Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Article

A Discrete Gamma Model Approach to Flexible Count Regression Analysis: Maximum Likelihood Inference

Chénangnon Frédéric Tovissodé [1] and Romain Glèlè Kakaï [1,*]

[1] Laboratoire de Biomathématiques et d'Estimations Forestières, Université d'Abomey-Calavi, Abomey-Calavi (Benin); romain.glelekakai@fsa.uac.bj (R.G.K)

* Correspondence: chenangnon@gmail.com (C.F.T.)

Abstract: Most existing flexible count regression models allow only approximate inference. Balanced discretization is a simple method to produce a mean-parametrizable flexible count distribution starting from a continuous probability distribution. This makes easy the definition of flexible count regression models allowing exact inference under various types of dispersion (equi-, under- and overdispersion). This study describes maximum likelihood (ML) estimation and inference in count regression based on balanced discrete gamma (BDG) distribution and introduces a likelihood ratio based latent equidispersion (LE) test to identify the parsimonious dispersion model for a particular dataset. A series of Monte Carlo experiments were carried out to assess the performance of ML estimates and the LE test in the BDG regression model, as compared to the popular Conway-Maxwell-Poisson model (CMP). The results show that the two evaluated models recover population effects even under misspecification of dispersion related covariates, with coverage rates of asymptotic 95% confidence interval approaching the nominal level as the sample size increases. The BDG regression approach, nevertheless, outperforms CMP regression in very small samples ($n = 15 - 30$), mostly in overdispersed data. The LE test proves appropriate to detect latent equidispersion, with rejection rates converging to the nominal level as the sample size increases. Two applications on real data are given to illustrate the use of the proposed approach to count regression analysis.

Keywords: Flexible count regression ; balanced discrete gamma distribution ; deviance statistic ; latent equidispersion ; likelihood ratio

1. Introduction

The development of flexible regression models for the analysis of counts has gained in popularity during the last decade [1–7]. Flexible regression models aim to relax the strong equidispersion (equal mean and variance) assumption of the Poisson regression model. Indeed, real count data are often overdispersed (variance larger than mean) or underdispersed (variance smaller than mean). In either case, the inability of a model to account for underdispersion or overdispersion can cause standard errors to be biased downward or upward, thus under or over estimating the statistical significance of associated explanatory variables [8,9].

Most existing approaches to flexible regression analysis bear one or more of the following three limitations. The first is the lack of full dispersion flexibility [10], *i.e.* the ability to unrestrictedly handle both underdispersed and overdispersed count data. For instance, the negative binomial regression [11] is a popular model which can only account for overdispersion, whereas the condensed Poisson regression model [12] can only address underdispersed counts. A second drawback is the violation of probability axioms in that the probability mass function (pmf) of the considered distribution does not sum up to one or is undefined for all or some parameter values. Examples of approaches with this limitation are the Quasi-Poisson regression [13], the Consul's generalized Poisson regression [14] and the extended Poisson-Tweedy [7] regression models for fitting underdispersed counts. A third limitation is the lack of a simple parametrization via the mean, leading to hardly interpretable regression fits [15]. The poisson polynomial regression [16], the

check for updates

Citation: Tovissodé, C.F.; Glèlè Kakaï, R. A Discrete Gamma Model Approach to Flexible Count Regression Analysis: Maximum Likelihood Inference. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: NA.

Conway-Maxwell-Poisson (CMP) regression [1,8], the Gamma-count regression [4] and the discrete Weibull regression [6] are examples of models bearing this drawback.

The failure to provide regression coefficients with a simple interpretation as in Poisson regression is the major limitation to the routine use of many flexible regression approaches, as applied scientists will likely sacrifice any perceived gains in model flexibility for a simpler, more easily interpretable approach [8]. One of the currently most attractive flexible count regression approach is the Mean-parametrized Conway-Maxwell-Poisson (MCMP) regression model [5,17]. Although this regression model is based on the CMP distribution [18] which does not have a simple parametrization via the mean, it tries to circumvent this limitation by directly modelling the mean of the count response using a computationally demanding procedure. Indeed, to compute the log-likelihood contribution of each observed count given the regression coefficients, this approach first replaces the series that defines the mean count by a truncated sum, and solves it for the natural location parameter of the distribution [5]. Then, the normalizing constant of the distribution is also approximated by the truncation of the series that defnes it. The hyper Poisson (HP) regression model [3], which is based on the HP distribution [19], also uses the same procedures to directly model the mean of the count response. Except for very big datasets, the huge computational effort required by these methods might not be a relevant problem in the context of high performance computers nowadays available to applied scientists. However, the involved procedures are proned to numerical problems related to the repeated search (root finding) of the natural location parameter of the distribution. These numerical problems reduce the attractivity of the methods for applied scientists and have yet motivated backup to approximate inference using the CMP distribution [20]. Furthermore, the numerical instability will likely be amplified when extending these regression approaches to the generalized linear mixed model framework [21] to handle multilevel count data.

This work considers an alternative flexible count regression approach based on the balanced discrete gamma (BDG) distribution [15]. Discretization is a method to produce, on purpose, count distributions, starting from continuous distributions. For instance, the discrete Weibull regression model [6] is based on the discrete Weibull distribution [22] which has been obtained from the continuous Weibull distribution. In fact, any count distribution is a discrete version of some continuous distributions [10]. The balanced discretization method [15] offers a simple route to start from a flexible continuous probability distribution and construct a mean-parametrizable count distribution that can be used for flexible count regression. Indeed, the BDG distribution, obtained by discretizing the continuous gamma distribution, is free of the three above listed limitations, namely, it allows to model any type of dispersion (under, equi, and overdispersion); it satisfies all probability axioms (in particular the pmf always sums up to one) and has a simple mean parametrization. Thus, the first advantage of the proposed BDG regression model over other flexible approaches (such as the HP and MCMP models) is its ability to directly model the mean of the count response without ressorting to approximations, thus reducing the related computational issues. A second advantage is the ability to easily simulate count outcomes [15], letting room for simulation and parametric bootstrapping-based inference in count models.

In this paper, we describe the BDG count regression and maximum likelihood inference in the model (Section 2). We have carried out Monte Carlo experiments to compare the finite sample performances of the BDG regression and the MCMP regression [5] approaches (Section 3). These two competing regression models are also compared based on two applications to real data (Section 4). Concluding remarks are given in Section 5.

2. Balanced Discrete Gamma Regression

We introduce the BDG regression model and describe the maximum likelihood estimation of the model parameters. We also introduce a latent equidispersion (LE) test to check the need of dispersion parameters and a deviance test to assess goodness-of-fit.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

2.1. The Balanced Discrete Gamma Distribution

A random variable Y follows a balanced discrete gamma (BDG) distribution with parameters $\mu \in \mathbb{R}_+$ and $a \in \mathbb{R}_+$ (with $\mathbb{R}_+ = (0, \infty)$), if the support of Y is $\mathbb{N} = \{0, 1, \cdots\}$ and its pmf (probability mass function) is given for any $y \in \mathbb{N}$ by [15]

$$\begin{aligned} f(y|\mu, a) &= (y-1)\gamma(a(y-1), b) - 2y\gamma(ay, b) + (y+1)\gamma(a(y+1), b) \\ &\quad - \mu[\gamma(a(y-1), b+1) - 2\gamma(ay, b+1) + \gamma(a(y+1), b+1)] \end{aligned} \tag{1}$$

where $b = a\mu$ and $\gamma(x, b) = \int_0^x u^{b-1} \exp(-u) du/\Gamma(b)$ is the lower incomplete gamma ratio with $\gamma(x, a) = 0$ if $x \notin \mathbb{R}_+$. We denote $\mathcal{BG}(\mu, a)$ a BDG distribution with parameters μ and a. A BDG variable $Y \sim \mathcal{BG}(\mu, a)$ has a simple representation in terms of a gamma variable as: $Y \overset{d}{=} \lfloor X \rfloor + U$ where X follows a continuous gamma distribution with mean μ and variance μ/a, $\lfloor x \rfloor$ denotes the integer part of x, U follows a Bernoulli distribution with success probability $(x - \lfloor x \rfloor)$ given that $X = x$, and $\overset{d}{=}$ means "equal in distribution". This representation as a simple probabilistic rounding of a continuous gamma variable is useful for generating random deviates from a BDG distribution. Interestingly, if $Y \sim \mathcal{BG}(\mu, a)$, then Y has expectation μ. The variance of Y is given by

$$\begin{aligned} \sigma^2 &= a^{-1}\mu + \zeta_0(\mu, a) \quad \text{with} \quad \zeta_0(\mu, a) = \sum_{z=0}^{\infty} \varrho(z, \mu, a), \\ \varrho(z, \mu, a) &= -\mu\left(\mu + a^{-1}\right)[\gamma(a(z+1), b+2) - \gamma(az, b+2)] \\ &\quad \mu(2z+1)[\gamma(a(z+1), b+1) - \gamma(az, b+1)] \\ &\quad -z(z+1)[\gamma(a(z+1), b) - \gamma(az, b)]. \end{aligned} \tag{2}$$

The term $\zeta_0(\mu, a)$ satisfies $0 < \zeta_0(\mu, a) < min\{\mu, 1/4\}$ and is well approximated by $\widehat{\zeta}_\alpha(\mu, a) = \sum_{z=z_i}^{z_f} \varrho(z, \mu, a)$ where z_i and z_f are the integer parts of respectively the $\alpha/2$ and $1 - \alpha/2$ quantiles of the continuous gamma distribution with mean μ and variance μ/a, and $\alpha \in (0, 1)$ is a tolerance value which controls the accuracy of $\widehat{\zeta}_\alpha(\mu, a)$ via $|\widehat{\zeta}_\alpha(\mu, a) - \zeta_0(\mu, a)| < 1 - \gamma\big(a(z_f + 1), b\big) + \gamma(az_i, b)$ [15].

From (2), it is seen that a is a scale parameter. When $a = 1$, the BDG gamma distribution corresponds to latent equidispersion (LE), *i.e.* the underlying continuous gamma distribution has equal mean and variance. Although this one-parameter distribution is slightly overdispersed, it turns to be very close to the Poisson distribution [15] and can thus serve for assessing the need of a flexible model in count regression analysis.

2.2. The Balanced Discrete Gamma Regression Model

Let us consider a population in which, an individual count response Y_i has mean value μ_i and a variance σ_i^2 depending on μ_i and a positive scalar a_i. The mean count μ_i is log-linearly related to a set of covariates $\boldsymbol{X}_i = (1, x_{1i}, x_{2i}, \cdots, x_{pi})^\top$ as $\mu_i = \exp\big(\boldsymbol{X}_i^\top \boldsymbol{\beta}\big)$ where the $p+1$ vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ contains regression coefficients, $\beta_j \in \mathbb{R}$ representing the marginal effect of the covariate X_j on the log-average count, $\log \mu_i$, so that a unit change in X_{ij} induces *ceteris paribus* a $\beta_j\%$ variation in the mean count μ_i as in Poisson regression. In addition, a_i has the log-linear form $a_i = \exp\big(\boldsymbol{Z}_i^\top \boldsymbol{\delta}\big)$ where $\boldsymbol{Z}_i = (1, z_{1i}, z_{2i}, \cdots, z_{qi})^\top$ and the $q+1$ vector $\boldsymbol{\delta} = (\delta_0, \delta_1, \cdots, \delta_q)^\top$, with $\delta_j \in \mathbb{R}$, controls the dispersion of Y_i via a_i. The set of dispersion covariates $\boldsymbol{Z}_i$ can be $\boldsymbol{Z}_i = 1$, *i.e.* $q = 0$, (we shall refer to this specification as "constant-dispersion"); a subset of $\boldsymbol{X}_i$ ("mean-related-dispersion", *e.g.* $\boldsymbol{Z}_i = \boldsymbol{X}_i$) or independent of $\boldsymbol{X}_i$ ("mean-free-dispersion").

For such a population, a balanced discrete gamma regression model is specified via

$$Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i \overset{ind}{\sim} \mathcal{BG}(\mu_i, a_i) \tag{3}$$

where $\overset{ind}{\sim}$ stands for "independently distributed as". The variance σ_i^2 of Y_i then depends on μ_i and a_i through (2). For a vector $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\top$ of n independant samples from the regression model (3), let $\mathbf{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)^\top$ and $\mathbf{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n)^\top$ denote the associated design matrices. Each of the matrices $\mathbf{X}$ and $\mathbf{Z}$ is assumed to be of full column rank. The design matrix $\mathbf{X}$ is directly available from the experimental or observational design of a study. On the contrary, a dispersion design matrix $\mathbf{Z}$ is not generally available. The researcher is thus required to decide which explanatory variables in $\mathbf{X}$ may affect the dispersion of the response variable Y_i. One can in practice start with $\mathbf{Z} = \mathbf{X}$ and use a model selection approach to drop some covariates, based on the significance of the associated dispersion parameter δ_j. When applicable, one may also try the transforms of some explanatory variables (*e.g.* quadratic term, interaction, square root, logarithm) in the selection of the dispersion covariates.

