Mathematical and Computational Applications

MDPI

Article

Prediction Interval for Compound Conway–Maxwell–Poisson Regression Model with Application to Vehicle Insurance Claim Data

Jahnavi Merupula [1], V. S. Vaidyanathan [1] and Christophe Chesneau [2,*]

[1] Department of Statistics, Pondicherry University, Puducherry 605014, India
[2] Department of Mathematics, LMNO, University of Caen, 14032 Caen, France
* Correspondence: christophe.chesneau@unicaen.fr

Abstract: Regression models in which the response variable has a compound distribution have applications in actuarial science. For example, the aggregate claim amount in a vehicle insurance portfolio can be modeled using a compound Poisson distribution. In this paper, we propose a regression model, wherein the response variable is assumed to have a compound Conway–Maxwell–Poisson (CMP) distribution. This distribution is a parsimonious two-parameter Poisson distribution that accounts for both over- and under-dispersed count data, making it more suitable for application in various fields. A two-part methodology in the framework of a generalized linear model is proposed to estimate the parameters. Additionally, a method to obtain the prediction interval of the response variable is developed. The workings of the proposed methodology are illustrated through simulated data. An application of the compound CMP regression model to real-life vehicle insurance claims data is presented.

Keywords: aggregate claims distribution; compound CMP regression model; generalized linear models; prediction intervals

Citation: Merupula, J.; Vaidyanathan, V.S.; Chesneau, C. Prediction Interval for Compound Conway–Maxwell–Poisson Regression Model with Application to Vehicle Insurance Claim Data. *Math. Comput. Appl.* **2023**, *28*, 39. https://doi.org/10.3390/mca28020039

Academic Editor: Sandra Ferreira

Received: 16 January 2023
Revised: 15 February 2023
Accepted: 27 February 2023
Published: 9 March 2023

1. Introduction

Compound regression models have applications in various research fields, including economics and finance. In economic consumer theory, for example, compound Poisson regression models are often used to examine the factors that account for the expenditures incurred by tourists during their stay at a location. The factors may include length of stay, type of holiday accommodations, age, occupation, socio-economic status of the tourist, etc. See Gómez-Déniz and Pérez-Rodríguez [1]. In actuarial risk theory, the aggregate claim amount incurred by the insurance company against the claims made by the policyholders is modeled using compound models. See Klugman et al. [2] and Bahnemann [3] for a detailed discussion on compound models, their distributional properties and applications in insurance claim modeling. Jørgensen and Paes De Souza [4] applied the compound Poisson regression model to determine the impact on the conditional mean of the aggregate claim amount caused by factors such as age and model of the vehicle, exposure, deductibles, etc., in the context of car insurance. In this paper, we propose a compound regression model using a two-parameter Poisson distribution. On this topic, some mathematical backgrounds are presented below in order to fix the notations. Let

$$S = \sum_{j=1}^{N} Y_j, \tag{1}$$

denote the random sum, where the distributions of the random variables N and $Y_1, Y_2, \ldots, Y_N$ are assumed to be discrete and continuous, respectively. Moreover, (Y_j)s are assumed to be independent and identically distributed. Therefore, in the sequel, we refer to Y_js as Y.

Further, N and Y in general are assumed to be independent. The above-mentioned S is a compound random variable. Suppose Y_j represents the claim amounts on an insurance portfolio, N denotes the number of claims made, then S represents the aggregate claim amount. When N has a Poisson distribution, the distribution of S is known as the compound Poisson distribution. Though the Poisson distribution is often used in constructing compound distributions, it is not suitable for modeling over- or under-dispersed count data. As an alternative to the Poisson distribution, one can use a generalized Poisson distribution (Consul and Jain [5]) to model count data that are either over- or under-dispersed. Recently, Shmueli et al. [6] studied a two-parameter Poisson distribution developed by Conway and Maxwell [7] known as the Conway–Maxwell–Poisson (CMP) distribution. This is a two-parameter flexible generalization of the Poisson distribution that can model both over- and under-dispersed data and has the feature to include the Poisson, geometric and Bernoulli distributions as special cases. A detailed discussion on the properties of this distribution and its applications can be found in Sellers et al. [8]. Also, Sellers and Premeaux [9] contains a detailed review on CMP regression models. In the context of compound distributions, assuming the CMP and binomial distributions for N and Y in Equation (1), a discrete compound CMP-binomial distribution is developed by Saavithri et al. [10].

Considering the Poisson distribution as the counting distribution, compound Poisson regression models are available in the literature. See Frees et al. [11], Andersen and Bonat [12], and Delong et al. [13]. However, its applicability is limited to data with equi-dispersed counts. To allow for flexibility in the compound regression models in terms of accommodating dispersed counts, a counting distribution that can model both over- and under-dispersed data should be considered. This serves as motivation to use the CMP distribution as the counting distribution to build a compound regression model.

The goal of this work is to create a regression model for S using a CMP distribution for N. The present work is novel because of the distribution used for N and its convolution with the distribution of Y. The problem of obtaining prediction intervals for the response variable S is also addressed. The parameters of the compound regression model are estimated using the generalized linear model (GLM) approach in two cases. In the first case, we assume that data on S are available but not on N and Y. We assume data on both N and Y are available in the latter case. For this case, a two-part likelihood-based estimation procedure is developed within the framework of the GLM. A methodology to obtain the prediction interval (PI) for the response variable of the proposed compound regression model is developed.

The rest of the paper is organized as follows: The compound CMP regression model is given in Section 2. In Section 3, the estimation of the parameters of the proposed regression model using the GLM approach is discussed. Section 4 deals with the suggested methodology for obtaining the prediction intervals for the compound CMP regression model. A numerical illustration of the estimation procedure using simulated data and an application to real-life vehicle insurance claims data is presented in Section 5. The conclusion of the paper is given in Section 6.

2. Compound CMP Regression Model

The probability mass function (pmf) of the random variable N having the CMP distribution is given by

$$P(N=n) = \frac{\lambda^n}{(n!)^\nu Z(\lambda,\nu)}, \qquad n = 0,1,2,\ldots,\ \lambda > 0,\ \nu \geq 0, \tag{2}$$

where $Z(\lambda,\nu) = \sum_{j=0}^{\infty} \lambda^j/(j!)^\nu$ is the normalizing constant. Some important remarks on this distribution are given below. The parameters λ and ν are the location and dispersion parameters, respectively. This pmf is not defined for $\lambda \geq 1$ and $\nu = 0$. The mean and variance of N are given by $E(N) = \lambda \frac{\partial \ln Z(\lambda,\nu)}{\partial \lambda}$ and $V(N) = \lambda \frac{\partial E(N)}{\partial \lambda}$, respectively. When

$\nu = 1$, the CMP distribution reduces to the Poisson distribution. For $\nu > 1$, the distribution is under-dispersed, and for $\nu < 1$, it is over-dispersed.

Since the location parameter λ of the CMP distribution does not represent its mean, a mean reparameterized form of the distribution is used in building the compound regression model. The pmf of N under the mean-reparametrization is given by

$$P(N = n) = \left(\mu_1 + \frac{e^{\phi} - 1}{2e^{\phi}}\right)^{ne^{\phi}} \frac{(n!)^{-e^{\phi}}}{Z(\mu_1, \phi)}, \quad n = 0, 1, 2, \ldots, \mu_1 > 0, \phi \in \mathbb{R}, \tag{3}$$

where $Z(\mu_1, \phi) = \sum_{j=0}^{\infty}\left(\mu_1 + \frac{e^{\phi} - 1}{2e^{\phi}}\right)^{je^{\phi}} \frac{1}{(j!)^{e^{\phi}}}$ is the normalizing constant. When $\phi = 0$, the distribution reduces to the Poisson distribution. For $\phi > 0$, the distribution is under-dispersed, and for $\phi < 0$, it is over-dispersed. See Ribeiro Jr et al. [14]. Here, $\mu_1 \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}$ corresponds to the mean of the distribution and $\phi = \ln(\nu)$. This approximation works reasonably well for $\nu \leq 1$ or $\lambda > 10^{\nu}$. The mean and variance of N are $E(N) = \mu_1$ and $V(N) = \mu_1 e^{-\phi}$, respectively.

