TWMS J. App. and Eng. Math. V.12, N.1, 2022, pp. 235-246

ALMOST UNBIASED RIDGE ESTIMATOR IN THE ZERO-INATED POISSON REGRESSION MODEL

YOUNUS AL-TAWEEL[1], ZAKARIYA ALGAMAL [2], §

Abstract. The zero-inflated Poisson regression (ZIP) model is a very popular model for count data that have extra zeros. In some situations, the count data are correlated and so multicollinearity exists among the explanatory variables. Thus, the traditional maximum likelihood estimator (MLE) becomes not a reliable estimator because the mean squared error (MSE) becomes inflated. The ridge estimator (RE) is used to overcome this problem. In this work, an almost unbiased ridge estimator for the ZIP model (AUZIPRE) is proposed to tackle the multicollinearity problem in count data. We investigate the behavior of the proposed estimator using a simulation study. Using the MSE measure, the results of the proposed estimator are compared with those of the RE and the MLE. Furthermore, we apply the proposed estimator on a real dataset. The results show that the performance of AUZIPRE outperforms for that of the RE and the MLE in the existing of the multicollinearity among the count data in the ZIP model.

Keywords: Count data, multicollinearity, zero-inflated Poisson regression, ridge estimator, almost unbiased ridge estimator.

AMS Subject Classification: 83-02, 99A00

1. Introduction

Regression models are commonly used in many disciplines of science, such as economic, biomedical, environment and so forth. Count data are usually analyzed using Poisson regression models. Suppose we have counts of events, $Y_i, i = \dots, n$, in a period of time. Thus, Y_i are random variables that have a Poisson distribution

$$p(y_i) = \frac{e^{-\pi_i}\pi_i^{y_i}}{y_i!}, \quad y_i = 0, 1, \dots \tag{1}$$

where $\pi_i > 0$ is s the average of events and it is equal to the mean and the variance of Y, $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = \pi_i$. The Poisson model is very common when the count data are unbounded. Now, let $\mathbf{x}_i = (x_1, \dots, x_p)$, be a vector of explanatory variables in the design matrix $\mathbf{X}$ and $\boldsymbol{\beta}$ be a vector of coefficient parameters. Using a link function, we have

[1] University of Mosul, College of Education for Pure Science, Department of Mathematics, Iraq.
e-mail: younus.altaweel@uomosul.edu.iq; ORCID: https://orcid.org/0000-0001-7167-8079.
[2] University of Mosul, College of Computers Sciences and Mathematics, Department of Statistics and Informatics, Iraq.
e-mail: zakariya.algamal@uomosul.edu.iq; ORCID: https://orcid.org/0000-0002-0229-7958.
§ Manuscript received: January 22, 2020; accepted: Febuary 28, 2020.
TWMS Journal of Applied and Engineering Mathematics, Vol.12, No.1

the Poisson regression model. The maximum likelihood estimator (MLE), $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ can be obtained using iterative weighted least square algorithm [1, 2].

In some cases, however, the count data may include a huge number of zeros or the count data exhibit overdispersion where the variance value of the response variable exceeds the value of the mean. In this case, if the standard Poisson regression model is applied, the variance of the estimated coefficients parameters will be underestimated. Hence, the zero-inflated Poisson (ZIP) model becomes more appropriate than the Poisson regression model for analyzing such kind of count data [3, 4, 5]. In the ZIP model, if the explanatory variables in the count data are highly correlated, the MLE may not perform well. This is because the variance of the estimated coefficient parameters will be high which introduces risks in their interpretation [6, 7]. This multicollinearity is often seen in count data models in applied economic studies where the explanatory variables are highly correlated. [8].

Ridge estimators (RE) are used to solve the multicollinearity problem in the correlated data for the ZIP model. For example, [8] used a RE for the ZIP model and they demonstrated their results with a simulation study and a real dataset. However, the RE may have a large bias. In order to overcome this problem, [9] proposed the almost unbiased ridge estimator (AURE) for linear regression model. Hence, we propose in this work the AURE for the ZIP model. We use a Monte Carlo simulation study to investigate the performance of the AURE for the ZIP model where we use the mean squared error (MSE) as a measure. In the Monte Carlo simulation study, we use different combinations of the sample size, different numbers of the explanatory variables, different levels of the correlation among the explanatory variables and different values of the intercept of the logit model. The results of the Monte Carlo simulation study show that the AURE estimator for the ZIP model outperforms the RE and the ML estimators. Moreover, a real dataset was also used to compare the behavior of the AURE with that of the RE and MLE. The results of the real dataset agree with those of the Monte Carlo simulation study.

