

COMPARISON OF THE PERFORMANCE OF THE GRADIENT AND NEWTON-RAPHSON METHOD TO ESTIMATE PARAMETERS IN SOME ZERO-INFLATED REGRESSION MODELS

Kim-Hung PHO¹, Buu-Chau TRUONG^{2,}*

¹Fractional Calculus, Optimization and Algebra Research Group, Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

²Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

*Corresponding Author: Buu-Chau TRUONG (Email: truongbuuchau@tdtu.edu.vn)

(Received: 20-Jun-2020; accepted: 25-Aug-2020; published: 31-Dec-2020)

DOI: <http://dx.doi.org/10.25073/jaec.202044.297>

Abstract. This paper compares the performance of the gradient and Newton-Raphson (N-R) method to estimate parameters in some zero-inflated (ZI) regression models such as zero-inflated Poisson (ZIP) model, zero-inflated Bell (ZIBell) model, zero-inflated binomial (ZIB) model and zero-inflated negative binomial (ZINB) model. In the present work, firstly, we briefly present the approach of the gradient and N-R method. We then introduce the origin, formulas and applications of the ZI models. Finally, we compare the performance of two investigated approaches for these models through the simulation studies with numerous sample sizes and several missing rates. A real data set is investigated in this study. Specifically, we compare the results and the execution time of the R code for two methods. Moreover, we provide some important notes on these two approaches and some scalable research directions for future work.

Keywords

Gradient method, Newton-Raphson method, optimization, statistics.

1. Introduction

In statistics and other sciences, estimating equations play a very essential and meaningful role in research because if there is an estimating equation, we can utilize estimating methods to find its solutions. The gradient and N-R method are the most accustomed approaches to address this problem. Terlaky [30] indicate that the gradient method is one of the widespread approach to find the optimal solution. Another regularly employed approach is the N-R method. This approach is named after two very renowned mathematicians in the world, Isaac Newton and Joseph Raphson. For the full details of these methods, see Pho et al. [24].

It will be known that, count data is a very conventional data in practice. It is often found in areas such as: transportation, education, economics, engineering, etc. In practice, the percentage of zero events in a count outcome variable is too excessive for the Poisson regression or negative binomial regression model to fit (Cameron and Trivedi [2], Chapter 6). Thus a zero-inflated (ZI) regression model was introduced and largely applied as a remedy to the zero-excess issue (Rideout et al. [27]).

Some common ZI regression models, for example: Hall [5] who proposed the ZIB regression model to deal with excess zeros in the binomial regression. Asymptotic properties of the maximum-likelihood estimator (MLE) of the ZIB model are discussed in Diallo et al. [3]. The ZIP regression model is proposed by Lambert [11]. Lemonte et al. [12] introduce the ZIBell regression models for count data, the ZINB regression model is stated in Ridout et al. [28], etc.

As we know, among all ZI regression models, zero-inflated Poisson (ZIP) models are the most widespread. This model has been developed and studied in a very diverse and plentiful manner including its theory and application. For instance, Lambert [11] presents the ZIP model with an application to defects in production. Li et al. [15] introduce the multivariate ZIP models and their applications. Xie et al. [39] provide the ZIP model in statistical process control. Jansakul et al. [10] propose the score tests for ZIP models. Li [13] develops a lack-of-fit test for the ZIP models. Long et al. [16] offers a marginalized ZIP regression model with overall exposure effects. Huang et al. [8] illustrate the ZIP model relied likelihood ratio test for drug safety signal detection, and so on.

In addition, we often encounter the factual data sets that have missing values. The problems about missing data play an extremely vital and significant role in the scientific research. Missing data is a very widespread and prevalent problem in several research disciplines, e.g., finances, engineering and medical, etc. This issue is derived by numerous reasons. For example: interviewees refuse to answer, incorrect or incomplete answers to questions, hypersensitive questions, etc. (see Schafer et al. [29]). Thus, it has been seen that addressing the missing values properly is a tremendously significant and crucial issue.

Therefore, it is extremely important and meaningful to research about the combination of the two most widespread methods for finding optimal solutions and the ZI of the regression models with missing data. Notwithstanding there are many papers present the gradient and N-R method as well as the ZI regression mod-

els with missing covariates, the issue about the comparison of performance of these approaches in the ZI regression models with missing covariates has not yet been studied. To remedy this handicap, in this study, our primary interest is to compare the performance of the gradient and N-R method in the ZI regression models with missing covariates. Specifically, we compare the results and the efficient computing of R code for two methods.

The residue of the paper is arranged as follows. In Section 2, we briefly present the approach of the gradient and N-R method. We then introduce the ZI regression models. The origin, formulas and applications of the ZIP model is offered in this section. The simulation study is introduced in Section 3. Section 4 provides a practical example. Some discussions and important notes on these two approaches is given in the next section. Finally, conclusions and future works are described in the last section.

2. Literature review

2.1. The gradient and Newton-Raphson method

These are two extremely famous approaches for finding optimal solutions of equations and systems of equations. These are also two fundamental methods and are programmed for the current computing software. Therefore, several scientists and researchers have presented and developed the theory and applications of these two methods. To save space, we do not provide them in this part. The detail of the gradient method can be found in Pho et al. [24]. The completed origin and how to build and illustrative the pictures of the N-R method can be seen in Truong et al. [33]. In addition, readers may refer in Wedderburn [38], Fischer [4], Ahmad et al. [1], Pho et al. [22], [25] and Tutunov et al. [34] and so on.

The formula for finding solutions to these two methods exists in one-dimension and multi-dimension cases. Besides there are many distortions for the formula to find solutions of these two approaches. In the current work, we only

provide the formulas employed in this paper. Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function, *the direction of steepest descent* is the vector $-\nabla g(t_0)$, where t_0 is a starting value, the sequence of iterates $\{t_k\}$ is computed by employing the following expression:

$$t_{k+1} = t_k - m_k \nabla g(t_k), \quad (1)$$

where $m_k > 0$ minimizes the following function

$$Q_k(m) = g(t_k - m \nabla g(t_k)). \quad (2)$$

The demonstration in (1) is the formula of the gradient approach in multi-dimension.

The formula root of the N-R approach in multi-dimension case is described as:

$$t_{n+1} = t_n - [Hg(t_n)]^{-1} \nabla g(t_n) \quad (3)$$

where

$$\nabla g(t) = \left[\frac{\partial g(t)}{\partial t_1}; \frac{\partial g(t)}{\partial t_2}; \dots; \frac{\partial g(t)}{\partial t_n} \right]^T$$

and

$$Hg(t) = \begin{bmatrix} \frac{\partial^2 g(t)}{\partial t_1^2} & \frac{\partial^2 g(t)}{\partial t_1 \partial t_2} & \dots & \frac{\partial^2 g(t)}{\partial t_1 \partial t_n} \\ \frac{\partial^2 g(t)}{\partial t_2 \partial t_1} & \frac{\partial^2 g(t)}{\partial t_2^2} & \dots & \frac{\partial^2 g(t)}{\partial t_2 \partial t_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g(t)}{\partial t_n \partial t_1} & \frac{\partial^2 g(t)}{\partial t_n \partial t_2} & \dots & \frac{\partial^2 g(t)}{\partial t_n^2} \end{bmatrix} \quad (4)$$

It should be borne in mind that: $\nabla g(t)$ and $Hg(t)$ are called the gradient and Hessian matrix of $g(t)$, respectively.

2.2. The zero-inflated (ZI) regression models

Count data is a very ubiquitous data set in practice. It is easily collected in many fields, e.g.; computer sciences, transportation, engineering, finance and so on. In practice, we often get the situations that the percentage of zero events in a count outcome is so excessive that the Poisson regression or negative binomial regression model is not adequate (Cameron and Trivedi [2], Chapter 6). Hence Rideout et al. [27] introduce a ZI regression model and it is largely applied as a remedy to the zero-excess issue. In statistics

and applied mathematics, a ZI model is a statistical model that relied on a ZI probability distribution. It means that a distribution recognizes frequent zero-valued observations.

The ZI regression models nowadays have numerous practical applications and it has been extensively researched and developed on theories and applications by scientists. An inclusive review of the ZI regression models with missing data is discussed in Lukusa et al. [18]. Because as we know, among all ZI regression models, zero-inflated Poisson (ZIP) models are the most widespread. It is therefore extremely meaningful and interesting to cover the ZIP model issues in the following sections.

2.3. The zero-inflated Poisson (ZIP) model

1) The origin of ZIP model

In the view of Lukusa et al. [17], the original idea of the ZIP model was proposed in econometric literature, see Mullahy [20]. Later this model has become very popular in Lambert's study [11] and was widely used in literature. For an overview of how to build and develop a ZIP model, see Lukusa et al. [17].

2) Recipe for ZIP model

Let Y be a count random variable, X and Z be covariates, $\mathcal{X} = (1, X, Z)^T$. Lambert [11] proposes the parametric ZIP regression model in which the non-susceptible probability (mixing weight) p is linked to \mathcal{X} via a logit-linear predictor, $p = H(\boldsymbol{\theta}^T \mathcal{X})$ with $H(u) = [1 + \exp(-u)]^{-1}$, and the Poisson mean λ is linked to \mathcal{X} via a log-linear predictor, $\lambda = \exp(\boldsymbol{\alpha}^T \mathcal{X})$, where $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are unknown parameters. The ZIP model can be illustrated as:

$$P(Y = y|X, Z) = H(\boldsymbol{\theta}^T \mathcal{X})I(y = 0) + [1 - H(\boldsymbol{\theta}^T \mathcal{X})] \frac{\exp[-\exp(\boldsymbol{\alpha}^T \mathcal{X})][\exp(\boldsymbol{\alpha}^T \mathcal{X})]^y}{y!}$$

for $y = 0, 1, 2, \dots$, where $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are called coefficients of zero-inflation model and count model, respectively. See more details in Lambert [11], Pho et al. [26] and Lukusa et al. [17].

