Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2021, 2021:55

https://doi.org/10.28919/cmbn/5658

ISSN: 2052-2541

PERFORMANCE OF ROBUST COUNT REGRESSION ESTIMATORS IN THE CASE OF OVERDISPERSION, ZERO INFLATED, AND OUTLIERS: SIMULATION STUDY AND APPLICATION TO GERMAN HEALTH DATA

MOHAMED R. ABONAZEL*, SAYED M. EL-SAYED, OMNIA M. SABER

Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University, Giza, Egypt

Abstract: This paper considers the count regression models in case of the dataset contains overdispersion and outliers. Seven robust and non-robust estimators are provided for four count regression (Poisson, negative binomial (NB), zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB)) models. The non-robust estimators were obtained by applying the maximum likelihood estimation on the four count models. While two robust estimators were obtained by applying the M-estimation on the Poisson and NB models (MP and MNB estimators), and the third robust estimator is the quantile regression of the count model (QRC estimator). Simulation study and empirical application were conducted to evaluate the performance and the efficiency for the robust and non-robust estimators of the four count regression models. The results showed that, in general, all robust estimators gave better performance than all non-robust estimators if the model contains outliers. And the QRC estimator reforms well even if the percent of the outlier values up to 25% when the sample size is large, dispersion value is small (less than or equal one). While when the dispersion value more than one, the MNB estimator is the efficient. The results of our application, which based on German health survey data in 1998, indicate that the significant variable that effect on the number of visits to doctor is the patient's condition (bad health or not in bad health), and the QRC estimator is the best for this data.

Keywords: count regression models; M-estimation; negative binomial regression; Poisson regression; zero inflated negative binomial regression; zero inflated Poisson regression; quantile regression.

2010 Subject Classification: 62J12, 62J99, 62P10.

*Corresponding author

E-mail address: mabonazel@cu.edu.eg.

Received March 8, 2021

1. INTRODUCTION

The most common probability distribution used to model count data is the Poisson distribution [1]. The Poisson distribution is favored because it accounts for the positive skewness inherent in count data, allows for zero counts, and has ease of use and interpretation. Count data often display substantial overdispersion with respect to the Poisson models. Overdispersion is occurs when the variance of the response variable is greater than the mean and may cause standard errors of the estimates to be deflated or underestimated, a variable may appear to be a significant predictor when it is in fact not significant. Therefore, the Poisson distribution is not very flexible; where it assumes equidispersion (i.e., the mean and variance are equal). On the other hand, the NB distribution employs an additional parameter that models overdispersion. That is, the negative binomial distribution as a Poisson (μ) distribution, where μ is itself a random variable that distributed as a gamma distribution. For a more discussion on both the Poisson and NB distributions, see, e.g., [1, 2, 3, 4, 5].

To deal with the overdispersion problem in count regression, some suggestions have been made in the statistical literature. Most of these suggestions are based on updating the count regression model itself or the estimation method used. Nelder and Wedderburn [6] suggested the use of the quasi-likelihood estimation method that define the relationship between mean and variance in the model. Ismail and Jemain [2] suggested Poisson-gamma mixture and generalized Poisson models as alternative models for Poisson model to deal with overdispersion problem. Also, Rahayu and Sadik [7] suggested zero inflated count regression models to detect zero inflated and overdispersion problems, they concluded that the ZINB model fits better than a standard ZIP in the presence of excess zeros and overdispersion problems.

It is well known that outlier values in the dataset can really mess up the analysis, since maximum likelihood (ML) estimators of generalized linear model (GLM) ate very sensitive to outliers [8, 9, 10]. Moreover, outliers may be a result of the overdispersion problem because outliers are observations which deviate from the common pattern of the data, see [11, 12, 13].

In several regression models, it is necessary to use robust regression or quantile regression to detect outliers and to provide resistant stable results in the presence of outliers, see, e.g., [14, 15, 16, 17, 18, 19, 20]. However, despite the fair amount of existing literature, robust inference for generalized linear models seems to be very limited [14, 21]. Specifically, robust estimators for Poisson regression model are proposed by [21, 22], the efficiency of the two robust estimators is studied by [14]. Machado and Santos Silva [23] developed the quantile regression estimator for count regression models. While the robust estimator for NB model is proposed by [24].

In this paper, we study the efficiency of robust and non-robust estimators for count regression models when the dataset contains overdispersion problem, many zeros, and outliers. To achieve this goal, a Monte Carlo simulation study and real data application have been performed.

This paper is organized as follows: Section 2 represents the four count regression (Poisson, NB, ZIP, and ZINB) models, their non-robust estimators, and how to choose the appropriate count regression model for any dataset. Three robust estimators are presented in Section 3. Section 4 shows the design and results of the simulation study. The results of our empirical application are discussed in Section 5. Section 6 offers some concluding remarks.

2. COUNT REGRESSION MODELS

2.1. Poisson Regression Model

The basic GLM for count data is the Poisson regression model with log link function, where the random component is specified by the Poisson distribution of the response variable which is a count. Consider the modeling framework of GLM where the response variable y_i, for $i = 1, \dots, n$ is drawn from a distribution belonging to the exponential family, such that $E(y_i|x_i) = Var(y_i|x_i) = \mu_i$ and linear predictor as: $g(\mu_i) = x_i^T\beta$, where $\beta = \left(\beta_1, \dots, \beta_p\right)^T \in \mathcal{R}^p$ is the vector of regression parameters, $x_i = \left(x_{i1}, \dots, x_{ip}\right) \in \mathcal{R}^p$ is a set of explanatory variables, and $g(\cdot)$ is the link function. Using the log link function, we can write the Poisson regression model in terms of the mean of the response, based on a sample $y_1, y_2, \dots, y_n$, as

$$y_i = \mu_i + \varepsilon_i = exp(x_i^T\beta) + \varepsilon_i; i = 1,2, \dots, n, \tag{1}$$

where ε_i is the error term. Then the log-likelihood function for n independent Poisson observations with probabilities given in the density function of Poisson distribution is [25]:

$$logL(\beta) = \sum_{i=1}^{n}\{y_i log(\mu_i) - \mu_i - log(y_i!)\}. \tag{2}$$

To get the ML estimates for the model in (1), we are maximizing the log-likelihood respect to β. However, maximizing the log-likelihood has not closed-form solution, so numerical search procedures (such as Fisher scoring [26]) are used to find the ML estimates. We will refer to the ML estimator of the Poisson model as the MLP estimator.

2.2. Negative Binomial Regression Model

The NB regression model is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model and an alternative approach to modeling overdispersion in count data is to start from a Poisson regression model and add a dispersion parameter $\alpha = \frac{1}{\phi}$. NB distribution can be viewed as a Poisson distribution according to a Gamma distribution. Thus, the NB distribution is known as a Poisson-Gamma mixture with the following formula [25]:

$$f(y_i;\mu_i,\alpha) = \frac{\Gamma(y_{i+}\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}, \quad y_i = 0,1,..n.$$

With $E(y_i)$ = μ_i and variance is $Var(y)$ = $\mu_i + \alpha^{-1}\mu_i^2 = \mu_i(1+\phi\mu_i)$.

This model is attractive because it manages to handle data that is overdispersed since it allows for random variation in the Poisson conditional mean by letting:

$$E(y_i|x_i) = Z_i\mu_i = Z_i exp\,(x_i^T\beta),\ i = 1,2,\dots,n,$$

where Z_i a random variable that is gamma distributed. This extra parameter α in the variance expression allows us to construct a more accurate model for certain count data, since now the mean and the variance do not need to be equal.