This research is organized as follows. Section 2 presents the methodology of the zero-inflated Poisson regression model. In Section 3, the ridge estimator is reviewed for the ZIP model. In Section 4, we present the almost unbiased ridge estimator for the ZIP model (AUZIPRE). In addition, several estimators of the ridge parameter are presented. In section 5, a Monte Carlo simulation is conducted to investigate the performance of the AUZIPRE in terms of the MSE. In Section 6, we apply the proposed AUZIPRE on a real dataset. Finally, in Section 7, the conclusion is given.

2. Zero-inflated Poisson regression model

The zero-inflated Poisson (ZIP) model was proposed by [4] for modeling zero-inflation in count data. The ZIP model can be seen as a mixture model for count data with extra zeros. The zeros in the count data for the ZIP model can be classified into two types. The first one comes from a non-susceptible group and it is known as structural zeros. The structural zero occurs with probability θ_i. The second type of zeros in the count data for the ZIP model comes from a susceptible group and it is known as random zeros. The random zero occurs with probability $(1-\theta_i)$ and has a Poisson distribution with mean μ_i [10].

The formula of the ZIP model is given by

$$p(Y=y) \quad = \quad \begin{cases} \theta_i + (1-\theta_i)e^{-\mu_i}, & \text{if } y_i = 0 \\ (1-\theta_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{if } y_i = 1, 2, \ldots,. \end{cases} \tag{2}$$

where the indicator function, $I_{\{\cdot\}}$, is for zero events, $\mu_i = \exp(\mathrm{x}_i\boldsymbol{\beta})$ represents the expected ith count for the ith observation and the probability $\theta_i \in [0, 1]$ is for the extra zeros [11].

The probability of extra zeros, θ_i, is given by

$$\theta_i = \frac{\exp(q_i\delta)}{1+\exp(q_i\delta)}, \tag{3}$$

where q_i is the ith row of the data logit matrix $\mathbf{Q}$.

The ZIP model reduces to a Poisson distribution when $\theta_i = 0$. When $\theta_i > 0$, however, there will be overdispersion in the distribution of Y_i which means there will be zero-inflation. The MLE of the ZIP model parameters can be obtained using Fisher scoring methods [12].

3. Zero-Inflated Poisson ridge estimator

In the presence of multicollinearity among the explanatory variables in the count data, the MLE may not be a reliable estimator. This is because the eigenvalues will be small for the explanatory variables that are highly correlated and so the MSE will be inflated [13, 14]. The ridge estimator (RE) is proposed by [15] to overcome this problem for linear regression where a positive amount is added to the diagonal of the matrix $\mathbf{X}^T\mathbf{X}$.

[8] proposed the RE for the ZIP model. The ZIP ridge estimator (ZIPRE) is defined by

$$\hat{\boldsymbol{\beta}}_{\text{ZIPRE}} = (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{MLE}}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is the MLE. The parameter $k \geq 0$ is called the ridge parameter. When the ridge parameter is zero, we have $\hat{\boldsymbol{\beta}}_{\text{ZIPRE}} = \hat{\boldsymbol{\beta}}_{\text{MLE}}$. However, we have $\|\hat{\boldsymbol{\beta}}_{\text{ZIPRE}}\| < \|\hat{\boldsymbol{\beta}}_{\text{MLE}}\|$ when $k > 0$ [16]. The non-diagonal elements of the matrix $\hat{\mathbf{W}}$ are zeros and the ith diagonal element equals to $\hat{\mu}_i$.

The mean squared error (MSE) of the MLE is defined by

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) &= \text{E}(\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta})^T \text{E}(\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta}) \\ &= \hat{\tau}\Sigma_{j=1}^{p}\frac{1}{\lambda_j}, \end{aligned} \tag{5}$$

where τ represents the dispersion parameter that is estimated by $\hat{\tau} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/(n - p - 1)$ [8]. The λ_j is the jth eigenvalue of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix [17]. The MSE of RE is obtained by

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{RE}}) &= \text{E}(\hat{\boldsymbol{\beta}}_{\text{RE}} - \boldsymbol{\beta})^T \text{E}(\hat{\boldsymbol{\beta}}_{\text{RE}} - \boldsymbol{\beta}) \\ &= \hat{\tau}\Sigma_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + k)^2} + k^2\Sigma_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j + k)^2}, \end{aligned} \tag{6}$$

where α_j is defined as the jth element of $\psi^T\boldsymbol{\beta}$ and ψ is the eigenvector of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix [17].