3) The likelihood function

It should be noted that $P(Y = 0|X, Z)$ can be rewritten as

$$\begin{aligned} P(Y = 0|X, Z) &= H(\boldsymbol{\theta}^T \mathcal{X}) \\ &\quad + [1 - H(\boldsymbol{\theta}^T \mathcal{X})] \exp[-\exp(\boldsymbol{\alpha}^T \mathcal{X})] \\ &= \frac{H(\boldsymbol{\theta}^T \mathcal{X})}{H(\boldsymbol{\theta}^T \mathcal{X} + \exp(\boldsymbol{\alpha}^T \mathcal{X}))} \end{aligned}$$

The likelihood function of the ZIP model is described as

$$\begin{aligned} L(\boldsymbol{\eta}) &= \prod_{i=1}^n \left[\frac{H(\boldsymbol{\theta}^T \mathcal{X}_i)}{H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))} \right]^{Y_i=0} \quad (5) \\ &\times \prod_{i=1}^n \left\{ \frac{[1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)]}{\exp[-\exp(\boldsymbol{\alpha}^T \mathcal{X}_i)][\exp(\boldsymbol{\alpha}^T \mathcal{X}_i)]^{Y_i}} \right\}^{Y_i>0} \end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T$ are parameters to be determined.

4) Applications of ZIP model to the factual data sets

Applications of the ZIP model to the factual data sets are extremely diverse and plentiful. These data sets can be easily explored in the fields of transportation, dental, hospital, computer company, etc. We only mentioned some of the actual data sets, which have been studied and developed by scientists with the ZIP model. For example, Xie [39] studies the ZIP models with the real data set is the number of read-write errors discovered in a computer hard disk used in a producing process (see Fig. 1). Lukusa et al. [17] consider the ZIP models with the practical data set is the number of motorcyclists who violated Taiwan's speed regulations was surveyed in 2007 (see Fig. 2). Unhappipat et al. [35] investigate the ZIP models with the factual data set is the number of major earthquakes every year in Thailand during 1933-2012 (see Fig. 3), and so on.

In Figs. 1-3, the proportion of zeros in the data set is very high, hence it is often referred to as "zero inflated" data. If encountering this type of data in the study, we have to espouse the appropriate regression models, the ZIP model is one of the models that is very concordant to research for count data with "zero inflated". It

is often considered as the first contender for this problem.

In fact, if we encounter the count data, which contains so many zeros as the graphs above, we can think of many other ZI regression models such as the ZIB model, the ZIBell model, the ZINB model, the ZIP model and so on. Because there are so many ZI regression models that can fit for data of this type, we can rely on a number of selection model criteria, to choose the best regression model for the dataset to be considered. Some common criteria are: Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC) and Vuong's test. The detailed formulas of these criteria can refer to Pho et al. [23]. In addition, readers may refer to Mahmoudi et al. [19], Tianhe et al. [31] and Wang et al. [37].

We present the main results of this paper in the next section.

3. Simulation study

We consider the simulation study to four ZI regression models: the ZIP model, the ZIBell model, the ZIB model and the ZINB model.

3.1. The ZIP model

Lukusa et al. [17] perform the N-R approach to estimate parameters for the ZIP model with missing covariates. Because our main objective in this article is to compare results and the operation time of R code of the N-R and gradient method. Thus in this simulation study, for sake of simplicity, we execute the estimation methods and the ways to create simulation data sets and establish parameters similar to the paper of Lukusa et al. [17].

Nevertheless, in this work we perform two approaches consist of the gradient and N-R methods to estimate parameters for the ZIP model with missing data. Three cases with missing data are investigated: low, medium and high missing rates. It should be remarked that albeit the considering model, the function that creates the simulation data and parameters are set in the same way as in Lukusa et al. [17]. The re-

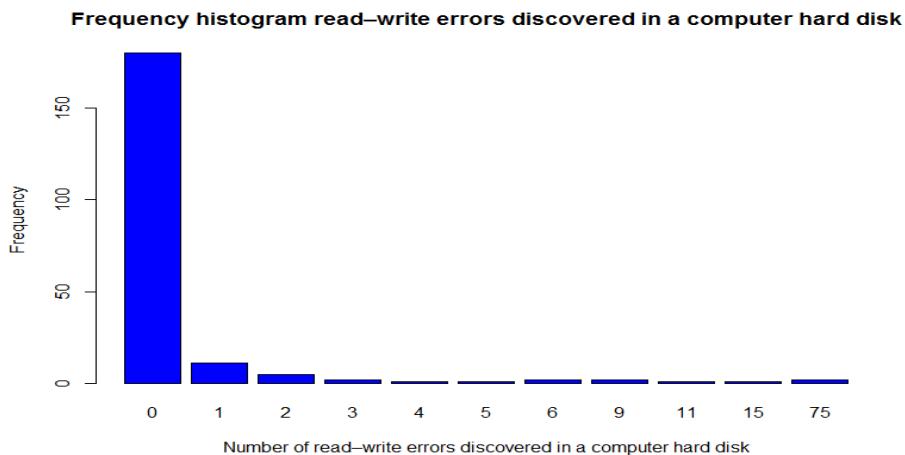


Fig. 1: The number of read–write errors discovered in a computer hard disk.

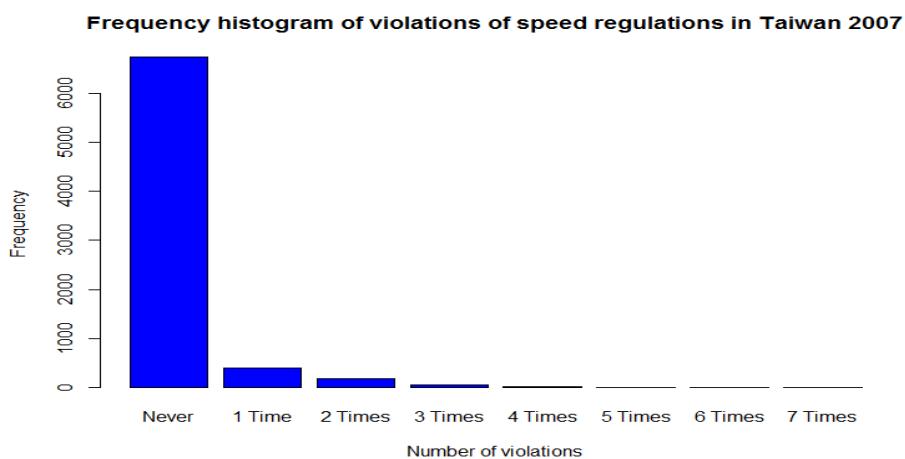


Fig. 2: The number of motorcyclists who violated Taiwan's speed regulations was surveyed in 2007.

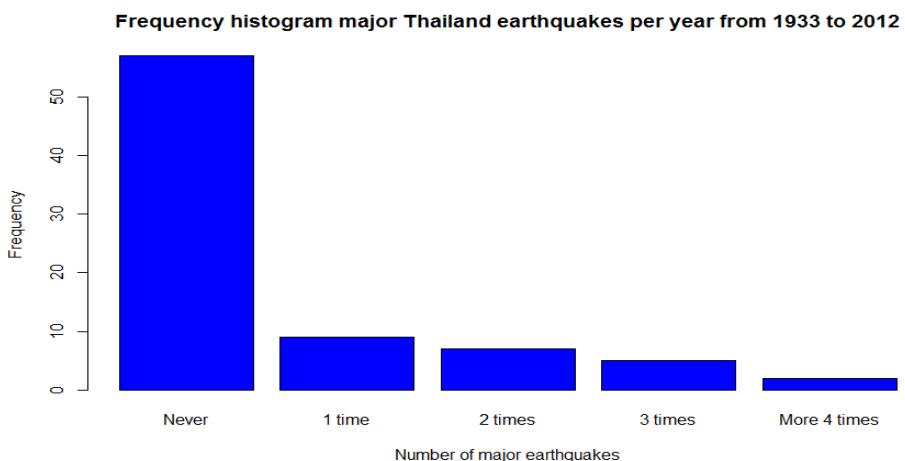


Fig. 3: The number of major earthquakes every year in Thailand during 1933-2012.

sults will be slightly different from this article, this reason is quite simple because each time R code is run, the data generation function will not be the same. In addition, different computers will offer slightly different values.

In the current work, the methods are used to estimate parameters for ZIP models with missing data are maximum likelihood estimation (MLE), complete case (CC), inverse probability weighting (IPW) and semi-parametric IPW (SIPW). These are very ubiquitous methods to address the missing data problems. Details of these approaches and their estimating functions can be found in Lukusa et al. [17]. It should be remarked that MLE method cannot address the data sets with missing values, we consider MLE as the benchmark for the sake of comparison.

The results of these three cases with two approaches consist of the gradient and N-R methods are offered in Tabs. 1-6, respectively. It needs to know that the symbols $\hat{\eta}_F$, $\hat{\eta}_{CC}$, $\hat{\eta}_W$, $\hat{\eta}_{WS}$ are estimator of MLE, CC, IPW and SIPW method, respectively. In addition, SD, ASE and CP are standard deviation, asymptotic standard error and coverage probability of a Wald-type 95% confidence interval, respectively.