As in Poisson model, the regression coefficients of NB model are estimated using the method of ML, by maximization the following log-likelihood function of NB model:

$$logL(\alpha,\beta) = \sum_{i=1}^{n}\left\{\begin{matrix}\left(\sum_{j=0}^{y_i-1} log(j+\alpha^{-1})\right) - log[\Gamma(y_i+1)] - (y_i+\alpha^{-1}) \\ log\left(1+\alpha\, exp(x_i^T\beta)\right) + y_i log(\alpha) + y_i\end{matrix}\right\}, \qquad (3)$$

where $log\left(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right) = \sum_{j=0}^{y_i-1} ln(j+\alpha^{-1})$ and $\mu_i = exp(x_i^T\beta)$. We will refer to the ML estimator of the NB model as the MLNB estimator.

2.3. Zero Inflated Poisson Model

A particular kind of overdispersion obtains when there are more zeros in the data than is consistent with a Poisson distribution, several statistical models have been proposed for count data with an excess of zeros, including the ZIP model introduced by [27]. It assumes that the sample is a mixture of two sorts of individuals: one group whose counts are generated by the standard Poisson regression model, and another group call them the absolute zero group who have zero probability of a count greater than zero.

In cases of overdispersion, the ZIP model typically fits better than a standard Poisson model [7]. The ZIP model consists of two components:

1. A binary logistic regression model for membership in the latent class of individuals for whom the response variable is necessarily zero.
2. A Poisson regression model for the latent class of individuals for whom the response may be or a positive count.

Suppose that case 1 occurs with probability π and case 2 occurs with probability $1-\pi$. Therefore, the probability distribution of the ZIP random variable y_i could be written:

$$f(y_i;\mu_i,\pi_i)=\begin{cases}\pi_i+(1-\pi_i)exp(-\mu_i) & if\ y_i=0;\\ (1-\pi_i)\left(\frac{\mu_i^{y_i}}{y_i!}\right)exp(-\mu_i) & if\ y_i>0,\end{cases} \tag{4}$$

where $\mu_i \geq 0, 0 \leq \pi_i \leq 1$, and the mean and variance for ZIP are $E(Y)=(1-\pi)\mu$; $Var(Y)=\mu(1-\pi)(1+\mu\pi)$. And $\pi_i=\frac{exp(z_i^T\gamma)}{1+exp(z_i^T\gamma)}$ $i=1,2,\dots n$, where z_i is the i^{th} row of Z which is the matrix for the logit model, and γ corresponds to a $(p_0\times 1)$ vector of coefficients. The model allows μ_i and $\pi_i$to depend on covariates through the relationships: $g(\pi_i)=log\left(\frac{\pi_i}{1-\pi_i}\right)=z_i^T\gamma$; $h(\mu_i)=x_i^T\beta$.

The regression coefficients are estimated using the method of maximum likelihood, where the log-likelihood function of ZIP model is defined as:

$$logL(\gamma,\beta)=\sum_{y_i=0}log\{exp(z_i^T\gamma)+exp[-exp(x_i^T\beta)]\}+\sum_{y_i>0}[y_ix_i^T\beta-exp(x_i^T\beta)-log(y_i!)]-\sum_{i=1}^{n}log[1+exp(z_i^T\gamma)]. \tag{5}$$

2.4. Zero Inflated Negative Binomial Model

The ZINB model is the most popular models to handle excess zeros and overdispersion problems introduced by [3]. It is formed of Poisson Gamma mixture distribution and has a dispersion parameter (α) that useful to describe the variation of the data. It assumes that there are two distinct data generation processes, the result of a Bernoulli trial is used to determine which of the two processes is used. For observation i, with probability φ_i the only possible response of the first process is zero counts, and with probability of $(1-\varphi_i)$ the response of the second process is governed by a negative binomial with mean φ_i. The ZINB distribution can be defined as:

$$f(y_i;\mu_i,\varphi_i,\alpha)=\begin{cases}\varphi_i(1-\varphi_i)\left(\frac{\alpha}{\alpha+\mu_i}\right)^{\alpha} & y_i=0;\\ (1-\varphi_i)\,\frac{\Gamma(y_i+\alpha)}{\Gamma(y_i+1)\Gamma(\alpha)}\left(\frac{\alpha}{\alpha+\mu_i}\right)^{\alpha}\left(\frac{\mu_i}{\alpha+\mu_i}\right)^{y_i} & y_i>0,\end{cases} \tag{6}$$

where $0\le\varphi_i\le 1$, $\mu_i\ge 0$, and $\Gamma(\cdot)$ is the gamma function. The mean and variance for the ZINB distribution are: $E(Y)=(1-\varphi)\mu$; $\text{Var }(Y)=(1-\varphi)[1+\mu(\varphi+\alpha)]$. And $\varphi_i=\frac{\exp(z_i^T\gamma)}{1+\exp(z_i^T\gamma)}$. As in ZIP model, the ZINB model allows μ_i and φ_i to depend on covariates through the relationships:

$$g(\varphi_i)=log\left(\frac{\varphi_i}{1-\varphi_i}\right)=z_i^T\gamma;\ h(\mu_i)=\log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)=x_i^T\beta.$$

The regression coefficients are estimated using the ML method. The log-likelihood function is given by:

$$logL(\gamma,\alpha,\beta)=\sum_{y_i=0}\log\left\{\varphi_i(1-\varphi_i)\left(\frac{\alpha}{\alpha+\mu_i}\right)^{\alpha}\right\}+\log\sum_{y_i>0}\Bigg[(1-\varphi_i)\left(\frac{\Gamma(y_i+\alpha)}{\Gamma(y_i+1)\Gamma(\alpha)}\left(\frac{\alpha}{\alpha+\mu_i}\right)^{\alpha}\left(\frac{\mu_i}{\alpha+\mu_i}\right)^{y_i}\right)\Bigg]. \tag{7}$$

Rahayu and Sadik [7] showed that the ZINB model fits better than the ZIP model in the presence of excess zeros and overdispersion problems.

Figure 1 shows how to choose the appropriate count regression model in the case of overdispersion and/or zero inflation.

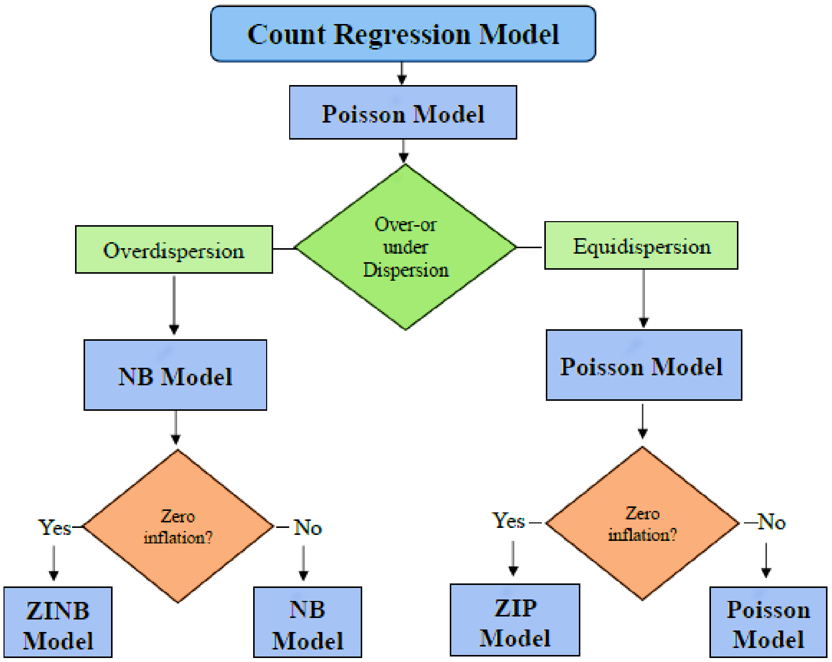

Figure 1: Selection of the appropriate count regression model.

3. ROBUST COUNT REGRESSION ESTIMATIONS

3.1. Robust Estimation of Poisson Model

In literature, there is no specific robust estimation for Poisson regression model. However, the general robust estimation methods for GLM can be applied to Poisson model [21, 28].