Convolutions can be used to obtain the probability density function (pdf) of the random sum S defined in Equation (1). In Equation (1), $N = 0$ implies $S = 0$. Let p_0 denote the probability mass at $S = 0$. Since S is not continuous at zero, the pdf of S is represented as a generalized pdf in terms of Dirac delta function as

$$f(s) = p_0\delta(s) + \sum_{i=1}^{\infty} g_Y^{*i}(s)P(N = i), \quad s \geq 0, \tag{4}$$

where $\delta(s)$ is the Dirac delta function such that $\int_0^{\infty} \delta(s)ds = 1$. Here, $P(N = i)$ denotes the pmf of the CMP distribution defined in Equation (3), and $g_Y^{*i}(.)$ denotes the pdf of the i-fold convolution of Y, whose distribution is assumed to be continuous with support in $\mathbb{R}^+$. Note that $p_0 = P(N = 0) = Z(\mu_1, \phi)^{-1}$. In this paper, the distribution of Y is considered to be a mean reparameterized gamma distribution. Based on Jorgensen [15] (Chapter 3), the pdf of Y is given by

$$g_Y(y; \mu_2, \psi) = \frac{1}{\Gamma(\psi)}\left(\frac{\psi}{\mu_2}\right)^{\psi} y^{\psi - 1} \exp\left(-\frac{\psi y}{\mu_2}\right), \quad y > 0, \mu_2 > 0, \psi > 0, \tag{5}$$

where μ_2 denotes the mean of Y, ψ denotes the dispersion parameter and $\Gamma(.)$ denotes the gamma function. This form is taken for mathematical convenience and to accommodate asymmetry in the distribution of Y. For example, in the context of insurance claim modeling, the individual claim amounts are always positive and often right-skewed. Since the gamma distribution is closed under convolution, we obtain

$$g_Y^{*i}(y) = \frac{1}{\Gamma(\psi)}\left(\frac{\psi}{i\mu_2}\right)^{\psi} y^{\psi - 1} \exp\left(-\frac{\psi y}{i\mu_2}\right), \quad y > 0, \mu_2 > 0, \psi > 0. \tag{6}$$

Using Equations (3) and (6) in Equation (4), we obtain

$$f(s) = p_0\delta(s) + \frac{s^{\psi - 1}\psi^{\psi}}{Z(\mu_1, \phi)\mu_2^{\psi}\Gamma(\psi)} \sum_{i=1}^{\infty}\left(\mu_1 + \frac{e^{\phi} - 1}{2e^{\phi}}\right)^{ie^{\phi}} \frac{(i!)^{-e^{\phi}}}{i^{\psi}} \exp\left(\frac{-\psi s}{i\mu_2}\right), \quad s \geq 0. \tag{7}$$

The pdf of S defined in Equation (7) is called the compound CMP gamma pdf. For the random sum defined in Equation (1), we have

$$\begin{cases} E(S) = E(N)E(Y), \\ V(S) = E(N)V(Y) + V(N)[E(Y)]^2. \end{cases} \tag{8}$$

See, for instance, Bahnemann [3] (Chapter 4). Using Equation (8), the mean and variance of the compound CMP gamma distribution given in Equation (7) are obtained as

$$\begin{cases} E(S) = \mu_1\mu_2, \\ V(S) = \mu_2^2\mu_1[\psi^{-1} + e^{-\phi}]. \end{cases} \tag{9}$$

To build a compound regression model for S, let $X = (\vec{1}, \vec{X}_1, \vec{X}_2, \ldots, \vec{X}_p)$ denote the design matrix where $\vec{X}_i, i = 1, 2, \ldots, p$ are the column vectors corresponding to the covariates $X_i, i = 1, 2, \ldots, p$ and $\vec{1}$ is the vector of $1's$. Following the GLM procedure given in De Jong et al. [16] (Chapter 5), the model is built by regressing S on X using the log-link function. This is because the log-link function guarantees that the expected value of the response variable is positive. Let μ denote the expected value of S. Then, the compound CMP gamma regression model is given by

$$\mu = \exp(X\delta), \tag{10}$$

where $\delta = (\delta_0, \delta_1, \ldots, \delta_p)'$ is a $(p+1) \times 1$ vector of regression parameters. In the context of modeling vehicle insurance claims data, S may denote the aggregate claim amount, and the covariates may denote the driver's age, vehicle type, and so on. In the sequel, the method of estimating the regression parameters using the likelihood approach is discussed.

3. Parameter Estimation

Consider a sample $\vec{s} = (s_1, s_2, \ldots, s_r)'$ of r observations on S. Let $D(> 0)$ positive values in $\vec{s}$ and $r - D$ zeros exist. Note that D can be assimilated to be random and $D \sim Binomial(r, 1 - p_0)$, where $p_0 = Z(\mu_1, \phi)^{-1}$. Therefore, the likelihood function L based on $\vec{s}$ and $D = d$ is

$$\begin{aligned} L &= \binom{r}{d} p_0^{r-d}(1-p_0)^d \prod_{k=1}^{d} f(s_k^+) \\ &= \binom{r}{d} \left(\frac{1}{Z(\mu_1,\phi)}\right)^{r-d} \left(1 - \frac{1}{Z(\mu_1,\phi)}\right)^d \prod_{k=1}^{d} f(s_k^+), \end{aligned} \tag{11}$$

where $f(s_k^+) = \dfrac{s_k^{\psi-1}\psi^\psi}{(Z(\mu_1,\phi)-1)\mu_2^\psi\Gamma(\psi)} \sum_{i=1}^{\infty} \left(\mu_1 + \dfrac{e^\phi - 1}{2e^\phi}\right)^{ie^\phi} \dfrac{(i!)^{-e^\phi}}{i^\psi} \exp\left(\dfrac{-\psi s_k}{i\mu_2}\right)$.
Thus, the log-likelihood function l based on $\vec{s}$ and $D = d$ is obtained as

$$\begin{aligned} l(\mu_1, \mu_2, \phi, \psi; \vec{s}) &= \ln\left(\binom{r}{d}\right) - r\ln(Z(\mu_1,\phi)) + (\psi-1)\sum_{k=1}^{d}\ln(s_k) - \sum_{k=1}^{d}\psi\ln(\mu_2) + d\psi\ln(\psi) \\ &\quad - d\ln(\Gamma(\psi)) + \sum_{k=1}^{d}\ln\left[\sum_{i=1}^{\infty}\left(\mu_1 + \frac{e^\phi-1}{2e^\phi}\right)^{ie^\phi} \frac{(i!)^{-e^\phi}}{i^\psi} \exp\left(\frac{-\psi s_k}{i\mu_2}\right)\right]. \end{aligned} \tag{12}$$