As seen in Tabs. 1-6, the results obtained from the two approaches are gradient and N-R method are almost the same. These results do not exhibit any significant bias, since the bias $< 5\%$, the value of SD is very close to ASE. The value of CP is very near 0.95 and the estimator is consistent.

About the operation time of R code of two methods, we performed the Dell desktop computer is configured with Intel Core i5, 8GB of RAM, 1TB of hard drive to check the operation time of R code of two approaches. The operation time of R code of the gradient method in case the low missing rate with $n = 500$ and $n = 1000$ is 25 and 42 minutes, respectively. In case of the medium missing rate with $n = 1000$ and $n = 2000$ is 45 and 78 minutes, respectively. In case of the high missing rate with $n = 1000$ and $n = 2000$ is 53 and 87 minutes, respectively.

Meanwhile, the operation time of R code of the N-R method in case the low missing rate with $n = 500$ and $n = 1000$ is 27 and 48 min-

utes, respectively. In case of the medium missing rate with $n = 1000$ and $n = 2000$ is 49 and 86 minutes, respectively. In case of the high missing rate with $n = 1000$ and $n = 2000$ is 58 and 96 minutes, respectively.

It has been seen that the operation time of R code of the gradient method is shorter than the N-R method in most cases. This is also in line with the theory that is discussed in detail in the next section.

3.2. The other ZI models

For comparison purposes, we also considered some other ZI regression models, for instance: the ZIBell, ZIB and ZINB model. The detailed expression of the likelihood and log-likelihood function of these models can refer in Lemonte et al. [12], Diallo et al. [3] and Li et al. [14], respectively. To perform the gradient method as well as the N-R method to find the estimating parameters of the regression models, we need to have a score function of these regression models. It should be noted that: the score function, obtained by differentiating the log-likelihood function of the regression models with respect to the unaware parameters. The detailed formula of score function of the ZIP, ZIBell, ZIB and ZINB model are offered in Appendix.

After getting the score function of regression models, one can easily apply the above two approaches: the gradient and N-R method to find the corresponding solution based on the algorithm is provided in Section 5. Note that, we used the R statistical software to write code for the ZIP model. For these regression models, we only need to replace the score function and the data generation function. Then we can easily get the results as the ZIP model presented above.

To save space, we do not offer results for this section. Our results gained show that the results of two approaches are almost the same and the operation time of R code of the gradient method is shorter than the N-R method in most cases. This conclusion is also consistent with the theory considered in detail in the next section.

Tab. 1: Simulation results by executing the gradient method for case 1 (low missing rate): $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T = (-1, -1, 0.5, 1, 0.7, 1)^T$.

		$n = 500$				$n = 1000$			
Parameter		$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$	$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$
Logistic portion									
θ_0	Bias	-0.0086	-0.4967	-0.0372	-0.0085	-0.0062	-0.5027	-0.0214	-0.0119
	SD	0.1745	0.2782	0.2934	0.1743	0.1276	0.1952	0.2049	0.1325
	ASE	0.1723	0.2750	0.2873	0.1796	0.1235	0.1941	0.1989	0.1288
	CP	0.9581	0.5462	0.9538	0.9624	0.9472	0.2351	0.9536	0.9472
θ_1	Bias	-0.0059	-0.5628	-0.0359	-0.0247	-0.0075	-0.5389	-0.0237	-0.0126
	SD	0.1543	0.2862	0.2971	0.2439	0.1148	0.2036	0.1985	0.1528
	ASE	0.1535	0.2794	0.2832	0.2371	0.1078	0.1916	0.1942	0.1545
	CP	0.9483	0.4912	0.9347	0.8962	0.9386	0.1761	0.9512	0.9472
θ_2	Bias	0.0063	-0.6068	-0.0105	0.0224	0.0095	-0.5913	0.0075	0.0154
	SD	0.2387	0.4132	0.4563	0.2625	0.1663	0.2758	0.3019	0.1723
	ASE	0.2385	0.4013	0.4317	0.2468	0.1659	0.2742	0.2982	0.1716
	CP	0.9532	0.6459	0.9347	0.9436	0.9518	0.3986	0.9532	0.9491
Poisson portion									
α_0	Bias	-0.0013	0.1231	-0.0025	0.0048	-0.0025	0.1224	-0.0028	0.0016
	SD	0.0446	0.0421	0.0536	0.0447	0.0285	0.0322	0.0349	0.0324
	ASE	0.0435	0.0417	0.0538	0.0439	0.0276	0.0319	0.0357	0.0323
	CP	0.9485	0.2549	0.9487	0.9373	0.9562	0.0431	0.9528	0.9391
α_1	Bias	0.0023	-0.0395	0.0015	-0.0047	0.0024	-0.0363	0.0026	-0.0037
	SD	0.0262	0.0274	0.0317	0.0283	0.0184	0.0182	0.0229	0.0195
	ASE	0.0271	0.0285	0.0283	0.0274	0.0192	0.0186	0.0228	0.0194
	CP	0.9524	0.7384	0.9422	0.9275	0.9586	0.5075	0.9563	0.9514
α_2	Bias	0.0014	-0.0652	-0.0034	0.0051	0.0026	-0.0635	0.0029	0.0047
	SD	0.0372	0.0384	0.0419	0.0386	0.0249	0.0265	0.0329	0.0276
	ASE	0.0368	0.0414	0.0415	0.0384	0.0252	0.0259	0.0314	0.0285
	CP	0.9522	0.6548	0.9507	0.9423	0.9452	0.3571	0.9452	0.9375

- On average 32% of X were missing in 1000 replications.
- The average rate of $Y = 0$ was 27% and 14% in 1000 full and validated simulated data sets, respectively.

Tab. 2: Simulation results by performing the N-R method for case 1 (low missing rate): $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T = (-1, -1, 0.5, 1, 0.7, 1)^T$.

		$n = 500$				$n = 1000$			
Parameter		$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$	$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$
Logistic portion									
θ_0	Bias	-0.0074	-0.5218	-0.0426	-0.0081	-0.0057	-0.5007	-0.0212	-0.0063
	SD	0.1672	0.2749	0.2935	0.1756	0.1275	0.1958	0.2049	0.1326
	ASE	0.1752	0.2719	0.2876	0.1819	0.1234	0.1927	0.1985	0.1274
	CP	0.9587	0.5386	0.9527	0.9634	0.9478	0.2278	0.9569	0.9425
θ_1	Bias	-0.0086	-0.5549	-0.0375	-0.0228	-0.0054	-0.5386	-0.0237	-0.0128
	SD	0.1534	0.2985	0.2973	0.2418	0.1145	0.2032	0.1986	0.1584
	ASE	0.1519	0.2776	0.2766	0.2009	0.1038	0.1927	0.1995	0.1574
	CP	0.9485	0.4826	0.9329	0.8963	0.9386	0.1749	0.9517	0.9365
θ_2	Bias	0.0073	-0.6096	-0.0028	0.0226	0.0095	-0.5917	0.0084	0.0127
	SD	0.2385	0.4163	0.4549	0.2617	0.1674	0.2739	0.3023	0.1735
	ASE	0.2376	0.3969	0.4258	0.2449	0.1694	0.2762	0.2948	0.1726
	CP	0.9549	0.6518	0.9342	0.9475	0.9568	0.3984	0.9537	0.9495
Poisson portion									
α_0	Bias	-0.0074	0.1236	-0.0058	0.0079	-0.0041	0.1248	-0.0036	0.0045
	SD	0.0426	0.0486	0.0532	0.0464	0.0285	0.0328	0.0349	0.0347
	ASE	0.0418	0.0459	0.0528	0.0461	0.0274	0.0319	0.0332	0.0350
	CP	0.9485	0.2547	0.9463	0.9372	0.9581	0.0459	0.9572	0.9327
α_1	Bias	0.0017	-0.0394	0.0042	-0.0067	0.0045	-0.0327	0.0035	-0.0043
	SD	0.0279	0.0241	0.0327	0.0285	0.0168	0.0194	0.0249	0.0176
	ASE	0.0267	0.0235	0.0219	0.0273	0.0157	0.0183	0.0235	0.0168
	CP	0.9531	0.7349	0.9502	0.9394	0.9518	0.5083	0.9537	0.9541
α_2	Bias	0.0012	-0.0695	-0.0032	0.0058	0.0049	-0.0696	0.0038	0.0061
	SD	0.0376	0.0385	0.0472	0.0348	0.0292	0.0237	0.0348	0.0276
	ASE	0.0349	0.0412	0.0467	0.0336	0.0285	0.0229	0.0328	0.0257
	CP	0.9508	0.6517	0.9503	0.9449	0.9485	0.3612	0.9428	0.9379

- On average 32% of X were missing in 1000 replications.
- The average rate of $Y = 0$ was 27% and 14% in 1000 full and validated simulated data sets, respectively.

Tab. 3: Simulation results by executing the gradient method for case 2 (medium missing rate): $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T = (0.1, -1, 0.5, 1, 0.7, 1)^T$.