Cantoni and Ronchetti [21] developed a robust estimation based on robust deviances that are natural generalization of quasi-likelihood functions, consider a general class of M-estimators of Mallows's type, where the influence of deviations on y and X are bounded separately. Then the robust estimator of the Poisson model (MP) is given by solving the following equations [14]:

$$\sum_{i=1}^{n}\left[\psi_c(r_i)w(x_i)\frac{\mu_i'}{\sqrt{Var(\mu_i)}}-\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\psi_c(r_i)]\,w(x_i)\frac{\mu_i'}{\sqrt{Var(\mu_i)}}\right]=0,$$

where $\mu_i' = \frac{\partial}{\partial\beta}\mu_i$, $r_i = \frac{y_i-\mu_i}{\sqrt{Var(\mu_i)}}$ are the Pearson residuals, $w(x_i)=\sqrt{1-h_i}$; h_i is the ith diagonal element of the hat matrix, and $\psi_c(\cdot)$ is the Huber function that defined by:

$$\psi_c(r_i)=\begin{cases} r_i & |r_i|\le c; \\ c\ sign(r_i) & |r_i|>c, \end{cases}$$

where the constant c is typically chosen to ensure a given level of asymptotic efficiency [21].

3.2. Robust Estimation of Negative Binomial Model

Following [21, 29], Aeberhard et al. [24] developed a M-estimator of the NB model (MNB), that yielding a robust estimator for β which achieve robustness on one hand in the response by bounding the Pearson residual that is the first appears naturally in the score function $\Psi_\beta(y_i,x_i,\beta,\alpha)$ and on the other hand on the design by introducing weights $w(x_i)$. Then the MNB estimator of β is given by solving the following equations:

$$\sum_{i=1}^{n}\frac{\psi_c(r_i)}{r_i}\Psi_\beta(y_i,x_i,\beta,\alpha)w(x_i)-a_i(\beta)=0, \tag{8}$$

where $\frac{\psi_c(r_i)}{r_i}\in[0,1]$, $\Psi_\beta(y_i,x_i,\beta,\alpha)=\frac{y_i-\mu_i}{\sqrt{Var(\mu_i)}}\ \mu_i'\ x_i$ and a Fisher consistency correction term $a_i(\beta)=E\left[\frac{\psi_c(r_i)}{r_i}\Psi_\beta(y_i,x_i,\beta,\alpha)w(x_i)\right]$.

For α, the MNB estimator for α is given by solving the following equations:

$$\sum_{i=1}^{n}\frac{\psi_c(r_i)}{r_i}\Psi_\alpha(y_i,x_i,\beta,\alpha)w(x_i)-b_i(\alpha)=0, \tag{9}$$

where $\Psi_\alpha(y_i, x_i, \beta, \alpha) = \left(\frac{-1}{\alpha^2}\right)\left(F(y_i + 1/\alpha) - F(1/\alpha) - log(\alpha\mu_i + 1) - \frac{\alpha(y_i - \mu_i)}{\alpha\mu_i + 1}\right)$; $F(u) = \frac{\partial log\,\Gamma(u)}{\partial u}$ denotes the digamma function, and $b_i(\alpha) = E\left[\frac{\psi_c(r_i)}{r_i}\Psi_\alpha(y_i, x_i, \beta, \alpha)w(x_i)\right]$ is another Fisher consistency term. Equations in (8) and (9) can be solved using Newton-Raphson or Fisher scoring algorithms to get the MNB estimator.

3.3. Quantile Regression Estimation of Count Models

Quantile regression is pure ducts of robust statistic. The basic idea is to estimate the conditional quantile of an outcome y given a vector of covariates x_i defined as for any pre-specified level $0 \le q \le 1$. Then the qth regression quantile is defined as any solution to the minimization problem [30]:

$$\sum_{y_i \ge x_i^T\beta} q\left|y_i - x_i^T\beta\right| + \sum_{y_i < x_i^T\beta}(1 - q)\left|y_i - x_i^T\beta\right|. \tag{10}$$

The sample median is the minimizer of the sum of absolute deviations. The model estimates the relationship between the qth quantile of a response distribution and the regression parameters, it was originally developed for continuous responses as count responses do not have continuous quantiles.

Quantile regression model for count is used when we simply cannot obtain a reasonably fitted Poisson, negative binomial, and zero count models, because mean regression models may be sensitive to response outliers and provide no information on factors affecting other distributional points (e.g., upper and lower 5% quantiles) of the response due to some data problems.

Researchers have attempted to design quantile count models according to two different methods. The first method is based on [31]. They estimated quantiles based on a semi-parametric modeling of the conditional mean of the count response, using a pseudo-likelihood algorithm. The problem with this approach is that the full range of the count distribution cannot be understood based on the predictors. The second general method was offered by [23], they suggested an artificial smoothing by adding a uniformly distributed noise to the count data. The act of smoothing the data is called ‘jittering’, which will convert the count data to a continuous variable that has a one-to-one relationship with the conditional quantiles of the counts. Machado and Santos Silva [23] replaced the response variable y with a jittered response variable:

$$Z = y + U, \tag{11}$$

Where $U\sim uniform[0,1)$, then Z is linearized at the conditional mean of each quantile as $exp(X^T\beta_q)$ and apply quantile regression of the form $Q_Z = (q|X) = exp\left(X^T\beta_q\right)$.

Machado and Santos Silva [23] establishing assumptions for deriving the approximate distribution of the quantile regression estimator for count model (QRC) as:

1. y is a discrete random variable and nonnegative integers, and x_i is a random vector in $\mathcal{R}^p$; the conditional probability function of y given x_i at $Q_y(q|x_i)$; $f_{y|x_i}\left(Q_y(q|x_i)\right)$ is uniformly bounded away from 0 for almost every realization of x_i.
2. $E(XX^T)$ is finite and nonsingular matrix, where $X = \left(x_1, \ldots, x_p\right)^T$.
3. With the achievement of Equation (11), for some known monotone transformation $T(\cdot;q)$, possibly depending on $q \in (0,1)$, so the following restriction on the quantile process of Z given X holds:

$$Q_{T(Z;q)}(q|X) = X^T\gamma(q) \tag{12}$$

where $\gamma(q) \in \Gamma$, a compact subset of $\mathcal{R}^p$.

Machado and Santos Silva [23] defined the QRC estimator $(\hat{\gamma}(q))$, for random sample of (Y, X, U), by:

$$\min_{c\in\mathcal{R}^p} \sum_{i=1}^{n} \rho_q(T(z_i;q) - x_i^T c), \tag{13}$$

where $\rho_q(\nu) = \nu\left(q - I(\nu < 0)\right)$. For more details about the properties of the QRC estimator, see [23].

4. Simulation Study

In this section, we investigate the performance of the above-mentioned estimators through a Monte Carlo simulation study.

4.1. Simulation Design

The simulation experiment has been designed to compare the performance of non-robust (MLP, MLNB, ZIP and ZINB) and robust estimators (MP, QRC and MNB) for different sample sizes, dispersion values, and percentages of outliers. R software is used to perform our Monte Carlo simulation study. For additional information on how to make a Monte Carlo simulation study using R, see [18, 32]. In the simulation study, the effective factors are chosen to be the dispersion values, the sample size, and the percentages of outliers. The response variable is generated using random numbers following the NB distribution with different dispersion parameters. The simulated model is carried with the following simulation settings:

1. The number of independent variables is two, where the independent variables are generated from uniform (-1, 1), and the value of true vector is one, i.e., $\beta = 1$.
2. The values of sample sizes were chosen to be 100, 200, 300,400, 500, and 1000 to represent moderate and large samples.
3. The percentages of outliers in the response variable, O%, were chosen to be $0\%, 5\%, 10\%, 15\%, 20\%$ and 25%. Following the work of Abonazel and Saber [14], the outliers generated from Poisson distribution with mean equal to $4 \times IQR(\mu)$; where IQR is the interquartile range.
4. The dispersion parameter, α , is chosen to be $0.5, 1,$ and 2.
5. The percentages of zeros (ZI%) in the response variable were chosen to be 0% and 50%.
6. Experiments are conducted on 500 repeats and all the results for the separate experiments are using the same series of random numbers.