Since $E(N) = \mu_1$ and $E(Y) = \mu_2$, from Equation (9), we obtain $\mu = \mu_1\mu_2$. Let the elements of the design matrix X be $x_{kl}, l = 0, 1, \ldots, p;\ k = 1, 2, \ldots, d$ with the k^{th} row given by $\boldsymbol{x_k} = (1, x_{k1}, x_{k2}, \ldots, x_{kp})$. Replacing μ_2 with $\dfrac{\mu}{\mu_1}$ and μ with $\exp(X\delta)$ in Equation (12), the log-likelihood function based on $\vec{s}$ and $D = d$ becomes

$$\begin{aligned} l(\delta, \mu_1, \phi, \psi; \vec{s}) &= \ln\left(\binom{r}{d}\right) - r\ln(Z(\mu_1,\phi)) + (\psi-1)\sum_{k=1}^{d}\ln(s_k) - \sum_{k=1}^{d}\psi\ln\left(\frac{e^{\sum_{l=0}^{p} x_{kl}\delta_l}}{\mu_1}\right) \\ &\quad + d\psi\ln(\psi) - d\ln(\Gamma(\psi)) + \sum_{k=1}^{d}\ln\left[\sum_{i=1}^{\infty}\left\{\left(\mu_1 + \frac{e^\phi-1}{2e^\phi}\right)^{ie^\phi} \frac{(i!)^{-e^\phi}}{i^\psi} \exp\left(\frac{-\psi s_k\mu_1}{ie^{\sum_{l=0}^{p} x_{kl}\delta_l}}\right)\right\}\right]. \end{aligned} \tag{13}$$

The maximum likelihood (ML) estimates of the parameters in Equation (13) can be obtained by solving the $(p+4)$ log-likelihood equations simultaneously. However, these equations are non-linear, and therefore closed-form solutions cannot be obtained. Hence, iterative algorithms based on numerical methods can be used to solve the equations to get the estimates for the parameters. Let $\hat{\delta}$ denote the ML estimate of δ. By the asymptotic property of the ML estimators, for large r, the following distribution approximation holds:

$$\Sigma_{\delta}^{1/2}(\hat{\delta} - \delta) \sim \mathcal{N}_{p+1}(\mathbf{0}, I),$$

where δ and Σ_{δ} denote the mean vector and the covariance matrix of $\hat{\delta}$, respectively. Using Equation (10), an estimate of the expected value of S given the covariates X can be obtained as $\hat{\mu} = \exp(X\hat{\delta})$.

Assume that data on S are unavailable, but data on N and Y are. This can happen in such situations as, for example, when modeling the aggregate claim amount when one has data on the claim frequency (N) and the individual claim amounts (Y). Using N and Y, we can compute the value of S and then build the regression model using the method described above. However, it is computationally more challenging to compute the estimates due to the presence of an infinite sum in the log-likelihood function. To reduce the computational difficulty, we can use N and Y to build two separate regression models to obtain $\hat{\mu}$. Towards this, a two-part GLM methodology is proposed to estimate μ assuming N and Y to be (1) independent and (2) dependent.

3.1. Independent Compound Regression Model

Using Equation (9), we have $\mu = \mu_1\mu_2$. The proposed two-part GLM method is implemented by building two separate regression models, namely, the CMP regression model and the gamma regression model, for the means of N and Y, respectively. Given the data on N, Y and X, the estimated mean of S is computed as $\hat{\mu} = \hat{\mu}_1\hat{\mu}_2$. Here, $\hat{\mu}_1$ and $\hat{\mu}_2$ are obtained by regressing N and Y separately on X. Using the log-link function, we have $\mu_1 = E(N) = e^{X\boldsymbol{\alpha}}, \mu_2 = E(Y) = e^{X\boldsymbol{\beta}}$, where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ denote the set of regression parameters.

Let $\vec{n} = (n_1, \ldots, n_m)'$ denote m observations on N. For each $n_k > 0$, let there be n_k observations on Y denoted by $y_{kj}, j = 1, 2, \ldots, n_k, k = 1, 2, \ldots, m$. Let $\vec{\bar{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)'$ where $\bar{y}_k = \begin{cases} \sum_{j=1}^{n_k} y_{kj}/n_k & \text{if } n_k > 0 \\ 0 & \text{if } n_k = 0. \end{cases}$

Let the design matrix X be of order $m \times (p+1)$ with elements $x_{kl}, k = 1, 2, \ldots, m;\ l = 0, 1, \ldots, p$. Since the distribution of Y has positive support, zeros in $\vec{\bar{y}}$, if any, are not to be considered. The corresponding sample observation in $\vec{\bar{y}}$ and the observed covariate matrix X are not included when building the gamma regression model. Let q denote the number of observations for which $\bar{y}_k = 0, k = 1, 2, \ldots, m$ and let $t = m - q$. Following Garrido et al. [17], the distribution of $Y \sim gamma(\mu_2, \psi)$ is equivalent to $\bar{Y}|N \sim gamma\left(\mu_2, \frac{\psi}{N}\right)$ for independently identically distributed $Y_1, \ldots, Y_N$. Using the pmf of N given in Equation (3) with $\mu_1 = e^{X\boldsymbol{\alpha}}$, the corresponding log-likelihood function is given by

$$l(\boldsymbol{\alpha}, \phi; \vec{n}) = \sum_{k=1}^{m} e^{\phi}\left[n_k \ln\left(e^{\sum_{l=0}^{p} x_{kl}\alpha_l} + \frac{e^{\phi}-1}{2e^{\phi}}\right) - \ln(n_k!)\right] - \sum_{k=1}^{m} \ln\left(Z(e^{\sum_{l=0}^{p} x_{kl}\alpha_l}, \phi)\right). \quad (14)$$

The ML estimates for the $(p+1)$ regression parameters are obtained by simultaneously solving the corresponding log-likelihood equations. Let $\hat{\boldsymbol{\alpha}} = (\hat{\alpha_0}, \hat{\alpha_1}, \ldots, \hat{\alpha_p})'$ denote the ML estimate of $\boldsymbol{\alpha}$. Then the ML estimate of μ_1 is obtained as $\hat{\mu}_1 = e^{X\hat{\boldsymbol{\alpha}}}$. In similar lines, the

ML estimate of $\boldsymbol{\beta}$, namely, $\hat{\boldsymbol{\beta}} = (\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_p})'$, is obtained using the likelihood function corresponding to the conditional pdf of $\bar{Y}$ given $N = n$. The conditional pdf is given by

$$f(\bar{y}|n; \mu_2, \psi) = \frac{1}{\Gamma(\psi/n)} \left(\frac{\psi/n}{\mu_2}\right)^{\psi/n} \bar{y}^{(\psi/n)-1} \exp\left(-\frac{\psi\bar{y}}{n\mu_2}\right), \quad \bar{y} > 0. \tag{15}$$

Taking $\mu_2 = e^{X\boldsymbol{\beta}}$ in Equation (15), the log-likelihood function is obtained as

$$l(\boldsymbol{\beta}, \psi; \vec{\bar{y}}) = -t \ln\left(\Gamma\left(\frac{\psi}{n}\right)\right) + \frac{t\psi}{n} \ln\left(\frac{\psi}{n}\right) + \sum_{k=1}^{t} \left[\left(\frac{\psi}{n} - 1\right) \ln(\bar{y}_k) - \frac{\psi\bar{y}_k}{ne^{\sum_{l=0}^{p} x_{kl}\beta_l}} - \frac{\psi}{n} \sum_{l=0}^{p} x_{kl}\beta_l\right]. \tag{16}$$