2.3. Estimating Balanced Discrete Gamma Regression Parameters

Given a sample vector $\boldsymbol{y}$ and design matrices $\mathbf{X}$ and $\mathbf{Z}$, the log-likelihood of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$ of the BDG regression model (3) is given by

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} \log f(y_i|\mu_i, a_i). \tag{4}$$

When $p = 0$ and $q = 0$ (both $\mathbf{X}_i$ and $\mathbf{Z}_i$ are scalar valued, and constant across all observations), $\boldsymbol{y}$ is an independent and identically distributed sample from a BGD variable $Y \sim \mathcal{BG}(\mu, a)$ where $\mu = \exp(\beta_0)$ and $a = \exp(\delta_0)$. This special case corresponds to BDG distribution fitting and is discussed in Appendix A. In the general BDG regression setup (3), the ML estimate $\widehat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$ can be obtained by using a numerical optimization routine to maximize the log-likelihood function (4). To this end, there is a need of an initial parameter value $\boldsymbol{\theta}^{(0)}$. We propose to tackle the initialization of $\boldsymbol{\beta}$ as in the generalized linear model (GLM) framework with the continuous gamma family [24]. Indeed, let us define the pseudo responses $\tilde{y}_i^{(0)} = \log(y_i + 0.1) + y_i/(y_i + 0.1) - 1$ and build the vector $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_n)^\top$. Solving the linear system

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}^{(0)} = \mathbf{X}^\top \tilde{\boldsymbol{y}} \tag{5}$$

for $\boldsymbol{\beta}^{(0)}$ provides an initial value for $\boldsymbol{\beta}$. To find an initial estimate of $\boldsymbol{\delta}$, we first approximate the variance of Y_i as $\sigma_i^{2(0)} = \left(y_i - \mu_i^{(0)}\right)^2$ with $\mu_i^{(0)} = \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}^{(0)}\right)$. We also consider the approximate variance expression $\sigma_i^{2(0)} = \mu_i^{(0)}/a_i^{(0)} + \zeta_{0i}^{(0)}$ where $\zeta_{0i}^{(0)} = min\left\{\mu_i^{(0)}/2, 1/8\right\}$ if $\sigma_i^{2(0)} > min\left\{\mu_i^{(0)}, 1/4\right\}$, *i.e.* ζ_{0i} is taken as the midpoint of its range $\left(0, min\left\{\mu_i^{(0)}, 1/4\right\}\right)$ [15]; and $\zeta_{0i}^{(0)} = \sigma_i^{2(0)}/2$ otherwise. If the dispersion covariate is constant across observations ($\boldsymbol{Z}_i = c$ with $c \in \mathbb{R}$), then $a_i^{(0)} = a^{(0)}$, $a^{(0)} = \sum_{i=1}^n \mu_i^{(0)} / \sum_{i=1}^n \left(\sigma_i^{2(0)} - \zeta_{0i}^{(0)}\right)$ and $\boldsymbol{\delta}^{(0)}$ is initialized to $\boldsymbol{\delta}^{(0)} = (\log a^{(0)})/c$. Otherwise, we compute the approximate scales $a_i^{(0)} = \mu_i^{(0)} / \left(\sigma_i^{2(0)} - \zeta_{0i}^{(0)}\right)$, define the pseudo scales $\tilde{a}_i^{(0)} = \log(a_i^{(0)} + 0.1) + a_i^{(0)}/(a_i^{(0)} + 0.1) - 1$ and build the vector $\tilde{\boldsymbol{a}}^{(0)} = (\tilde{a}_1^{(0)}, \tilde{a}_2^{(0)}, \cdots, \tilde{a}_n^{(0)})^\top$. Solving the linear system

$$\mathbf{Z}^\top \mathbf{Z} \boldsymbol{\delta}^{(0)} = \mathbf{Z}^\top \tilde{\boldsymbol{a}}^{(0)} \tag{6}$$

for $\boldsymbol{\delta}^{(0)}$ then provides an initial value for $\boldsymbol{\delta}$. The starting value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)\top}, \boldsymbol{\delta}^{(0)\top})^\top$ can then be supplied to a numerical optimization routine (*e.g.* the *optim* function in R freeware [23]) to maximize the log-likelihood function (4). Interestingly, most of optimization routines available in common statistical softwares compute, on request, the hessian of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

log-likelihood function, *i.e.* the negative observed information matrix $-\boldsymbol{I}_{obs}(\boldsymbol{\theta})$, evaluated at the ML estimate $\widehat{\boldsymbol{\theta}}$. This provides an approximate covariance matrix for the maximum likelihood estimator: $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = [\boldsymbol{I}_{obs}(\boldsymbol{\theta})]^{-1}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. In other words, after the optimization step, $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ is an approximate covariance matrix for $\widehat{\boldsymbol{\theta}}$. Then, statistical hypotheses of the form $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ can be tested assuming asymptotic normality for $\boldsymbol{\theta}$ with mean vector $\widehat{\boldsymbol{\theta}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}$.

2.4. Testing for Latent Equidispersion

In a parsimony perspective, it is important to check the usefulness of a two or more parameter model against a one–parameter model. We consider, to this end, a latent equidispersion (LE) test to assess the need of a dispersion parameter $\boldsymbol{\delta}$ in the balanced discrete gamma regression model. The null hypothesis corresponding to LE is H_0: $\boldsymbol{\delta} = \mathbf{0}$ (equidispersion), and the two-sided alternative is H_1: $\boldsymbol{\delta} \neq \mathbf{0}$. To test H_0, we consider the likelihood ratio (LR) statistic [25]

$$LR = 2\Big[\ell\Big(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}, \mathbf{X}, \mathbf{Z}\Big) - \ell\Big(\widetilde{\boldsymbol{\theta}}|\boldsymbol{y}, \mathbf{X}, \mathbf{0}\Big)\Big]. \tag{7}$$

where $\widehat{\boldsymbol{\theta}}$ is the ML estimate and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}^\top, \mathbf{0})^\top$ with $\widetilde{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ under H_0. Thanks to the regularity conditions indicated in Appendix A, if H_0 is true, then the statistic LR converges in distribution to the $\chi^2_{(k)}$ law with $k = q + 1$ degrees of freedom as $n \to \infty$ [26].

2.5. Diagnosing a Balanced Discrete Gamma Regression Model Fit

To assess the goodness-of-fit of a balanced discrete gamma regression, we consider the deviance statistic proposed for general count regression models by [27]. The deviance is defined for the regression covariates $\mathbf{X}$ conditional on the ML estimates of the dispersion parameters $a_i = \exp\big(\mathbf{Z}_i^\top \boldsymbol{\delta}\big)$. Specifically, a *saturated model* fit is defined as the model fit with a separate intercept for each observation, estimated with the restriction $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}$ (*i.e.* the ML estimate of $\boldsymbol{\delta}$ under the regression model (3)). The log-likelihood ℓ_{s} of the saturated model fit is thus given by $\ell_{\mathrm{s}} = \ell\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{s}}|\boldsymbol{y}, \mathbf{I}_n, \mathbf{Z}\Big)$ where $\widehat{\boldsymbol{\theta}}_{\mathrm{s}} = \Big(\widehat{\boldsymbol{\beta}}_{\mathrm{s}}^\top, \widehat{\boldsymbol{\delta}}^\top\Big)^\top$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{s}}$ is the ML estimate of $\boldsymbol{\beta}$ using $\mathbf{X} = \mathbf{I}_n$ (the $n \times n$ identity matrix) with the constraint $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}$. The *deviance* of the ML fit of the model (3) is the LR statistic given by

$$D \;=\; 2\Big[\ell\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{s}}|\boldsymbol{y}, \mathbf{I}_n, \mathbf{Z}\Big) - \ell\Big(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}, \mathbf{X}, \mathbf{Z}\Big)\Big]. \tag{8}$$

For goodness-of-fit evaluation, the statistic D has an asymptotic $\chi^2_{(k)}$ distribution with $k = n - p - 1$ degrees of freedom, if the assumed balanced discrete gamma regression model is consistent with the observed data $\boldsymbol{y}$ (*i.e.* no important covariate is missing in $\mathbf{X}$).

The deviance statistic (8) actually expresses the information (as measured by Kullback–Leibler divergence) recovered through the use of the covariates $\mathbf{X}$. The overall recoverable information by inclusion of regressors is obtained considering an intercept-only mean model, *i.e.* using $\mathbf{X} = \mathbf{J}_n$ with $\mathbf{J}_n$ the n-vector of all ones. The log-likelihood of the resulting *null model* fit is given by $\ell_{\mathrm{n}} = \ell\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{n}}|\boldsymbol{y}, \mathbf{J}_n, \mathbf{Z}\Big)$ where $\widehat{\boldsymbol{\theta}}_{\mathrm{n}} = \Big(\widehat{\beta}_{\mathrm{n}}, \widehat{\boldsymbol{\delta}}^\top\Big)^\top$ and $\widehat{\beta}_{\mathrm{n}}$ is the ML estimate of the model intercept β_0 using $\mathbf{X} = \mathbf{J}_n$ under the constraint $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}$. The *null deviance* (potentially recoverable information) is then given by

$$D_{\mathrm{n}} \;=\; 2\Big[\ell\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{s}}|\boldsymbol{y}, \mathbf{I}_n, \mathbf{Z}\Big) - \ell\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{n}}|\boldsymbol{y}, \mathbf{J}_n, \mathbf{Z}\Big)\Big]. \tag{9}$$

The *adjusted deviance* R^2 (deviance reduction ratio) given by [27]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

$$R^2_{dev} = 1 - \frac{D}{D_{\rm n}} \frac{n-1}{n-p-1} \tag{10}$$

expresses the percentage of information explained by the ML fit based on $\mathbf{X}$, adjusted for the number p of predictors in $\mathbf{X}$. Note that $R^2_{dev} = 0$ if $p = 0$. When $p > 0$, the overall significance of the ML fit can further be tested considering the LR statistic

$$G = D_{\rm n} - D. \tag{11}$$

If the deviance reduction by $\mathbf{X}$ is not significant (*i.e.* there is no important covariate in $\mathbf{X}$), then the statistic G has an asymptotic $\chi^2_{(k)}$ distribution with $k = p$ degrees of freedom.

3. Performance Studies

We carried out three simulation experiments. The first experiment assesses the empirical distribution of the likelihood ratio (LR) statistic (7) to test latent equidispersion (LE). The second experiment evaluates the power of the LE test in both balanced discrete gamma (BDG) and Conway-Maxwell-Poisson (CMP) samples. The last experiment compares the abilities of the BDG and the Mean-parametrized CMP (MCMP) [5] regression models to recover population effects of covariates and their relative robustness to the misspecification of dispersion covariates. We consider the following data models:

$$\text{M1:} \qquad Y_i \overset{ind}{\sim} \mathcal{BG}(\mu_i, a_i), \tag{12}$$

$$\text{M2:} \qquad Y_i \overset{ind}{\sim} \mathcal{CMP}_{\mu}(\mu_i, a_i), \tag{13}$$

where, $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, $a_i = \exp(\delta_0 + \delta_1 z_i)$, and $\mathcal{CMP}_{\mu}(\mu_i, a_i)$ denotes a CMP distribution with mean μ_i and dispersion parameter a_i [5]. The model parameter is denoted $\boldsymbol{\theta} = (\beta_0, \beta_1, \delta_0, \delta_1)^\top$. The population effects were arbitrarily set to $\beta_0 = \beta_1 = 1$. The values of δ_0 and δ_1 vary between simulation settings. The covariates x_i and z_i are independent, each following a uniform distribution in the interval $(-1, 1)$.

All computations were performed in R freeware [23] on a Windows plateform Intel(R) Core(TM) i7-7500U CPU running at 2.9 GHz and 12 GB of RAM. BDG regression models were fitted as described in Section 2 and MCMP regression models were fitted using the routine *glm.cmp* of the package *mpcmp* [28].

3.1. Empirical Distribution of the Likelihood Ratio Statistic

This experiment aims to compare, in finite sample ($n = 15, 30, 100, 500$), the empirical distribution of the likelihood ratio (LR) statistic (7) for latent equidispersion (LE) to the theoretical χ^2_k distribution when the data is generated from the BDG regression model. We thus consider the data model M1 in (12) with $a_i = 1$. For each sample size n, the covariates x_i and z_i $(i = 1, 2, \cdots, n)$ were generated once and kept fixed through the following steps:

- generate $B = 1000$ samples $\boldsymbol{y}^{(r)}$ $(r = 1, 2, \cdots, B)$ from the target data model under the assumption $H_0 \colon \delta_0 = \delta_1 = 0$ (*i.e.* using $\boldsymbol{\theta} = (\beta_0, \beta_1, 0, 0)^\top$ so that $a_i = 1$);
- for each sample $\boldsymbol{y}^{(r)}$, find the ML estimates

  a.) $\tilde{\boldsymbol{\theta}}^{(r)} = \left(\tilde{\beta}_0^{(r)}, \tilde{\beta}_1^{(r)}, 0, 0\right)^\top$ under H_0;

  b.) $\widehat{\boldsymbol{\theta}}^{(r,c)} = \left(\widehat{\beta}_0^{(r,c)}, \widehat{\beta}_1^{(r,c)}, \widehat{\delta}_0^{(r,c)}, 0\right)^\top$ under constant-dispersion specfication $H_{\rm c}$: $\delta_1 = 0$, *i.e.* using $\mathbf{Z} = \mathbf{J}_n$ (a column of n ones);

  c.) $\widehat{\boldsymbol{\theta}}^{(r,z)} = \left(\widehat{\beta}_0^{(r,z)}, \widehat{\beta}_1^{(r,z)}, \widehat{\delta}_0^{(r,z)}, \widehat{\delta}_1^{(r,z)}\right)^\top$ under mean-free-dispersion specfication $H_{\rm z}$ (*i.e.* using $\mathbf{Z}_i = (1, z_i)^\top$);

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

d.) $\widehat{\boldsymbol{\theta}}^{(r,x)} = \left(\widehat{\beta}_0^{(r,x)}, \widehat{\beta}_1^{(r,x)}, \widehat{\delta}_0^{(r,x)}, \widehat{\delta}_1^{(r,x)}\right)^\top$ under mean-related-dispersion specfication H_{x} (*i.e.* using $\mathbf{Z} = \mathbf{X}$);

- compute the LR statistics $LR^{(r,c)} = 2\left[\ell\left(\widehat{\boldsymbol{\theta}}^{(r,c)}|\boldsymbol{y}^{(r)}, \mathbf{X}, \mathbf{J}_n\right) - \ell\left(\widetilde{\boldsymbol{\theta}}^{(r)}|\boldsymbol{y}^{(r)}, \mathbf{X}, \mathbf{0}\right)\right]$ related to step b.), $LR^{(r,\mathrm{z})} = 2\left[\ell\left(\widehat{\boldsymbol{\theta}}^{(r,z)}|\boldsymbol{y}^{(r)}, \mathbf{X}, \mathbf{Z}\right) - \ell\left(\widetilde{\boldsymbol{\theta}}^{(r)}|\boldsymbol{y}^{(r)}, \mathbf{X}, \mathbf{0}\right)\right]$ related step c.) and $LR^{(r,x)} = 2\left[\ell\left(\widehat{\boldsymbol{\theta}}^{(r,x)}|\boldsymbol{y}^{(r)}, \mathbf{X}, \mathbf{X}\right) - \ell\left(\widetilde{\boldsymbol{\theta}}^{(r)}|\boldsymbol{y}^{(r)}, \mathbf{X}, \mathbf{0}\right)\right]$ related to step d.).

Figure 1 compares the empirical distribution function of the LR statistic for constant-dispersion ($k = 1$), mean-free-dispersion ($k = 2$) and mean-related-dispersion fits ($k = 2$) with the corresponding asymptotic χ^2_k distributions for data generated from the BDG regression model under latent equidispersion. It can be seen that there are important deviations from the theoretical distribution in small samples ($n = 15$). However, as the sample size n increases (30–500), the empirical cdf approaches the theoretical cdf, irrespective of the misspecification or not of some dispersion covariates in the regression model.