To compare the performance of non-robust and robust estimators, the average of mean squared error (MSE) and mean absolute error (MAE) of each estimation are computed:

$$MSE = \frac{1}{500}\sum_{l=1}^{500}\left(\hat{\beta}_l - \beta\right)^2 ; \; MAE = \frac{1}{500}\sum_{l=1}^{500}\left|\hat{\beta}_l - \beta\right|, \quad (14)$$

where $\hat{\beta}_l$ is the vector of estimated values at l^{th} experiment of 500 Monte Carlo experiments, while β is the vector of true coefficients.

4.2. Simulation Results

The results of the Monte Carlo simulation study have been provided in Tables 1 to 8. Specifically, Tables 1 to 5 presented the MSE and MAE values of all estimators (non-robust and robust) when the percentage of zeros in the response variable is zero (ZI = 0%), while the simulation results in the case of the zero inflated (ZI = 50%) are presented in Tables 6 to 8. We can summarize the simulation resalts as follows.

Generally, as n increases, MSE and MAE values of all estimators (non-robust and robust) decrease for different values of O%, ZI%, and α.

Table 1: Simulation resalts of robust and non-robust estimators when $n = 100, \mathrm{ZI} = 0\%$

Estimator	**MSE**						**MAE**					
	O%						**O%**					
	0	**5**	**10**	**15**	**20**	**25**	**0**	**5**	**10**	**15**	**20**	**25**
$\boldsymbol{\alpha = 0.5}$												
MLP	0.089	0.437	0.906	1.412	1.820	2.292	0.397	0.966	1.470	1.881	2.167	2.462
ZIP	0.107	0.787	1.559	2.230	2.722	3.256	0.456	1.334	1.977	2.423	2.702	2.984
MLNB	0.070	0.483	0.978	1.455	1.852	2.312	0.354	1.008	1.521	1.896	2.175	2.466
ZINB	0.071	0.485	0.979	1.456	1.854	2.312	0.358	1.010	1.522	1.896	2.176	2.466
QRC	0.238	0.187	0.166	0.155	0.159	0.191	0.696	0.611	0.563	0.533	0.536	0.598
MP	0.119	0.109	0.109	0.134	0.179	0.263	0.476	0.449	0.441	0.498	0.601	0.757
MNB	0.086	0.097	0.107	0.127	0.181	0.312	0.390	0.419	0.448	0.490	0.591	0.760
$\boldsymbol{\alpha = 1}$												
MLP	0.151	0.435	0.759	1.198	1.608	2.011	0.519	0.932	1.308	1.689	2.000	2.278
ZIP	0.244	0.924	1.522	2.107	2.661	3.126	0.716	1.435	1.926	2.312	2.634	2.894
MLNB	0.111	0.452	0.809	1.228	1.647	2.016	0.453	0.933	1.344	1.698	2.019	2.270
ZINB	0.116	0.459	0.806	1.228	1.651	2.024	0.468	0.945	1.346	1.699	2.022	2.275
QRC	0.525	0.442	0.390	0.342	0.280	0.309	1.054	0.949	0.897	0.820	0.726	0.752
MP	0.268	0.239	0.230	0.244	0.237	0.316	0.741	0.691	0.668	0.660	0.635	0.766
MNB	0.162	0.169	0.174	0.222	0.268	0.420	0.534	0.541	0.566	0.652	0.713	0.910
$\boldsymbol{\alpha = 2}$												
MLP	0.267	0.469	0.844	1.140	1.538	1.906	0.682	0.932	1.334	1.590	1.898	2.138
ZIP	0.522	1.151	1.860	2.347	2.892	3.304	1.070	1.603	2.096	2.394	2.683	2.890
MLNB	0.194	0.452	0.822	1.132	1.539	1.924	0.592	0.911	1.314	1.579	1.892	2.144
ZINB	0.210	0.471	0.841	1.144	1.563	1.941	0.625	0.942	1.336	1.593	1.909	2.155
QRC	1.413	1.239	1.066	1.063	0.915	0.844	1.685	1.597	1.491	1.490	1.383	1.323
MP	0.746	0.664	0.604	0.599	0.553	0.576	1.266	1.199	1.148	1.121	1.061	1.058
MNB	0.319	0.345	0.343	0.414	0.488	0.631	0.751	0.780	0.788	0.872	0.952	1.104

Table 2: Simulation resalts of robust and non-robust estimators when $n = 200, \mathrm{ZI} = 0\%$

Estimator	**MSE**						**MAE**					
	O%						**O%**					
	0	**5**	**10**	**15**	**20**	**25**	**0**	**5**	**10**	**15**	**20**	**25**
$\boldsymbol{\alpha = 0.5}$												
MLP	0.042	0.239	0.578	0.971	1.334	1.697	0.271	0.737	1.213	1.614	1.911	2.166
ZIP	0.064	0.515	1.089	1.680	2.133	2.535	0.358	1.125	1.707	2.162	2.452	2.678
MLNB	0.035	0.276	0.639	1.033	1.382	1.744	0.249	0.788	1.272	1.658	1.942	2.196
ZINB	0.036	0.276	0.639	1.033	1.382	1.744	0.251	0.789	1.272	1.658	1.942	2.196
QRC	0.158	0.121	0.090	0.078	0.078	0.112	0.582	0.503	0.422	0.384	0.379	0.462
MP	0.065	0.053	0.067	0.083	0.126	0.200	0.361	0.316	0.340	0.400	0.522	0.685
MNB	0.042	0.045	0.057	0.071	0.103	0.176	0.272	0.288	0.329	0.378	0.462	0.596
$\boldsymbol{\alpha = 1}$												
MLP	0.073	0.339	0.757	1.202	1.623	2.062	0.363	0.858	1.363	1.771	2.081	2.361
ZIP	0.153	0.852	1.573	2.206	2.743	3.236	0.576	1.435	2.026	2.448	2.743	2.991
MLNB	0.055	0.385	0.831	1.275	1.694	2.111	0.322	0.905	1.425	1.816	2.129	2.390
ZINB	0.057	0.386	0.831	1.275	1.694	2.111	0.327	0.907	1.425	1.816	2.129	2.390
QRC	0.377	0.306	0.238	0.188	0.168	0.150	0.875	0.798	0.702	0.624	0.578	0.519
MP	0.166	0.137	0.127	0.132	0.155	0.201	0.592	0.537	0.499	0.491	0.528	0.626
MNB	0.076	0.075	0.097	0.118	0.156	0.227	0.369	0.365	0.424	0.479	0.560	0.678
$\boldsymbol{\alpha = 2}$												
MLP	0.131	0.340	0.723	1.045	1.370	1.740	0.482	0.818	1.269	1.596	1.866	2.130
ZIP	0.392	1.085	1.807	2.309	2.755	3.197	0.937	1.599	2.121	2.449	2.695	2.925
MLNB	0.097	0.361	0.753	1.088	1.391	1.754	0.422	0.836	1.299	1.625	1.872	2.132
ZINB	0.106	0.369	0.754	1.091	1.392	1.755	0.450	0.851	1.301	1.628	1.873	2.133
QRC	1.136	0.995	0.853	0.741	0.647	0.564	1.476	1.387	1.305	1.225	1.156	1.090
MP	0.536	0.488	0.446	0.414	0.394	0.387	1.060	1.027	0.996	0.968	0.935	0.892
MNB	0.157	0.168	0.187	0.197	0.268	0.342	0.536	0.546	0.590	0.612	0.732	0.836

In case of the percentage of outliers equal to zero ($\mathrm{O} = 0\%$), the non-robust estimators are more efficient than robust estimators. Specifically, the MLNB estimator has the smallest values of MSE and MAE, compared to other estimators, if $\mathrm{O} = \mathrm{ZI} = 0\%$ and any value of α and n (as shown in Tables 1 to 5). However, when $\mathrm{ZI} = 0\%$ and $\mathrm{O} \geq 10\%$, the non-robust estimators are extremely sensitive to the presence of outliers, this is confirmed by the simulation results; where all non-robust estimators have the largest values of MSE and MAE, on the other hand MSE and

MAE values of all robust estimators is the smallest. Specifically, the best two robust estimators are QRC and MNB.