The likelihood equations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are, respectively, given by

$$\sum_{k=1}^{m} x_{kl}(n_k - e^{\sum_{l=0}^{p} x_{kl}\alpha_l}) = 0 \tag{17}$$

and

$$\sum_{k=1}^{t} \frac{x_{kl} n_k}{e^{\sum_{l=0}^{p} x_{kl}\beta_l}} (\bar{y}_k - e^{\sum_{l=0}^{p} x_{kl}\beta_l}) = 0, \quad l = 0, 1, \ldots, p. \tag{18}$$

Since Equations (17) and (18) are non-linear, iterative procedures can be used to solve them. As an alternate, one can use the in-built functions `cmp()` and `glm(., family='gamma')` available in R to obtain $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Using $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, the ML estimate of the expected value of S, namely, $\hat{\mu} = \hat{\mu}_1\hat{\mu}_2$, can be computed. By the asymptotic property of the ML estimators, we have

$$\Sigma_{\boldsymbol{\alpha}}^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \sim \mathcal{N}_{p+1}(\mathbf{0}, I)$$

and

$$\Sigma_{\boldsymbol{\beta}}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}_{p+1}(\mathbf{0}, I).$$

Here, $\boldsymbol{\alpha}$ and $\Sigma_{\boldsymbol{\alpha}}$ denote the mean vector and covariance matrix of $\hat{\boldsymbol{\alpha}}$, respectively. Similarly, $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{\beta}}$ denote the mean vector and covariance matrix of $\hat{\boldsymbol{\beta}}$, respectively. The standard errors of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the square root of the diagonal elements of the corresponding covariance matrices. Since $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ do not have closed-form expressions, their standard errors can be obtained using the sample Hessian matrix. The sample Hessian matrices of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, namely, $H_{\hat{\boldsymbol{\alpha}}}$ and $H_{\hat{\boldsymbol{\beta}}}$, are given by $H_{\hat{\boldsymbol{\alpha}}} = e^{\hat{\phi}} e^{X\hat{\boldsymbol{\alpha}}} XX'$ and $H_{\hat{\boldsymbol{\beta}}} = \hat{\psi} XX'$, respectively. Since the expressions of the standard errors of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ contain the dispersion parameters ϕ and ψ, respectively, they may be estimated using the following formulas:

$$\hat{\phi} = \ln\left\{(m - (p+1)) \sum_{k=1}^{m} \frac{\hat{\mu}_{1k}}{(n_k - \hat{\mu}_{1k})^2}\right\} \tag{19}$$

and

$$\hat{\psi} = \frac{1}{(t - (p+1))} \sum_{k=1}^{t} \left(\frac{\bar{y}_k - \hat{\mu}_{2k}}{\hat{\mu}_{2k}}\right)^2, \tag{20}$$

where $\hat{\mu}_{1k}$ and $\hat{\mu}_{2k}$ are the estimated values of μ_1 and μ_2, respectively, corresponding to the k^{th} observation.

3.2. Dependent Compound Regression Model

Although independence between N and Y is commonly assumed in compound regression models, it is rarely observed in practice. For instance, in the framework of modeling the aggregate claim amounts, it is typical to observe that the claim amounts depend on the claim frequency as well. See, for example, the work of Garrido et al. [17]. As a result, N is included as a covariate in the regression model of $\bar{Y}$. Let θ represent the regression

parameter associated with N. Since S denotes a random sum, it can be written as $S = N\bar{Y}$. The GLM of S through the log-link function is given by Garrido et al. [17] as

$$\mu = e^{X\beta} M'_N(\theta),$$

where $M'_N(\theta)$ represents the derivative of the moment generating function of N with respect to θ. Taking N as CMP, $M'_N(\theta)$ is obtained as

$$M'_N(\theta) = \sum_{n=0}^{\infty} n e^{\theta n} \left(\mu_1 + \frac{e^{\phi}-1}{2e^{\phi}}\right)^{ne^{\phi}} \frac{(n!)^{-e^{\phi}}}{Z(\mu_1,\phi)}.$$

Note that if $\theta = 0$, i.e., when N is independent of $\bar{Y}$, $M'_N(\theta) = E(N)$, and thus the dependent compound regression model will coincide with the independent compound regression model. The pdf of S under dependent case is given by

$$f_S(s) = f_{\bar{Y}|N}(\bar{y}|n) f_N(n),$$

where $f_{\bar{Y}|N}(\bar{y}|n)$ is indicated in Equation (15) with $\mu_2 = \mu_\theta$ and $\psi = \psi_\theta$. The corresponding log-likelihood function is

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi, \psi, \theta) = l(\boldsymbol{\alpha}, \phi; \vec{n}) + l(\boldsymbol{\beta}, \psi, \theta; \vec{\bar{y}}|\vec{n}),$$

where $l(\boldsymbol{\alpha}, \phi; \vec{n})$ corresponds to Equation (14). Let the ML estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and θ be denoted as $\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}$ and $\tilde{\theta}$, where $\tilde{\boldsymbol{\alpha}}$ is obtained using Equation (17). The function $l(\boldsymbol{\beta}, \psi, \theta; \vec{\bar{y}}|\vec{n})$ corresponds to Equation (16) with μ_2 replaced with μ_θ. To obtain the estimates of $\boldsymbol{\beta}$ and θ, the GLM of $E(\bar{Y}|N, X)$ is used with the log-link function and is defined by $\mu_\theta = e^{X\boldsymbol{\beta}+\theta N}$. The corresponding likelihood equations of the regression parameters are

$$\sum_{k=1}^{t} \frac{n_k x_{kl}}{e^{\sum_{l=0}^{p} x_{kl}\beta_l + \theta n_k}} (\bar{y_k} - e^{\sum_{l=0}^{p} x_{kl}\beta_l + \theta n_k}) = 0 \tag{21}$$

and

$$\sum_{k=1}^{t} \frac{n_k^2}{e^{\sum_{l=0}^{p} x_{kl}\beta_l + \theta n_k}} (\bar{y_k} - e^{\sum_{l=0}^{p} x_{kl}\beta_l + \theta n_k}) = 0, \quad l = 0, 1, \ldots, p. \tag{22}$$

The dispersion parameter ψ_θ can be estimated using

$$\hat{\psi}_\theta = \frac{1}{(t-(p+1))} \sum_{k=1}^{t} \left(\frac{\bar{y}_k - \hat{\mu}_{\theta k}}{\hat{\mu}_{\theta k}}\right)^2,$$

where $\hat{\mu}_{\theta k}$ is the estimated value of μ_θ corresponding to the k^{th} observation. In addition, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\theta}$ can be obtained by solving Equations (21) and (22) through iterative algorithms. Thus, the estimate of μ is given by $\tilde{\mu} = e^{X\tilde{\beta}} M'_N(\tilde{\theta})$. Denote $\boldsymbol{\beta}_\theta = \begin{bmatrix} \boldsymbol{\beta} \\ \theta \end{bmatrix}_{(p+2)\times 1}$ and its ML estimate as $\tilde{\boldsymbol{\beta}}_\theta = \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\theta} \end{bmatrix}_{(p+2)\times 1}$. By the asymptotic property of the ML estimators, we have

$$\Sigma_{\boldsymbol{\beta}_\theta}^{1/2} (\tilde{\boldsymbol{\beta}_\theta} - \boldsymbol{\beta}_\theta) \sim \mathcal{N}_{p+2}(\mathbf{0}, I).$$

Here, $\boldsymbol{\beta}_\theta$ and $\Sigma_{\boldsymbol{\beta}_\theta}$ denote the mean vector and covariance matrix of $\tilde{\boldsymbol{\beta}_\theta}$, respectively. The standard error of $\tilde{\boldsymbol{\beta}}_\theta$ corresponds to the square root of the diagonal elements of the sample Hessian matrix, which is given by $H_{\tilde{\boldsymbol{\beta}}_\theta} = \hat{\psi}_\theta X^{*\prime} A X^*$, where X^* is a matrix of order $t \times (p+2)$ that denotes the design matrix which includes $\vec{n}$. A is a $t \times t$ diagonal matrix with positive elements of $\vec{n}$. Note that $H_{\tilde{\boldsymbol{\alpha}}} = H_{\hat{\boldsymbol{\alpha}}}$.

