Heliyon 10 (2024) e24225

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A new modified biased estimator for Zero inflated Poisson regression model

Muhammad Zeeshan [a], Aamna Khan [a,*], Muhammad Amanullah [a], M.E. Bakr [b], Arwa M. Alshangiti [b], Oluwafemi Samson Balogun [c], M. Yusuf [d]

[a] *Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan*
[b] *Department of Statistics and Operations Research, College of Science, King Saud University, P.O. Box 2455, Riyadh, 11451, Saudi Arabia*
[c] *Department of Computing, Faculty of Forestry and Technology, University of Eastern Finland, FI-70211, Kuopio, Finland*
[d] *Helwan University, Faculty of Science, Departement of Mathematics, Cairo, Egypt*

ARTICLE INFO

Keywords:
Count data
Zero-inflated Poisson ridge regression
Excessive zeroes
Ridge estimators
Multicollinearity

ABSTRACT

Zero-inflated Poisson (ZIP) model is widely used for counting data with excessive zeroes. The multicollinearity is the common factor in the explanatory variables of the count data. In this context, typically, maximum likelihood estimation (MLE) generates unsatisfactory results due to inflation of mean square error (MSE). In the solution of this problem usually, ridge parameters are used. In this study, we proposed a new modified zero-inflated Poisson ridge regression model to reduce the problem of multicollinearity. We experimented within the context of a specified simulation strategy and recorded the behavior of proposed estimators. We also apply our proposed estimator to the real-life data set and explore how our proposed estimators perform well in the presence of multicollinearity with the help of ZIP model for count data.

1. Introduction

The count data is widely used in many fields, such as insurance, public health, epidemiology, psychology and many other kinds of research work. Poisson is commonly used in count data to evaluate that mean and variance are the same. Still, this assumption is restricted when data is overdispered (variance of response variable more than mean value). However, the probability density function (PDF) of the model can be expressed in the following sense:

$$f(y_i) = \frac{e^{\lambda_i}{\lambda_i}^{y_i}}{y_i!}, y_i = 0, 1, \dots, n.$$

Where y_i is the response variable and belongs to Poisson regression with mean value λ_i.in the Poisson regression model, $\ln(\lambda_i) = X'\beta$ is the linear combination of the explanatory variables $X_i = (X_{i1}\dots X_{ip})'$. While $\ln(\lambda_i)$ is defined as a canonical link function between explanatory variables and response variables in linear form.

By incorporating both a degenerated distribution at zero and Poisson regression, ZIP models offer better fitting capabilities,

* Corresponding author.
E-mail addresses: muhammadzesshan74@gmail.com (M. Zeeshan), aamnaa@bzu.edu.pk (A. Khan), aman_stat@yahoo.com (M. Amanullah), mmibrahim@ksu.edu.sa (M.E. Bakr), Arwash@ksu.edu.sa (A.M. Alshangiti), samson.balogun@uef.fi (O.S. Balogun), mohammed.yousof@yahoo.com (M. Yusuf).

https://doi.org/10.1016/j.heliyon.2024.e24225
Received 10 October 2023; Received in revised form 15 December 2023; Accepted 4 January 2024
Available online 12 January 2024

capturing the excess zeroes more effectively. This leads to enhanced model accuracy, especially in situations where the excess zeroes significantly impact the data distribution. The significance of ZIP models lies in their ability to effectively handle situations where the prevalence of zero counts challenges traditional regression models, thereby improving the accuracy and reliability of statistical analyses in various domains.

Lambert [1] proposed this model to handle the presence of excess zeros in count data. The ZIPRM is a mixture of a zero-part component distribution and a non-zero-part component distribution. That, the first takes zero with a high probability. π_i that follows the Poisson distribution, whereas the second takes values with a low probability $(1 - \pi_i)$.

$$y_i \sim \begin{cases} 0 \quad with\ probability\ \pi_i \\ Poisson\ (\mu)\ with\ probability\ (1-\pi_i) \end{cases}$$

The probability density function of ZIP can be written as below.