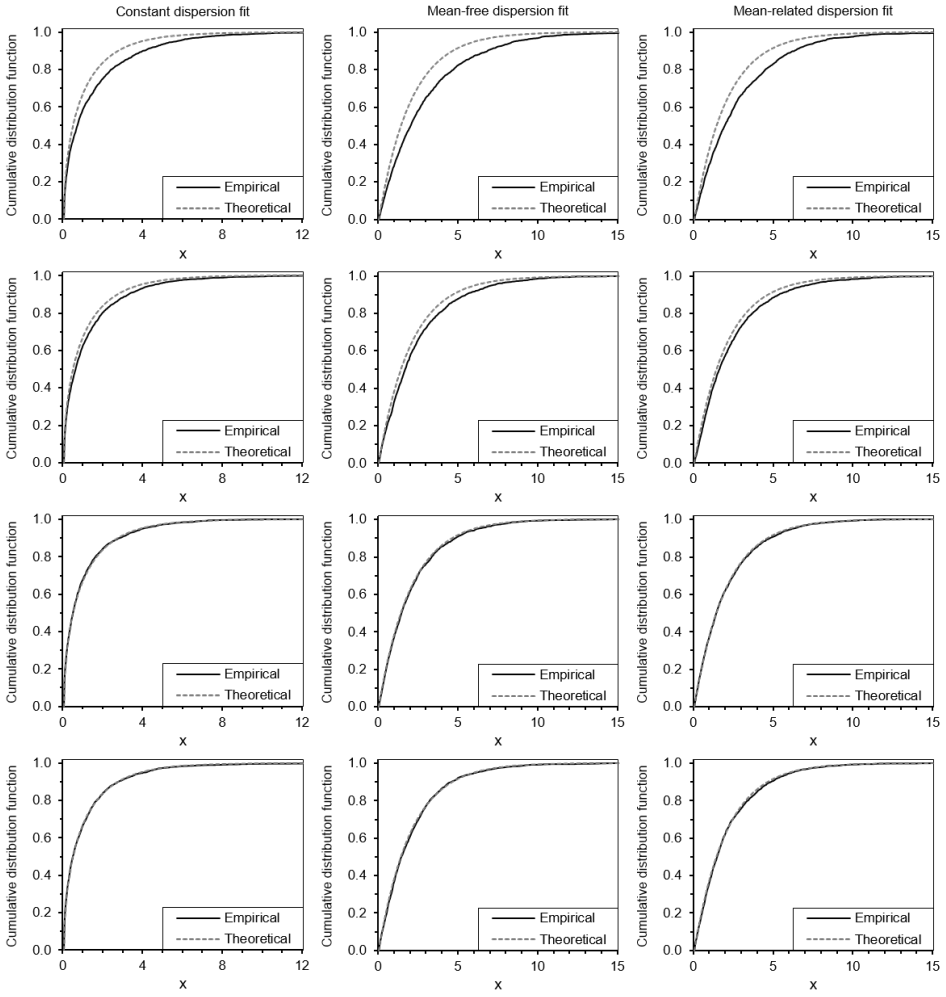

Figure 1. Distribution function of the likelihood ratio statistic obtained from 1000 Monte Carlo repetitions (Empirical) and χ^2_k distribution function (Theoretical) for latent equidispersion test in balanced discrete gamma regression with constant-dispersion (left panel, $k = 1$), mean-free-dispersion (central panel, $k = 2$) and mean-related-dispersion (right panel, $k = 2$) using balanced discrete gamma samples of sizes of $n = 15$ (first row) $n = 30$ (second row), $n = 100$ (third row) and $n = 500$ (last row).

3.2. Power of the Likelihood Ratio Test for Equidispersion

This experiment aims to evaluate the power of the LR test in BDG regression analysis. We consider BDG data from model M1 (12) and CMP data from model M2 (13). Here, the dispersion parameters have values $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$ with sample sizes $n = 15, 30, 50, 100, 250, 500$. For each combination of data model, sample size n, and dispersion parameters δ_0 and δ_1, the BDG regression model was fitted first under LE restriction (H_0) and then under constant (H_c), mean-free (H_z) and mean-related (H_x) dispersion specifications as described in steps a.) – d.) in 3.1. For each specification, the rejection rate of the hypothesis H_0 at the nominal level $\alpha = 5\%$ was computed as the proportion of the $B = 1000$ replicates $LR^{(r)}$ which exceeded the upper $100(1-\alpha)\%$ quantile of $\chi^2_{(k)}$, with $k = 1$ under H_c and $k = 2$ under H_z or H_x. From this experiment, we expect rejection rates close to 0.05 when H_0 is true (BDG data with $a_i = 1$) or approximately true (Poisson data, *i.e.* CMP data with $a_i = 1$) and rejection rates close to 0.95 as δ_0 or δ_1 differs from zero ($a_i \neq 1$).

Figures 2 and 3 present the rejection rates of the LE test as a function of the sample size n. For BDG data with $a_i = 1$ (H_0 is true), the rejection rate of the LE test approaches the nominal level $\alpha = 5\%$ (Figure 2 A and C, $\delta_1 = 0$). Indeed, for small samples ($n = 15$), the rejection rates varies from 7.54% (observed for mean-related-dispersion fits) to 10.69% (observed for mean-free-dispersion fits). For larger samples ($n = 30 - 500$), the rejection rate ranges from 5.26% to 6.75%. Similar results were obtained for Poisson data, with a rejection rate ranging from 3.80% to 11.87% (see Figure A1 in Appendix B.1). However, it appears that the rejection rate under Poisson data does not approach 5%, but shows an increasing trend for $n \in [100, 500]$ (Figure A1). In other words, for sufficiently large Poisson samples ($n >> 500$), the LE test will be more powerful than desired.

For BDG data with non null δ_0 or δ_1 (LE does not hold), the rejection rate of the LE test increases with n and tends to one when there is no important missing dispersion covariate in the fitted model (Figure 2). Similar trends were observed for CMP data with non null δ_0 or δ_1 (Figure 3). However, despite the similarity of the trends, when $\delta_1 \neq 1$, there are important differences between the rejection rates under BDG and CMP data for $n \leq 100$. The differences (between the rate for CMP data and the rate for BDG data) mostly occur when $\delta_1 = -0.5$ and amounts on average to -12.65%, ranging from -17.69% to -7.45%.

Under the scenarios of misspecification of dispersion covariates (fitting a constant-dispersion or a mean-related-dispersion model to BDG data with $\delta_1 \neq 0$), the rejection rate also increases with n, but at a much lower rate, which reaches 33.70% for constant-dispersion fits and 30.70% for mean-related-dispersion fits at $n = 500$ (Figures 2 A and C, $\delta_1 = 1$). Similar trends were observed for CMP data (Figure 3 A and C, $\delta_1 = 1$), the rejection rate reaching 21.90% for constant-dispersion fits and 32.80% for mean-related-dispersion fits at $n = 500$.

3.3. Properties of Maximum Likelihood Estimates

This experiment aims to compare the abilities of the BDG regression and the MCMP regression models to recover population effects of covariates. We consider the data models M1 (12) and M2 (13), dispersion parameter values $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$, and sample sizes $n = 15, 30, 50, 100, 250, 500$. For each combination of data model, sample size n, and dispersion parameters δ_0 and δ_1, the BDG and MCMP regression models were fitted under constant-dispersion (H_c) and mean-free-dispersion (H_z) specifications. Here, the ML estimates $\hat{\theta}_j^{(r)}$ of model parameters and their asymptotic 95% confidence intervals $CI_j^{(r)}$ were recorded from each model fit. The performance of the ML estimators of the population effects is assessed, for each β_j ($j = 0, 1$), by computing the relative bias (%) in the estimate $\hat{\beta}_j$ of β_j, the root mean square error (RMSE) of $\hat{\beta}_j$; and the coverage rate (CR) of the asymptotic confidence interval CI_j, *i.e.* the percentage of the $B = 1000$ confidence interval $CI_j^{(r)}$ which contain β_j (Table 1). For the dispersion parameters δ_j, the average

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

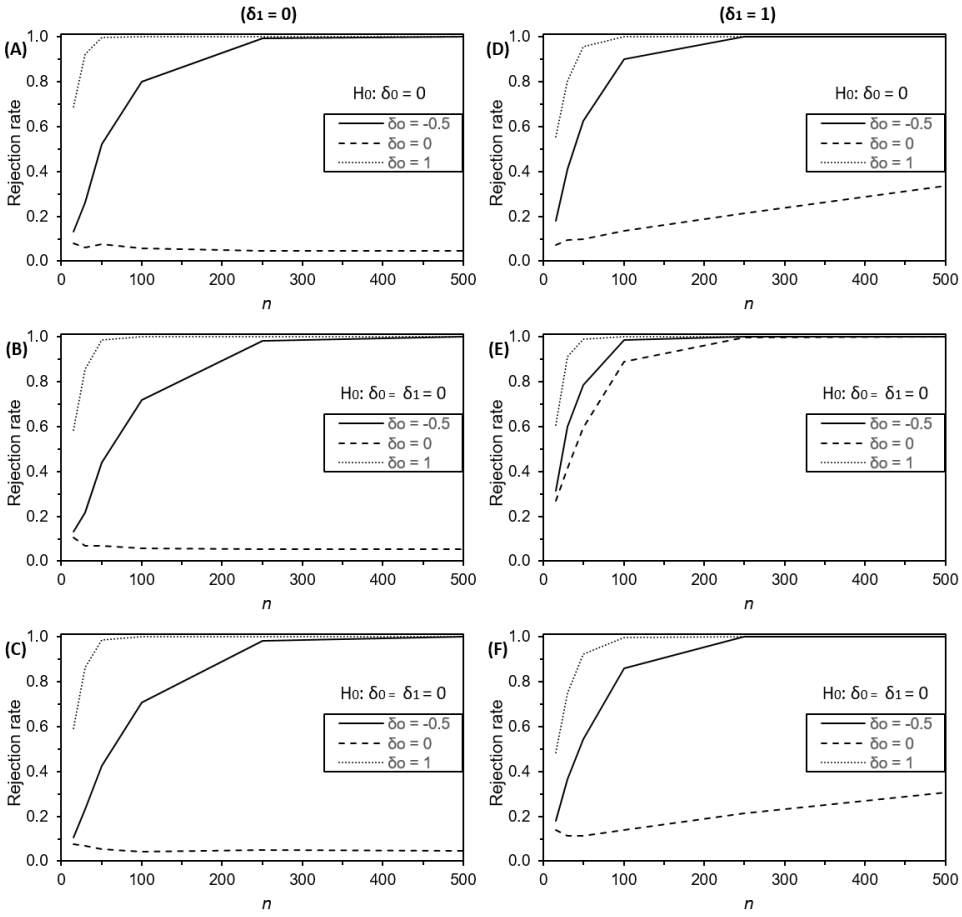

Figure 2. Power of the likelihood ratio test to detect equidispersion at the nominal level $\alpha = 5\%$ for balanced discrete gamma regression fits under constant-dispersion (first row), mean-free-dispersion (second row) and mean-related-dispersion (last row) in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

estimate, the RMSE and the CR were computed for fits using the right data model. From this experiment, we expect on one hand bias and RMSE close to zero, and on the other hand, CR close to 95%.

Table 1: Performance measures for the maximum likelihood estimator of a count regression parameter β_j

Performance measure	Formula
Relative bias (%) in $\hat{\beta}_j$	$100 \times \lvert\beta_j\rvert^{-1}\left(\overline{\hat{\beta}}_j - \beta_j\right)$
Root mean square error of $\hat{\beta}_j$	$\sqrt{B^{-1}\sum_{r=1}^{B}\left(\hat{\beta}_j^{(r)} - \beta_j\right)^2}$
Coverage rate (%) of the asymptotic CI for β_j	$100 \times B^{-1}\sum_{r=1}^{B} I_{CI_j^{(r)}}(\beta_j)$

Table notes: $\overline{\hat{\beta}}_j$ is the average estimate of β_j: $\overline{\hat{\beta}}_j = B^{-1}\sum_{r=1}^{B}\hat{\beta}_j^{(r)}$; $B = 1000$ replications; CI = confidence interval; $I_{\mathbb{A}}(x)$ is the indicator function which returns one if $x \in \mathbb{A}$ and zero otherwise.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

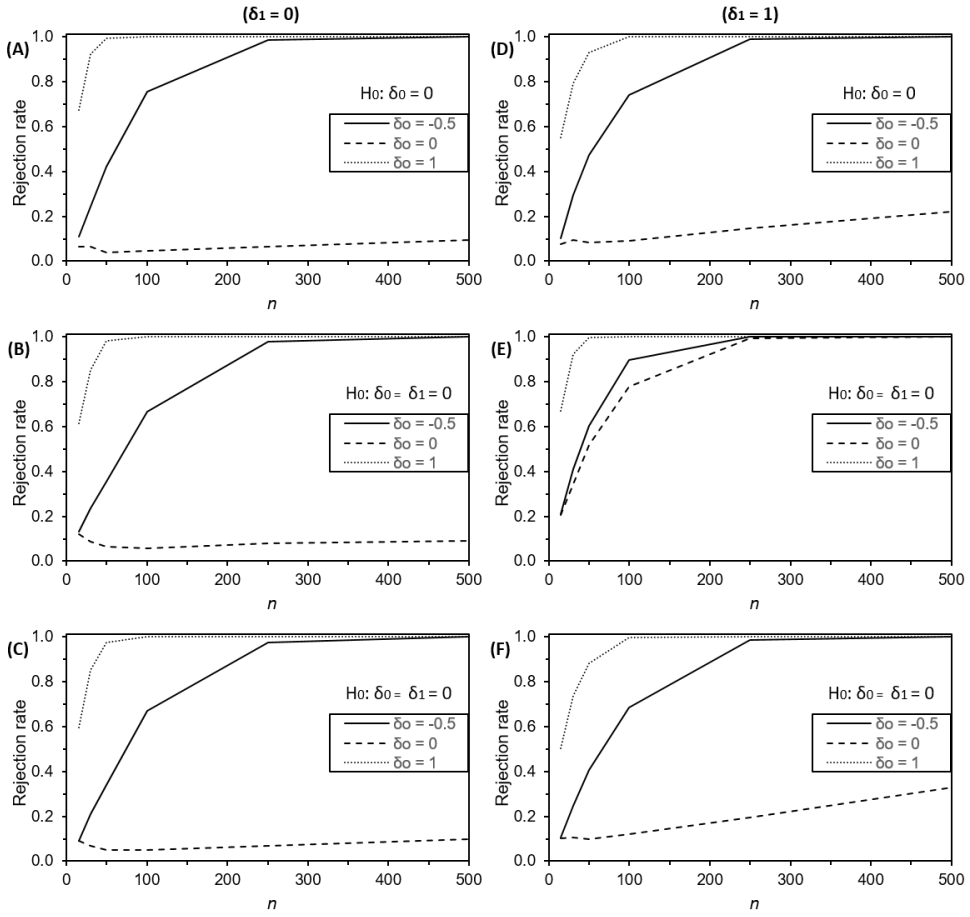

Figure 3. Power of the likelihood ratio test to detect equidispersion at the nominal level $\alpha = 5\%$ for balanced discrete gamma regression fits under constant-dispersion (first row), mean-free-dispersion (second row) and mean-related-dispersion (last row) in Coway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

3.3.1. Performance under Constant Dispersion Specification

This section presents the simulation results from the fits without a dispersion covariate. This corresponds to a dispersion model with either a correct specification (when $\delta_1 = 0$) or a misspecification scenario ($\delta_1 = 1$).

Figure 4 depicts the relative bias in the ML estimates of regression parameters (β_0 and β_1) under BDG data. It appears that the estimates of the parameters are on average very close to their true values. Indeed, in samples of size $n = 15$, irrespective of the fitting model, the biases in the estimates of the regression parameters are on average very low, with generally negative biases less than 5% in $\hat{\beta}_0$ and positive biases less than 5% in $\hat{\beta}_1$. Moreover, as n increases, the absolute value of the bias decreases and falls under 0.5% at $n = 500$. Similar results are observed under the scenarios using CMP data (see Figure A2 in Appendix B.2).