4. The almost unbiased Zero-Inflated Poisson ridge estimator

The RE for tackling the multicollinearity problem may have a large bias when the value of the ridge parameter is large. [9], [21], [23] and [22] proposed the almost unbiased ridge estimator (AURE) for linear regression model to solve the multicollinearity problem. Hence, we present in this work the AURE for the ZIP model. The almost unbiased ridge estimator for the ZIP (AUZIPRE) model can overcome the multicollinearity problem and is able to decrease the bias of the ZIPRE. The AUZIPRE is defined by

$$\hat{\boldsymbol{\beta}}_{\text{AUZIPRE}} = (\mathbf{I} - (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X} + k\mathbf{I})^{-2}k^2)\hat{\boldsymbol{\beta}}_{\text{MLE}}. \tag{7}$$

By taking the expectation of equation (7), we have

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}}) &= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\mathrm{E}(\mathbf{y}) \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\boldsymbol{\beta} \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)\boldsymbol{\beta}. \qquad (8)
\end{aligned}$$

The bias of the AUZIPRE is given by

$$\begin{aligned}
\mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}}) &= \mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}})-\boldsymbol{\beta} \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)\boldsymbol{\beta}-\boldsymbol{\beta} \\
&= -k^2(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}\boldsymbol{\beta} \\
&= -k^2\sum_{j=1}^{p}\frac{\alpha_j}{(\lambda_j+k)^2}. \qquad (9)
\end{aligned}$$

The variance of the AUZIPRE is given by

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}}) &= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)\mathrm{Var}(\hat{\boldsymbol{\beta}}) \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)^T \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\hat{\tau} \\
&= (\mathbf{I}-(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+k\mathbf{I})^{-2}k^2)^T \\
&= \hat{\tau}\sum_{j=1}^{p}\frac{1}{\lambda_j}\left(1-\frac{k^2}{(\lambda_j+k)^2}\right)^2. \qquad (10)
\end{aligned}$$

The MSE of the AUZIPRE is found by using equations (9) and (10)

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}}) &= \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}})+\left(\mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{AUZIPRE}})\right)^2 \\
&= \hat{\tau}\sum_{j=1}^{p}\frac{1}{\lambda_j}\left(1-\frac{k^2}{(\lambda_j+k)^2}\right)^2 \\
&+ \left(-k^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j+k)^2}\right)^2 \\
&= \hat{\tau}\sum_{j=1}^{p}\frac{(\lambda_j^2+2\lambda_j k)^2}{\lambda_j(\lambda_j+k)^4}+k^4\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j+k)^4} \\
&= \hat{\tau}\sum_{j=1}^{p}\frac{(\lambda_j+2k)^2\lambda_j+k^4\alpha_j^4}{(\lambda_j+k)^4}. \qquad (11)
\end{aligned}$$

Theorem 4.1. *In the ZIP model, we have* $\|bias(\hat{\boldsymbol{\beta}}_{AUZIPRE})\|^2 < \|bias(\hat{\boldsymbol{\beta}}_{ZIPRE})\|^2$ *for* $k > 0$.

Proof. Let $D_1 = \|\text{bias}(\hat{\boldsymbol{\beta}}_{\text{ZIPRE}})\|^2 - \|\text{bias}(\hat{\boldsymbol{\beta}}_{\text{AUZIPRE}})\|^2$. Hence, we have

$$\begin{aligned} D_1 &= \sum_{j=1}^{p} \frac{k^2\alpha_j^2}{(\lambda_j+k)^2} - \sum_{j=1}^{p} \frac{k^4\alpha_j^2}{(\lambda_j+k)^4} \\ &= \sum_{j=1}^{p} \frac{\lambda_j^2 k^2 \alpha_j^2 + 2k^3\lambda_j\alpha_j^2}{(\lambda_j+k)^4} \\ &= \sum_{j=1}^{p} k^2 \left\{ \frac{\lambda_j\alpha_j^2(\lambda_j+2k)}{(\lambda_j+k)^4} \right\}. \end{aligned}$$