$$Pr(y) = \begin{cases} \pi + (1 - \pi) exp\ (-\mu) \quad if\ k = 0 \\ (1 - \pi) \dfrac{exp(-\mu)\ \mu^k}{k!} \quad if\ k > 0 \end{cases} \tag{1}$$

Where $\mu \geq 0$, $0 \leq \pi \leq 1$, It has two parameters π and μ. It indicates as ZIP (π, μ). The mean of the ZIPRM and the variance is given as $E(Y) = (1-\pi)\mu$, $V(Y) = \mu(1-\pi)(1-\mu\pi)$. With

When we add the link function into a probability distribution, then the ZIPRM is obtained as

$$log(\mu_i) = \acute{Z}\beta$$

$$logit(\pi_i) = \{\pi_i / (1 + \pi_i)\} = \gamma$$

Where Z represents the vector of the independent variable whose value is zero processes and γ is the regression coefficient vector.

In the realm of biostatistics, Lindsey [2] employed the Generalized Linear Model (GLM). In real-life scenarios, there often exist instances where response variables yield two outcomes, representing success or failure. For example, in the business industry, linear regression serves as a tool to assess the correlation between advertising expenditure and revenue generation. When relationships involve more than two variables, the application of the GLM becomes essential to observe the effects of these variables. Consider a scenario where an agricultural scientist investigates the correlation between water and fertilizer application on corn crops. The scientist employs varying levels of water and fertilizer to assess their combined impact on crop yield.

Among the primary types of zero-inflated count (ZIC) models, prominent ones include zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), and zero-inflated generalized Poisson (ZIGP).

Poisson regression, estimates the probability of events occurring within a specified time or space. Zero-inflated Poisson (ZIP) stands out as a specialized form of Poisson regression designed to handle situations where excessive zeroes are observed. Lambert introduced ZIP in 1992 as an alternative technique precisely for dealing with an abundance of zeroes, noticeable when the observed zeroes far exceed those predicted by the fitted Poisson regression model. The ZIP model comprises two main components: a degenerated distribution at the zero point and Poisson regression.

Zero-inflated Poisson (ZIP) models hold significant importance in statistical modeling and analysis due to several key reasons. ZIP models are known for their ease of fitting and their capacity to yield superior data analysis outcomes. Maximum Likelihood Estimates (MLE) are generally considered approximately normal for large samples, and constructing their confidence intervals often involves inverting likelihood ratio tests or relying on the approximate normality of the MLE. The method based on the likelihood ratio test is typically favored over MLE for constructing confidence intervals due to its superior performance. Recent studies and current research, such as the ones already mentioned, validate and expand the use of ZIP models in a variety of domains, demonstrating their continued significance and relevance in modern statistical analysis.

Multicollinearity in generalized linear models (GLMs) poses similar challenges as in traditional linear regression models. GLMs extend regression analysis to handle non-normally distributed response variables and nonlinear relationships between predictors and the response.

In GLMs, multicollinearity occurs when predictor variables are highly correlated. Thus, numerous issues occur, and for reliable and interpretable results, it is essential to manage multicollinearity.

Ridge regression is a method for reducing the issues brought on by multicollinearity. A penalty term is added to the regression model as part of this regularisation technique in order to reduce the coefficients, especially those of highly correlated variables. The ridge penalty helps lower the variance of the coefficient estimates by shrinking the coefficients towards zero but without completely eliminating them. Multicollinearity can be effectively handled with ridge regression. Ridge regression reduces the influence of multicollinearity on the estimate procedure by penalizing the regression coefficients. By balancing the trade-off between variance and bias it enhances the prediction performance of the model and permits more stable and trustworthy coefficient estimations.

By limiting the variance inflation of coefficients brought on by highly correlated variables, ridge regression essentially functions as a remedy for multicollinearity, improving the stability and precision of the regression model. When working with datasets where multicollinearity is a common problem, it is especially helpful.

However, different techniques of the biased estimators are constructed to overcome the problem of multicollinearity, such as Al-Hassan Y.M. [3], Mansson and Shukur [4], Kibria et al. [5], Chang [6], Asar and Genc [7], Rashad et al. [8], Akram [9], Lukman [10], Ertan et al. [11].