		$n = 1000$				$n = 2000$			
Parameter		$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$	$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$
Logistic portion									
θ_0	Bias	-0.0108	-0.7549	0.0126	0.0084	-0.0027	-0.6891	-0.0105	0.0094
	SD	0.0912	0.1796	0.1995	0.1084	0.0684	0.1286	0.1319	0.0752
	ASE	0.0946	0.1802	0.1992	0.1057	0.0648	0.1271	0.1307	0.0746
	CP	0.9496	0.0528	0.9513	0.9478	0.9521	0.0157	0.9464	0.9475
θ_1	Bias	-0.0074	-0.7513	-0.0349	-0.0298	-0.0112	-0.684	-0.0324	-0.0279
	SD	0.0853	0.1865	0.1984	0.1765	0.0574	0.1315	0.1294	0.1132
	ASE	0.0846	0.1822	0.1935	0.1695	0.0546	0.1295	0.1287	0.1095
	CP	0.9464	0.0208	0.9472	0.9279	0.9501	0.0124	0.9496	0.9513
θ_2	Bias	0.0084	-0.8713	-0.0097	0.0107	0.0094	-0.8469	-0.0114	0.0109
	SD	0.1307	0.2765	0.3258	0.1541	0.0972	0.1963	0.2184	0.1047
	ASE	0.1302	0.2763	0.3260	0.1532	0.0956	0.1959	0.2145	0.1032
	CP	0.9578	0.1354	0.9476	0.9481	0.9479	0.0147	0.9505	0.9482
Poisson portion									
α_0	Bias	-0.0136	0.1764	-0.0103	0.0099	-0.0018	0.1766	0.0028	0.0032
	SD	0.0332	0.0412	0.0462	0.0379	0.0231	0.0285	0.0334	0.0270
	ASE	0.0322	0.0409	0.0453	0.0375	0.0229	0.0284	0.0347	0.0261
	CP	0.9475	0.0184	0.9426	0.9023	0.9518	0.0127	0.9485	0.9458
α_1	Bias	0.0012	-0.0546	0.0019	-0.0054	0.0035	-0.0549	0.0004	-0.0015
	SD	0.0217	0.0224	0.0294	0.0236	0.0154	0.0162	0.0209	0.0165
	ASE	0.0215	0.0302	0.0288	0.0225	0.0152	0.0160	0.0204	0.0154
	CP	0.9476	0.3291	0.9510	0.8946	0.9474	0.0745	0.9513	0.9246
α_2	Bias	0.0013	-0.0849	0.0045	0.0042	0.0025	-0.0953	-0.0021	0.0029
	SD	0.0293	0.0293	0.0352	0.0324	0.0208	0.0217	0.0265	0.0233
	ASE	0.0285	0.0291	0.0350	0.0318	0.0205	0.0212	0.0263	0.0229
	CP	0.9423	0.1561	0.9412	0.9426	0.9428	0.0086	0.9402	0.9451

- On average 41% of X were missing in 1000 replications.
- The average rate of $Y = 0$ was 47% and 19% in 1000 full and validated simulated data sets.

Tab. 4: Simulation results by utilizing the N-R method for case 2 (medium missing rate): $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T = (0.1, -1, 0.5, 1, 0.7, 1)^T$.

Parameter		$n = 1000$				$n = 2000$			
		$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$	$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$
Logistic portion									
θ_0	Bias	-0.0089	-0.5812	0.0024	0.0057	-0.0061	-0.6582	-0.0114	0.0101
	SD	0.0931	0.1794	0.1963	0.1152	0.0648	0.1329	0.1357	0.0774
	ASE	0.0927	0.1792	0.1955	0.1095	0.0631	0.1296	0.1328	0.0762
	CP	0.9516	0.0676	0.9505	0.9457	0.9522	0.0154	0.9471	0.9467
θ_1	Bias	-0.0057	-0.6942	-0.0416	-0.0357	-0.0084	-0.6984	-0.0195	-0.0192
	SD	0.0851	0.1962	0.1969	0.1784	0.0549	0.1372	0.1355	0.1186
	ASE	0.0846	0.1937	0.1952	0.1663	0.0541	0.1355	0.1336	0.1176
	CP	0.9457	0.0184	0.9385	0.9364	0.9456	0.0032	0.9471	0.9409
θ_2	Bias	0.0076	-0.8846	-0.0214	0.0196	0.0058	-0.8978	-0.0119	0.0106
	SD	0.1357	0.2843	0.3276	0.1582	0.0953	0.1942	0.2147	0.1046
	ASE	0.1452	0.2819	0.3270	0.1574	0.0947	0.1940	0.2148	0.1029
	CP	0.9594	0.1347	0.9464	0.9489	0.9443	0.0065	0.9523	0.9394
Poisson portion									
α_0	Bias	-0.0082	0.1763	-0.0049	0.0052	-0.0047	0.1769	0.0092	0.0075
	SD	0.0335	0.0361	0.0436	0.0384	0.0249	0.0276	0.0352	0.0254
	ASE	0.0327	0.0359	0.0428	0.0379	0.0233	0.0272	0.0350	0.0256
	CP	0.9437	0.0189	0.9432	0.9247	0.9513	0.0029	0.9481	0.9419
α_1	Bias	0.0025	-0.0549	0.0013	-0.0086	0.0068	-0.0574	0.0032	-0.0045
	SD	0.0224	0.0236	0.0294	0.0239	0.0153	0.0167	0.0209	0.0173
	ASE	0.0219	0.0232	0.0288	0.0234	0.0151	0.0163	0.0204	0.0170
	CP	0.9475	0.3613	0.9458	0.8943	0.9459	0.0674	0.9484	0.9194
α_2	Bias	0.0032	-0.0843	0.0025	0.0041	0.0038	-0.0946	-0.0022	0.0042
	SD	0.0294	0.0296	0.0376	0.0331	0.0209	0.0217	0.0264	0.0238
	ASE	0.0287	0.0295	0.0372	0.0325	0.0206	0.0214	0.0261	0.0232
	CP	0.9453	0.1456	0.9471	0.9383	0.9429	0.0156	0.9415	0.9396

- On average 41% of X were missing in 1000 replications.
- The average rate of $Y = 0$ was 47% and 19% in 1000 full and validated simulated data sets.

Tab. 5: Simulation results by executing the gradient method for case 3 (high missing rate): $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T = (0.1, -1, 0.5, 1, 0.7, 1)^T$.

		$n = 1000$				$n = 2000$			
Parameter		$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$	$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$
Logistic portion									
θ_0	Bias	-0.0074	-0.6917	0.0035	0.0182	-0.0075	-0.6927	-0.0086	0.0079
	SD	0.1043	0.1990	0.2149	0.1238	0.0743	0.1405	0.1493	0.0870
	ASE	0.1028	0.1973	0.2148	0.1195	0.0749	0.1386	0.1485	0.0827
	CP	0.9572	0.0658	0.9519	0.9379	0.9568	0.0013	0.9481	0.9459
θ_1	Bias	-0.0073	-0.7347	-0.0456	-0.0571	-0.0084	-0.7168	-0.0236	-0.0259
	SD	0.0942	0.2056	0.2148	0.1872	0.0689	0.1458	0.1469	0.1249
	ASE	0.0939	0.1973	0.2048	0.1664	0.0695	0.1372	0.1426	0.1179
	CP	0.9484	0.0172	0.9384	0.9095	0.9461	0.0072	0.9496	0.9387
θ_2	Bias	0.0098	-0.9263	-0.0195	0.0283	0.0019	-0.9176	-0.0084	0.0108
	SD	0.1486	0.2987	0.3452	0.1669	0.1085	0.2076	0.2374	0.1182
	ASE	0.1471	0.2952	0.3395	0.1648	0.1079	0.2068	0.2350	0.1137
	CP	0.9609	0.1329	0.9471	0.9458	0.9431	0.0062	0.9509	0.9382
Poisson portion									
α_0	Bias	-0.0061	0.1894	-0.0034	0.0107	-0.0064	0.1895	0.0081	0.0095
	SD	0.0386	0.0494	0.0576	0.0479	0.0285	0.0247	0.0385	0.0279
	ASE	0.0371	0.0485	0.0549	0.0365	0.0271	0.0336	0.0369	0.0265
	CP	0.9428	0.0186	0.9408	0.8931	0.9549	0.0034	0.9484	0.9338
α_1	Bias	0.0035	-0.0649	0.0065	-0.0138	0.0095	-0.0679	0.0041	-0.0095
	SD	0.0284	0.0276	0.0369	0.0294	0.0185	0.0149	0.0271	0.0163
	ASE	0.0259	0.0248	0.0257	0.0285	0.0176	0.0132	0.0280	0.0157
	CP	0.9472	0.3464	0.9485	0.8896	0.9473	0.0685	0.9476	0.9248
α_2	Bias	0.0085	-0.0994	0.0027	0.0019	0.0064	-0.0985	-0.0036	0.0028
	SD	0.0318	0.0386	0.0373	0.0349	0.0285	0.0269	0.0295	0.0274
	ASE	0.0297	0.0371	0.0364	0.0336	0.0272	0.0258	0.0284	0.0267
	CP	0.9473	0.1508	0.9451	0.9379	0.9472	0.0085	0.9417	0.9398

- On average 53% of X were missing in 1000 replications.
- The average rate of $Y = 0$ was 51% and 23% in 1000 full and validated simulated data sets.

Tab. 6: Simulation results by utilizing the N-R method for case 3 (high missing rate): $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T = (0.1, -1, 0.5, 1, 0.7, 1)^T$.