Hence, for $k > 0$, the proof is completed. □

Theorem 4.2. *For the ZIP model, if* $k > \left(3\hat{\tau} - \lambda_j\alpha_j^2 + \sqrt{\lambda_j^2\alpha_j^4 + 9\hat{\tau}^4 + 10\lambda_j\alpha_j^2\hat{\tau}}\right)/4\alpha_j^2$, *for* $j = 1, \ldots, p$, *then the AUZIPRE is superior to the ZIPRE in terms of the MSE.*

Proof. Let $D_2 = \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{ZIPRE}}) - \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{AUZIPRE}})$. Hence, we have

$$\begin{aligned} D_2 &= \frac{\hat{\tau}\lambda_j}{(\lambda_j+k)^2} + \frac{k^2\alpha_j^2}{(\lambda_j+k)^2} - \frac{\hat{\tau}(\lambda_j^2+2\lambda_j k)^2}{\lambda_j(\lambda_j+k)^4} - \frac{k^4\alpha_j^2}{(\lambda_j+k)^4} \\ &= \sum_{j=1}^{n} \left(\frac{\lambda_j \left\{ (2\alpha_j^2)k^2 + (\lambda_j\alpha_j^2 - 3\hat{\tau})k - 2\hat{\tau}\lambda_j \right\} k}{(\lambda_j+k)^4} \right). \end{aligned}$$

The D_2 is a positive definite for $k > 0$, if and only if $\left\{ (2\alpha_j^2)k^2 + (\lambda_j\alpha_j^2 - 3\hat{\tau})k - 2\hat{\tau}\lambda_j \right\} > 0$. Thus, this function is quadratic of k and has the following root

$$k = \frac{\left(3\hat{\tau} - \lambda_j\alpha_j^2 + \sqrt{\lambda_j^2\alpha_j^4 + 9\hat{\tau}^4 + 10\lambda_j\alpha_j^2}\right)}{4\alpha_j^2\hat{\tau}}.$$

Hence, the AUZIPRE is superior to the ZIPRE in terms of the MSE for the ZIP model, the proof is completed. □

Theorem 4.3. *For the ZIP model, the AUZIPRE is superior to the MLE.*

Proof. Let $D_3 = \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{ML}}) - \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{AUZIPRE}})$. Hence, we have

$$\begin{aligned} D_3 &= \sum_{j=1}^{p} \frac{\hat{\tau}}{\lambda_j} - \sum_{j=1}^{p} \frac{\hat{\tau}\lambda_j^2(\lambda_j+2k)^2}{\lambda_j(\lambda_j+k)^4} - \sum_{j=1}^{p} \frac{k^4\alpha_j^2}{(\lambda_j+k)^4} \\ &= \sum_{j=1}^{p} k^2 \frac{\left\{ (\hat{\tau} - \lambda_j\alpha_j^2)k^2 + 4\hat{\tau}\lambda_j k + 2\hat{\tau}\lambda_j^2 \right\}}{\lambda_j(\lambda_j+k)^4}. \end{aligned} \tag{12}$$

From equation (12), it can be shown that D_3 is a positive definite if and only if $\left\{ (\hat{\tau} - \lambda_j\alpha_j^2)k^2 + 4\hat{\tau}\lambda_j k + 2\hat{\tau}\lambda_j^2 \right\} > 0$. Hence, the AUZIPRE is superior to the MLE in terms of the MSE for the ZIP model, the proof is completed. □

4.1. **Estimating the ridge parameter k.** In order to obtain values of the ridge estimator, k, several methods have been proposed by authors as there is no specific rule for obtaining the value of k. In this study, values of the ridge estimator, k, for the AUZIPRE in the ZIP model were proposed from the work of [18] and [19]. The estimators for the ridge parameter, k are as follows

$$k_1 = \frac{\hat{\tau}}{(\prod_{i=1}^{p} \hat{\alpha}_j^2)^{1/p}}, \quad k_2 = \text{median}(m_j^2),$$

$$k_3 = \frac{p\hat{\tau}^2}{\hat{\alpha}_j^T \hat{\alpha}_j} + \frac{1}{(\lambda_{\max} \hat{\alpha}_j^T \hat{\alpha}_j)}, \quad k_4 = \frac{p\hat{\tau}^2}{\hat{\alpha}_j^T \hat{\alpha}_j} + \frac{1}{2\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}},$$

where $\hat{\alpha}_j$ is the jth element of $\psi\hat{\boldsymbol{\beta}}_{\text{MLE}}$, ψ is the eigenvector of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix, $m_j = \sqrt{\frac{\hat{\tau}^2}{\hat{\alpha}_j^2}}$, $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix.