But if $\mathrm{ZI} = 50\%$ and $\mathrm{O} = 0\%$ (as shown in Tables 6 to 8), the ZINB is the best estimator. However, when $\mathrm{ZI} = 50\%$ and the percentage of outliers increases ($\mathrm{O} \geq 10\%$ and $\alpha > 1$), the best robust estimator is MNB.

Table 3: Simulation resalts of robust and non-robust estimators when $n = 300, \mathrm{ZI} = 0\%$

Estimator	**MSE**						**MAE**					
	O%						**O%**					
	0	**5**	**10**	**15**	**20**	**25**	**0**	**5**	**10**	**15**	**20**	**25**
$\boldsymbol{\alpha = 0.5}$												
MLP	0.028	0.232	0.590	0.973	1.401	1.744	0.224	0.742	1.247	1.632	1.975	2.209
ZIP	0.050	0.494	1.106	1.664	2.205	2.591	0.326	1.122	1.741	2.164	2.505	2.718
MLNB	0.024	0.290	0.671	1.047	1.481	1.814	0.206	0.825	1.328	1.693	2.033	2.258
ZINB	0.024	0.290	0.671	1.047	1.481	1.814	0.208	0.826	1.328	1.693	2.033	2.258
QRC	0.144	0.099	0.064	0.050	0.057	0.082	0.552	0.454	0.358	0.306	0.322	0.403
MP	0.052	0.041	0.045	0.069	0.113	0.183	0.333	0.278	0.285	0.374	0.506	0.671
MNB	0.029	0.032	0.038	0.051	0.084	0.135	0.233	0.244	0.267	0.327	0.421	0.522
$\boldsymbol{\alpha = 1}$												
MLP	0.049	0.309	0.662	1.124	1.527	1.924	0.293	0.841	1.301	1.742	2.043	2.303
ZIP	0.135	0.821	1.450	2.107	2.611	3.042	0.544	1.442	1.978	2.419	2.700	2.918
MLNB	0.037	0.364	0.746	1.207	1.586	1.990	0.259	0.908	1.379	1.805	2.082	2.346
ZINB	0.038	0.365	0.746	1.207	1.586	1.990	0.267	0.909	1.380	1.805	2.082	2.346
QRC	0.348	0.253	0.193	0.137	0.125	0.112	0.846	0.727	0.634	0.532	0.502	0.453
MP	0.140	0.118	0.094	0.102	0.136	0.178	0.546	0.505	0.435	0.434	0.503	0.605
MNB	0.049	0.057	0.063	0.081	0.133	0.197	0.298	0.322	0.350	0.406	0.524	0.634
$\boldsymbol{\alpha = 2}$												
MLP	0.088	0.285	0.658	1.001	1.417	1.761	0.392	0.767	1.257	1.590	1.935	2.178
ZIP	0.348	1.029	1.751	2.283	2.839	3.252	0.875	1.572	2.124	2.450	2.763	2.970
MLNB	0.064	0.324	0.727	1.069	1.487	1.810	0.341	0.812	1.323	1.643	1.987	2.211
ZINB	0.069	0.326	0.728	1.070	1.487	1.810	0.358	0.818	1.325	1.644	1.987	2.211
QRC	1.036	0.881	0.729	0.632	0.525	0.426	1.367	1.282	1.162	1.101	1.026	0.940
MP	0.478	0.423	0.371	0.350	0.328	0.315	0.994	0.948	0.914	0.896	0.865	0.823
MNB	0.101	0.114	0.111	0.150	0.187	0.273	0.430	0.446	0.461	0.538	0.615	0.754

If the dispersion parameter is less than or equal one ($\alpha \leq 1$) and the percentage of outliers is more than or equal twenty ($\text{O} \geq 20\%$ and $\text{ZI} = 0\%$), the QRC estimator give the smallest values of MSE and MAE, especially when the sample size is increased ($n > 200$). On the other hand, the MNB estimator has the smallest values of MSE and MAE when the percentage of outliers is less than twenty ($O < 20\%$ and $\text{ZI} = 0\%$) for different values of $n\%$ and α.

Table 4: Simulation resalts of robust and non-robust estimators when $n = 400, \text{ZI} = 0\%$

Estimator	**MSE**						**MAE**					
	O%						**O%**					
	0	**5**	**10**	**15**	**20**	**25**	**0**	**5**	**10**	**15**	**20**	**25**
$\boldsymbol{\alpha = 0.5}$												
MLP	0.020	0.240	0.613	1.033	1.434	1.814	0.189	0.775	1.291	1.699	2.007	2.257
ZIP	0.039	0.529	1.169	1.774	2.278	2.696	0.284	1.185	1.807	2.249	2.552	2.775
MLNB	0.016	0.290	0.698	1.121	1.513	1.885	0.173	0.848	1.379	1.775	2.066	2.307
ZINB	0.016	0.290	0.698	1.121	1.513	1.885	0.174	0.848	1.379	1.775	2.066	2.307
QRC	0.130	0.088	0.055	0.041	0.043	0.069	0.523	0.433	0.338	0.280	0.275	0.377
MP	0.042	0.032	0.038	0.064	0.104	0.172	0.299	0.248	0.259	0.363	0.495	0.655
MNB	0.020	0.023	0.030	0.045	0.068	0.115	0.191	0.207	0.244	0.305	0.376	0.479
$\boldsymbol{\alpha = 1}$												
MLP	0.036	0.291	0.700	1.143	1.575	2.011	0.251	0.831	1.357	1.765	2.086	2.368
ZIP	0.117	0.813	1.532	2.160	2.699	3.175	0.513	1.457	2.050	2.455	2.752	2.990
MLNB	0.027	0.348	0.774	1.233	1.655	2.076	0.221	0.905	1.426	1.839	2.144	2.411
ZINB	0.028	0.348	0.774	1.233	1.655	2.076	0.225	0.905	1.426	1.839	2.144	2.411
QRC	0.314	0.241	0.173	0.122	0.095	0.102	0.789	0.703	0.591	0.508	0.443	0.432
MP	0.124	0.099	0.082	0.089	0.111	0.168	0.511	0.466	0.415	0.408	0.449	0.595
MNB	0.037	0.040	0.047	0.072	0.109	0.169	0.256	0.273	0.305	0.387	0.477	0.589
$\boldsymbol{\alpha = 2}$												
MLP	0.064	0.275	0.611	0.999	1.389	1.764	0.334	0.774	1.227	1.623	1.931	2.199
ZIP	0.320	1.052	1.701	2.314	2.841	3.281	0.843	1.613	2.100	2.496	2.773	2.997
MLNB	0.045	0.317	0.675	1.068	1.458	1.837	0.284	0.827	1.290	1.679	1.982	2.251
ZINB	0.049	0.318	0.676	1.068	1.458	1.837	0.300	0.830	1.290	1.679	1.982	2.251
QRC	0.981	0.836	0.704	0.581	0.493	0.410	1.305	1.215	1.132	1.045	0.981	0.926
MP	0.451	0.390	0.355	0.306	0.300	0.297	0.949	0.906	0.901	0.845	0.839	0.817
MNB	0.072	0.075	0.093	0.117	0.165	0.226	0.362	0.375	0.426	0.483	0.583	0.693

It notes that when $\text{ZI} = 0\%$, as in Tables 1 to 5, the ZIP and ZINB estimators have larger values of MSE and MAE than MLP and MLNB, respectively, even if the percentage of outliers increases. This means that the ZIP and ZINB estimators are not suitable as robust estimators if the data nonzero inflated. But if $\text{ZI} = 50\%$ and $\text{O} = 0\%$, as in Tables 6 to 8, the ZIP and ZINB estimators are more efficiency than MLP and MLNB, respectively, even if α up to 2.