Parameter		$n = 1000$				$n = 2000$			
		$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$	$\hat{\boldsymbol{\eta}}_F$	$\hat{\boldsymbol{\eta}}_{CC}$	$\hat{\boldsymbol{\eta}}_W$	$\hat{\boldsymbol{\eta}}_{Ws}$
Logistic portion									
θ_0	Bias	-0.0072	-0.6859	0.0062	0.0109	-0.0104	-0.6846	-0.0074	0.0085
	SD	0.1076	0.1985	0.2153	0.1274	0.0759	0.1449	0.1486	0.0892
	ASE	0.1052	0.1976	0.2142	0.1149	0.0728	0.1351	0.1472	0.0885
	CP	0.9514	0.0652	0.9536	0.9395	0.9549	0.0064	0.9459	0.9473
θ_1	Bias	-0.0085	-0.7396	-0.0472	-0.0574	-0.0063	-0.7175	-0.0262	-0.0253
	SD	0.0963	0.2095	0.2162	0.1895	0.0693	0.1458	0.1486	0.1275
	ASE	0.0955	0.1986	0.2130	0.1728	0.0672	0.1386	0.1469	0.1185
	CP	0.9485	0.0209	0.9386	0.9149	0.9495	0.0212	0.9485	0.9408
θ_2	Bias	0.0076	-0.9317	-0.0186	0.0265	0.0085	-0.9276	-0.0086	0.0182
	SD	0.1481	0.2996	0.3487	0.1686	0.1095	0.2079	0.2386	0.1194
	ASE	0.1472	0.2964	0.3373	0.1648	0.1087	0.2062	0.2229	0.1167
	CP	0.9631	0.1412	0.9495	0.9473	0.9485	0.0632	0.9509	0.9485
Poisson portion									
α_0	Bias	-0.0095	0.1913	-0.0072	0.0159	-0.0071	0.1924	0.0075	0.0068
	SD	0.0394	0.0461	0.0573	0.0486	0.0292	0.0281	0.0373	0.0294
	ASE	0.0385	0.0452	0.0563	0.0374	0.0285	0.0373	0.0361	0.0285
	CP	0.9419	0.0185	0.9476	0.8917	0.9508	0.0031	0.9485	0.9394
α_1	Bias	0.0076	-0.0649	0.0036	-0.0149	0.0019	-0.0621	0.0034	-0.0068
	SD	0.0259	0.0296	0.0376	0.0285	0.0147	0.0165	0.0283	0.0195
	ASE	0.0247	0.0282	0.0245	0.0272	0.0138	0.0158	0.0275	0.0184
	CP	0.9495	0.3449	0.9485	0.8892	0.9476	0.0659	0.9431	0.9248
α_2	Bias	0.0079	-0.0985	0.0058	0.0063	0.0027	-0.0982	-0.0091	0.0073
	SD	0.0359	0.0384	0.0375	0.0369	0.0252	0.0248	0.0283	0.0279
	ASE	0.0246	0.0351	0.0341	0.0353	0.0247	0.0242	0.0286	0.0280
	CP	0.9490	0.1472	0.9459	0.9332	0.9473	0.0095	0.9386	0.9417

- On average 53% of X were missing in 1000 replications.
- The average rate of $Y = 0$ was 51% and 23% in 1000 full and validated simulated data sets.

4. A practical example

The data set performed in this study is from the number of motorcyclists who violated Taiwan's speed regulations was surveyed in 2007. This data set has been utilized in the paper of Lukusa et al. [17]. There are four interesting variables in this analysis: the number of transgressed speed regulation (Y) with some covariates including Motorcycle-engine (Z), Distance-covered (X), and Age of respondents (W). This data set includes 7,386 respondents with 1,122 missing values. The detailed description can refer in Lukusa et al. [17] and Pho et al. (2019b). Similarly in Lukusa et al. [17], we also use dummy variables for Z, X and W.

It should be noted that MLE method cannot address the data sets with missing values, we consider MLE as the benchmark for the sake of comparison in the simulation study. Hence the MLE method will not be executed in the analysis for the practical data set. The diagram of Y is offered in Fig. 2 (**Section 2.3.4**). It can be seen that the data of Y is a count data. Thus some of regression models may be suitable for this data set as the ZIP model, Poisson model and negative binomial model. Likewise Lukusa et al. [17], we also used the Vuong's test (1989) to check this suitability. The results of Vuong's test show that the ZIP model is the most suitable one in comparison to the other models (Poisson and negative binomial model, p value < 0.001).

Because our main objective in this article is to compare results and the operation time of R code of the N-R and gradient method. Hence to save space, we only provide the results of ZIP regression analysis for this data set. The results of the N-R and gradient approach are offered in Tab. 7 and Tab. 8, respectively.

As seen from Tab. 7 and Tab. 8 that, the results obtained from the two approaches are N-R and gradient method are almost the same. The operation time of R code of the N-R and gradient method in this case is 8 and 6 minutes, respectively. Hence the operation time of R code of the gradient method is shorter than the N-R method in this analysis. This result is also consistent with the theory that is presented in detail in the next section.

5. Discussions

Firstly, we offer a general approach to apply the N-R method as well as the gradient method, it is required to execute the following five-steps algorithm:

- Step 1: Generate the data set which is based on the properties and characteristics of the regression model under consideration,
- Step 2: Find the score function of the regression model under investigation,
- Step 3: Obtain the first derivative of score function (hessian matrix) if using the N-R method, in the case of the gradient method, this step is not needed.
- Step 4: Calculate the N-R (the gradident) formula of the regression model under consideration.
- Step 5: Obtain the result from the program.

Next, we discuss the memory and computation of the gradient and N-R method. For the memory, each iteration of the N-R approach needs to have $O(p^2)$ storage ($p \times p$ Hessian), while each iteration of the gradient approach requires storage $O(p)$ (p -dimensional gradient). About computation, each loop of the N-R method needs to have $O(p^3)$ flops (solving a ponderous $p \times p$ linear system), meanwhile each loop of the gradient method requires $O(p)$ flops (scaling/adding n -dimensional vectors). Thus, it has been seen that, about theoretically if comparing the operation time of R code of two approaches, the gradient approach provides the results faster than the N-R method.

It should be remarked that, based on the formula (3) then their differences are not significant, the only difference is m_k and the inverse of the Hessian matrix. The simple way to understand, it can be said that if m_k and the inverse of the Hessian matrix exchange positions with each other, then formula (1) becomes formula (3) and vice versa.

In addition, the general formula of the Hessian matrix is fixed. For each successive iteration, we only need to substitute the specific number into

Tab. 7: Results of the practical data set by using the N-R method.

Variable	Parameter	CC method		SIPW method	
		$\hat{\theta}_{cc}$	ASE	$\hat{\theta}_{Ws}$	ASE
Poisson portion					
Intercept	β_0	-0.5613	0.2813	-0.5834	0.4221
DX_2	β_1	0.0870	0.1643	0.0911	0.1868
DX_3	β_2	0.1285	0.1619	0.1230	0.1902
DX_4	β_3	0.3525	0.1730	0.3532	0.2142
DZ_2	β_4	0.3112	0.2796	0.3126	0.4662
DZ_3	β_5	0.5268	0.2993	0.5564	0.4725
DZ_4	β_6	0.4705	0.2875	0.4898	0.4542
Logistic portion					
Intercept	γ_0	3.2557	0.2972	3.4536	0.3853
DX_2	γ_1	-0.5426	0.2070	-0.5391	0.2150
DX_3	γ_2	-0.8094	0.2043	-0.8036	0.2161
DX_4	γ_3	-0.9797	0.2194	-0.9701	0.2385
DZ_2	γ_4	-0.7468	0.2838	-0.8366	0.3979
DZ_3	γ_5	-2.3844	0.3446	-2.4186	0.4357
DZ_4	γ_6	-2.9797	0.3313	-3.0548	0.4230

Tab. 8: Results of the practical data set by using the gradient method.

Variable	Parameter	CC method		SIPW method	
		$\hat{\theta}_{cc}$	ASE	$\hat{\theta}_{Ws}$	ASE
Poisson portion					
Intercept	β_0	-0.5609	0.2821	-0.5841	0.4236
DX_2	β_1	0.0873	0.1638	0.0920	0.1872
DX_3	β_2	0.1276	0.1635	0.1219	0.1913
DX_4	β_3	0.3531	0.1724	0.3526	0.2139
DZ_2	β_4	0.3108	0.2785	0.3131	0.4673
DZ_3	β_5	0.5270	0.2996	0.5571	0.4732
DZ_4	β_6	0.4712	0.2883	0.4887	0.4538
Logistic portion					
Intercept	γ_0	3.2563	0.2969	3.4543	0.3842
DX_2	γ_1	-0.5418	0.2086	-0.5375	0.2134
DX_3	γ_2	-0.8085	0.2057	-0.8061	0.2179
DX_4	γ_3	-0.9784	0.2176	-0.9768	0.2392
DZ_2	γ_4	-0.7449	0.2857	-0.8384	0.3953
DZ_3	γ_5	-2.3867	0.3472	-2.4195	0.4361
DZ_4	γ_6	-2.9783	0.3346	-3.0552	0.4252

the general formula of the Hessian matrix to find the desired value. Meanwhile, the expression of m_k is offered in (2) , it will be changed after every repetition. Therefore, if done manually, it will take time to calculate each iteration. Furthermore, the gradient method depends on its direction, there are two accustomed approaches: the direction of steepest descent and the direction of steepest ascent. So sometimes we need to designate the appropriate approach to address the problem in the best way. These are two very serious disadvantages of the gradient method.

To our knowledge, most the current computing softwares for finding solutions of equations and systems of equations are programmed on the basis of N-R method. For example, the optim function is introduced by Nash [21]. The nleqslv function is presented by Hasselman [6]. The maxLik function is developed by Henningsen and Toomet [7], and so on. Truong et al. [32] compare results and the operation time of R code of these three functions. They showed that the results of the three approaches are almost the same and the operation time of R code to get the result of the optim function is the fastest. We now turn on present the conclusions and future works in the last section.

6. Conclusions and future works

In this study, we briefly presented the approach of the gradient and N-R method, introduced the origin, formulas and applications of the ZIP model. We compared the performance of two investigated approaches in this article for the ZI regression models such as the ZIP model, the ZIBell model, the ZIB model and the ZINB model with missing data through the simulation studies with numerous sample sizes and several missing rates. A real data set is investigated in this study. Specifically, we compare results and the operation time of R code of two methods. Both the theory and the results obtained from simulation studies with numerous sample sizes and consider several missing rates proved that the operation time of R code of gradient method is faster than N-R method. Moreover, we provided

some important notes on these two approaches and mention some scalable research directions in this paper.