Hoerl and Kennard [12] proposed the concept of ridge estimator (RE) to analyze the statistical data in the presence of multicollinearity. In the ridge estimator, it has a small positive biased amount into the diagonal element of the $X'VX$ matrix. The ridge estimator is written as below,

$$\beta_{RE} = (ZVZ + KI)^{-1}(ZVZ)\beta_{ML} \tag{2}$$

Where K considers the biased or ridge parameter, which is selected according to the required biasedness to the estimator, in the case of K = 0, the ridge estimator (RE) becomes the Ordinary Least Square (OLS) estimator. The amount of ridge parameter related to the (ZVZ) matrix. RE failed to identify the ill-conditioning in the presence of multicollinearity.

There are several empirical studies presented in the literature to estimate and compare the performance of the ridge estimator, many of them are as follows: Algamal [13], Lukman et al. [14], Younus et al. [15], Arum et al. [16] and Abonazel et al. [17]. Therefore, we proposed the new modified biased estimator for zero-inflated Poisson regression model.

1.1. *Proposed biased estimators*

We followed the scheme of Alkhamisi et al. [18] and proposed three biased estimators on the basis of the defined literature. Our proposed estimators are given below:

$$\widehat{k}_1 = max\left\{\frac{\lambda}{(n-p) + \lambda\widehat{\alpha}^2}\right\} \tag{3}$$

$$\widehat{k}_2 = median\left\{\frac{\lambda}{(n-p) + \lambda\widehat{\alpha}^2}\right\} \tag{4}$$

$$\widehat{k}_3 = \frac{1}{p}\sum_{i=0}^{n}\frac{\lambda}{(n-p) + \lambda\widehat{\alpha}^2} \tag{5}$$

1.2. *Simulation design*

In this section, we analyze the Mean Squared Errors (MSEs) of estimators for the Zero-Inflated Poisson Regression Model (ZIPRM). We assess these estimators based on various approaches, including the traditional Maximum Likelihood (ML) method, pre-existing biased estimators, and newly proposed estimators. This evaluation is conducted through the implementation of a Monte Carlo simulation scheme.

For evaluating the different biased estimators, we primarily focus on the MSE criterion.

$$MSE = \frac{\sum_{i}^{R}(\widehat{\beta}_i - \beta)^T(\ \widehat{\beta}_i -\ \beta)}{R} \tag{6}$$

Where R is the total number of repetitions (5000) in a scheme of simulation and the coefficients of the parameter that are selected by setting $\sum_{i=1}^{n}\beta_i^2\ = 1$.

Following McDonald (1975), we generated the simulated data using the equation.

$$z_i = \left(\sqrt{1-\rho^2} * w_{ij}\right) + \rho^2 * w_{ip};\ j = 1, 2, ..., p \ \ i = 1, 2, ..., n \tag{7}$$

Where ρ^2 indicates the correlation among explanatory variables and w_{ij} the pseudo-random number. There are four number of ρ^2 in the simulation, 0.85, $0.90, 0.95,$ and 0.99 respectively. These levels of correlations are used in the data set for simulation. It is also observed that dependent variables are also generated from above defined statistical equation, and binary variables are generated from pseudo-random numbers from the binomial distribution, where $\pi_i = \frac{\exp\ (z_i{'}\gamma)}{1+\exp\ (z_i{'}\gamma)}$ Having binary variable 1 and intercept belongs to $\gamma,$ and the binary variable with response one is generated by using $\mu_i\ = \exp\ (\beta_0\ + \beta_1 x_1\ + \beta_2 x_2\ +\ \dots\ + \beta_p x_p$. The explanatory variable, called regressor considered as 2 & 3, and the slop of the parameters are estimated by using $\sum_{j=1}^{p}{\beta_j}^2\ = 1$ so in this context Poisson regression intercept is generally equal to zero. However, in the ZIP regression model, the probability of both zeroes and once is possible; therefore, the intercept of the logit model is observed differently. In case, when the intercept is equal to 0, both events (zeroes & once) have an equal chance of occurrence. In this context, when the intercept is observed as positive, the probability of zeroes inflated is higher than once, while in this study number of intercept are set as 0, 1, 2 and 3. The sample size of the simulation is fitted in specific order represented as 25, 50, 100, 150 & 200.