Table 5: Simulation resalts of robust and non-robust estimators when $n = 500, \text{ZI} = 0\%$

Estimator	**MSE**						**MAE**					
	O%						**O%**					
	0	**5**	**10**	**15**	**20**	**25**	**0**	**5**	**10**	**15**	**20**	**25**
$\boldsymbol{\alpha = 0.5}$												
MLP	0.016	0.219	0.590	1.001	1.372	1.755	0.167	0.746	1.277	1.677	1.968	2.227
ZIP	0.034	0.499	1.127	1.708	2.181	2.598	0.272	1.156	1.785	2.212	2.502	2.731
MLNB	0.013	0.276	0.683	1.088	1.456	1.830	0.154	0.836	1.376	1.752	2.034	2.281
ZINB	0.013	0.276	0.683	1.088	1.456	1.830	0.156	0.836	1.376	1.752	2.034	2.281
QRC	0.119	0.078	0.047	0.034	0.038	0.062	0.504	0.406	0.315	0.255	0.261	0.359
MP	0.036	0.029	0.033	0.059	0.096	0.161	0.276	0.238	0.236	0.354	0.484	0.645
MNB	0.016	0.020	0.026	0.040	0.066	0.115	0.170	0.190	0.227	0.290	0.365	0.473
$\boldsymbol{\alpha = 1}$												
MLP	0.029	0.267	0.653	1.093	1.533	1.904	0.226	0.797	1.316	1.733	2.065	2.304
ZIP	0.121	0.763	1.453	2.072	2.625	3.022	0.527	1.414	2.002	2.411	2.720	2.917
MLNB	0.022	0.315	0.726	1.179	1.608	1.965	0.198	0.866	1.391	1.806	2.121	2.345
ZINB	0.023	0.316	0.726	1.180	1.608	1.965	0.204	0.866	1.391	1.807	2.121	2.345
QRC	0.302	0.219	0.159	0.108	0.084	0.081	0.764	0.665	0.567	0.478	0.417	0.384
MP	0.117	0.089	0.075	0.078	0.103	0.149	0.490	0.440	0.401	0.379	0.438	0.566
MNB	0.031	0.035	0.042	0.062	0.101	0.173	0.239	0.254	0.290	0.360	0.455	0.583
$\boldsymbol{\alpha = 2}$												
MLP	0.052	0.261	0.576	0.989	1.337	1.751	0.306	0.775	1.207	1.617	1.912	2.198
ZIP	0.327	1.046	1.651	2.300	2.752	3.254	0.852	1.621	2.091	2.491	2.742	2.987
MLNB	0.039	0.317	0.649	1.068	1.412	1.815	0.265	0.842	1.286	1.686	1.969	2.244
ZINB	0.042	0.317	0.649	1.069	1.413	1.815	0.281	0.844	1.286	1.686	1.969	2.244
QRC	0.946	0.803	0.671	0.556	0.460	0.387	1.266	1.166	1.086	1.018	0.941	0.898
MP	0.427	0.382	0.321	0.296	0.282	0.287	0.909	0.907	0.845	0.836	0.829	0.824
MNB	0.060	0.063	0.076	0.109	0.141	0.224	0.329	0.345	0.381	0.474	0.541	0.686

Table 6: Simulation resalts of robust and non-robust estimators when $n = 300,\ \text{ZI} = 50\%$

Estimator	**MSE**			**MAE**		
	O%			**O%**		
	0	**10**	**20**	**0**	**10**	**20**
$\boldsymbol{\alpha = 1}$						
MLP	0.650	0.449	0.692	1.102	0.981	1.247
ZIP	0.203	0.935	1.574	0.662	1.543	2.070
MLNB	0.613	0.490	0.759	1.041	1.029	1.313
ZINB	0.101	0.478	0.967	0.432	1.013	1.561
QRC	6.205	4.985	3.710	3.380	3.117	2.709
MP	2.916	2.183	1.675	2.383	2.183	2.026
MNB	0.406	0.272	0.330	0.909	0.735	0.802
$\boldsymbol{\alpha = 2}$						
MLP	0.747	0.567	0.618	1.212	1.119	1.176
ZIP	0.449	0.831	1.189	1.000	1.419	1.755
MLNB	0.674	0.558	0.659	1.126	1.116	1.222
ZINB	0.196	0.366	0.622	0.596	0.852	1.155
QRC	7.799	6.961	6.223	3.776	3.714	3.624
MP	4.209	3.558	3.066	2.853	2.747	2.674
MNB	0.657	0.525	0.561	1.151	1.034	1.051

Table 7: Simulation resalts of robust and non-robust estimators when $n = 500,\ \text{ZI} = 50\%$

Estimator	**MSE**			**MAE**		
	O%			**O%**		
	0	**10**	**20**	**0**	**10**	**20**
$\boldsymbol{\alpha = 1}$						
MLP	0.586	0.400	0.667	1.022	0.950	1.229
ZIP	0.157	0.969	1.616	0.589	1.612	2.134
MLNB	0.562	0.454	0.736	0.966	1.014	1.295
ZINB	0.057	0.466	0.974	0.330	1.031	1.606
QRC	6.082	4.966	3.563	3.395	3.154	2.651
MP	2.805	2.105	1.579	2.333	2.169	1.999
MNB	0.455	0.261	0.486	0.901	0.756	1.029
$\boldsymbol{\alpha = 2}$						
MLP	0.629	0.463	0.487	1.090	1.022	1.086
ZIP	0.400	0.766	1.056	0.952	1.392	1.687
MLNB	0.595	0.471	0.539	1.027	1.039	1.155
ZINB	0.117	0.291	0.501	0.463	0.764	1.064
QRC	7.639	6.789	6.114	3.822	3.734	3.654
MP	4.115	3.445	2.970	2.844	2.741	2.675
MNB	0.500	0.383	0.435	0.973	0.916	1.005

Table 8: Simulation resalts of robust and non-robust estimators when $n = 1000$, ZI $= 50\%$

Estimator	**MSE**			**MAE**		
	O%			**O%**		
	0	**10**	**20**	**0**	**10**	**20**
$\boldsymbol{\alpha = 1}$						
MLP	0.528	0.344	0.623	0.915	0.915	1.215
ZIP	0.134	0.969	1.636	0.555	1.641	2.162
MLNB	0.518	0.413	0.700	0.883	1.004	1.289
ZINB	0.029	0.446	0.950	0.234	1.060	1.619
QRC	6.106	4.965	3.485	3.482	3.203	2.617
MP	2.777	2.067	1.549	2.366	2.184	2.004
MNB	0.407	0.201	0.441	0.805	0.684	1.006
$\boldsymbol{\alpha = 2}$						
MLP	0.554	0.368	0.419	0.960	0.921	1.052
ZIP	0.334	0.689	1.024	0.862	1.353	1.701
MLNB	0.537	0.392	0.494	0.917	0.985	1.148
ZINB	0.051	0.205	0.448	0.308	0.667	1.026
QRC	7.546	6.731	6.099	3.821	3.736	3.689
MP	4.053	3.385	2.946	2.827	2.721	2.698
MNB	0.430	0.308	0.384	0.853	0.817	0.995

5. APPLICATION TO GERMAN HEALTH DATA

In this section, we study the efficiency of the non-robust and robust estimators of the count regression model based on real dataset. We use the set of data "*badhealth*" available in the R package $\{COUNT\}$. They were obtained from the German health survey for the year 1998 only and consist of 1127 observations on three variables. This data used by [25, 33]. In our application, the sample size used is 500 observations: $n = 500$.