For future works, we can employ the gradient as well as N-R method to investigate and study to some of the regression models that have not yet been considered with missing data, for instance: probit, binomial, negative binomial, Poisson, zero-inflated Bernoulli (ZIBer), zero-inflated negative binomial (ZINB), zero-inflated generalized Poisson (ZIGP), zero-inflated power series (ZIPS) or multivariate zero-inflated models, etc. These are very new, potential, interesting and meaningful research directions in the near future.

References

- [1] Ahmad, F., Tohidi, E., & Carrasco, J. A. (2016). A parameterized multi-step Newton method for solving systems of nonlinear equations. *Numerical Algorithms*, 71(3), 631-653.
- [2] Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (Vol. 53). Cambridge university press.
- [3] Diallo, A. O., Diop, A., & Dupuy, J. F. (2017). Asymptotic properties of the maximum-likelihood estimator in zero-inflated binomial regression. *Communications in Statistics-Theory and Methods*, 46(20), 9930-9948.
- [4] Fischer, A. (1992). A special Newton-type optimization method. *Optimization*, 24(3-4), 269-284.
- [5] Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4), 1030-1039.
- [6] Hasselman, B. (2009). nleqslv: Solve systems of non linear equations. R package version 2.1. 1.
- [7] Henningsen, A., & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3), 443-458.

- [8] Huang, L., Zheng, D., Zalkikar, J., & Tiwari, R. (2017). Zero-inflated Poisson model based likelihood ratio test for drug safety signal detection. *Statistical methods in medical research*, 26(1), 471-488.
- [9] Ismail, N., & Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. In *Casualty Actuarial Society E-Forum*, 41(20), 1-18.
- [10] Jansakul, N., & Hinde, J. P. (2002). Score tests for zero-inflated Poisson models. *Computational statistics & data analysis*, 40(1), 75-96.
- [11] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1-14.
- [12] Lemonte, A. J., Moreno-Arenas, G., & Castellares, F. (2020). Zero-inflated Bell regression models for count data. *Journal of Applied Statistics*, 47(2), 265-286.
- [13] Li, C. S. (2011). A lack-of-fit test for parametric zero-inflated Poisson models. *Journal of Statistical Computation and Simulation*, 81(9), 1081-1098.
- [14] Li, C. S., Lee, S. M., & Yeh, M. S. (2019). A test for lack-of-fit of zero-inflated negative binomial models. *Journal of Statistical Computation and Simulation*, 89(7), 1301-1321.
- [15] Li, C. S., Lu, J. C., Park, J., Kim, K., Brinkley, P. A., & Peterson, J. P. (1999). Multivariate zero-inflated Poisson models and their applications. *Technometrics*, 41(1), 29-38.
- [16] Long, D. L., Preisser, J. S., Herring, A. H., & Golin, C. E. (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in medicine*, 33(29), 5151-5165.
- [17] Lukusa, T. M., Lee, S. M., & Li, C. S. (2016). Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, 79(4), 457-483.
- [18] Lukusa, T. M., Lee, S. M., & Li, C. S. (2017). Review of Zero-Inflated Models with Missing Data. *Current Research in Biostatistics*, 7(1), 1-12.
- [19] Mahmoudi, M. R., Heydari, M. H., Avazzadeh, Z., & Pho, K. H. (2020). Goodness of fit test for almost cyclostationary processes. *Digital Signal Processing*, 96, 102597.
- [20] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3), 341-365.
- [21] Nash, J. C. (1979). Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation. Adam Hilger, Bristol. 1990. Chap, 3, 30-48.
- [22] Pho, K. H., Nguyen, V. T. (2018). Comparison of Newton-Raphson Algorithm and Maxlik Function. *Journal of Advanced Engineering and Computation*, 2(4), 281-292.
- [23] Pho, K. H., Ly, S., Ly, S., Lukusa, T. M. (2019). Comparison among Akaike Information Criterion, Bayesian Information Criterion and Vuong's test in Model Selection: A Case Study of Violated Speed Regulation in Taiwan. *Journal of Advanced Engineering and Computation*, 3(1), 293-303.
- [24] Pho, K. H., Tran T.K., Ho T.D.C., Wong, W.K. (2019). Optimal Solution Techniques in Decision Sciences: A Review, *Advances in Decision Sciences*, 23(1), 1-47.
- [25] Pho, K. H., Heydari, M. H., Tuan, B. A., & Mahmoudi, M. R. (2020). Numerical study of nonlinear 2D optimal control problems with multi-term variable-order fractional derivatives in the Atangana-Baleanu-Caputo sense. *Chaos, Solitons & Fractals*, 134, 109695.
- [26] Pho, K. H., Fu, Z., Mahmoudi, M. R., & Tuan, B. A. (2020). Prediction for the processes with almost cyclostationary structure. *Journal of Statistical Computation and Simulation*, 1-9.
- [27] Ridout, M., Demétrio, C. G., & Hinde, J. (1998). Models for count data with many

- zeros. In Proceedings of the XIXth international biometric conference, Cape Town, South Africa: International Biometric Society Invited Papers, 19, 179-192.
- [28] Ridout, M., Hinde, J., & Demetrio, C. G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1), 219-223.
- [29] Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2), 147.
- [30] Terlaky, T. (2013). Interior point methods of mathematical programming (Vol. 5). Springer Science & Business Media.
- [31] Tianhe, Y., Mahmoudi, M. R., Qasem, S. N., Tuan, B. A., & Pho, K. H. (2020). Numerical function optimization by condition-alized PSO algorithm. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-21.
- [32] Truong, B. C., Nguyen, V. B., Truong, H. V., & Ho, T. D. C. (2019). Comparison of Optim, Nleqslv and MaxLik to Estimate Parameters in Some of Regression Models. *Journal of Advanced Engineering and Computation*, 3(4), 532-550.
- [33] Truong, B. C., Van Thuan, N., Hau, N. H., & McAleer, M. (2019). Applications of the Newton-Raphson Method in Decision Sciences and Education. *Advances in Decision Sciences*, 23(4), 1-28.
- [34] Tutunov, R., Bou-Ammar, H., & Jadbabaie, A. (2019). Distributed Newton Method for Large-Scale Consensus Optimization. *IEEE Transactions on Automatic Control*, 64(10), 3983-3994.
- [35] Unhapipat, S., Pal, N., & Tiensuwan, M. (2017). Modeling Rare Events using a Zero-Inflated Poisson (ZIP) Distribution: Some New Results on Point Estimation. *The Philippine Statistician*, 66(2), 47-69.
- [36] Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.
- [37] Wang, Z., Parvin, H., Qasem, S. N., Tuan, B. A., & Pho, K. H. (2020). Cluster ensemble selection using balanced normalized mutual information. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-23.
- [38] Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3), 439-447.
- [39] Xie, M., He, B., & Goh, T. N. (2001). Zero-inflated Poisson model in statistical process control. *Computational statistics & data analysis*, 38(2), 191-201.

About Authors

Kim-Hung PHO is a lecturer and researcher of Fractional Calculus, Optimization and Algebra Research Group, Faculty of Mathematics and Statistics in Ton Duc Thang University, Ho Chi Minh City, Vietnam. His current research interests are regression models with missing data, randomized response technique, copula, applied Statistics, computational Statistics, applied Probability, Numerical analysis, Optimal problems, Mathematics education models and financial Mathematics.

Buu-Chau TRUONG is a lecturer at Faculty of Mathematics and Statistics in Ton Duc Thang University, Ho Chi Minh City, Vietnam. Her currently research interests include Time Series Analysis, Bayesian Inference, Data Analysis, Applied Statistics, Financial Risk Management and Financial Mathematics.

Appendix

1. The score function of the ZIP regression model

Executing the natural logarithm $L(\boldsymbol{\eta})$ in (5), one has the log-likelihood function of the ZIP model can then be presented as follows:

$$\begin{aligned}\ell(\boldsymbol{\eta}) = \log L(\boldsymbol{\eta}) &= \sum_{i=1}^n \ell_i(\boldsymbol{\eta}) \\ &= \sum_{i=1}^n \log \left\{ I(Y_i = 0) \left\{ \log [H(\boldsymbol{\theta}^T \mathcal{X}_i)] - \log [H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \right\} \right\} \\ &\quad + \sum_{i=1}^n \log \left\{ I(Y_i > 0) \left\{ \log [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] + [Y_i \boldsymbol{\alpha}^T \mathcal{X}_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) - \log(Y_i!)] \right\} \right\},\end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T$ are parameters to be determined. It should be remarked that $I(Y_i > 0) = 1 - I(Y_i = 0)$ and $Y_i I(Y_i = 0) = 0$, $i = 1, \dots, n$. We have

$$\frac{\partial}{\partial u} H(u) = H(u)[1 - H(u)] \frac{\partial u}{\partial \boldsymbol{\eta}} = H^{(1)}(u) \frac{\partial u}{\partial \boldsymbol{\eta}}$$

and

$$\frac{\partial}{\partial u} \log[H(u)] = \frac{\frac{\partial}{\partial u} H(u)}{H(u)} = [1 - H(u)] \frac{\partial u}{\partial \boldsymbol{\eta}},$$