4. Prediction Intervals

From the estimates of the regression parameters, we can obtain an estimate of the expected value of S for some fixed values of the covariates. Given the covariates, it is frequently useful to predict the actual value of S. In a regression setup, the actual value of S is related to its expected value as

$$S = \hat{E}(S|X) + \epsilon,$$

where ϵ is the error term. Since ϵ is unobserved, it is not possible to predict the actual S. In contrast, the prediction interval is a constructed interval that contains the predicted value of actual S. In this section, a method for calculating the PI for S is proposed. Let S_0 denote the response given the covariate $\boldsymbol{x_0} = (1, x_{01}, \ldots, x_{0p})$. Thus, we have $S_0 = \hat{E}(S_0|\boldsymbol{x_0}) + \epsilon$, where $\hat{E}(S_0|\boldsymbol{x_0}) = \exp(\boldsymbol{x_0}\hat{\boldsymbol{\delta}}) = \hat{\mu}_0$ (say). Assuming $E(\epsilon) = 0$, we get, $E(S_0) = \hat{\mu}_0$. Additionally, we have $V(S_0) = V(\hat{\mu}_0) + V(\epsilon)$. Hence, the $100(1-\alpha)\%$ PI for S_0 is given by $[k_1, k_2]$, such that

$$P[k_1 \leq S_0 \leq k_2] = 1 - \alpha, \tag{23}$$

where $\alpha \in (0, 1)$. Here, k_1 and k_2 correspond, respectively, to the lower $\left(\frac{\alpha}{2}\right)^{th}$ and upper $\left(\frac{\alpha}{2}\right)^{th}$ percentiles of the distribution of S_0, which is the compound CMP gamma distribution with mean $E(S_0)$ and variance $V(S_0)$. Since $V(S_0)$ depends on $V(\hat{\mu}_0)$, we proceed as below to obtain an expression for $V(\hat{\mu}_0)$. To begin, consider

$$\hat{\mu}_0 = \exp(\boldsymbol{x_0}\hat{\boldsymbol{\delta}}) \implies \ln(\hat{\mu}_0) = \boldsymbol{x_0}\hat{\boldsymbol{\delta}}. \tag{24}$$

Using the Taylor series expansion of $\ln(A)$ at $E(A)$, we have

$$\ln(A) \approx \ln(E(A)) + (A - E(A))\frac{1}{E(A)}.$$

Thus, we have

$$E(\ln(A)) \approx \ln(E(A)) \tag{25}$$

and

$$V(\ln(A)) \approx \frac{V(A)}{E(A)^2}. \tag{26}$$

Taking A to be $\hat{\mu}_0$ in Equations (25) and (26), we obtain $E(\ln(\hat{\mu}_0)) \approx \ln E(\hat{\mu}_0)$ and $V(\ln(\hat{\mu}_0)) \approx \frac{V(\hat{\mu}_0)}{E(\hat{\mu}_0)^2}$. From Equation (24), we establish that

$$\begin{aligned} E(\ln(\hat{\mu}_0)) &\approx E(\boldsymbol{x_0}\hat{\boldsymbol{\delta}}) = \boldsymbol{x_0}E(\hat{\boldsymbol{\delta}}) \\ \implies E(\hat{\mu}_0) &\approx \exp(\boldsymbol{x_0}E(\hat{\boldsymbol{\delta}})) = \exp(\boldsymbol{x_0}\boldsymbol{\delta}) = \mu_0. \end{aligned}$$

In a similar manner, we obtain

$$\begin{aligned} V(\hat{\mu}_0) &\approx V(\ln(\hat{\mu}_0))E(\hat{\mu}_0)^2 = V(\boldsymbol{x_0}\hat{\boldsymbol{\delta}})\mu_0^2 = \boldsymbol{x_0}V(\hat{\boldsymbol{\delta}})\boldsymbol{x_0'}\mu_0^2 \\ &= \boldsymbol{x_0}\text{diag}(\Sigma_\delta)\boldsymbol{x_0'}\mu_0^2. \end{aligned}$$

An estimate of $V(\epsilon)$, namely, $\hat{V}(\epsilon)$, can be obtained by dividing the residual sum of squares (RSS) of the compound CMP regression model by $m - (p+1)$. Using $V(\hat{\mu}_0)$ and $\hat{V}(\epsilon)$, we obtain $V(S_0)$. However, obtaining the values of k_1 and k_2 from Equation (23) is not easy since the cumulative distribution function of the compound CMP gamma distribution is not invertible. One may use bootstrap procedures to identify k_1 and k_2. We propose

below a heuristic method to obtain the PI using the two-part GLM methodology given in the previous section.

The PI for S_0 is obtained using the PIs of N_0 and $\bar{Y}_0$, where $N_0 = \hat{E}(N_0|\boldsymbol{x_0}) + \epsilon$ and $\bar{Y}_0 = \hat{E}(\bar{Y}_0|\boldsymbol{x_0}) + \epsilon$. Note that $\hat{E}(N_0|\boldsymbol{x_0})$ is obtained from the GLM of N on X and $\hat{E}(\bar{Y}_0|\boldsymbol{x_0})$ is obtained using the GLM of $\bar{Y}$ on X. Denoting $\hat{E}(N_0|\boldsymbol{x_0}) = \hat{\mu}_{01}$ and $\hat{E}(\bar{Y}_0|\boldsymbol{x_0}) = \hat{\mu}_{02}$, we have, $\hat{\mu}_{01} = \exp(\boldsymbol{x_0}\hat{\boldsymbol{\alpha}})$ and $\hat{\mu}_{02} = \exp(\boldsymbol{x_0}\hat{\boldsymbol{\beta}})$. Proceeding along similar lines for obtaining the PI for S_0, the PIs for N_0 and $\bar{Y}_0$ can be obtained, respectively, as $[a_1, a_2]$ and $[b_1, b_2]$, such that

$$P[a_1 \leq N_0 \leq a_2] = 1 - \alpha$$

and

$$P[b_1 \leq \bar{Y}_0 \leq b_2] = 1 - \alpha,$$

where $\alpha \in (0, 1)$. Since N_0 has a mean reparameterized CMP distribution given in Equation (3), a_1 and a_2 are respectively, the lower $\left(\frac{\alpha}{2}\right)^{th}$ and upper $\left(\frac{\alpha}{2}\right)^{th}$ percentiles of the CMP distribution with mean $\hat{\mu}_{01}$ and dispersion parameter $\phi = \dfrac{\hat{\mu}_{01}}{V(\hat{\mu}_{01}) + \hat{V}(\epsilon)}$, where $V(\hat{\mu}_{01}) = \boldsymbol{x_0}\text{diag}(\Sigma_{\boldsymbol{\alpha}})\boldsymbol{x_0'}\mu_{01}^2$. Likewise, b_1 and b_2 correspond respectively, to the lower $\left(\frac{\alpha}{2}\right)^{th}$ and upper $\left(\frac{\alpha}{2}\right)^{th}$ percentiles of the mean reparameterized gamma distribution given in Equation (15) with mean $\hat{\mu}_{02}$ and dispersion parameter $\psi = \dfrac{V(\hat{\mu}_{02}) + \hat{V}(\epsilon)}{\hat{\mu}_{02}^2}$, where $V(\hat{\mu}_{02}) = \boldsymbol{x_0}\text{diag}(\Sigma_{\boldsymbol{\beta}})\boldsymbol{x_0'}\mu_{02}^2$. Supposing $\Sigma_{\boldsymbol{\alpha}}$ and $\Sigma_{\boldsymbol{\beta}}$ are not known, the corresponding sample Hessian matrices can be used to compute $V(\hat{\mu}_{01})$ and $V(\hat{\mu}_{02})$. The values of $\hat{V}(\epsilon)$ of the CMP and gamma regression models can be obtained by dividing the RSS of the corresponding regression models by $m - h$ and $t - h$, where h denotes the number of regression parameters in the model.