2. Results discussion

Table 1 through 8 present simulation results that assess the performance of our recently suggested estimators in various settings.

Table 1
Estimated MSE's when $p = 2$ & $logit = 0$.

n	ρ	MLE	K1	K2	K3
50	0.85	78.237	6.8469	4.3987	2.2819
	0.9	131.4851	6.9993	5.4137	2.2971
	0.95	362.9292	6.9726	5.2619	2.1294
	0.99	232.3076	6.8973	4.2675	2.1398
25	0.85	755.555	6.2724	4.1446	2.1446
	0.9	322.021	4.0221	4.0118	2.0172
	0.95	149.023	4.0113	4.2196	2.0042
	0.99	44.409	4.3035	3.1848	1.1848
100	0.85	63.3826	4.2547	2.1182	1.1182
	0.9	95.4922	3.2545	2.1174	1.1174
	0.95	99.3634	3.3444	2.2118	1.2118
	0.99	99.8491	3.3501	1.2498	0.2188
150	0.85	265.4028	2.2547	0.1182	0.1852
	0.9	282.4036	1.2545	0.1174	0.1161
	0.95	313.9426	0.3444	0.2118	0.1873
	0.99	394.2804	0.3501	0.2188	0.1165
200	0.85	98.2162	0.2522	0.1145	0.1145
	0.9	56.2416	0.3241	0.1692	0.1692
	0.95	51.7444	0.2987	0.1621	0.1621
	0.99	49.5699	0.2934	0.1541	0.1541

Table 2
Estimated MSE's when $p = 2$ & $logit = 1$.

n	ρ	MLE	K1	K2	K3
25	0.85	737.294	0.3203	0.2086	0.2086
	0.9	1002.41	0.3224	0.2087	0.2087
	0.95	698.254	0.3559	0.2554	0.2181
	0.99	350.108	0.3761	0.2761	0.2761
50	0.85	442.051	0.2843	0.2262	0.2514
	0.9	482.126	0.2644	0.2644	0.5995
	0.95	852.011	0.2541	0.2448	0.6620
	0.99	553.287	0.2263	0.2167	0.6642
100	0.85	326.450	0.3852	0.2545	0.2545
	0.9	221.5563	0.2576	0.1264	0.1264
	0.95	546.2347	0.4062	0.2892	0.2892
	0.99	987.565	0.3986	0.2696	0.2696
150	0.85	47.5642	0.2539	0.1189	0.4689
	0.9	42.7418	0.2535	0.1186	0.4186
	0.95	32.997	0.3472	0.2211	0.3311
	0.99	17.2527	0.2549	0.2617	0.1217
200	0.85	216.1638	0.2529	0.1493	0.1173
	0.9	180.662	0.2532	0.1252	0.1182
	0.95	131.1501	0.3169	0.1258	0.1918
	0.99	125.356	0.3273	0.1460	0.0146

The sample size 'n,' number of explanatory variables 'p,' and collinearity levels varied in these cases. When we evaluated the performance of the estimators, we found a pattern: our suggested methods' Mean Squared Errors (MSEs) were consistently smaller than those of the Maximum Likelihood Estimators (MLEs) and of the estimators that were already in use. With regard to Table 1, when 'p = 2,' it was clear that our suggested estimators performed much better than MLEs when high degrees of collinearity were present. Analyzing the effects of different correlation degrees on the computed MSEs was part of this examination. A pattern of declining MLE values across various correlation levels (ρ) is seen as sample size (n) grows, indicating greater estimation accuracy with bigger sample sizes (see Table 9).

Similarly, the MSE values of proposed estimators fluctuate as the correlation level rises (ρ goes from 0.85 to 0.99), demonstrating the influence of correlation on estimator accuracy.

Notably, bigger MSE values are typically produced by higher correlation levels for smaller sample sizes (n = 25, for example), indicating higher estimate errors as a result of increasing collinearity among variables.