The definition of the three variables is shown in Table 9. While Table 10 presents some descriptive statistics of these variables. Table 10 shows that the response variable (NUMVISIT) have a large dispersion, since the value of variance more than the value of mean. Moreover, there is large difference between the values of mean and median, it is indicator of outliers in the data. We will check the outliers by plot the boxplot of NUMVISIT variable. Figure 2 shows that the response variable (NUMVISIT) contains large number of zeros (based on frequency distribution plot) and some outlier values (based on boxplot).

Table 9: Definition of variables

Variable	Definition
NUMVISIT	Number of visits to doctor during 1998. NUMVISIT is the response variable.
BADH	The patient's condition **(**dummy variable): 1 if the patient claims to be in bad health, or 0 if is not in bad health.
AGE	The age of patient: from 20 to 60 years old.

Table 10: Descriptive Statistics (sample size = 500)

Variable	Minimum	Maximum	Mean	Median	Variance
NUMVISIT	0	40	2.41	1.00	14.54
AGE	20	60	37.43	35.00	111.83
BADH	**Coding**	**Level**		**Count**	**%**
	0	Not in bad health		460	92
	1	Bad health		40	8

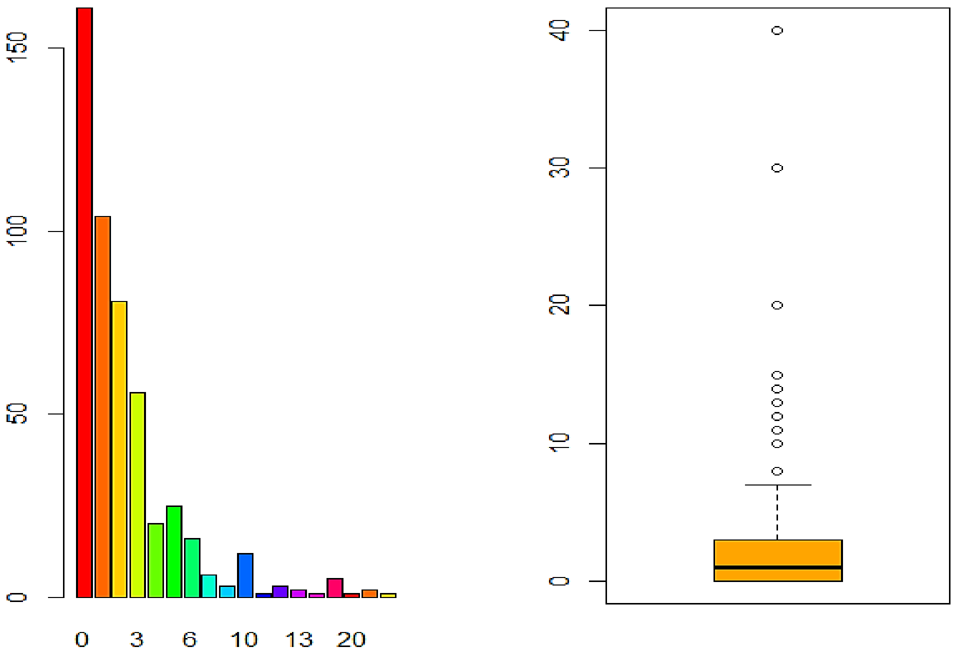

Figure 2: Frequency distribution and boxplot of NUMVISIT variable.

To check multicollinearity problem, we calculate the Pearson correlation coefficient between BADH and AGE, the coefficient is 0.156, this is a weak positive correlation; it indicates that is no multicollinearity problem. See e.g., [4, 5, 34, 35, 36] for handling and solving this problem in GLM if this problem exists. Moreover, we applied the Z-score test to check overdispersion: H_0: Equidispersion vs. H_1: True dispersion is greater than one. The Z-score statistic 3.94 with P-value < 0.001, then we can reject H_0, this means that the true dispersion is greater than one. And the estimated value of the dispersion of the Z-score test is 4.26. This indicates that the dataset has overdispersion problem.

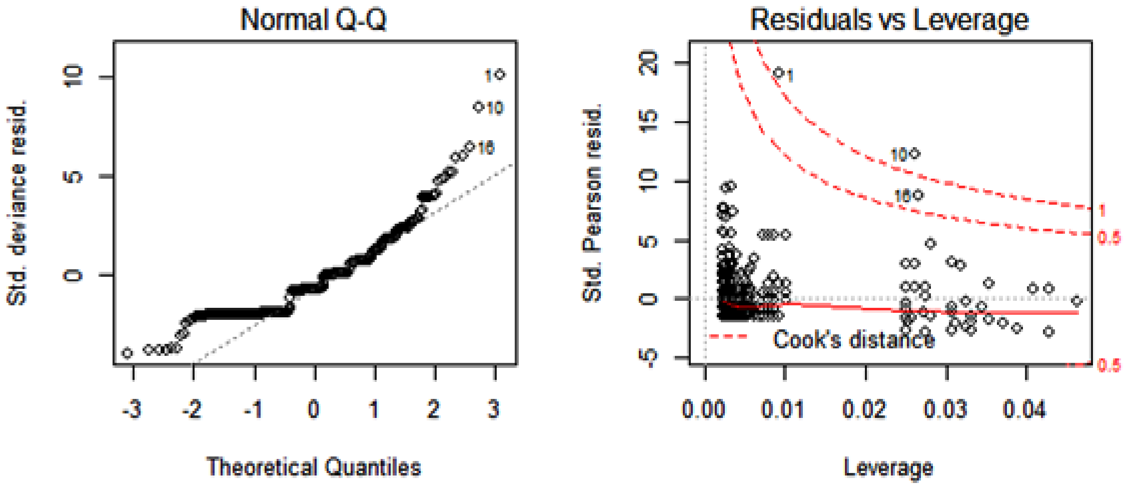

Figure 3: Residual diagnostics

Figure 3 indicates that the residuals are not distributed normal, and the residuals contains some outlier values (based on boxplot). This is confirmed by Shapiro-Wilk test results of the residuals: W-statistic = 0.693 with P-value < 0.0001. Since the P-value less then 0.05, then reject H_0, this means that the residuals are not assumed to be normally distributed.

It is verified that there are overdispersion problem, zero inflated, and outlier values in this data. Next, we will estimate the coeffects of the count regression model using non-robust and robust estimators. The estimation results are presented in Table 11.

Table 11: Non-robust and Robust Count Estimation Results

Estimator	**Intercept**	**BADH**	**AGE**
Non-robust			
MLP	0.542*** (0.108)	1.262*** (0.069)	0.0039 (0.003)
ZIP	0.681*** (0.118)	1.035*** (0.072)	0.009** (0.003)
MLNB	0.525** (0.206)	1.26*** (0.187)	0.0043 (0.005)
ZINB	0.377· (0.215)	1.19*** (0.187)	0.010· (0.005)
Robust			
QRC	0.542 (0.290)	1.26*** (0.263)	-0.009 (0.007)
MP	0.603*** (0.130)	1.22*** (0.086)	-0.005· (0.003)
MNB	0.922* (0.219)	1.37*** (0.192)	-0.009 (0.005)

Note: *** if P-value < 0.001, **if P-value < 0.01 and * if P-value < 0.05. And standard error of the coefficient in parentheses.

From Table 11, it is indicated that all values of estimates are close together. And the BADH variable is significant in all estimators, but the AGE variable is not significant in all estimators, except ZIP estimator. To select the suitable estimator for this data, we check the values of MAE, MSE and root mean standard error (RMSE) of all estimators that listed in Table 12.

Table 12: Goodness of fit criterion

Estimator	**MAE**	**MSE**	**RMSE**
Non-robust			
MLP	2.406	13.61	3.69
ZIP	2.404	7.778	2.788
MLNB	2.406	13.62	3.691
ZINB	2.407	7.854	2.802
Robust			
QRC	**0.289**	**0.1999**	**0.447**
MP	1.922	7.004	2.646
MNB	2.279	12.775	3.574

Table 12 shows that the QRC estimator has the smallest MAE, MSE, and RMSE compared to all estimators, this is due to large sample size (500) and small dispersion value (the estimated value of the dispersion of MLNB is 0.918), note that the difference between this value and the estimated value of the dispersion from the Z-score test can be explained by presence of outliers in the data.