The k_1 and k_2 estimators were proposed by [18] for the ridge estimator in the multiple linear regression model. The k_3 and k_4 estimators were proposed by [19] for the ridge estimator in the multiple linear regression model. Hence, we will use these four estimators for the AUZIPRE using the MSE measure and compare the results with those of the ZIPRE and MLE.

5. Monte Carlo Simulation study

In this section, we investigate the performance of the AUZIPRE. This investigation is achieved by comparing the estimated MSE of AUZIPRE with the ZIPRE and MLE using a Monte Carlo simulation experiment with several different levels of multicollinearity.

The MSE measure was calculated using the following formula

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{AUZIPRE}}) = \sum_{i=1}^{R} \frac{(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})}{R}. \tag{13}$$

where $\hat{\boldsymbol{\beta}}_i$ is the ith simulated value of $\boldsymbol{\beta}$. The number of the replications in the Monte Carlo simulation is set to be $R = 1000$.

5.1. **The simulation design of experiment.** In order to generate the design of the experiment, we generated explanatory variables $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{in})$ using the following formula

$$x_{ij} = (1-\rho^2)^{1/2}\vartheta_{ij} + \rho\vartheta_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p. \tag{14}$$

where ρ is the correlation coefficient between the explanatory variables and ϑ_{ij}'s are independent pseudo-random variables. The ϑ_{ij} were simulated from the standard normal distribution. For the explanatory variables, we set $p = 2$ and $p = 4$. The level of the correlation is the key point in the design, so we considered four different values of ρ, 0.85, 0.90, 0.95 and 0.99.

Then, a binary variable was generated from the binomial distribution using pseudo-random numbers where $\theta_i = \frac{\exp(q_i\delta)}{1+\exp(q_i\delta)}$. The q_i have the value of 1 and δ consists of the intercept term only. Since the intercept of the logit model δ affects probability of obtaining zeros and ones, its value is set to be 0, 1 and 2 [8]. Then, we obtained the binary variables that have values of one from the Poisson distribution with $\mu_i = \exp(\beta_0 + x_1\beta_1 + \ldots, x_p\beta_p)$. The sum of the coefficient regression parameters $\boldsymbol{\beta}$ was assumed to be 1 and the intercept of the Poisson model was always set to be zero. The response variable, y, of the ZIP model was generated using equation (2) with different sample sizes $n = 50, 100, 150$ and 200.

5.2. **The discussion of simulation results.** This section presents the results of the MSE for the simulation experiment. Using equation (13), the MSE was calculated. Tables 1-6 show the MSE values that were calculated for the different estimators, k_1, k_2, k_3 and k_4 under various combinations of n, p and ρ for the AUZIPRE, ZIPRE and MLE.

TABLE 1. Estimated MSE when $p = 2$ and the intercept of the logit = 0.

			ZIPRE				AUZIPRE			
n	ρ	MLE	k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
50	0.85	1.540	0.834	0.651	0.571	0.534	0.688	0.561	0.519	**0.489**
	0.90	1.602	0.838	0.661	0.561	0.522	0.678	0.548	0.484	**0.449**
	0.95	1.854	0.899	0.724	0.588	0.548	0.721	0.579	0.474	**0.433**
	0.99	4.087	1.145	0.968	0.819	0.795	0.977	0.822	0.685	**0.650**
100	0.85	2.130	1.220	0.968	0.773	0.774	0.863	0.628	**0.459**	0.460
	0.90	2.440	1.350	1.140	0.910	0.911	1.020	0.804	**0.571**	0.572
	0.95	3.220	1.490	1.380	1.160	1.160	1.230	1.110	**0.839**	**0.839**
	0.99	7.462	1.669	1.623	1.520	1.514	1.561	1.521	1.381	**1.373**
150	0.85	2.110	1.500	1.270	1.080	1.080	1.180	0.909	**0.690**	0.691
	0.90	2.344	1.575	1.385	1.174	1.175	1.298	1.060	**0.803**	0.804
	0.95	2.895	1.598	1.514	1.321	1.321	1.355	1.254	**0.997**	0.998
	0.99	5.856	1.694	1.651	1.544	1.542	1.565	1.523	1.377	**1.375**
200	0.85	1.539	0.938	0.666	0.577	0.577	0.585	0.380	**0.323**	0.324
	0.90	1.582	0.998	0.733	0.625	0.626	0.649	0.433	**0.351**	0.352
	0.95	1.805	1.138	0.891	0.747	0.748	0.809	0.579	**0.446**	0.447
	0.99	3.571	1.473	1.362	1.202	1.202	1.253	1.141	0.944	**0.943**

The best values are in bold font.