Table 1 offers insights into how different combinations of sample size and correlation levels affect the accuracy of estimators in a Zero-Inflated Poisson Regression Model when the number of explanatory variables is fixed at 2 and the logit value is set at 0. The performance of biasing parameters k1is good in most of the scenarios and k2 shows a near-equal trend like k3 across all simulation scheme criteria but k3 proved to be the best out of all proposed estimators. However, with the introduction of explanatory variable 3, the outcomes vary significantly concerning these criteria.

Table 3
Estimated MSE's when $p = 2$ & $logit = 2$.

n	ρ	MLE	K1	K2	K3
25	0.85	465.028	0.4821	0.3811	0.4123
	0.9	229.007	0.3703	0.2622	0.3702
	0.95	328.054	0.3244	0.2416	0.3197
	0.99	163.342	0.3218	0.2232	0.2938
50	0.85	25.017	0.4155	0.3657	0.3144
	0.9	69.442	0.4027	0.3211	0.3022
	0.95	67.549	0.4083	0.3205	0.3009
	0.99	210.488	0.4011	0.3168	0.2987
100	0.85	42.0347	0.4057	0.2786	0.2786
	0.9	51.0088	0.4195	0.3025	0.3025
	0.95	68.2222	0.3963	0.2634	0.2634
	0.95	9963.75	0.4124	0.2874	0.2874
150	0.85	540.577	0.2544	0.1207	0.1207
	0.9	701.7697	0.2538	0.119	0.119
	0.95	611.9216	0.2545	0.1209	0.1209
	0.99	820.637	0.3646	0.2394	0.2394
200	0.85	713.2106	2.4596	0.3331	0.2077
	0.9	409.4623	2.2517	0.2531	0.1183
	0.95	541.2113	2.2727	0.3295	0.2036
	0.99	576.4957	2.4098	0.2546	0.1224

Table 4
Estimated MSE's when $p = 2$ & $logit = 3$.

n	ρ	MLE	K1	K2	K3
25	0.85	226.365	0.4521	0.3199	0.2419
	0.9	772.325	0.4366	0.2962	0.2562
	0.95	655.021	0.4180	0.2726	0.1856
	0.99	692.150	0.4017	0.2499	0.1649
50	0.85	482.372	0.3854	0.2252	0.1343
	0.9	433.028	0.3211	0.2025	0.1137
	0.95	323.351	0.2899	0.1696	0.1696
	0.99	134.396	0.2141	0.2526	0.2085
100	0.85	552.029	0.2562	0.3227	0.1827
	0.9	361.567	0.2575	0.3269	0.1289
	0.95	102.708	0.4288	0.3012	0.1018
	0.99	576.723	0.4247	0.2982	0.1182
150	0.85	52.6247	0.2544	0.2202	0.1522
	0.9	48.7975	0.2539	0.2196	0.1166
	0.95	49.0234	0.3643	0.2843	0.2323
	0.99	32.3549	0.3574	0.1372	0.2278
200	0.85	83.5286	0.3387	0.1178	0.1165
	0.9	76.6578	0.2534	0.2199	0.1199
	0.95	90.4620	0.3375	0.2129	0.1125
	0.99	124.9344	0.2536	0.1282	0.0202

2.1. Real data application

We apply our proposed estimators to the real-life data set consisting of biochemist information. The data set has 915 observations. The dependent variable is the number of published articles, and there are five independent variables explained as follows.

- Gender: Coded as 0 for male and 1 for female
- MentorArts: The number of published articles within the last three years of Ph.D. Study
- Prestige: prestige time of Ph.D. student
- Marriage: Marriage status is 0 for single and 1 for married
- Children: The number of children up to 5 years of age

The performance evaluation of our proposed estimators is detailed in the given below table.