The PI for S_0 given $\boldsymbol{x_0}$ can be constructed using the PIs of N_0 and $\bar{Y}_0$. By virtue of equality $S = N\bar{Y}$, a trivial PI for S_0 given $\boldsymbol{x_0}$ can be taken to be $[k_1, k_2] = [a_1 b_1, a_2 b_2]$. When N is large, it may be useful to know the PI for S_0. For example, in modeling aggregate claim amounts from insurance data, the company may want to know the PI for the aggregate claim amount for high claim frequencies so that enough funds can be maintained. In this case, the PI for S_0 given $\boldsymbol{x_0}$ can be defined as $[a_2 b_1, a_2 b_2]$. This definition of PI is used in the remaining part.

5. Numerical Illustration

5.1. Simulation Study

This section provides a numerical illustration of how to compute the PI for S using simulated data for the independent and dependent compound regression models. To generate random samples from the CMP and gamma regression models with a single covariate $\vec{X}_1 = (x_{11}, x_{21}, \ldots, x_{m1})'$, generated from a standard normal distribution, the following steps are implemented:

1. Generate $n_k, k = 1, 2, \ldots, m$, from the CMP distribution given in Equation (3) with mean $\mu_{1k} = \exp(\alpha_0 + \alpha_1 x_{k1})$ by fixing α_0, α_1 and ϕ. Obtain $\vec{n} = (n_1, n_2, \ldots, n_m)'$.
2. For each $n_k > 0$, generate $y_{kj}, j = 1, 2, \ldots, n_k$ from the gamma distribution given in Equation (5) with mean μ_{2k} by fixing ψ, β_0, β_1, and θ, where $\mu_{2k} = \exp(\beta_0 + \beta_1 x_{k1})$ for the independent compound regression model and $\exp(\beta_0 + \beta_1 x_{k1} + \theta n_k)$ for the dependent compound regression model. Compute $\bar{y}_k$ and obtain $\vec{\bar{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)'$.

For simulation, the values of the regression parameters are taken as $\alpha_0 = 0.5, \alpha_1 = 0.3, \beta_0 = 1, \beta_1 = 0.5$ and $\theta = 0.5$. The dispersion parameter ψ of the gamma distribution is set to 1.5. To accommodate over-, equi- and under-dispersion in N, three choices of the dispersion parameter ϕ, namely, $\phi = -1.6, 0$, and 1.6, are considered. The CMP and gamma GLMs are fitted to the generated $\vec{n}$ and $\vec{\bar{y}}$ values, using their respective log-link functions for both the independent and dependent compound regression models.

All the computations are carried out in R (version 4.1.1). The `cmp()` function in `cmpreg` package (Ribeiro Jr [18]) and the `glm()` function are used to carry out the CMP and gamma regression, respectively. To compute the value of $M'_N(\hat{\theta})$ in the dependent compound regression model, the `com.expectation()` function in `compoisson` package is employed. `qcom()` function in the `compoisson` package is used to determine the quantile values from the CMP distribution and the function `qgammaAlt()` in the `EnvStats` package is used to determine quantile values from the gamma distribution. For the above choices of the parameters, the 95% PI for S is obtained for the independent and dependent compound regression models under three choices of sample size (m), namely, $m = 25, 50$ and 100. The actual S observations, denoted by $\vec{s} = (s_1, s_2, \ldots, s_m)'$, are computed by $s_k = n_k \bar{y}_k, k = 1, 2, \ldots, m$.

The proportion of $\vec{s}$ lying within its PI is presented in Table 1 for the various choices of m and ϕ. Additionally, the plots of the corresponding prediction bands are displayed in Tables 2 and 3. From Table 1, it can be observed that, for the choices of the covariate and coefficients considered, the proportion is large for $\phi = 1.6$ in the independent compound regression model and for $\phi = -1.6$ in the dependent compound regression model.

Table 1. Proportion of S lying in its respective PIs.

m	ϕ	Independent Model	Dependent Model
25	−1.6	0.6667	0.9444
	0	0.7777	0.8333
	1.6	0.8400	0.8400
50	−1.6	0.7353	0.8529
	0	0.6500	0.7000
	1.6	0.7656	0.8297
100	−1.6	0.6615	0.9077
	0	0.7088	0.9493
	1.6	0.7777	0.9393

Table 2. Prediction bands of independent compound regression model for over-, equi- and under-dispersed data.

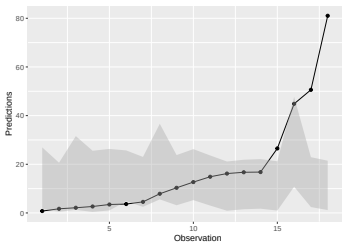
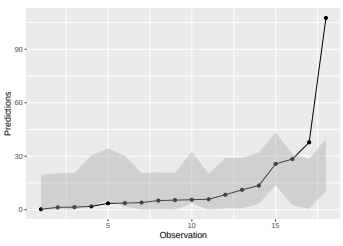
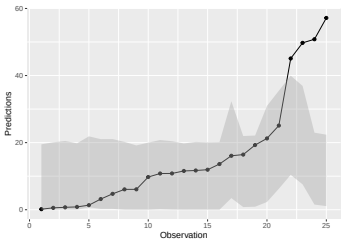
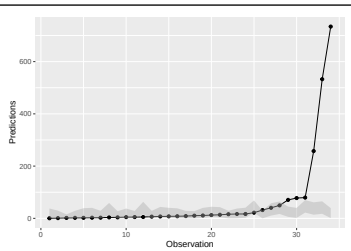
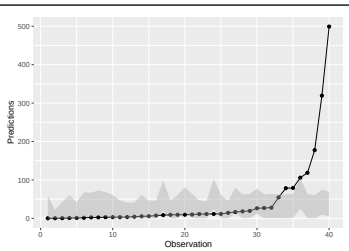
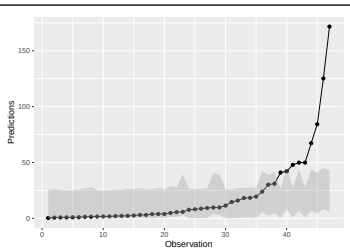
m	$\phi = -1.6$	$\phi = 0$	$\phi = 1.6$
25			
50			