TABLE 2. Estimated MSE when $p = 4$ and the intercept of the logit = 0.

			ZIPRE				AUZIPRE			
n	ρ	MLE	k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
50	0.85	2.347	2.841	2.610	1.877	1.868	2.407	2.126	1.371	**1.364**
	0.90	2.789	2.792	2.517	1.807	1.794	2.349	2.023	1.290	**1.278**
	0.95	4.094	2.638	2.358	1.671	1.658	2.155	1.850	1.171	**1.152**
	0.99	13.738	2.223	1.992	1.401	1.354	1.720	1.477	1.004	**0.933**
100	0.85	1.610	2.644	2.379	1.703	1.698	2.180	1.894	1.252	**1.250**
	0.90	1.831	2.692	2.355	1.639	1.635	2.217	1.851	1.163	**1.160**
	0.95	2.475	2.701	2.333	1.508	1.495	2.228	1.823	1.008	**0.997**
	0.99	6.892	2.505	2.142	1.205	1.157	2.028	1.657	0.755	**0.682**
150	0.85	2.438	2.628	2.170	1.764	1.759	2.102	1.522	1.174	**1.170**
	0.90	2.825	2.768	2.394	1.898	1.892	2.282	1.770	1.307	**1.302**
	0.95	3.869	2.883	2.520	1.950	1.940	2.414	1.915	1.357	**1.348**
	0.99	10.925	2.557	2.237	1.597	1.567	1.960	1.581	0.981	**0.939**
200	0.85	1.539	1.706	1.399	1.012	1.010	1.209	0.941	0.670	**0.669**
	0.90	1.745	1.840	1.489	1.057	1.055	1.257	0.967	0.637	**0.636**
	0.95	2.233	2.105	1.727	1.214	1.210	1.442	1.119	0.676	**0.673**
	0.99	5.335	2.352	1.922	1.335	1.333	1.654	1.225	0.704	**0.699**

The best values are in bold font.

TABLE 3. Estimated MSE when $p = 2$ and the intercept of the logit = 1.

			ZIPRE				AUZIPRE			
n	ρ	MLE	k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
50	0.85	2.487	1.697	1.615	1.500	1.501	1.531	1.445	1.301	**1.300**
	0.90	2.787	1.690	1.608	1.479	1.479	1.518	1.429	**1.268**	**1.268**
	0.95	3.712	1.706	1.629	1.469	1.470	1.533	1.452	**1.245**	1.246
	0.99	11.065	1.781	1.737	1.582	1.583	1.644	1.601	**1.394**	1.395
100	0.85	2.382	1.577	1.485	1.304	1.305	1.305	1.205	**0.986**	0.987
	0.90	2.698	1.607	1.518	1.313	1.314	1.352	1.242	**0.990**	0.991
	0.95	3.549	1.684	1.608	1.376	1.377	1.459	1.366	**1.060**	1.061
	0.99	9.483	1.838	1.804	1.661	1.661	1.717	1.677	**1.466**	1.467
150	0.85	2.389	1.714	1.646	1.493	1.493	1.496	1.405	**1.187**	1.188
	0.90	2.647	1.740	1.671	1.501	1.502	1.540	1.442	**1.197**	1.198
	0.95	3.288	1.778	1.719	1.530	1.531	1.600	1.518	**1.237**	1.238
	0.99	7.306	1.864	1.830	1.692	1.692	1.753	1.706	**1.491**	1.492
200	0.85	1.850	1.648	1.528	1.405	1.405	1.399	1.248	**1.083**	1.084
	0.90	1.937	1.667	1.563	1.425	1.426	1.425	1.291	**1.104**	1.105
	0.95	2.262	1.696	1.612	1.451	1.452	1.464	1.356	**1.130**	1.131
	0.99	4.742	1.779	1.748	1.566	1.566	1.605	1.565	**1.290**	**1.290**

The best values are in bold font.

TABLE 4. Estimated MSE when $p = 4$ and the intercept of the logit = 1.