The RMSE of the ML estimates of β_0 and β_1 under BDG data are shown in Figure 5. It appears that in samples of size $n = 15$, for BDG regerssion fits, the RMSE of the estimates amounts on average to 0.43 when $\delta_0 = -0.5$, and falls to 0.27 for $\delta_0 = 0$, and to 0.16 for $\delta_0 = 1$. For CMP fits, the RMSE of the estimates amounts on average to 0.32 when $\delta_0 = -0.5$, and falls to 0.25 for $\delta_0 = 0$, and to 0.17 for $\delta_0 = 1$. Irrespective of the fitting model, the RMSE decreases as n increases, reaching about 0.04 at $n = 500$. It is also apparent from Figure 5 that a higher δ_0 (lower variance-to-mean ratio) implies a lower RMSE (more accurate estimates). Similar results are observed under CMP data (see Figure A3 in Appendix B.2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

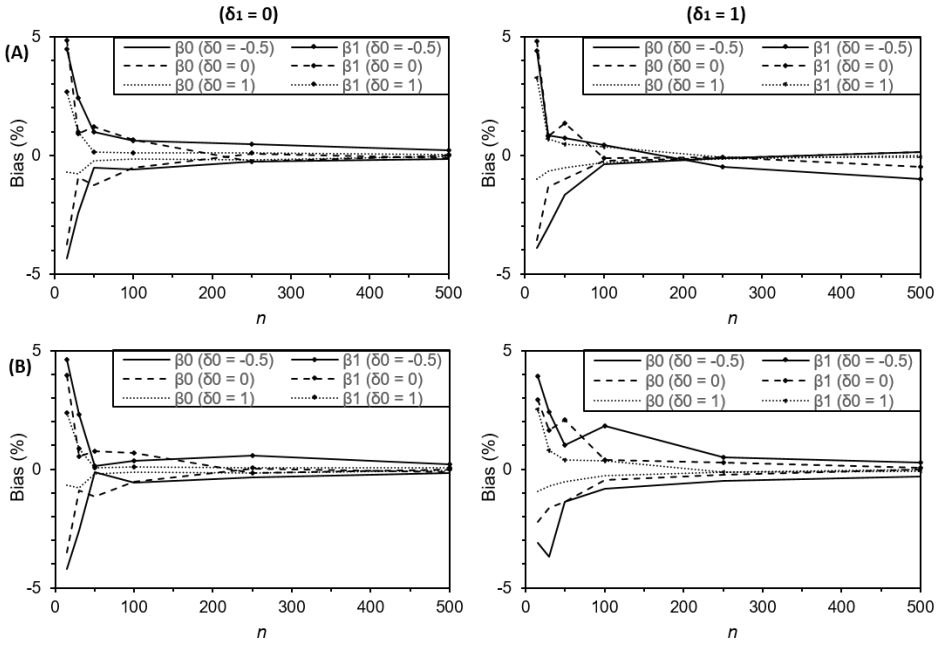

Figure 4. Percent bias in the estimates of regression parameters under constant-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

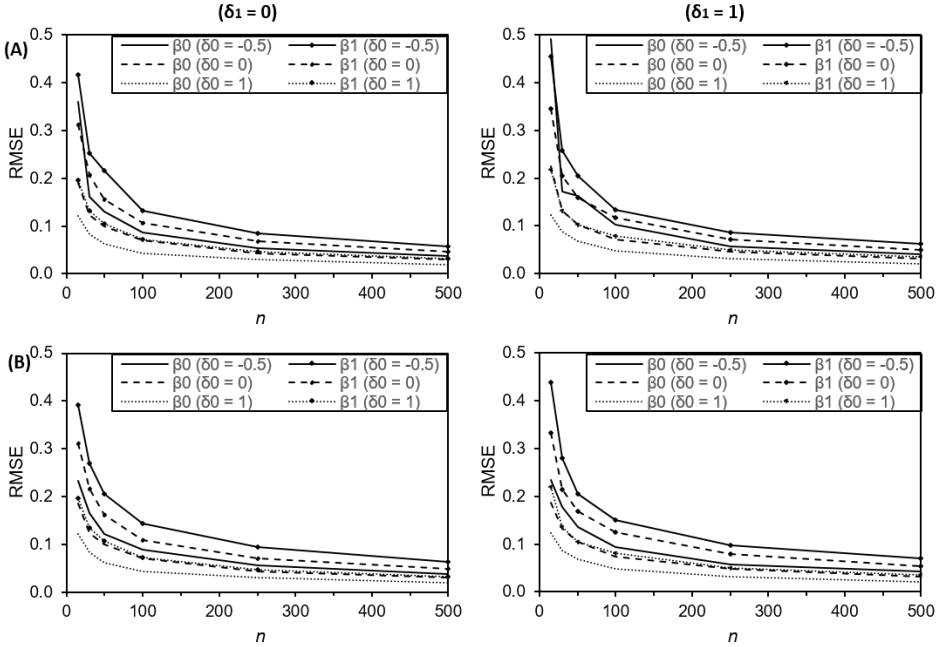

Figure 5. Root mean square error (RMSE) of regression parameter estimates under constant-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

Figure 6 presents the CR of the ML estimates of regression parameters (β_0 and β_1) under BDG data. It can be observed that, in samples of size $n = 15$, the CR is lower than the nominal level (95%), but generally stays above 90%. Moreover, the CR increases toward the nominal level as n increases. However, for MCMP regerssion fits in data generated with $\delta_0 = -0.5$ and $\delta_1 = 1$, as n increases from 250 to 500, the CR of the intercept β_0

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

decreases from 92.80% to 91.60%, and the CR of the slope β_1 decreases from 92.10% to 91.60%. Similar results are observed under the scenarios using CMP data, except that, for MCMP regerssion fits in data generated with $\delta_0 = -0.5$ and $\delta_1 = 1$, as n increases from 250 to 500, the CR does not decreases as under BDG data (see Figure A4 in Appendix B.2).

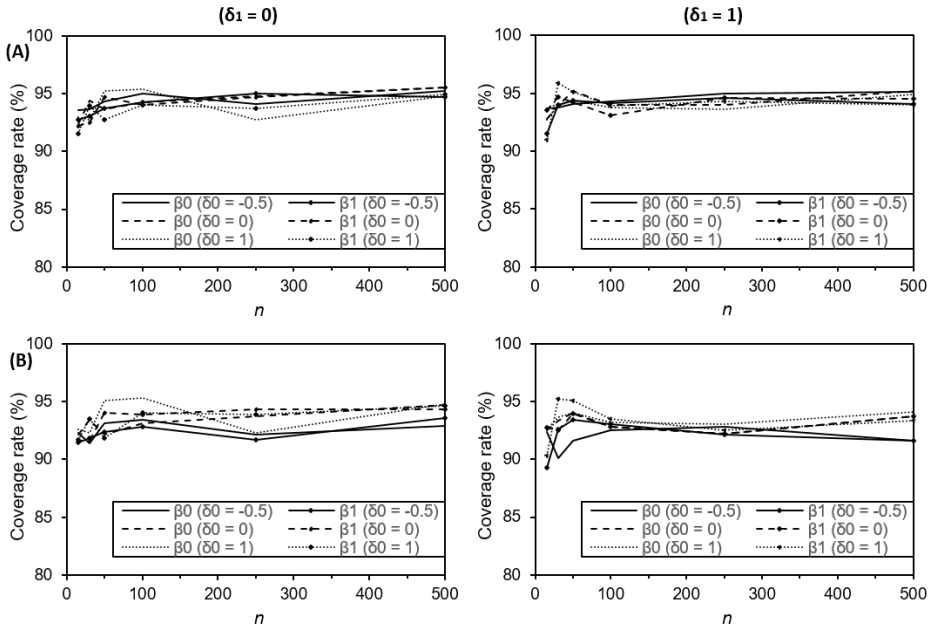

Figure 6. Coverage rate of asymptotic confidence interval for regression parameters under constant-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

It worth noting that the simulation results related to population effects are not generally sensitive to the misspecification of the dispersion model, *i.e.* when a dispersion covariate is missing ($\delta_1 = 1$). Indeed, under this scenario, slight differences are noted only when fitting MCMP regression model to BDG data (Figure 4 B) and when fitting BDG regression model to CMP data (Figure A2).

The averages of the maximum likelihood estimate of the dispersion parameter δ_0 in both BDG regression and MCMP regression are displayed in Table 2 (see the related bias in Table 7 in Appendix B.3). It can be observed that, for both BDG regression and MCMP regression fits, the bias in the estimate $\hat{\delta}_0$ decreases with the sample size. Moreover, the bias generally decreases as the true value δ_0 increases, *i.e.* estimates from data with a lower variance-to-mean ratio is less biased. It also appears that, for $n \leq 50$, $\hat{\delta}_0$ is generally less biased under the misspecification scenarios ($\delta_1 = 1$) than under the correct specification scenario ($\delta_1 = 0$). For $n \geq 250$ however, the estimate $\hat{\delta}_0$ is generally less biased under correct specification scenario than under the misspecification scenarios.

The RMSE for $\hat{\delta}_0$, under constant-dispersion fit, is generally under 0.75 for both BDG regression and MCMP regression fits, and approaches zero as the sample size n increases (see Figure A5 in Appendix B.3). The CR is above 86% at $n = 15$. Moreover, the CR increases toward the nominal level 95% as n increases, except when $\delta_0 = 1$ and $\delta_1 = 1$, in which case, the CR decreases and reaches about 60% at $n = 500$.

3.3.2. Performance under Variable Dispersion Specification

This section presents the simulation results from the fits under mean-free-dispersion specification, *i.e.* when the dispersion covariate z_i is included in the fit. This corresponds to a dispersion model with either a correct specification (when $\delta_1 = 1$) and an overfit scenario ($\delta_1 = 0$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Table 2: Maximum likelihood estimates of the dispersion parameter (δ_0) under constant-dispersion fit of balanced discrete gamma (BDG) regression model in BDG samples and Mean-parametrized Conway-Maxwell-Poisson (MCMP) regression model in Conway-Maxwell-Poisson (CMP) samples

δ_0	δ_1	BDG regression in BDG data						MCMP regression in CMP data					
		15	30	50	100	250	500	15	30	50	100	250	500
-0.5	0	-0.3124	-0.3636	-0.4585	-0.4733	-0.4814	-0.4911	-0.2631	-0.3884	-0.4279	-0.4661	-0.4830	-0.4915
	1	-0.4258	-0.4791	-0.5079	-0.5078	-0.5091	-0.5088	-0.2435	-0.4164	-0.4672	-0.4827	-0.5097	-0.5089
0	0	0.2341	0.1137	0.0690	0.0330	0.0172	0.0078	0.2421	0.1198	0.0805	0.0376	0.0145	0.0035
	1	0.1041	-0.0230	-0.0641	-0.0838	-0.1040	-0.1063	0.1533	0.0280	-0.0159	-0.0157	-0.0538	-0.0535
1	0	1.2880	1.1496	1.0826	1.0463	1.0195	1.0095	1.1980	1.1003	1.0546	1.0341	1.0098	1.0060
	1	1.1302	1.0055	0.9399	0.9055	0.8709	0.8723	1.0569	0.9807	0.9303	0.9220	0.8896	0.8954

Table notes: Est. = Estimate of dispersion parameter.

Figure 7 depicts the relative bias in the ML estimates of regression parameters (β_0 and β_1) under BDG data. It appears that the estimates of the parameters are on average very close to their true values. Indeed, in samples of size $n = 15$, irrespective of the fitting model, the biases in the estimates of the regression parameters are on average very low, with generally negative biases less than 5% in $\hat{\beta}_0$ and positive biases less than 5% in $\hat{\beta}_1$. Moreover, as n increases, the absolute value of the bias decreases and falls under 0.5% at $n = 500$. Similar results are observed under the scenarios using CMP data (see Figure A6 in Appendix B.4).

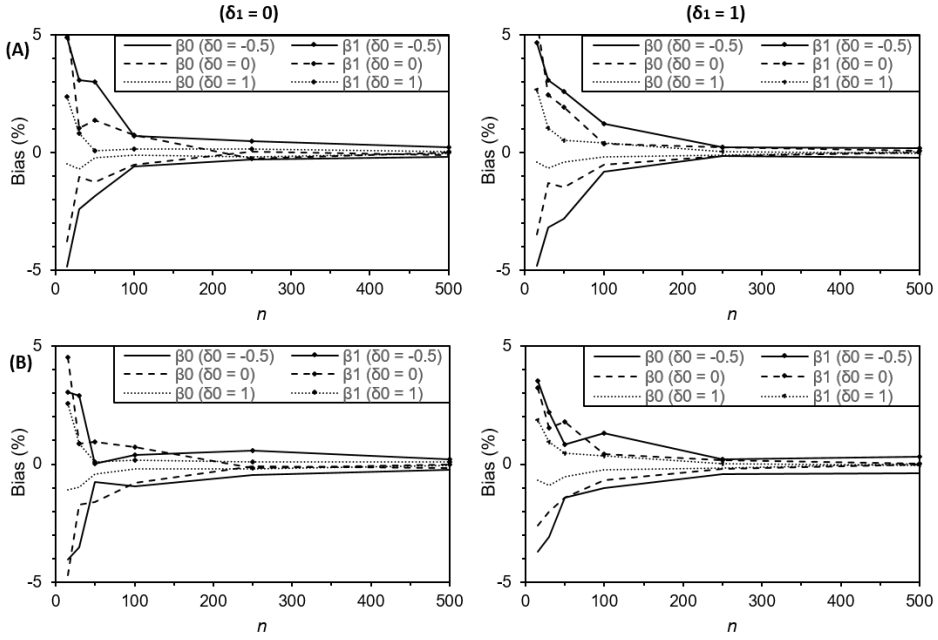

Figure 7. Percent bias in the estimates of regression parameters under mean-free-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

The RMSE of the ML estimates of regression parameters (β_0 and β_1) under BDG data are shown in Figure 8. It appears that in samples of size $n = 15$, for BDG regerssion fits, the RMSE of the estimates amounts on average to 0.48 when $\delta_0 = -0.5$, and falls to 0.29 for $\delta_0 = 0$, and to 0.17 for $\delta_0 = 1$. For CMP fits, the RMSE of the estimates amounts on average to 0.33 when $\delta_0 = -0.5$, and falls to 0.23 for $\delta_0 = 0$, and to 0.17 for $\delta_0 = 1$. Irrespective of the fitting model, the RMSE decreases as n increases, reaching about 0.04 at $n = 500$. It is also apparent from Figure 8 that a higher δ_0 (lower variance-to-mean ratio)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

implies a lower RMSE (more accurate estimates). Similar results are observed under CMP data (see Figure A7 in Appendix B.4).