Table 2. *Cont.*

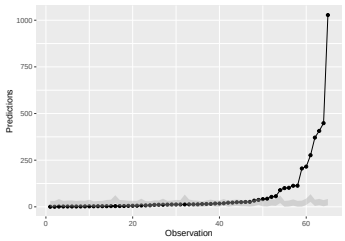
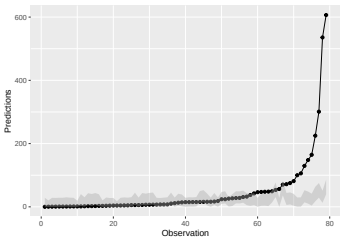
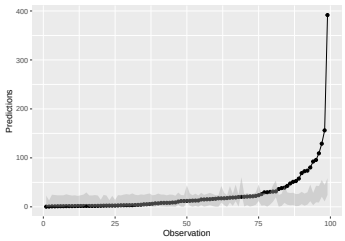
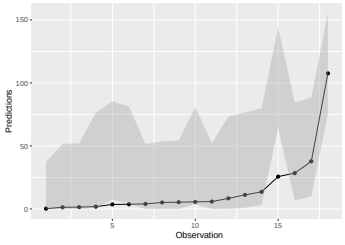
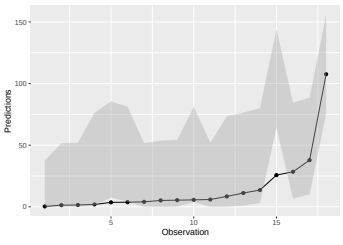
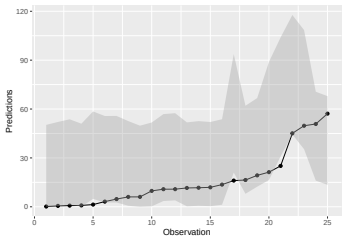
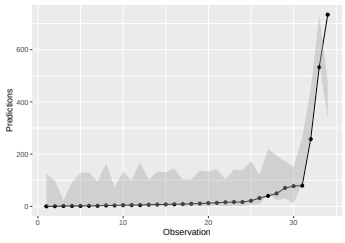
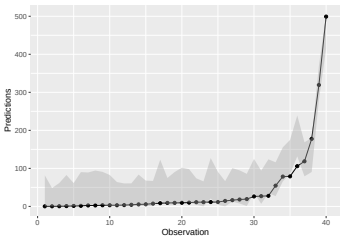
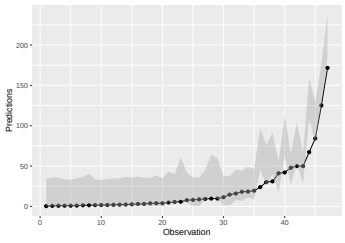
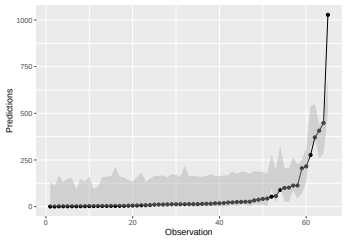
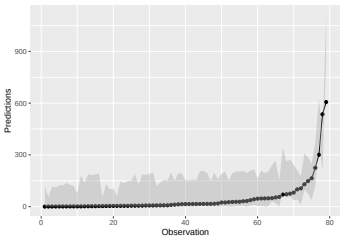
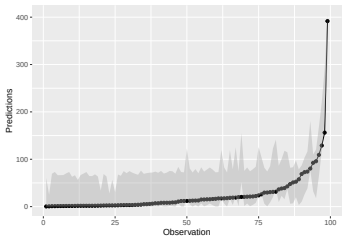
m	$\phi = -1.6$	$\phi = 0$	$\phi = 1.6$
100			

Table 3. Prediction bands of dependent compound regression model for over-, equi- and under-dispersed data.

m	$\phi = -1.6$	$\phi = 0$	$\phi = 1.6$
25			
50			
100			

5.2. Real-Life Application

In this section, the proposed two-part methodology to obtain the PI for the compound CMP gamma regression is applied to real-life vehicle insurance claims data. The dataset pertains to the average damage claims for privately owned and insured vehicles in Britain in the year 1975. See Dutang and Charpentier [19]. It consists of 128 observations on five variables, namely, the owner's age (X_1), car age (X_2), model (X_3), number of claims (N) and average claim amount ($\bar{Y}$) in pounds. The variable X_1 consists of eight categories of age group; the variable X_2, four categories of car age; and the variable X_3, four categories of model. The aggregate claim amount (S) for each observation is obtained by multiplying the average claim amount by the number of claims. A dispersion test on N, performed using the function `dispersiontest()` available in R under `AER` package, resulted in a dispersion index of 119.8246 and a p-value of 2.091×10^{-6}, indicating that N is over-dispersed. Similarly, the Kolmogorov–Smirnov test on $\bar{Y}$ yielded a p-value of 0.7191 to assess the goodness-of-fit of the gamma distribution. As a result, the CMP distribution can be used

to model N, whereas the gamma distribution can be used to model $\bar{Y}$. To implement the proposed estimation methodology and validate its performance, 80% of the observations are randomly chosen as training data and the rest 20% as test data. The observations in the training data are used to fit the independent and dependent compound regression models. The owner's age, car age and car model are the considered covariates in the model. The in-built functions `cmp()` function in `cmpreg` package and the `glm()` function are used to obtain the estimates of CMP and gamma regression models, respectively. The estimates of the regression parameters, their corresponding p-values (in parenthesis) and the AIC values are given in Table 4. Using the AIC values for the CMP and gamma regression models, the combined AIC values for the compound regression models are obtained as 2110.31 and 2108.31, respectively. For each observation in the test data, the PI for S is computed using the estimates of the fitted model. The corresponding prediction band of the independent and dependent compound regression model is displayed in Figure 1. From this figure, it can be noted that some observations do not fall within the prediction band. One reason for this is that these observations have large claim frequencies when compared with the other observations, and the corresponding limits of the PI based on the CMP regression are also large. As a result, the limits of the PI of such observations deviate from their observed values. The proportion of observed S in the test data lying within its PI is found to be 0.4782 and 0.6956 for the independent and dependent compound regression models, respectively. Based on the combined AIC values and the proportions, it can be inferred that the dependent compound regression model provides a relatively better fit for modeling the aggregate claim amount.

Table 4. Parameter estimates, p-values and AIC for the CMP and gamma regression models for the real-life data.

Covariates	CMP Regression Model	Gamma Regression Model (Independent Case)	Gamma Regression Model (Dependent Case)
(Intercept)	1.5007 ($< 2 \times 10^{-16}$)	5.7421 ($< 2 \times 10^{-16}$)	5.7754 ($< 2 \times 10^{-16}$)
OwnerAge21–24	1.5885 ($< 2 \times 10^{-16}$)	−0.2010 (0.0670)	−0.1800 (0.0964)
OwnerAge25–29	2.6237 ($< 2 \times 10^{-16}$)	−0.1129 (0.2705)	−0.0497 (0.6357)
OwnerAge30–34	2.7585 ($< 2 \times 10^{-16}$)	−0.3276 (0.0034)	−0.2542 (0.0262)
OwnerAge35–39	2.8854 ($< 2 \times 10^{-16}$)	−0.3150 (0.0047)	−0.2271 (0.0496)
OwnerAge40–49	3.5362 ($< 2 \times 10^{-16}$)	−0.2722 (0.0081)	−0.1140 (0.3528)
OwnerAge50–59	3.3678 ($< 2 \times 10^{-16}$)	−0.1854 (0.0843)	−0.0590 (0.6219)
OwnerAge60+	3.0280 ($< 2 \times 10^{-16}$)	−0.3054 (0.0036)	−0.2120 (0.0553)
ModelB	1.0255 ($< 2 \times 10^{-16}$)	0.0584 (0.4260)	0.1414 (0.0877)
ModelC	0.6930 ($< 2 \times 10^{-16}$)	0.1083 (0.1387)	0.1500 (0.0450)
ModelD	−0.1889 (0.00485)	0.4041 (6.01×10^{-7})	0.3762 (2.40×10^{-6})
CarAge10+	−1.9174 ($< 2 \times 10^{-16}$)	−0.8138 ($< 2 \times 10^{-16}$)	−0.9494 (5.87×10^{-16})
CarAge4–7	−0.1558 (6.65×10^{-5})	−0.0615 (0.3959)	−0.0727 (0.3089)