These coefficients and associated statistics are crucial in determining the impact of each predictor variable on the response variable in the count model. The estimates, along with their standard errors and significance levels, help assess the significance and magnitude of each predictor's effect on the count response. Afterward, the MSE of the ML estimator and the three proposed estimators that

Table 5
Estimated MSE's when $p = 3$ & $logit = 0$.

n	ρ	MLE	K_1	K_2	K_3
25	0.85	57.2927	0.3055	0.3123	0.146
	0.9	1600.92	0.2997	0.2282	0.1322
	0.95	626.698	0.3157	0.2423	0.1599
	99	585.984	0.3599	0.2899	0.1756
50	0.85	32.659	0.2607	0.2506	0.0824
	0.9	68.412	0.2985	0.2441	0.1365
	0.95	56.234	0.2606	0.2255	0.0814
	0.99	48.0766	0.2624	0.1248	0.0845
100	0.85	52.332	0.2544	0.2297	0.0719
	0.9	98.248	0.2122	0.1458	0.0125
	0.95	229.49	0.3796	0.4155	0.1982
	0.99	587.246	0.3747	0.3165	0.1848
150	0.85	32.5404	0.3248	0.3144	0.1417
	0.9	43.7688	0.3286	0.3051	0.1353
	0.95	64.0837	0.3143	0.3247	0.1459
	0.99	311.572	0.3334	0.2716	0.1406
200	0.85	92.3138	0.2896	0.3077	0.1054
	0.9	15.1034	0.2914	0.2776	0.1073
	0.95	10.9194	0.2517	0.2292	0.0649
	0.99	4.6654	0.2518	0.2223	0.0652

Table 6
Estimated MSE's when $p = 3$ & $logit = 1$.

n	ρ	MLE	K1	K2	K3
25	0.85	3205.59	0.3516	0.4121	0.2161
	0.9	495.422	0.3954	0.3947	0.2633
	0.95	254.296	0.2956	0.2669	0.2658
	0.99	681.236	0.2134	0.2245	0.2345
50	0.85	438.279	0.2569	0.3334	0.2661
	0.9	617.479	0.2817	0.2572	0.1035
	0.95	683.294	0.2765	0.2146	0.1125
	0.99	924.321	0.2458	0.2265	0.1254
100	0.85	970.887	0.4059	0.6689	0.2389
	0.9	629.706	0.4069	0.4485	0.2237
	0.95	2002.16	0.4079	0.4341	0.2325
	0.99	732.149	0.4142	0.4009	0.2155
150	0.85	523.124	0.2527	0.2084	0.0684
	0.9	245.231	0.2528	0.2394	0.0686
	0.95	94.3439	0.3388	0.3058	0.1406
	0.99	401.48	0.3547	0.3057	0.1552
200	0.85	631.081	0.3201	0.3247	0.1284
	0.9	214.738	0.2512	0.2114	0.0671
	0.95	158.333	0.2351	0.2429	0.0666
	0.99	33.3708	0.3155	0.2046	0.1174

occurred are given below, suggesting the superiority of our proposed criteria.

MSE:	MLE: 28.264	k_1 : 8.005	k_2 : 4.231	k_3 : 2.219

3. Conclusion

In this section, we focus on our findings from exploring Zero-Inflated Poisson regression. We investigate both simulated scenarios and the application of this model in real-life datasets. Our exploration delves into the influence of various factors on result inflation and deeply analyzes ZIP ridge regression behavior when faced with different levels of multicollinearity among explanatory variables. Within our study, we introduce and assess a variety of ridged biasing estimators under varying degrees of multicollinearity, using Mean Squared Error (MSE) as our benchmark for performance evaluation. We systematically varied correlation levels at 0.85, 0.90, 0.95, and 0.99.

Throughout our investigation, we varied intercept counts from 0 to 3 for the logit value of ZIP and maintained a single count of 0 for Poisson. By utilizing simulation techniques, we formulated two sets of explanatory variables: one containing two variables and another comprising three. For estimating regression coefficients within the ZIP regression model, we opted for the GLM. We employed MLE as our primary statistical tool to manage unknown coefficients. Each simulation was conducted with 5000 replications to ensure robustness.