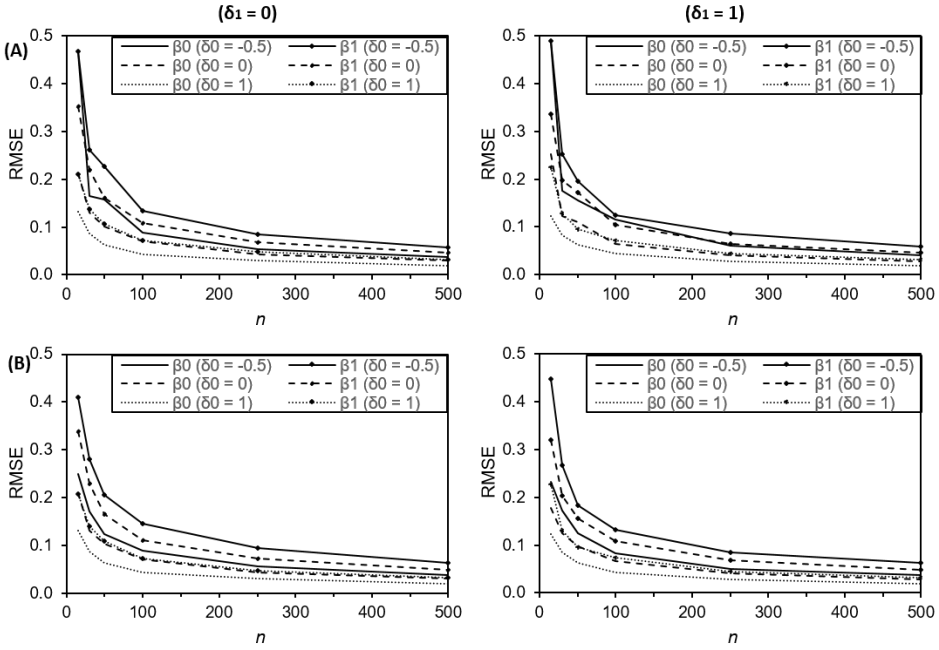

Figure 8. Root mean square error (RMSE) of regression parameter estimates under mean-free-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

Figure 9 presents the CR of the ML estimates of regression parameters (β_0 and β_1) under BDG data. It can be observed that, in samples of size $n = 30$, the CR is lower than the nominal level (95%), but generally stays above 90%. Moreover, the CR increases toward the nominal level as n increases. However, for MCMP regerssion fits to data generated with $\delta_0 = -0.5$ and $\delta_1 = 1$, as n increases from 250 to 500, the CR of the intercept β_0 decreases from 92.90% to 91.80%, and the CR of the slope β_1 decreases from 93.00% to 90.70%. For BDG regerssion fits (Figure 9 A) in small samples ($n = 15$), the CR ranges between 87.58% and 89.54% for $\delta_1 = 0$, and between 86.62% and 89.58% for $\delta_1 = 1$. The CR is slightly lower for MCMP regerssion fits (Figure 9 B) in smaller samples ($n = 15$), ranging between 82.57% and 85.04% for $\delta_1 = 0$, and between 80.20% and 86.45% for $\delta_1 = 1$. Similar results are observed under the scenarios using CMP data, except that, for MCMP regerssion fits to data generated with $\delta_0 = -0.5$ and $\delta_1 = 1$, as n increases from 250 to 500, the CR does not decreases as under BDG data (see Figure A8 in Appendix B.4).

It worth noting that, under variable dispersion specification, the simulation results related to population effects are not sensitive to an overfit, when a dispersion covariate is wrongly included ($\delta_1 = 0$). The maximum likelihood estimates of dispersion parameters in BDG regression are displayed in Table 3. In samples of size $n < 50$ ($n = 15, 30$), the bias in the estimate $\hat{\delta}_0$ is positive and high (22% to 57%), and shows a decreasing trend as δ_0 increases, with an average which drops from 42% when $\delta_0 = -0.5$ to 34% when $\delta_0 = 1$ (see the related bias in Table 8 in Appendix B.5). As the sample size increases, the bais in $\hat{\delta}_0$ decreases and falls under 13% at $n = 50$, and under 2% at $n = 500$. In regards to δ_1, irrespective of the sample size, the average estimate is in absolute value lower than 0.01 when $\delta_1 = 0$ (Table 3). For $\delta_1 = 1$, in samples of sizes $n < 50$ ($n = 15, 30$), the bias in the estimate $\hat{\delta}_1$ is positive and shows an increasing trend as δ_0 increases, from 14% when $\delta_0 = -0.5$ to on average 18% when $\delta_0 = 1$. As the sample size increases, the bais in $\hat{\delta}_1$ decreases and falls under 10% at $n = 50$, and under 2% at $n = 500$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

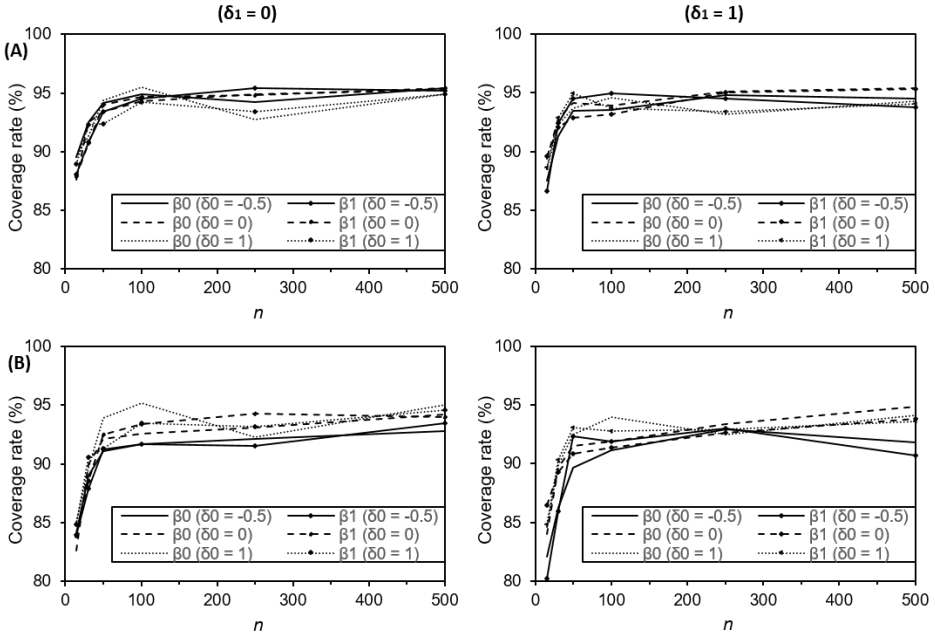

Figure 9. Coverage rate of asymptotic confidence interval for regression parameters under mean-free-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in balanced discrete gamma samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$

Table 3: Maximum likelihood estimates of dispersion parameters under mean-free-dispersion balanced discrete gamma regression model specification in balanced discrete gamma samples

δ_0	Est.	$\delta_1 = 0$						$\delta_1 = 1$					
		15	30	50	100	250	500	15	30	50	100	250	500
-0.5	$\hat{\delta}_0$	-0.2164	-0.3247	-0.4571	-0.4637	-0.4782	-0.4898	-0.3078	-0.3206	-0.5209	-0.4675	-0.4835	-0.4902
	$\hat{\delta}_1$	0.0088	0.0237	-0.0059	0.0011	-0.0060	0.0021	1.1482	1.1372	1.0707	1.0366	1.0093	1.0046
0	$\hat{\delta}_0$	0.3747	0.1631	0.0931	0.0443	0.0212	0.0096	0.3515	0.1625	0.0742	0.0489	0.0192	0.0081
	$\hat{\delta}_1$	0.0346	0.0018	-0.0004	0.0162	-0.0014	0.0018	1.1632	1.1568	1.0898	1.0433	1.0104	1.0090
1	$\hat{\delta}_0$	1.4670	1.2241	1.1184	1.0609	1.0254	1.0123	1.4228	1.2439	1.1248	1.0561	1.0216	1.0128
	$\hat{\delta}_1$	-0.0026	0.0309	0.0231	-0.0017	0.0033	0.0003	1.1877	1.1641	1.0922	1.0260	1.0137	1.0123

Table notes: Est. = Estimate of dispersion parameter.

The maximum likelihood estimates of dispersion parameters in MCMP regression are displayed in Table 4 (see the related bias in Table 9 in Appendix B.5). In small samples ($n = 15$), the bias in the estimate $\hat{\delta}_0$ is negative when $\delta_0 = -0.5$, varying from -6% ($\delta_1 = 0$) to -63% ($\delta_1 = 1$). As δ_0 increases, the bias in $\hat{\delta}_0$ increases and becomes positive, with an average of 24% when $\delta_0 = 1$. For $n \geq 30$, the bias is positive and decreases with n, from on average 11% at $n = 30$ to less than 2% at $n = 500$.

In regards to δ_1, the estimate $\hat{\delta}_1$, from small samples ($n = 15$) generated using $\delta_1 = 0$, is positive (Table 4), decreasing from $\hat{\delta}_1 = 0.2370$ when $\delta_0 = -0.5$ to $\hat{\delta}_1 = 0.0431$ when $\delta_0 = 1$. However, as n increases, the estimate drops, in absolute value, from about 0.014 at $n = 30$ to less than 0.004 at $n = 500$. For data generated using $\delta_1 = 1$, in samples of size $n = 15$, the bias in the estimate $\hat{\delta}_1$ is -19% when $\delta_0 = -0.5$, 17% when $\delta_0 = 0$ and 8% when $\delta_0 = 1$ (see Tables 8 and 9 in Appendix B.5). As n increases, the absolute value of the bias drops, on average, from 14% ($n = 30$) to 0.9% ($n = 500$).

The RMSE of dispersion parameter estimates are generally under 1.5 for BDG regression and under 2 for MCMP regression, and approach zero as the sample size n increases (see Figure A9 in Appendix B.5). Moreover, the CR of dispersion parameter estimates are generally above 86% at $n = 15$, and increases toward the nominal level 95% as n increases

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Table 4: Maximum likelihood estimates of dispersion parameters under mean-free-dispersion Mean-parametrized Conway-Maxwell-Poisson regression model specification in Conway-Maxwell-Poisson samples

δ_0	Est.	$\delta_1 = 0$						$\delta_1 = 1$					
		15	30	50	100	250	500	15	30	50	100	250	500
-0.5	$\hat{\delta}_0$	-0.5308	-0.4619	-0.4668	-0.4618	-0.4820	-0.4908	-0.8162	-0.4491	-0.4837	-0.4754	-0.4878	-0.4942
	$\hat{\delta}_1$	0.2370	-0.0064	0.0088	0.0038	0.0044	-0.0025	0.8085	1.1826	1.1504	1.0413	1.0096	1.0068
0	$\hat{\delta}_0$	0.2008	0.1196	0.0903	0.0431	0.0167	0.0047	0.1776	0.0928	0.0558	0.0399	0.0066	0.0000
	$\hat{\delta}_1$	0.1575	-0.0252	-0.0189	-0.0058	-0.0009	0.0040	1.1650	1.1506	1.1366	1.0180	1.0118	1.0160
1	$\hat{\delta}_0$	1.2500	1.1441	1.0754	1.0425	1.0134	1.0076	1.2292	1.1331	1.0772	1.0477	1.0152	1.0100
	$\hat{\delta}_1$	0.0431	0.0097	-0.0018	-0.0038	0.0020	0.0026	1.0811	1.0893	1.0764	1.0334	1.0179	1.0049

Table notes: Est. = Estimate of dispersion parameter.

(see Figure A10). Very high RMSE (up to 14) and low CR (down to 81%) are nevertheless observed for MCMP regression fit when $\delta_0 = -0.5$.

4. Applications

To demonstrate the use of the proposed approach to flexible count modelling, we revisit two datasets popular in the count modelling literature. We compare the BDG model fits to competing model fits including the MCMP model and HP model fits. BDG models were fitted in R [23] using procedures described in Section 2, MCMP models were fitted using the routine *glm.cmp* of the package *mpcmp* [28] and HP models were fitted using the routine *glm.hP* of the package *DGLMExtPois* [29]. The two datasets are available in the R package *mpcmp*.

4.1. The Class Attendance Dataset

The class attendance dataset was built to assess the effect of the gender, academic programme ("General", "Academic" and "Vocational") and math score (standardized score out of 100) on the class non-attendance (number of days absent) of 314 students from two urban high schools [5]. Some summary statistics of the data are given in Appendix C. Huang et al. [5] compared the constant-dispersion fits of the negative binomial [11], Consul's Generalized Poisson [14], HP [3], CMP and MCMP regression models to this dataset. Their results show that, after accounting for the explanatory variables, the data exhibits overdispersion. In terms of both interpretability and parsimony, it also appears that, among the competing models, the HP and MCMP regression models provide the best fits to the data, the HP model giving a slightly parsimonious fit.

Here, for comparison purposes, we fit the BDG regression under constant-dispersion specification. The results, displayed in Table 5 aside HP and MCMP regression fits, indicate that the BDG regression fit has the lowest AIC value (or equivalently the highest maximized likelihood value) and is thus superior to the two competing models for these data. However, it appears that the data do not agree with either the HP, MCMP or BDG regression fit, as indicated by the deviance based goodness-of-fit test results. The BDG regression fit has the lowest residual deviance ($D = 362.77$), but still too larger than the expected value (DF = 308). The lack of fit of the models is apparent from the low proportion of information explained by the included covariates ($R^2_{dev} < 20\%$). Some important explanatory variables for class attendance (*e.g.* the appartenance to groups of disadvantaged backgrounds, family education and parental income [30]) are obviously missing in the fitted models. The estimated regression coefficients are nevertheless indicative of the association between the considered covariates and class attendance. For instance, the BDG fit indicates that (i) female students missed on average $\exp(+0.24) = 1.27$ times more days of school as compared to male students; (ii) students in the General programme missed on average $\exp(+1.27) = 3.56$ times more days of school as compared to students in the Vocational programme; (iii) a 10-point increase in math score is associated with a $100 \times 10 \times (+0.006) = 6\%$ decrease in the expected number of days of absence from school.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Table 5: Estimated coefficients, standard errors (SE), dispersion parameter (δ_0), residual deviance (D), deviance R^2 (R^2_{dev}) and Akaike's Information Criterion (AIC) for hyper Poisson (HP), Mean-parametrized Conway-Maxwell-Poisson (MCMP), and balanced discrete gamma (BDG) regression model fits to the attendance dataset

Parameter	HP				MCMP				BDG			
	Estimate	SE	z-value	p-value	Estimate	SE	z-value	p-value	Estimate	SE	z-value	p-value
Intercept	2.73	0.18	15.15	<0.001	2.71	0.19	14.24	<0.001	2.84	0.14	20.10	<0.001
Gender = Male	-0.22	0.12	-1.87	0.084	-0.21	0.12	-1.83	0.067	-0.24	0.10	-2.38	0.017
Prog. = Academic	-0.43	0.16	-2.68	0.019	-0.43	0.17	-2.51	0.012	-0.60	0.12	-4.86	0.000
Prog. = Vocational	-1.26	0.19	-6.82	<0.001	-1.25	0.19	-6.61	<0.001	-1.27	0.15	-8.22	<0.001
Math score	-0.01	0.002	-3.00	0.003	-0.01	0.002	-2.65	0.008	-0.006	0.002	-3.14	0.002
δ_0	6.15	0.52	11.87	<0.001	-3.91	0.86	-4.53	<0.001	-1.95	0.10	-18.77	<0.001
D (DF = 308)	-			-	377.08			0.004	362.77			0.017
R^2_{dev}	-				0.158				0.195			
AIC	1739.80				1741.03				1724.65			

Table notes: Prog. = programme; - = not available.