where $H^{(1)}(u) = H(u)[1 - H(u)]$. It has been seen that,

$$\begin{aligned}S_{i1}(\boldsymbol{\eta}) &= \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ I(Y_i = 0) \left\{ \log [H(\boldsymbol{\theta}^T \mathcal{X}_i)] - \log [H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \right\} \right\} \\ &\quad + \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ I(Y_i > 0) \left\{ \log [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] + [Y_i \boldsymbol{\alpha}^T \mathcal{X}_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) - \log(Y_i!)] \right\} \right\} \\ &= I(Y_i = 0) \mathcal{X}_i \left\{ [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \right\} \\ &\quad - [1 - I(Y_i = 0)] H(\boldsymbol{\theta}^T \mathcal{X}_i) \mathcal{X}_i \\ &= I(Y_i = 0) \mathcal{X}_i \left[H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) - H(\boldsymbol{\theta}^T \mathcal{X}_i) \right] \\ &\quad - H(\boldsymbol{\theta}^T \mathcal{X}_i) \mathcal{X}_i + I(Y_i = 0) H(\boldsymbol{\theta}^T \mathcal{X}_i) \mathcal{X}_i \\ &= \mathcal{X}_i \left[I(Y_i = 0) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) - H(\boldsymbol{\theta}^T \mathcal{X}_i) \right] \\ &= \mathcal{X}_i H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \left[I(Y_i = 0) - \frac{H(\boldsymbol{\theta}^T \mathcal{X}_i)}{H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))} \right].\end{aligned}$$

and

$$\begin{aligned}
S_{i2}(\boldsymbol{\eta}) &= \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{\alpha}} \left\{ I(Y_i = 0) \left\{ \log [H(\boldsymbol{\theta}^T \mathcal{X}_i)] - \log [H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \right\} \right\} \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}} \left\{ I(Y_i > 0) \left\{ \log [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] + [Y_i \boldsymbol{\alpha}^T \mathcal{X}_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) - \log(Y_i!)] \right\} \right\} \\
&= -I(Y_i = 0) \mathcal{X}_i \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \\
&\quad + [1 - I(Y_i = 0)] \mathcal{X}_i [Y_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)] \\
&= -\mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \\
&\quad + \mathcal{X}_i [Y_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)] - \mathcal{X}_i Y_i I(Y_i = 0) + \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) \\
&= -\mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))] \\
&\quad + \mathcal{X}_i [Y_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)] + \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) \\
&= -\mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) + \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \\
&\quad + \mathcal{X}_i [Y_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)] + \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) \\
&= \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \\
&\quad + \mathcal{X}_i [Y_i - \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)] \\
&= \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \\
&\quad + \mathcal{X}_i \{Y_i - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i) + H(\boldsymbol{\theta}^T \mathcal{X}_i)] \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)\} \\
&= \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \\
&\quad + \mathcal{X}_i \{Y_i - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) - H(\boldsymbol{\theta}^T \mathcal{X}_i) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)\} \\
&= \mathcal{X}_i \{Y_i - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)\} \\
&\quad + \mathcal{X}_i I(Y_i = 0) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) - \mathcal{X}_i H(\boldsymbol{\theta}^T \mathcal{X}_i) \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) \\
&= \mathcal{X}_i \{Y_i - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)\} \\
&\quad + \mathcal{X}_i \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \left[I(Y_i = 0) - \frac{H(\boldsymbol{\theta}^T \mathcal{X}_i)}{H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))} \right] \\
&= \mathcal{X}_i \{Y_i - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)\} + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) S_{i1}(\boldsymbol{\eta}).
\end{aligned}$$

Let $S_i(\boldsymbol{\eta}) = \partial \ell_i(\boldsymbol{\eta}) / \partial \boldsymbol{\eta} = (S_{i1}^T(\boldsymbol{\eta}), S_{i2}^T(\boldsymbol{\eta}))^T$, ($i = 1, \dots, n$) be the score function of the ZIP model. Then $S_{i1}(\boldsymbol{\eta}) = \partial \ell_i(\boldsymbol{\eta}) / \partial \boldsymbol{\theta}$ and $S_{i2}(\boldsymbol{\eta}) = \partial \ell_i(\boldsymbol{\eta}) / \partial \boldsymbol{\alpha}$ are respectively given by

$$\begin{aligned}
S_{i1}(\boldsymbol{\eta}) &= \mathcal{X}_i H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)) \left[I(Y_i = 0) - \frac{H(\boldsymbol{\theta}^T \mathcal{X}_i)}{H(\boldsymbol{\theta}^T \mathcal{X}_i + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i))} \right], \\
S_{i2}(\boldsymbol{\eta}) &= \mathcal{X}_i \{Y_i - [1 - H(\boldsymbol{\theta}^T \mathcal{X}_i)] \exp(\boldsymbol{\alpha}^T \mathcal{X}_i)\} + \exp(\boldsymbol{\alpha}^T \mathcal{X}_i) S_{i1}(\boldsymbol{\eta}).
\end{aligned}$$

2. The score function of the ZIB regression model

The log-likelihood of $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ of the ZIB model, based on the observations (Z_i, X_i, W_i) , $i = 1, \dots, n$, is given by (see, Diallo et al. [3]):

$$\begin{aligned}
\ell(\boldsymbol{\psi}) &= \sum_{i=1}^n \ell_i(\boldsymbol{\psi}) \\
&= \sum_{i=1}^n \left\{ J_i \log \left(e^{\boldsymbol{\gamma}^T W_i} + \left(1 + e^{\boldsymbol{\beta}^T X_i} \right)^{-m_i} \right) - \log \left(1 + e^{\boldsymbol{\gamma}^T W_i} \right) + (1 - J_i) \left[Z_i \boldsymbol{\beta}^T X_i - m_i \log \left(1 + e^{\boldsymbol{\beta}^T X_i} \right) \right] \right\}
\end{aligned}$$

Let $S_i(\psi) = \partial\ell_i(\psi)/\partial\psi^T = (S_{i1}^T(\psi), S_{i2}^T(\psi))^T$, ($i = 1, \dots, n$) be the score function of the ZIB model. Then $S_{i1}(\psi) = \partial\ell_i(\psi)/\partial\beta^T$ and $S_{i2}(\psi) = \partial\ell_i(\psi)/\partial\gamma^T$ are respectively provided by:

$$\begin{aligned} S_{i1}(\psi) &= \frac{\partial\ell_i(\psi)}{\partial\beta^T} \\ &= J_i \frac{X_i(-m_i)e^{\beta^T X_i} \left(1 + e^{\beta^T X_i}\right)^{-m_i-1}}{e^{\gamma^T W_i} + (1 + e^{\beta^T X_i})^{-m_i}} + (1 - J_i) \left[Z_i X_i - m_i \frac{X_i e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} \right] \\ &= X_i \left\{ -J_i \frac{m_i e^{\beta^T X_i} \left(1 + e^{\beta^T X_i}\right)^{-m_i-1}}{e^{\gamma^T W_i} + (1 + e^{\beta^T X_i})^{-m_i}} + (1 - J_i) \left[Z_i - \frac{m_i e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} \right] \right\} \\ &= X_i \left\{ -J_i \frac{m_i e^{\beta^T X_i}}{e^{\gamma^T W_i} (1 + e^{\beta^T X_i})^{m_i+1}} + (1 - J_i) \left[Z_i - \frac{m_i e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} \right] \right\} \end{aligned}$$

and

$$\begin{aligned} S_{i2}(\psi) &= \frac{\partial\ell_i(\psi)}{\partial\gamma^T} = J_i \frac{W_i e^{\gamma^T W_i}}{e^{\gamma^T W_i} + (1 + e^{\beta^T X_i})^{-m_i}} - \frac{W_i e^{\gamma^T W_i}}{1 + e^{\gamma^T W_i}} \\ &= W_i \left\{ \frac{J_i e^{\gamma^T W_i}}{e^{\gamma^T W_i} + (1 + e^{\beta^T X_i})^{-m_i}} - \frac{e^{\gamma^T W_i}}{1 + e^{\gamma^T W_i}} \right\} \\ &= W_i \left\{ \frac{J_i e^{\gamma^T W_i} \left(1 + e^{\beta^T X_i}\right)^{m_i}}{e^{\gamma^T W_i} (1 + e^{\beta^T X_i})^{m_i} + 1} - \frac{e^{\gamma^T W_i}}{1 + e^{\gamma^T W_i}} \right\} \end{aligned}$$

3. The score function of the ZINB regression model

Let $\{(y_i, \mathcal{X}_{i1}, \mathcal{X}_{i2}) : i = 1, \dots, n\}$ be the data set and $\theta = (\beta^T, \gamma^T, \frac{1}{\alpha})^T$ be the unknown ZINB parameter vector of interest to be estimated. Then the likelihood function of the ZINB distribution is expressed as (see, Ismail and Zamani [9])

$$L(\theta) = \prod_{i=1}^n \left[p_i + (1 - p_i) \left(\frac{1}{\mu_i + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \right]^{I(y_i=0)} \prod_{i=1}^n \left[(1 - p_i) \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{\mu_i}{\mu_i + \frac{1}{\alpha}} \right)^{y_i} \left(\frac{1}{\mu_i + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \right]^{I(y_i>0)},$$

where $p_i = H(\beta^T \mathcal{X}_{i1})$ and where $H(u) = \{1 - \exp(-u)\}^{-1}$, $\mu_i = \exp(\gamma^T \mathcal{X}_{i2})$, $\mathcal{X}_i = (1, X_i^T, Z_i^T)^T$, $i = 1, \dots, n$ and $I(\cdot)$ is an indicator function.