Table 4. *Cont.*

Covariates	CMP Regression Model	Gamma Regression Model (Independent Case)	Gamma Regression Model (Dependent Case)
CarAge8–9	-1.4876 ($< 2 \times 10^{-16}$)	-0.4188 (8.64×10^{-8})	-0.5323 (2.02×10^{-8})
NClaims	-	-	-0.0010 (0.0301)
$\hat{\phi}$	-0.8374 ($< 2 \times 10^{-16}$)	-	-
$\hat{\psi}$	-	0.0667	0.0644
AIC	984.7148	1125.6	1123.6

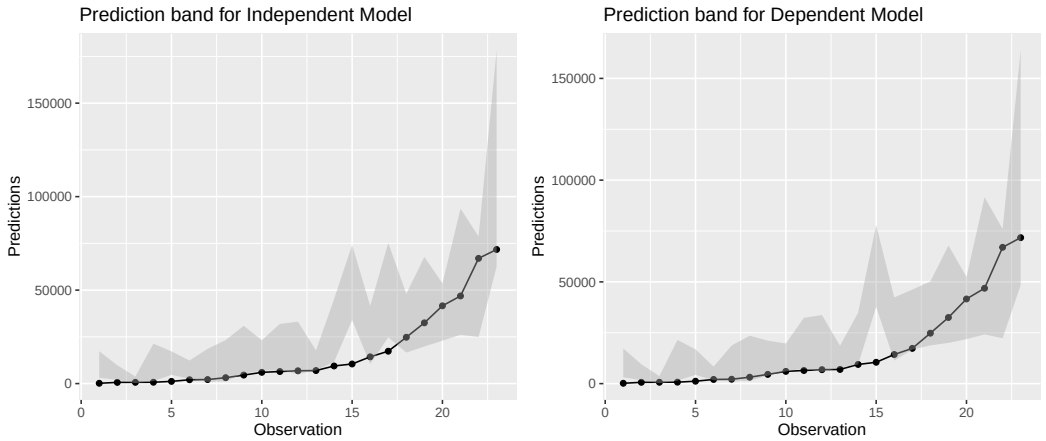


Figure 1. Prediction band for the test data under independent model and dependent model.

6. Conclusions

The Poisson distribution is generally used in compound regression models as the counting distribution. In practice, the Poisson distribution's equi-dispersion assumption is frequently violated. The methodology presented in this paper provided a way to handle non-equi-dispersed count data in the context of compound regression models by using the CMP distribution. The proposed compound regression model can be used when the count data are over- or under-dispersed. The estimation of the parameters was carried out using a two-part GLM approach for the independent and dependent compound regression models. This approach is less complex and provides separate estimates for the count and the continuous distribution involved in the model. Since, in practice, knowledge of the actual value of the response variable rather than its predicted value is more useful, a methodology to obtain the prediction interval of the response variable was proposed. An application of the two-part GLM method to real-life data revealed that the dependent compound regression model performs relatively better than the independent compound regression model. Thus, in practice, one can start with the dependent compound regression model and look for the significance of the count variable in the model. If the count variable is found to be not significant, then the independent compound regression model can be used. To conclude, the proposed compound CMP regression model could be an alternative to modeling a compound random variable when the count data are not equi-dispersed.

Author Contributions: J.M. has contributed to the conceptualization, methodology, mathematical derivation and simulation. V.S.V. and C.C. have contributed equally to mathematical derivation and original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gómez-Déniz, E.; Pérez-Rodríguez, J.V. Modelling distribution of aggregate expenditure on tourism. *Econ. Model.* **2019**, *78*, 293–308. [CrossRef]
2. Klugman, S.A.; Panjer, H.H.; Willmot, G.E. *Loss Models: From Data to Decisions*; John Wiley & Sons: New York, NY, USA, 2012; Volume 715.
3. Bahnemann, D. *Distributions for Actuaries*; Casualty Actuarial Society: Arlington, VA, USA, 2015; Volume 2.
4. Jørgensen, B.; Paes De Souza, M.C. Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.* **1994**, *1994*, 69–93. [CrossRef]
5. Consul, P.C.; Jain, G.C. A generalization of the Poisson distribution. *Technometrics* **1973**, *15*, 791–799. [CrossRef]
6. Shmueli, G.; Minka, T.P.; Kadane, J.B.; Borle, S.; Boatwright, P. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *J. R. Stat. Soc. Ser. (Appl. Stat.)* **2005**, *54*, 127–142. [CrossRef]
7. Conway, R.W.; Maxwell, W.L. A queuing model with state dependent service rates. *J. Ind. Eng.* **1962**, *12*, 132–136.
8. Sellers, K.F.; Borle, S.; Shmueli, G. The COM-Poisson model for count data: A survey of methods and applications. *Appl. Stoch. Model. Bus. Ind.* **2012**, *28*, 104–116. [CrossRef]
9. Sellers, K.F.; Premeaux, B. Conway-Maxwell-Poisson regression models for dispersed count data. *Wiley Interdiscip. Rev. Comput. Stat.* **2021**, *13*, e1533. [CrossRef]
10. Saavithri, V.; Priyadharshini, J.; Banu, Z.P. Compound COM-Poisson Distribution with Binomial Compounding Distribution. Acailable online: https://www.internationaljournalssrg.org/uploads/specialissuepdf/ICRMIT/2018/MTT/ICRMIT-P122.pdf (accessed on 15 January 2023).
11. Frees, E.W.; Gao, J.; Rosenberg, M.A. Predicting the frequency and amount of health care expenditures. *N. Am. Actuar. J.* **2011**, *15*, 377–392. [CrossRef]
12. Andersen, D.A.; Bonat, W.H. Double generalized linear compound Poisson models to insurance claims data. *Electron. J. Appl. Stat. Anal.* **2017**, *10*, 384–407.
13. Delong, Ł.; Lindholm, M.; Wüthrich, M.V. Making Tweedie's compound Poisson model more accessible. *Eur. Actuar. J.* **2021**, *11*, 185–226. [CrossRef]
14. Ribeiro, E.E., Jr.; Zeviani, W.M.; Bonat, W.H.; Demétrio, C.G.; Hinde, J. Reparametrization of COM-Poisson regression models with applications in the analysis of experimental data. *Stat. Model.* **2020**, *20*, 443–466. [CrossRef]
15. Jorgensen, B. *The Theory of Dispersion Models*; CRC Press: Boca Raton, FL, USA, 1997.
16. De Jong, P.; Heller, G.Z. *Generalized Linear Models for Insurance Data*; Cambridge University Press: Cambridge, UK, 2008.
17. Garrido, J.; Genest, C.; Schulz, J. Generalized linear models for dependent frequency and severity of insurance claims. *Insur. Math. Econ.* **2016**, *70*, 205–215. [CrossRef]
18. Ribeiro, E.E., Jr. *Cmpreg: Reparametrized COM-Poisson Regression Models*; R Package Version 0.0.1. Available online: https://rdrr.io/github/JrEduardo/cmpreg/ (accessed on 15 January 2023).
19. Dutang, C.; Charpentier, A. *CASdatasets: Insurance Datasets*; 2019. R Package Version 1.0-11. Available online: http://cas.uqam.ca/ (accessed on 15 January 2023).