4.2. The Cottonbolls Dataset

The cottonbolls study [4] examines the association between the number of cotton bolls produced by a plant and the level of defoliation (def) (0%, 25%, 50%, 75% and 100%) and the growth stages (vegetative, flower-bud, blossom, fig and cotton boll). We refer to Figure 2 of [4] and Figure 3 of [5] for an exploratory analysis of the data which exibits strong underdispersion. Huang et al. [5] compared the fits of the Gamma-Count (GC) [4], Consul's Generalized Poisson [14], HP [3], CMP and MCMP regression models to this dataset, considering five different sets of linear predictors introduced by [4] and constant-dispersion specification. It appears that, among the competing models, the GC, CMP and MCMP regression models provide the best fits to the data with the following set of linear predictors for the mean μ:

$$\log(\mu) = \gamma_0 + \gamma_{1j}\text{def} + \gamma_{2j}\text{def}^2$$

where, the index $j \in \{\text{vegetative, flower-bud, blossom, fig, cotton boll}\}$ indicates the growth stage. Although the CMP (AIC = 440.50) and GC (AIC = 440.77) model fits give a slightly parsimonious fit as compared to the MCMP fit (AIC = 440.82), the latter is the best in terms of interpretability (in the GC model, μ is the mean waiting time between successive counts rather than the mean count).

Here, for comparison purposes, we fit the BDG regression under constant-dispersion specification. The results are displayed in Table 6 aside the MCMP fit results. The deviance based goodness-of-fit test results indicate that the data agree with both the MCMP and BDG models. It appears that, in terms of explanatory power ($R^2_{dev} = 63\%$), interpretability and parsimony, BDG regression provides the best fit for these data (AIC = 437.87).

5. Conclusion

In the analysis of count outcomes, it is now common sens to check the appropriateness of the basic Poisson model against some flexible count models. The major limitation of popular flexible count models is the inability to directly model the mean of a count response. With the HP regression, the MCMP regression constitutes the state-of-art in flexible count modelling, but still relies on approximations. We have proposed a new approach using BDG distributions, exempt of approximations and computationally demanding procedures. For parsimony, we have introduced a latent equidispersion (LE) test based on the likelihood ratio (LR) statistic. The latter converges in distribution to the theoretical χ^2 law, and the power of the LE test tends to one as the sample size increases. Our simulation results also show that, in small to moderate samples ($n = 15 - 500$), the LE test does not differentiate BDG and Poisson data, making the LE test usable to check the appropriateness of the Poisson model in small to moderate samples. Larger samples, however, contain enough information for the test to differentiate Poisson from BDG distributions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Table 6: Estimated coefficients, standard errors (SE), dispersion parameter (δ_0), residual deviance (D), deviance R^2 (R^2_{dev}) and Akaike's Information Criterion (AIC) for Mean-parametrized Conway-Maxwell-Poisson (MCMP) and balanced discrete gamma (BDG) regression model fits to the cottonbolls dataset

Parameter	MCMP				BDG			
	Estimate	SE	z-value	p-value	Estimate	SE	z-value	p-value
γ_0	2.19	0.03	74.56	<0.001	2.19	0.03	74.33	<0.001
$\gamma_{1\text{vegetative}}$	0.44	0.24	1.82	0.069	0.46	0.24	1.93	0.053
$\gamma_{2\text{vegetative}}$	-0.80	0.27	-2.96	0.003	-0.82	0.27	-3.05	0.002
$\gamma_{1\text{flowerbud}}$	0.29	0.24	1.22	0.222	0.29	0.23	1.23	0.220
$\gamma_{2\text{flowerbud}}$	-0.49	0.26	-1.85	0.064	-0.48	0.26	-1.83	0.067
$\gamma_{1\text{blossom}}$	-1.25	0.28	-4.42	<0.001	-1.28	0.28	-4.54	<0.001
$\gamma_{2\text{blossom}}$	0.68	0.32	2.13	0.033	0.72	0.32	2.27	0.023
$\gamma_{1\text{fig}}$	0.35	0.26	1.33	0.184	0.29	0.26	1.11	0.267
$\gamma_{2\text{fig}}$	-1.29	0.32	-4.08	<0.001	-1.21	0.31	-3.92	<0.001
$\gamma_{1\text{cottonboll}}$	0.01	0.23	0.03	0.974	0.001	0.23	0.00	0.997
$\gamma_{2\text{cottonboll}}$	-0.02	0.26	-0.07	0.941	-0.004	0.26	-0.02	0.986
δ_0	1.58	0.13	12.39	<0.001	1.63	0.14	11.52	<0.001
D (DF = 123)	125.51			0.420	124.62			0.442
R^2_{dev}	0.606				0.631			
AIC	440.82				437.87			

The results from simulation experiments also show that BDG regression outperforms MCMP regression in small samples ($n = 15 - 50$) in terms of coverage rates of asymptotic 95% confidence intervals for regression parameters. Moreover, in small samples, BDG regression outperforms MCMP regression in estimating dispersion parameters, in terms of bias ($n = 15 - 30$), RMSE ($n = 15$) and coverage of asymptotic 95% confidence intervals ($n = 15 - 50$), mostly under severe overdispersion scenarios. Overall, estimates from BDG regression are, at the least, as good as estimates from MCMP regression, in all tested simulation scenarios. Furthermore, the data analysis examples show that the proposed approach is at least comparable to existing models and can provide substantial gain in parsimony.

Apart from dispersion, another frequent characteristic of count data is zero modification (inflation or deflation). Future works will tackle the extension of BDG regression to control zero modification while ensuring full dispersion flexibility. Finally, since count data are often grouped by some sampling units (individual, geographical area, time), we envisage the extension of the BDG regression model to the analysis of multivariate count outcomes.

Author Contributions: Conceptualization, C.F.T.; methodology, C.F.T. and R.G.K.; software, C.F.T.; validation, C.F.T. and R.G.K.; formal analysis, C.F.T.; resources, R.G.K.; writing—original draft preparation, C.F.T.; writing—review and editing, C.F.T. and R.G.K.; visualization, C.F.T.; supervision, R.G.K.; project administration, R.G.K.; funding acquisition, R.G.K. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The authors confirm that the data supporting the findings of this work are available within the article.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Appendix A. Fitting the Balanced Discrete Gamma Distribution

When $\boldsymbol{X}_i = 1$ and $\boldsymbol{Z}_i = 1$ (for all $i = 1, 2, \cdots, n$) in the model (3), $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\top$ is an independent and identically distributed sample from a BGD variable $Y \sim \mathcal{BG}(\mu, a)$ where $\mu = \exp(\beta_0)$ and $a = \exp(\delta_0)$. In this case, the log-likelihood (4) of $\boldsymbol{\theta} = (\beta_0, \delta_0)^\top$ simplifies to

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i|\mu, a). \tag{A1}$$

Recall that for $x > 0$ and $b > 0$, the incomplete gamma ratio $\gamma(x, b)$ and its successive derivatives with respect to x or b are continuous functions of both x and b. The pmf in (1) and its successive derivatives with respect to μ or a are thus continuous with respect to both μ and a, since the pmf only involves the incomplete gamma ratio. As a result, the log-likelihood function $\ell(\cdot|\boldsymbol{y})$ is regular and its first and second derivatives with respect to β_0 and δ_0 are continuous functions of $\boldsymbol{\theta} \in \mathbb{R}^2$ for any possible sample $\boldsymbol{y}$.

The ML estimate of $\boldsymbol{\theta}$ can be obtained using a numerical optimization routine to maximize the log-likelihood function in (A1), *e.g.* the *optim* function in R freeware [23]. Most optimization routines require starting values for parameters to be found. We suggest approximate moment estimates as starting values for μ and a. Since the BDG distribution have been given under mean parametrization, the sample mean provides a moment estimate for μ: $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Then, by equalizing the unbiased sample variance estimate $\sigma^{2(0)} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu^{(0)})^2$ to an approximation of the variance in (2) as $\sigma^{2(0)} = \mu^{(0)}/a^{(0)} + \zeta_0^{(0)}$ (with $\zeta_0^{(0)} = min\{\mu^{(0)}/2, 1/8\}$ if $\sigma^{2(0)} > min\{\mu^{(0)}, 1/4\}$ and $\zeta_0^{(0)} = \sigma^{2(0)}/2$ otherwise), one obtains the approximate moment estimate

$$a^{(0)} = \frac{\mu^{(0)}}{\sigma^{2(0)} - \zeta_0^{(0)}}. \tag{A2}$$

The parameters β_0 and δ_0 can thus be initialized with $\beta_0^{(0)} = \log \mu^{(0)}$ and $\delta_0^{(0)} = \log a^{(0)}$. After the log-likelihood maximization to find $\widehat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\delta}_0)^\top$, the ML estimate of μ and a are recoverable through $\hat{\mu} = \exp(\hat{\beta}_0)$ and $\hat{a} = \exp(\hat{\delta}_0)$. After the optimization step, the inverse $\widehat{\boldsymbol{\Sigma}}$ of the negative hessian of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, provides an approximate covariance matrix for $\widehat{\boldsymbol{\theta}}$. Then, for asymptotic inference, $\boldsymbol{\theta}$ is assumed to be normally distributed with mean $\widehat{\boldsymbol{\theta}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$.

However, inference on the original parameter $\boldsymbol{\xi} = (\mu, a)^\top$ is of greater importance. This can be achieved by translating restrictions on $\boldsymbol{\xi}$ to the logarithmic scale parameter $\boldsymbol{\theta} = (\beta_0, \delta_0)^\top$. Thus, the null assumption $\mu = \mu_0$ for $\mu_0 > 0$ is equivalent to $\beta_0 = \log \mu_0$ and the assumption $a = 1$ (LE) is equivalent to $\delta_0 = 0$. In addition, assuming that elements of $\boldsymbol{\xi}$ follow log-normal distributions, asymptotic confidence bounds can be obtained for μ and a following [32]. An asymptotic covariance matrix can also be obtained for the original parameter estimate $\widehat{\boldsymbol{\xi}}$ as

$$\mathrm{Var}\left[\widehat{\boldsymbol{\xi}}\right] = \left[\boldsymbol{J}(\widehat{\boldsymbol{\xi}})\right]\widehat{\boldsymbol{\Sigma}}\left[\boldsymbol{J}(\widehat{\boldsymbol{\xi}})\right]^\top \tag{A3}$$

where $\boldsymbol{J}(\boldsymbol{\xi}) = diag(\mu, a)$ (Jacobian matrix of the back parameter transformation).

Appendix B. Additional Simulation Results

Appendix B.1. Rejection Rate of the Latent Equidispersion Test in Poisson Data

Figure A1 presents the rejection rate of the likelihood ratio test to detect equidispersion in balanced discrete gamma regression using Poisson samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

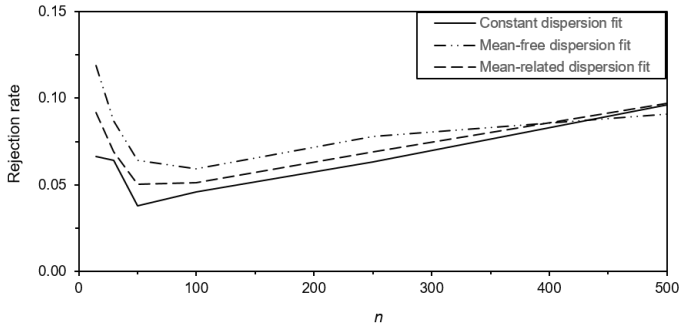

Figure A1. Power of the likelihood ratio test to detect equidispersion at the nominal level $\alpha = 5\%$ for balanced discrete gamma regression fits in Poisson samples of various sizes n in the range [15, 500].

Appendix B.2. Properties of Regression Parameters Under Constant-Dispersion

Figures A2–A4 present the bias, coverage rate (CR) and root mean square error (RMSE) of maximum likelihood estimates of regression parameters under constant-dispersion specification in Conway-Maxwell-Poisson samples.

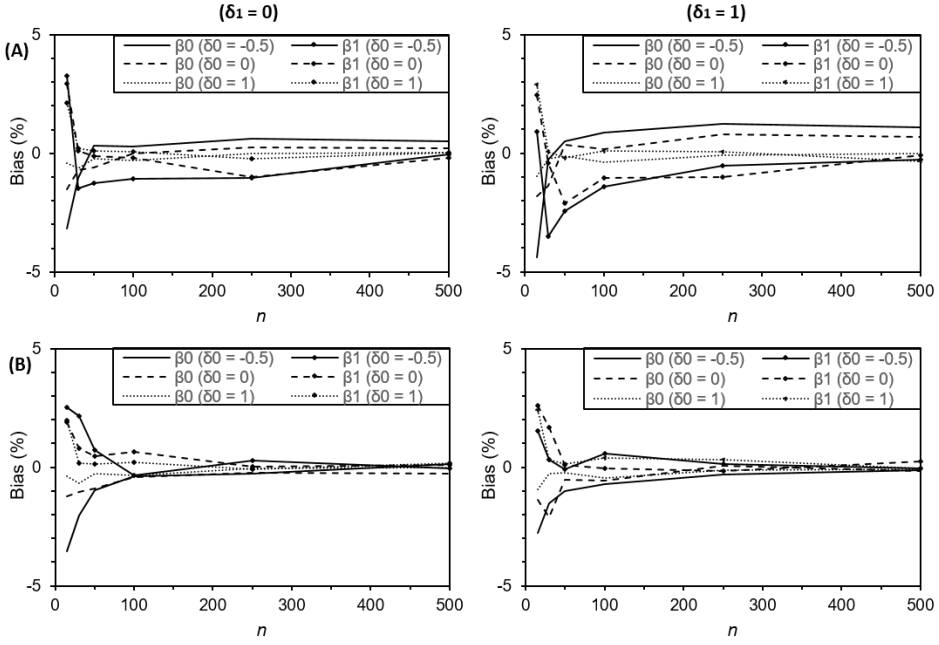

Figure A2. Percent bias in the estimates of regression parameters under constant-dispersion balanced discrete gamma (first row) and Mean-parametrized Conway-Maxwell-Poisson (last row) regression model specifications in Conway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

Appendix B.3. Bias in Dispersion Parameter Estimates under Constant-Dispersion

Table 7 displays the bias in the dispersion parameter estimate under constant-dispersion specification in balanced discrete gamma and Conway-Maxwell-Poisson samples.