Table 7

Estimated MSE's when $p = 3$ & $logit = 2$.

n	ρ	MLE	K_1	K_2	K_3
25	0.85	605.593	0.2111	0.4251	0.2115
	0.9	321.111	0.2089	0.2154	0.2321
	0.95	298.632	0.2045	0.2115	0.2265
	0.99	211.114	0.2001	0.2001	0.2111
50	0.85	6.6335	0.2734	0.2586	0.1092
	0.9	14.2135	0.2014	0.2011	0.1452
	0.95	19.5844	0.2717	0.2153	0.1021
	0.99	49.2012	0.1485	0.2058	0.2011
100	0.85	63.2021	0.1145	0.1589	0.1963
	0.9	58.7986	0.2545	0.2445	0.0725
	0.95	34.9705	0.2559	0.2305	0.0779
	0.99	95.857	0.4207	0.3462	0.2213
150	0.85	58.7211	0.3633	0.4147	0.1742
	0.9	1892.92	0.3607	0.3797	0.1637
	0.95	144.317	0.3055	0.2228	0.1567
	0.99	100.997	0.3562	0.2635	0.1593
200	0.85	231.557	0.2521	0.2017	0.0676
	0.9	256.222	0.2521	0.2168	0.0167
	0.95	244.385	0.2522	0.2432	0.0677
	0.99	150.89	0.3238	0.2198	0.1582

Table 8

Estimated MSE's when $p = 3$ & $logit = 3$.

n	ρ	MLE	K_1	K_2	K_3
25	0.85	56.233	0.2355	0.2224	0.0158
	0.9	32.014	0.2215	0.2104	0.0114
	0.95	28.715	0.2114	0.2118	0.0025
	0.99	26.229	0.2018	0.2226	0.0189
50	0.85	81.023	0.2009	0.2548	0.1156
	0.9	62.001	0.3256	0.208	0.1025
	0.95	19.0372	0.2648	0.2559	0.0879
	0.99	118.316	0.2839	0.1508	0.1187
100	0.85	23.586	0.2558	0.2534	0.0785
	0.9	6872.37	0.4395	0.5235	0.2616
	0.95	233.001	0.2558	0.2181	0.2078
	0.99	291.361	0.4326	0.4631	0.2376
150	0.85	48.2604	0.3641	0.4456	0.1718
	0.9	81.7675	0.3698	0.4028	0.1745
	0.95	53.002	0.2153	0.2473	0.0695
	0.99	650.634	0.3658	0.2658	0.1655
200	0.85	22.025	0.2522	0.1974	0.0168
	0.9	49.7968	0.3359	0.3323	0.1411
	0.95	84.8502	0.3411	0.3095	0.1506
	0.99	437.615	0.3321	0.2413	0.1237

Table 9

Count model coefficients.

- The estimate of β_1 is 23.1303, with a standard error of 12.582. The Z-value is 2.460, corresponding to a p-value of 0.043, suggesting statistical significance.
- The estimate of β_2 is 2.1921, with a standard error of 4.2602. The Z-value is 0.996, corresponding to a highly significant p-value of 0.003.
- Each coefficient estimates of $\beta_3, \beta_{4,}\beta_5$, along with its associated standard error, Z-value, and p-value, is provided. Notably, they all have low p-values (0.00), indicating high statistical significance.

Coefficients	Estimates	Std.Error	Z Value	Pr (>|z|)
β_0	23.1303	12.582	2.460	0.043
β_1	2.1921	4.2602	0.996	0.003
β_2	−4.1798	3.6109	−1.601	0.00
β_3	3.2409	6.0154	0.362	0.00
β_4	16.994	9.5301	0.284	0.00
β_5	3.1940	2.0089	0.315	0.00

Consistently across our diverse scenarios, our observations underscored the superior performance of proposed estimators k1, k2, and k3 across various aspects. Notably, when dealing with simulations involving two explanatory variables, both k3 and k2 displayed equally superior performance compared to k1. However, in scenarios with three explanatory variables, k6 exhibited notably commendable performance, followed by k1 and k2. Our comprehensive analysis ultimately led us to the decisive conclusion that k3 surpasses all other estimators, emerging as the optimal choice for robust regression estimation. This conclusion stems from a thorough evaluation of their performance across diverse scenarios, firmly establishing k3 as the most reliable estimator for precise regression estimations.