Let $\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \ell_i(\theta)$ and $S_i(\psi) = \partial\ell_i(\theta)/\partial\theta^T = (S_{i1}^T(\theta), S_{i2}^T(\theta), S_{i3}^T(\theta))^T$, ($i = 1, \dots, n$) be the score function of the ZINB model. Then $S_{i1}(\psi) = \partial\ell_i(\theta)/\partial\beta^T$, $S_{i2}(\psi) =$

$\partial\ell_i(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}^T$ and $S_{i3}(\boldsymbol{\psi}) = \partial\ell_i(\boldsymbol{\theta})/\partial 1/\alpha$ are respectively provided by:

$$\begin{aligned}
 S_{i1}(\boldsymbol{\theta}) &= \frac{\partial\ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}} \\
 &= \frac{\partial\ell_i(\beta, \boldsymbol{\gamma}, \frac{1}{\alpha})}{\partial\boldsymbol{\beta}} \\
 &= I(y_i = 0) \frac{\partial}{\partial\boldsymbol{\beta}} \log \left[H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \right] \\
 &\quad + I(y_i > 0) \frac{\partial}{\partial\boldsymbol{\beta}} \left[\log(1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) + \log \left(\Gamma \left(y_i + \frac{1}{\alpha} \right) \right) - \log(\Gamma(y_i + 1)) - \log \left(\Gamma \left(\frac{1}{\alpha} \right) \right) \right] \\
 &\quad + I(y_i > 0) \frac{\partial}{\partial\boldsymbol{\beta}} \left[y_i(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) - y_i \log(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}) + \frac{1}{\alpha} \log \left(\frac{1}{\alpha} \right) - \frac{1}{\alpha} \log(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}) \right] \\
 &= \frac{I(y_i = 0) \left[\mathcal{X}_{i1} H^{(1)}(\boldsymbol{\beta}^T \mathcal{X}_{i1}) - \mathcal{X}_{i1} H^{(1)}(\boldsymbol{\beta}^T \mathcal{X}_{i1}) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \right]}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}}} \\
 &\quad - I(y_i > 0) \frac{\mathcal{X}_{i1} H^{(1)}(\boldsymbol{\beta}^T \mathcal{X}_{i1})}{1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})} \\
 &= \frac{I(y_i = 0) \mathcal{X}_{i1} H^{(1)}(\boldsymbol{\beta}^T \mathcal{X}_{i1}) \left[1 - \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \right]}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}}} \\
 &\quad - I(y_i > 0) \mathcal{X}_{i1} H(\boldsymbol{\beta}^T \mathcal{X}_{i1})
 \end{aligned}$$

$$\begin{aligned}
S_{i2}(\boldsymbol{\theta}) &= \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \frac{1}{\alpha})}{\partial \boldsymbol{\gamma}} = \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \frac{1}{\alpha})}{\partial \boldsymbol{\gamma}} \\
&= I(y_i = 0) \frac{\partial}{\partial \boldsymbol{\gamma}} \log \left[H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^d \right] \\
&\quad + I(y_i > 0) \frac{\partial}{\partial \boldsymbol{\gamma}} \left[\log(1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) + \log \left(\Gamma \left(y_i + \frac{1}{\alpha} \right) \right) - \log(\Gamma(y_i + 1)) - \log \left(\Gamma \left(\frac{1}{\alpha} \right) \right) \right] \\
&\quad + I(y_i > 0) \frac{\partial}{\partial \boldsymbol{\gamma}} \left[y_i (\boldsymbol{\gamma}^T \mathcal{X}_{i2}) - y_i \log \left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right) + \frac{1}{\alpha} \log \left(\frac{1}{\alpha} \right) - \frac{1}{\alpha} \log \left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right) \right] \\
&\quad (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}-1} \left(-\frac{\frac{1}{\alpha} \mathcal{X}_{i2} \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right)^2} \right) \\
&= I(y_i = 0) \frac{1}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}}} \\
&\quad + I(y_i > 0) \frac{\partial}{\partial \boldsymbol{\gamma}} \left[y_i \mathcal{X}_{i2} - y_i \frac{\mathcal{X}_{i2} \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} - \frac{1}{\alpha} \frac{\mathcal{X}_{i2} \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right] \\
&\quad (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \left(-\frac{\mathcal{X}_{i2} \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right) \\
&= I(y_i = 0) \frac{1}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}}} \\
&\quad + I(y_i > 0) \frac{\partial}{\partial \boldsymbol{\gamma}} \left[y_i \mathcal{X}_{i2} \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right) - \frac{1}{\alpha} \frac{\mathcal{X}_{i2} \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right] \\
&\quad (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \left(-\frac{\mathcal{X}_{i2} \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right) \\
&= I(y_i = 0) \frac{1}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i1}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}}} \\
&\quad + I(y_i > 0) \frac{\partial}{\partial \boldsymbol{\gamma}} \left[\left(\frac{\frac{1}{\alpha} \mathcal{X}_{i2}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right) (y_i - \exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})) \right]
\end{aligned}$$

$$\begin{aligned}
S_{i3}(\boldsymbol{\theta}) &= \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial 1/\alpha} = \frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \frac{1}{\alpha})}{\partial 1/\alpha} \\
&= I(y_i = 0) \frac{\partial}{\partial 1/\alpha} \log \left[H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \right] \\
&\quad + I(y_i > 0) \frac{\partial}{\partial 1/\alpha} \left[\log(1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) + \log \left(\Gamma \left(y_i + \frac{1}{\alpha} \right) \right) - \log(\Gamma(y_i + 1)) - \log \left(\Gamma \left(\frac{1}{\alpha} \right) \right) \right] \\
&\quad + I(y_i > 0) \frac{\partial}{\partial 1/\alpha} \left[y_i (\boldsymbol{\gamma}^T \mathcal{X}_{i2}) - y_i \log \left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right) + \frac{1}{\alpha} \log \left(\frac{1}{\alpha} \right) - \frac{1}{\alpha} \log \left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right) \right] \\
&= I(y_i = 0) \frac{\partial}{\partial 1/\alpha} \frac{\left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}-1} \frac{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} - \frac{1}{\alpha}}{\left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right)^2}}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}}} \\
&\quad + I(y_i > 0) \left[\frac{\Gamma \left(y_i + \frac{1}{\alpha} \right) \psi_0 \left(y_i + \frac{1}{\alpha} \right)}{\Gamma \left(y_i + \frac{1}{\alpha} \right)} - \frac{\Gamma \left(\frac{1}{\alpha} \right) \psi_0 \left(\frac{1}{\alpha} \right)}{\Gamma \left(\frac{1}{\alpha} \right)} \right] \\
&\quad + I(y_i > 0) \frac{\partial}{\partial 1/\alpha} \left[\frac{y_i}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} + \log \left(\frac{1}{\alpha} \right) + 1 - \log \left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right) - \frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right] \\
&= I(y_i = 0) \frac{\partial}{\partial 1/\alpha} \frac{\left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}-1} \frac{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2})}{\left(\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha} \right)^2}}{H(\boldsymbol{\beta}^T \mathcal{X}_{i1}) + (1 - H(\boldsymbol{\beta}^T \mathcal{X}_{i1})) \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right)^d} + I(y_i > 0) \left[\psi_0 \left(y_i + \frac{1}{\alpha} \right) - \psi_0 \left(\frac{1}{\alpha} \right) \right] \\
&\quad + I(y_i > 0) \frac{\partial}{\partial 1/\alpha} \left[\frac{y_i - \frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} + \log \left(\frac{\frac{1}{\alpha}}{\exp(\boldsymbol{\gamma}^T \mathcal{X}_{i2}) + \frac{1}{\alpha}} \right) + 1 \right]
\end{aligned}$$

4. The score function of the ZIBell regression model

The log-likelihood of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T)^T$ of the ZIBell model is given by (see, Lemonte et al. [12])

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^n \ell_i(\boldsymbol{\theta}) \\
&= \sum_{y_i:y_i=0} \log \left[e^{\eta_{2i}} + \exp \left(1 - e^{W(\mu_i)} \right) \right] - \sum_{i=1}^n \log(1 - e^{\eta_{2i}}) + \sum_{y_i:y_i>0} y_i \log[W(\mu_i)] - \sum_{y_i:y_i>0} e^{W(\mu_i)}
\end{aligned}$$

where $\mu_i = e^{\eta_{1i}} = \exp(x_i^T \boldsymbol{\beta})$, $\pi_i = e^{\eta_{2i}} = \exp(s_i^T \boldsymbol{\tau})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$, $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_q)^T$, $\mathbf{X} = [x_1, x_2, \dots, x_n]^T$, and $\mathbf{S} = [s_1, s_2, \dots, s_n]^T$. Let $S(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T = (S_1^T(\boldsymbol{\theta}), S_2^T(\boldsymbol{\theta}))^T$,

$(i = 1, \dots, n)$ be the score function of the ZIBell model. Then $S_1(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\beta}^T = \mathbf{X}^T\boldsymbol{\gamma}$ and $S_2(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\tau}^T = \mathbf{S}^T\boldsymbol{\delta}$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_n)^T$, and $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_n)^T$, with

$$\gamma_i = \begin{cases} -\frac{\exp(1 + \eta_{1i} - e^{W(\mu_i)})}{[e^{\eta_{2i}} + \exp(1 - e^{W(\mu_i)})][1 + W(\mu_i)]}, & y_i = 0, \\ \frac{y_i - \mu_i}{1 + W(\mu_i)}, & y_i > 0, \end{cases}$$

$$\delta_i = \frac{e^{\eta_{2i}} I(y_i = 0)}{e^{\eta_{2i}} + \exp(1 - e^{W(\mu_i)})} - \frac{e^{\eta_{2i}}}{1 + e^{\eta_{2i}}},$$

where $I(\cdot)$ denotes an indicator function.