Figure A5 presents RMSE and CR for $\hat{\delta}_0$ under constant-dispersion BDG regression and MCMP regression fits.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

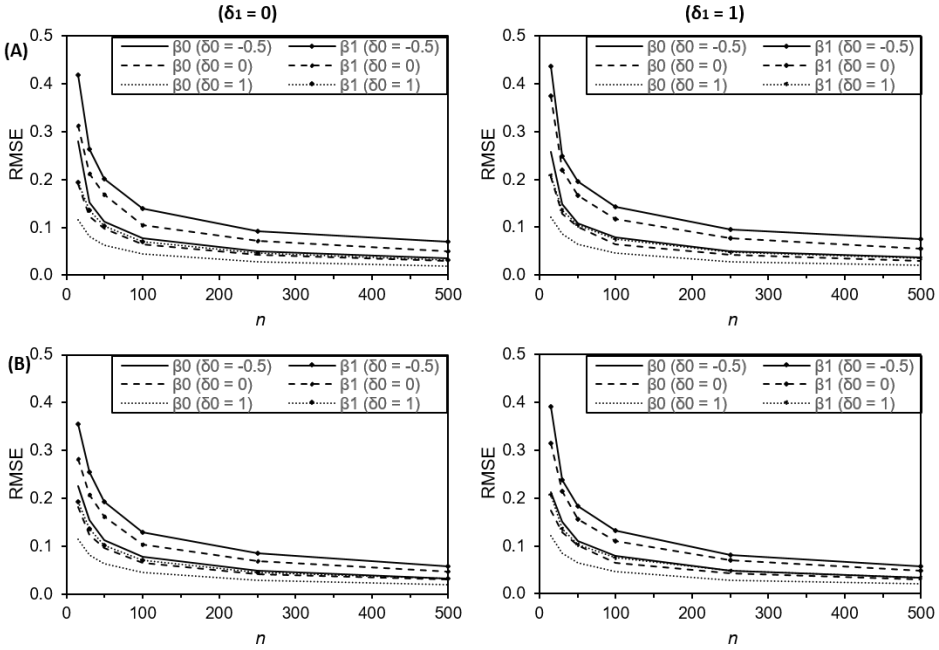

Figure A3. Root mean square error (RMSE) of regression parameter estimates under constant-dispersion balanced discrete gamma (first row) and Mean-parametrized Conway-Maxwell-Poisson (last row) regression model specifications in Mean-parametrized Conway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

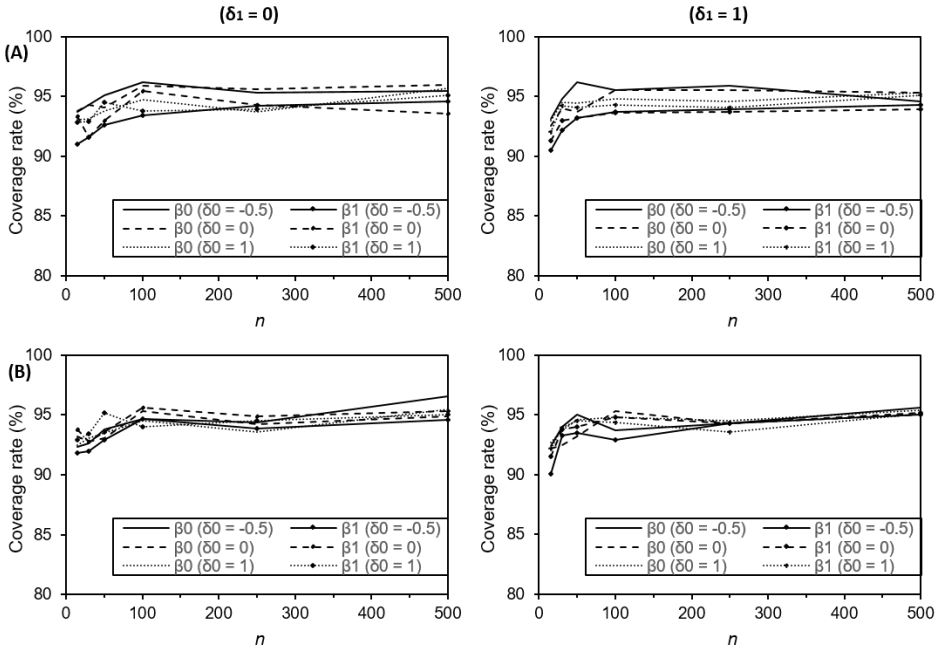

Figure A4. Coverage rate of asymptotic confidence interval for regression parameters under constant-dispersion balanced discrete gamma (first row) and Mean-parametrized Conway-Maxwell-Poisson (last row) regression model specifications in Conway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

Table 7: Bias in the maximum likelihood estimates ($\hat{\delta}_0$) of the dispersion parameter (δ_0) under constant-dispersion specification in balanced discrete gamma (BDG) and Conway-Maxwell-Poisson (CMP) samples

δ_0	δ_1	BDG						CMP					
		15	30	50	100	250	500	15	30	50	100	250	500
-0.5	0	0.3752	0.2728	0.0830	0.0534	0.0372	0.0178	0.4738	0.2232	0.1442	0.0678	0.0340	0.0170
	1	0.1484	0.0418	-0.0158	-0.0156	-0.0182	-0.0176	0.5130	0.1672	0.0656	0.0346	-0.0194	-0.0178
0	0	0.2341	0.1137	0.0690	0.0330	0.0172	0.0078	0.2421	0.1198	0.0805	0.0376	0.0145	0.0035
	1	0.1041	-0.0230	-0.0641	-0.0838	-0.1040	-0.1063	0.1533	0.0280	-0.0159	-0.0157	-0.0538	-0.0535
1	0	0.2880	0.1496	0.0826	0.0463	0.0195	0.0095	0.1980	0.1003	0.0546	0.0341	0.0098	0.0060
	1	0.1302	0.0055	-0.0601	-0.0945	-0.1291	-0.1277	0.0569	-0.0193	-0.0697	-0.0780	-0.1104	-0.1046

Table note: the table displays the absolute bias ($\hat{\delta}_0 - \delta_0$) if the true parameter value is zero ($\delta_0 = 0$) and the relative bias ($100 \times (\hat{\delta}_0 - \delta_0)/|\delta_0|$) if the true parameter value is not zero ($\delta_0 = 0, 1$).

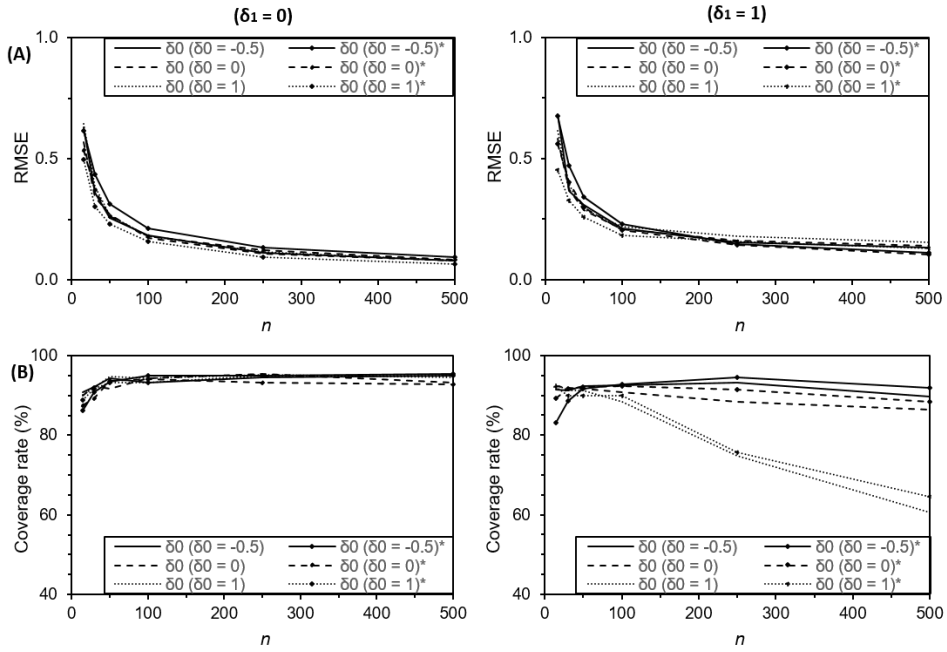

Figure A5. (**A**) Root mean square error (RMSE) and (**B**) Coverage rate of asymptotic confidence interval for dispersion parameter estimates under constant-dispersion balanced discrete gamma (BDG) regression fitted to BDG samples and Mean-parametrized Conway-Maxwell-Poisson (MCMP) regression fitted to CMP samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$. The curves with a start (*) at the end of the legend are for MCMP regression fits whereas the other curves are for BDG regression fits.

Appendix B.4. Properties of Regression Parameters Under Mean-free-Dispersion

Figures A6–A8 present the bias, coverage rate (CR) and root mean square error (RMSE) of maximum likelihood estimates of regression parameters under constant-dispersion specification in Conway-Maxwell-Poisson samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

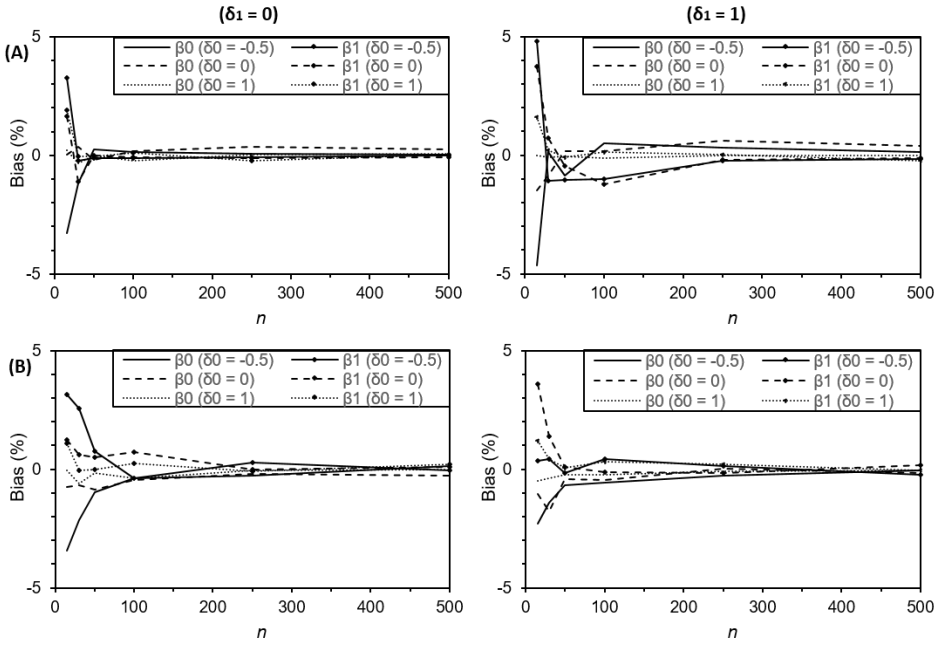

Figure A6. Percent bias in the estimates of regression parameters under mean-free-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in Conway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

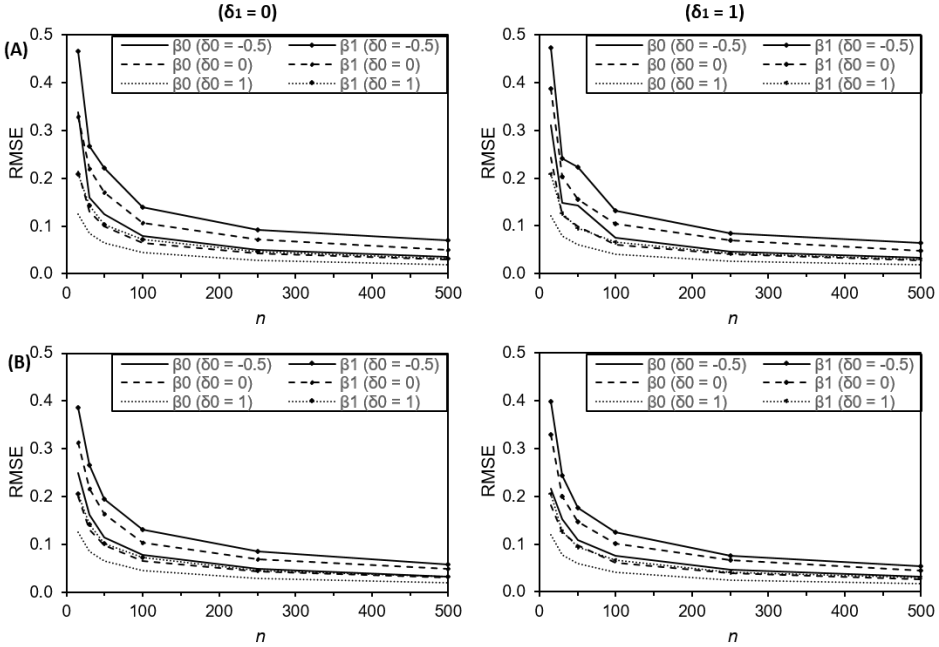

Figure A7. Root mean square error (RMSE) of regression parameter estimates under mean-free-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in Mean-parametrized Conway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

Appendix B.5. Properties of Dispersion Parameter Estimates under Mean-free-Dispersion

Tables 8 and 9 display the bias in dispersion parameter estimate under mean-free-dispersion specification, respectively, in balanced discrete gamma and Conway-Maxwell-Poisson samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

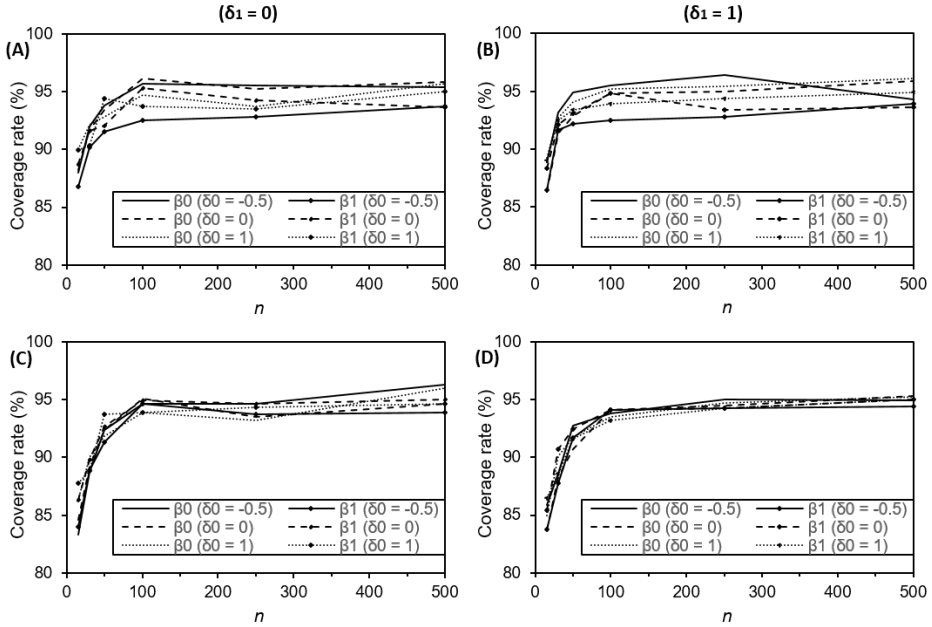

Figure A8. Coverage rate of asymptotic confidence interval for regression parameters under mean-free-dispersion balanced discrete gamma (**A**) and Mean-parametrized Conway-Maxwell-Poisson (**B**) regression model specifications in Conway-Maxwell-Poisson samples of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$.