4. Future recommendation

- Undertake comprehensive validation investigations utilizing heterogeneous datasets from distinct fields to assess the resilience and applicability of the adjusted estimator in a range of real-world contexts.
- Examine other methods or improvements to minimise bias and improve estimating efficiency, perhaps by using different weighting schemes or modifying the model.
- Examine ways to improve the updated estimator's prediction ability and suitability for use with intricate datasets by adding more variables or predictors.
- Evaluate the practical utility and impact of the updated estimator by applying it to certain businesses or professions (e.g., healthcare, ecology, finance, or social sciences) where zero-inflated data is common.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Muhammad Zeeshan: Writing – original draft. **Aamna Khan:** Conceptualization. **Muhammad Amanullah:** Supervision. **M.E. Bakr:** Software. **Arwa M. Alshangiti:** Resources. **Oluwafemi Samson Balogun:** Project administration. **M. Yusuf:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research project was supported by the Researchers Supporting Project Number (RSPD2024R1004), King Saud University, Riyadh, Saudi Arabia.

References

[1] Diane Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, Technometrics 34 (1) (1992) 1–14.
[2] James K. Lindsey, Applying Generalized Linear Models, Springer Science & Business Media, 2000.
[3] Al-Hassan, M. Yazid, Performance of a new ridge regression estimator, J. Assoc. Arab Univ. Basic and Appl. Sci. 9 (1) (2010) 23–26.
[4] Kristofer Månsson, Ghazi Shukur, A Poisson ridge regression estimator, Econ. Modell. 28 (4) (2011) 1475–1481.
[5] BM Golam Kibria, Kristofer Månsson, Ghazi Shukur, Performance of some logistic ridge regression estimators, Comput. Econ. 40 (2012) 401–414.
[6] Xinfeng Chang, On the almost unbiased ridge and Liu estimator in the logistic regression model, in: International Conference on Social Science, Education Management and Sports Education, vol. 2015, Atlantis Press, 2015, pp. 1658–1660.
[7] Yasin Asar, Aşır Genç, Two-parameter ridge estimator in the binary logistic regression, Commun. Stat. Simulat. Comput. 46 (9) (2017) 7088–7099.
[8] Nadwa Khazaal Rashad, Nawal Mahmood Hammood, Zakariya Yahya Algamal, Generalized ridge estimator in negative binomial regression model, in: Journal of Physics: Conference Series, vol. 1897, IOP Publishing, 2021 012019 no. 1.
[9] M.N. Akram, M.R. Abonazel, M. Amin, B.G. Kibria, N. Afzal, A New Stein Estimator for the Zero-inflated Negative Binomial Regression Model, 2022.
[10] Adewale F. Lukman, Mohammad Arashi, Vilmos Prokaj, Robust biased estimators for Poisson regression model: simulation and applications, Concurrency Comput. Pract. Ex. 35 (7) (2023) e7594.
[11] Esra Ertan, Kadri Ulaş Akay, A New Class of Poisson Ridge-type Estimator, 2023.
[12] Hoerl, Robert W. Kannard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", 2000.
[13] Zakariya Yahya Algamal, A new method for choosing the biasing parameter in ridge estimator for generalized linear model, Chemometr. Intell. Lab. Syst. 183 (2018) 96–101.
[14] Adewale F. Lukman, BM Golam Kibria Issam Dawoud, Zakariya Y. Algamal, Benedicta Aladeitan, A new ridge-type estimator for the gamma regression model, Sci. Tech. Rep. 2021 (2021) 1–8.
[15] Farah Abdul Ghani Younus, Rafal Adeeb Othman, Zakariya Yahya Algamal, Modified Ridge Estimator in Zero-Inflated Poisson Regression Model, 2022.
[16] Kingsley C. Arum, Fidelis I. Ugwuowo, Henrietta E. Oranye, Robust modified jackknife ridge estimator for the Poisson regression model with multicollinearity and outliers, Sci. Afr. (2022) e01386.
[17] Mohamed R. Abonazel, Fuad A. Awwad, Elsayed Tag Eldin, BM Golam Kibria, and Ibrahim G. Khattab. "Developing a Two-Parameter Liu Estimator for the COM–Poisson Regression Model: Application and Simulation, 2023.
[18] Khalaf Alkhamisi, Ghazi Shukur, "Some Modifications for Chosing Ridge Parameter", 2006.