606-618.
16. Lord, D., S.R. Geedipally, and S.D. Guikema, *Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting underdispersion.* Risk Anal, 2010. **30**(8): p. 1268-76.
17. Lord, D., S.D. Guikema, and S.R. Geedipally, *Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes.* Accid Anal Prev, 2008. **40**(3): p. 1123-34.
18. Santarelli, M.F., et al., *A Conway-Maxwell-Poisson (CMP) model to address data dispersion on positron emission tomography.* Comput Biol Med, 2016. **77**: p. 90-101.
19. Chatla, S.B. and G. Shmueli, *Efficient estimation of COM–Poisson regression and a generalized additive model.* Computational Statistics & Data Analysis, 2018. **121**: p. 71-88.
20. Shmueli, G., et al., *A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution.* 2005. **54**(1): p. 127-142.
21. Guikema, S.D. and J.P. Coffelt, *A flexible count data regression model for risk analysis.* Risk Anal, 2008. **28**(1): p. 213-23.
22. Kibria, B.M.G., *Performance of some new ridge regression estimators.* Communications in Statistics - Simulation and Computation, 2003. **32**(2): p. 419-435.