6. Conclusion

In this paper, seven robust and non-robust estimators (three robust estimators and four non-robust estimators) of four count regression models (two Poisson models and two NB models) have been studied, under the assumption that the dataset contains overdispersion problem, many zeros, and outliers. We studied the efficiency of these estimators under this assumption by making the Monte Carlo simulation study for different sample sizes, percentages of outlier values, dispersion values, and zero values. Moreover, we used the German health survey data as an empirical application.

Simulation results showed that all robust estimators perform well against all non-robust estimators when the dataset contains outliers, regardless of the sample size and the dispersion value. QRC and MNB are the best two robust estimators, specifically when the dataset contains the overdispersion problem (dispersion parameter more than one), many zeros (the half of the data is zeros), and many outliers (the percentage of outlier values more than 15%). The results of the application indicate that the QRC estimator is the best for this data, and the significant variable that effect on the number of visits to doctor is the patient's condition (bad health or not in bad health), however the patient's age variable is not significant. As a future work, one may extend the robust count estimators mentioned in this paper to longitudinal and panel data models, see [37, 38, 39].

Conflict of Interests

The authors declare that there is no conflict of interests.

REFERENCES

[1] R. Winkelmann, Econometric analysis of count data, 5th Edition, Springer Verlag, Berlin, (2008).

[2] N. Ismail, A.A. Jemain, Handling overdispersion with negative binomial and generalized Poisson regression models. Casualty Actuarial Society Forum, (2007), p 103–158.

[3] W.H. Greene, Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, NYU Working Paper No. EC-94-10, New York University, USA, (1994).

[4] E.A. Rady, M.R. Abonazel, I. M. Taha, A new biased estimator for zero-inflated count regression models, J. Mod. Appl. Stat. Meth. In press.

[5] A.F. Lukman, B. Aladeitan, K. Ayinde, M.R. Abonazel, Modified ridge-type for the Poisson regression model: simulation and application, J. Appl. Stat. (2021), https://doi.org/10.1080/02664763.2021.1889998.

[6] J.A. Nelder, R.M. Wedderburn, generalized linear models, J. R. Stat. Soc. 135 (1972), 370-384.

[7] L.P. Rahayu, K. Sadik, Overdispersion study of Poisson and zero-inflated Poisson regression for some characteristics of the data on lamda, n, p, Int. J. Adv. Intell. Inform. 2 (2016), 140-148.

[8] A. Agresti, Categorical data analysis, 2nd Edition, Wiley, New York, USA, (2001).

[9] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust statistics: the approach based on influence functions, John Wiley & Sons, New York, (2011).

[10] A. Cameron, P.K. Trivedi, Regression analysis of count data. 2nd Edition, Cambridge University Press, New York, USA, (2013).

[11] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions. Wiley, New York, USA, (1986).

[12] A.M. Leroy, P.J. Rousseeuw, Robust regression and outlier detection. Wiley Series in Probability and Mathematical Statistics, Wiley, New York, USA, (1987).

[13] J. Fox, Applied regression analysis and generalized linear models. Sage Publications, USA, (2015).

[14] M.R. Abonazel, O.M. Saber, A comparative study of robust estimators for Poisson regression model with outliers. J. Stat. Appl. Probab. 9 (2020), 279-286.

[15] M.R. Abonazel, A. Rabie, The impact of using robust estimations in regression models: An application on the Egyptian economy, J. Adv. Res. Appl. Math. Stat. 4 (2019), 8-16.

[16] M.R. Abonazel, A.A. Gad, Robust partial residuals estimation in semiparametric partially linear model, Commun. Stat.-Simul. Comput. 49 (2020), 1223-1236.

[17] M.M. Elgohary, M.R. Abonazel, N.M. Helmy, A.R. Azazy, New robust-ridge estimators for partially linear model, Int. J. Appl. Math. Res. 8 (2019), 46-52

[18] M.R. Abonazel, Handling outliers and missing data in regression models using R: simulation examples, Acad. J. Appl. Math. Sci. 6 (2020), 187-203.

[19] A.H. Youssef, A.R. Kamel, M.R. Abonazel, Robust SURE Estimates of Profitability in the Egyptian Insurance Market, Stat. J. IAOS, (2021), https://doi.org/10.3233/SJI-200734

[20] A.H. Youssef, M.R. Abonazel, A.R. Kamel, Efficiency comparisons of robust and non-robust estimators for seemingly unrelated regressions model, J. Stat. Appl. Probab. In press.

[21] E. Cantoni, E. Ronchetti, Robust inference for generalized linear models. J. Amer. Stat. Assoc. 96 (2001), 1022-1030.

[22] S. Hosseinian, S. Morgenthaler, Weighted maximum likelihood estimates in Poisson regression. International Conference on Robust Statistics. A Universidade de Valladolid, Italy, (2011).

[23] J.A.F. Machado, J.M. Santos Silva, Quantiles for counts. J. Amer. Stat. Assoc. 100 (2005), 1226-1237.

[24] W.H. Aeberhard, E. Cantoni, S. Heritier, Robust inference in the negative binomial regression model with an application to falls data. Biometrics, 70 (2014), 920-931.

[25] J.M. Hilbe, Negative binomial regression, 2nd Edition, Cambridge University Press, Cambridge UK, (2011).

[26] P. McCullagh, J.A. Nelder, Generalized linear models. Chapman and Hall, London, (1989).

[27] D. Lambert, Zero-inflated Poisson regression with an application to defects in manufacturing, Technometrics, 34 (1992), 1-14.

[28] H.R. Kunsch, L.A. Stefanski, R.J. Carroll, conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. J. Amer. Stat. Assoc. 84 (1989), 460-466.

[29] E. Cantoni, E. Ronchetti, A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. J. Health Econ. 25 (2006), 198-213.

[30] R. Koenker, G. Basset, Regression Quantiles, Econometrica, 46 (1978), 33-50.

[31] C. Gourieroux, A. Monfort, A. Trognon, Pseudo maximum likelihood methods: Theory, Econometrica, 52 (1984), 681-700.

[32] M.R. Abonazel, A practical guide for creating Monte Carlo simulation studies using R, Int. J. Math. Comput. Sci. 4 (2018), 18-33.

[33] S. Vílchez-López, A. Sáez-Castillo, M. Olmo-Jiménez, GWRM: An R Package for identifying sources of variation in overdispersed count data, PloS one, 11 (2016), e0167570.

[34] M.R. Abonazel, R.A. Farghali, Liu-type multinomial logistic estimator, Sankhya B, 81 (2019), 203-225.

[35] E.A. Rady, M.R. Abonazel, I.M. Taha, New shrinkage parameters for Liu-type zero inflated negative binomial estimator. In The 54th Annual Conference on Statistics, Computer Science, and Operation Research, Cairo University, Egypt, (2019).

[36] R.A. Farghali, M. Qasim, B.M.G. Kibria, M.R. Abonazel, Generalized two-parameter estimators in the multinomial logit regression model: methods, simulation and application, Commun. Stat. – Simul. Comput. (2021), https://doi.org/10.1080/03610918.2021.1934023.

[37] A.H. Youssef, M.R. Abonazel, E.G. Ahmed, Estimating the number of patents in the world using count panel data models, Asian J. Probab. Stat. 6 (2020), 24-33.

[38] A.H. Youssef, E.G. Ahmed, M.R. Abonazel, The performance of count panel data estimators: a simulation study and application to patents in Arab countries, J. Math. Comput. Sci. In press.

[39] A.M. El-Masry, A.H. Youssef, M.R. Abonazel, Using logit panel data modeling to study important factors affecting delayed completion of adjuvant chemotherapy for breast cancer patients, Commun. Math. Biol. Neurosci. 2021 (2021), Article ID 48.