Table 8: Bias in maximum likelihood estimates of dispersion parameters ($\hat{\delta}_0$ and $\hat{\delta}_1$) under mean-free-dispersion balanced discrete gamma regression model specification in balanced discrete gamma samples

δ_0	Est.	$\delta_1 = 0$						$\delta_1 = 1$					
		15	30	50	100	250	500	15	30	50	100	250	500
-0.5	$\hat{\delta}_0$	0.5672	0.3506	0.0858	0.0726	0.0436	0.0204	0.3844	0.3588	-0.0418	0.0650	0.0330	0.0196
	$\hat{\delta}_1$	0.0088	0.0237	-0.0059	0.0011	-0.0060	0.0021	0.1482	0.1372	0.0707	0.0366	0.0093	0.0046
0	$\hat{\delta}_0$	0.3747	0.1631	0.0931	0.0443	0.0212	0.0096	0.3515	0.1625	0.0742	0.0489	0.0192	0.0081
	$\hat{\delta}_1$	0.0346	0.0018	-0.0004	0.0162	-0.0014	0.0018	0.1632	0.1568	0.0898	0.0433	0.0104	0.0090
1	$\hat{\delta}_0$	0.4670	0.2241	0.1184	0.0609	0.0254	0.0123	0.4228	0.2439	0.1248	0.0561	0.0216	0.0128
	$\hat{\delta}_1$	-0.0026	0.0309	0.0231	-0.0017	0.0033	0.0003	0.1877	0.1641	0.0922	0.0260	0.0137	0.0123

Table note: Est. = Dispersion parameter estimate for which the bias is computed; the table displays the absolute bias if the true parameter value is zero (*e.g.* $(\hat{\delta}_0 - \delta_0)$ when $\delta_0 = 0$) and the relative bias if the true parameter value is not zero (*e.g.* $100 \times (\hat{\delta}_0 - \delta_0)/|\delta_0|$ when $\delta_0 \neq 0$).

Table 9: Bias in maximum likelihood estimates of dispersion parameters ($\hat{\delta}_0$ and $\hat{\delta}_1$) under mean-free-dispersion Mean-parametrized Conway-Maxwell-Poisson regression model specification in Conway-Maxwell-Poisson samples

δ_0	Est.	$\delta_1 = 0$						$\delta_1 = 1$					
		15	30	50	100	250	500	15	30	50	100	250	500
-0.5	$\hat{\delta}_0$	-0.0616	0.0762	0.0664	0.0764	0.0360	0.0184	-0.6324	0.1018	0.0326	0.0492	0.0244	0.0116
	$\hat{\delta}_1$	0.2370	-0.0064	0.0088	0.0038	0.0044	-0.0025	-0.1915	0.1826	0.1504	0.0413	0.0096	0.0068
0	$\hat{\delta}_0$	0.2008	0.1196	0.0903	0.0431	0.0167	0.0047	0.1776	0.0928	0.0558	0.0399	0.0066	0.0000
	$\hat{\delta}_1$	0.1575	-0.0252	-0.0189	-0.0058	-0.0009	0.0040	0.1650	0.1506	0.1366	0.0180	0.0118	0.0160
1	$\hat{\delta}_0$	0.2500	0.1441	0.0754	0.0425	0.0134	0.0076	0.2292	0.1331	0.0772	0.0477	0.0152	0.0100
	$\hat{\delta}_1$	0.0431	0.0097	-0.0018	-0.0038	0.0020	0.0026	0.0811	0.0893	0.0764	0.0334	0.0179	0.0049

Table note: Est. = Dispersion parameter estimate for which the bias is computed; the table displays the absolute bias if the true parameter value is zero (*e.g.* $(\hat{\delta}_0 - \delta_0)$ when $\delta_0 = 0$) and the relative bias if the true parameter value is not zero (*e.g.* $100 \times (\hat{\delta}_0 - \delta_0)/|\delta_0|$ when $\delta_0 \neq 0$).

Figures A9 and A10 present the RMSE and the CR for dispersion parameter estimates under mean-free-dispersion BDG and MCMP regerssion fits.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2021 doi:10.20944/preprints202104.0592.v1

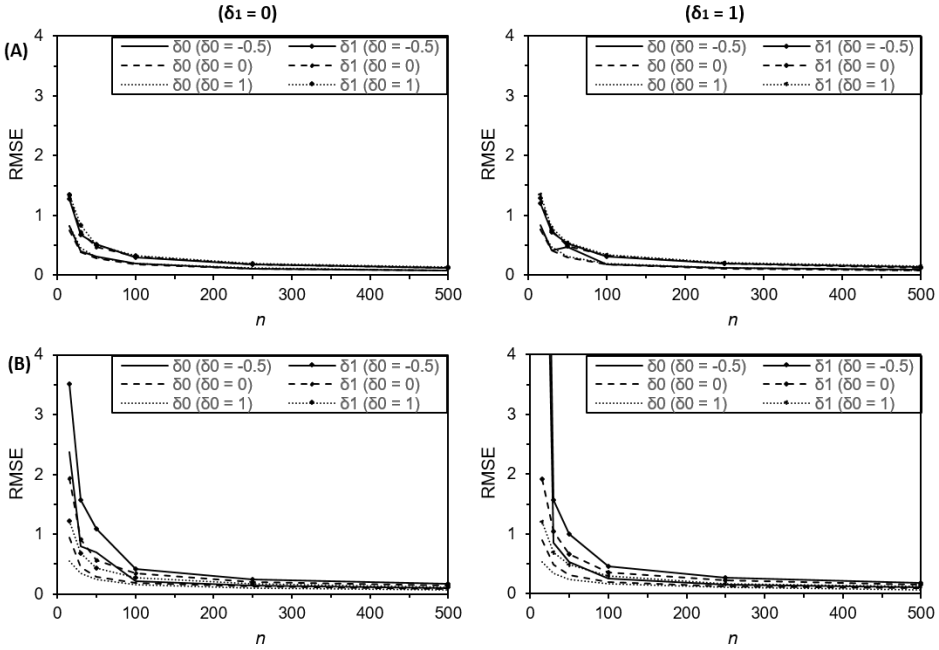

Figure A9. Root mean square error (RMSE) of dispersion parameter estimates under mean-free-dispersion balanced discrete gamma (BDG) regression fitted to BDG samples (**A**) and Mean-parametrized Conway-Maxwell-Poisson (CMP) regression fitted to CMP samples (**B**) of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$. The curves $\delta_0(\delta_0 = -0.5)$ and $\delta_1(\delta_0 = -0.5)$ on graphic (**B**, $\delta_1 = 1$) have been truncated for a better visualization: the respective RMSEs at $n = 15$ are 10.54 and 14.07.

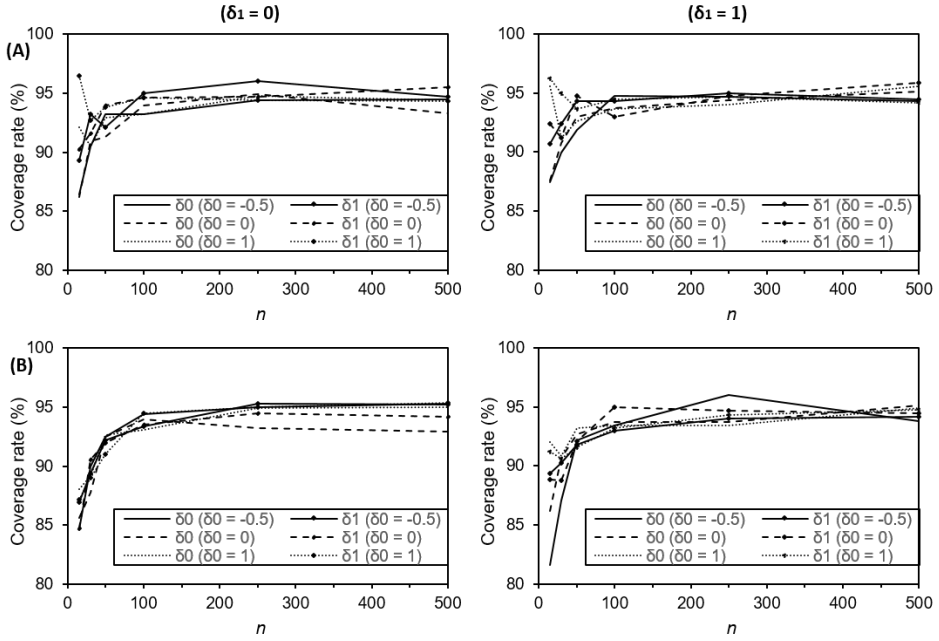

Figure A10. Coverage rate of asymptotic 95% confidence interval for dispersion parameter estimates under mean-free-dispersion balanced discrete gamma (BDG) regression fitted to BDG samples (**A**) and Mean-parametrized Conway-Maxwell-Poisson (CMP) regression fitted to CMP samples (**B**) of various sizes n in the range [15, 500] with dispersion model coefficients $\delta_0 = -0.5, 0, 1$ and $\delta_1 = 0, 1$

Appendix C. Summary Statistics for the Class Attendance Dataset

Table 10 presents some summary statistics for the class attendance dataset.

Table 10: Summary statistics for the class attendance dataset ($n = 314$)

Numerical variables						Categorical variables		
Variable	Min	Max	Median	Mean	SD	Variable	Category	Percentage (%)
Ndays	0	35	4	5.96	7.04	Gender	Female	50.96
Math	1	99	48	48.3	25.4		Male	49.04
						Programme	General	12.74
							Academic	53.18
							Vocational	34.08

Table notes: Ndays = number of days absent from high school; Math = standardized math score out of 100; Min = minimum; Max = maximum, SD = standard deviation.

References

1. Sellers, K.F.; Shmueli, G. A flexible regression model for count data. *The Annals of Applied Statistics* **2010**, pp. 943–961.
2. Zou, Y.; Geedipally, S.R.; Lord, D. Evaluating the double Poisson generalized linear model. *Accident Analysis & Prevention* **2013**, *59*, 497–505.
3. Sáez-Castillo, A.; Conde-Sánchez, A. A hyper-Poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis* **2013**, *61*, 148–157.
4. Zeviani, W.M.; Ribeiro Jr, P.J.; Bonat, W.H.; Shimakura, S.E.; Muniz, J.A. The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics* **2014**, *41*, 2616–2626.
5. Huang, A. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling* **2017**, *17*, 359–380.
6. Klakattawi, H.; Vinciotti, V.; Yu, K. A simple and adaptive dispersion regression model for count data. *Entropy* **2018**, *20*, 142.
7. Bonat, W.H.; Jørgensen, B.; Kokonendji, C.C.; Hinde, J.; Demétrio, C.G. Extended Poisson–Tweedie: properties and regression models for count data. *Statistical Modelling* **2018**, *18*, 24–49.
8. Sellers, K.F.; Premeaux, B. Conway–Maxwell–Poisson regression models for dispersed count data. *Wiley Interdisciplinary Reviews: Computational Statistics* **2020**, pp. 1–13.
9. Forthmann, B.; Gühne, D.; Doebler, P. Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology* **2020**, *73*, 32–50.
10. Hagmark, P.E. On construction and simulation of count data models. *Mathematics and Computers in Simulation* **2008**, *77*, 72–80.
11. Hilbe, J.M. *Negative binomial regression*; Cambridge University Press, 2011; p. 573.
12. Sellers, K.F.; Morris, D.S. Underdispersion models: Models that are "under the radar". *Communications in Statistics-Theory and Methods* **2017**, *46*, 12075–12086.
13. Wedderburn, R.W. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika* **1974**, *61*, 439–447.
14. Consul, P.; Famoye, F. Generalized Poisson regression model. *Communications in Statistics-Theory and Methods* **1992**, *21*, 89–109.
15. Tovissodé, C.F.; Honfo, S.H.; Doumatè, J.T.; Glèlè Kakaï, R. On the Discretization of Continuous Probability Distributions Using a Probabilistic Rounding Mechanism. *Mathematics* **2021**, *9*, 555.
16. Cameron, A.C.; Johansson, P. Count data regression using series expansions: with applications. *Journal of Applied Econometrics* **1997**, *12*, 203–223.
17. Huang, A.; Kim, A. Bayesian Conway–Maxwell–Poisson regression models for overdispersed and underdispersed counts. *Communications in Statistics-Theory and Methods* **2019**, pp. 1–12.
18. Conway, R.W.; Maxwell, W.L. A queuing model with state dependent service rates. *Journal of Industrial Engineering* **1962**, *12*, 132–136.
19. Johnson, N.L.; Kemp, A.W.; Kotz, S. *Univariate discrete distributions*; Vol. 444, John Wiley & Sons, 2005.
20. Ribeiro Jr, E.E.; Zeviani, W.M.; Bonat, W.H.; Demetrio, C.G.; Hinde, J. Reparametrization of COM–Poisson regression models with applications in the analysis of experimental data. *Statistical Modelling* **2020**, *20*, 443–466.
21. McCulloch, C.E.; Neuhaus, J.M. Generalized linear mixed models. *Encyclopedia of biostatistics* **2005**, *4*.
22. Nakagawa, T.; Osaki, S. The discrete Weibull distributiona. *IEEE* **1975**, *24*, 300–301.
23. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.
24. de Jong, P.; Heller, G.Z. *Generalized linear models for insurance data*; International series on actuarial science, Cambridge University Press, 2008.

 doi:10.20944/preprints202104.0592.v1

25. Neyman, J.; Pearson, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* **1928**, pp. 175–240.
26. Wilks, S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics* **1938**, *9*, 60–62.
27. Cameron, A.C.; Windmeijer, F.A. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics* **1996**, *14*, 209–220.
28. Fung, T.; Alwan, A.; Wishart, J.; Huang, A. *mpcmp: Mean-Parametrizied Conway-Maxwell Poisson Regression*, 2020. R package version 0.3.5.
29. Saez-Castillo, A.J.; Conde-Sanchez, A.; Martinez, F. *DGLMExtPois: Double Generalized Linear Models Extending Poisson Regression*, 2020. R package version 0.1.3.
30. Gasparini, L.C. On the measurement of unfairness An application to high school attendance in Argentina. *Social Choice and Welfare* **2002**, *19*, 795–810.
31. McCullagh, P.; Nelder, J. *Generalized linear models*; Chapman and Hall, 1989.
32. Dahiya, R.C.; Guttman, I. Shortest confidence and prediction intervals for the log-normal. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **1982**, pp. 277–291.