			ZIPRE				AUZIPRE			
n	ρ	MLE	k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
50	0.85	42.202	3.890	3.862	3.991	3.746	3.843	3.801	4.433	**3.647**
	0.90	17.265	3.859	3.828	3.744	3.702	3.799	3.754	3.650	**3.591**
	0.95	24.055	3.447	3.389	3.285	3.210	3.343	3.288	3.186	**3.096**
	0.95	63.632	3.826	3.769	3.492	3.492	3.705	3.626	**3.233**	**3.233**
100	0.85	2.774	3.983	3.977	3.950	3.950	3.968	3.959	**3.912**	**3.912**
	0.90	3.226	3.980	3.972	3.940	3.940	3.962	3.950	**3.895**	**3.895**
	0.95	4.457	3.961	3.950	3.903	3.903	3.933	3.914	**3.833**	**3.833**
	0.95	13.956	3.768	3.716	3.535	3.531	3.643	3.576	3.321	**3.314**
150	0.85	2.912	3.621	3.578	3.530	3.530	3.518	3.478	3.421	**3.420**
	0.90	3.262	3.543	3.502	3.433	3.433	3.419	3.379	3.303	**3.302**
	0.95	4.416	3.548	3.492	3.397	3.396	3.389	3.324	3.214	**3.213**
	0.95	13.975	3.425	3.405	3.127	3.127	3.159	3.115	**2.815**	**2.815**
200	0.85	2.123	3.987	3.983	3.970	3.970	3.975	3.967	**3.942**	**3.942**
	0.90	2.360	3.981	3.974	3.956	3.956	3.963	3.951	**3.917**	**3.917**
	0.95	3.006	3.949	3.933	3.895	3.895	3.906	3.881	**3.818**	**3.818**
	0.95	7.797	3.427	3.397	3.303	3.294	3.307	3.278	3.151	**3.139**

The best values are in bold font.

TABLE 5. Estimated MSE when $p = 2$ and the intercept of the logit = 2.

			ZIPRE				AUZIPRE			
n	ρ	MLE	k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
50	0.85	3.210	1.824	1.791	1.710	1.711	1.712	1.676	**1.562**	1.563
	0.90	3.648	1.816	1.784	1.691	1.692	1.698	1.663	**1.533**	**1.533**
	0.95	4.976	1.813	1.788	1.670	1.670	1.687	1.662	**1.496**	**1.496**
	0.99	16.425	1.850	1.842	1.730	1.731	1.744	1.744	**1.580**	**1.580**
100	0.85	5.578	1.978	1.981	1.970	1.970	1.958	1.964	**1.944**	**1.944**
	0.90	6.266	1.972	1.977	1.960	1.960	1.946	1.955	**1.924**	1.925
	0.95	8.343	1.958	1.970	1.942	1.942	1.920	1.943	**1.893**	1.894
	0.99	25.334	1.963	1.977	1.950	1.950	1.931	1.956	**1.908**	**1.908**
150	0.85	5.330	1.986	1.988	1.982	1.982	1.974	1.977	**1.965**	**1.965**
	0.90	5.778	1.983	1.986	1.977	1.977	1.966	1.972	**1.955**	**1.955**
	0.95	7.077	1.975	1.981	1.966	1.966	1.951	1.964	**1.935**	**1.935**
	0.99	17.264	1.973	1.982	1.964	1.964	1.949	1.965	**1.932**	**1.932**
200	0.85	4.657	1.989	1.991	1.985	1.985	1.979	1.982	**1.971**	**1.971**
	0.90	4.881	1.987	1.989	1.982	1.982	1.974	1.978	**1.966**	**1.966**
	0.95	5.588	1.981	1.985	1.976	1.976	1.963	1.971	**1.953**	**1.953**
	0.99	11.360	1.969	1.979	1.960	1.960	1.940	1.958	**1.923**	**1.923**

The best values are in bold font.

TABLE 6. Estimated MSE when $p = 4$ and the intercept of the logit = 2.

			ZIPRE				AUZIPRE			
n	ρ	MLE	k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
50	0.85	36.625	3.930	3.902	3.852	3.841	3.895	3.853	3.781	**3.770**
	0.90	42.902	3.906	3.880	3.839	3.803	3.861	3.826	3.789	**3.722**
	0.95	33.365	3.787	3.774	3.690	3.676	3.718	3.703	3.593	**3.572**
	0.95	170.307	3.564	3.554	3.378	3.351	3.450	3.445	3.240	**3.195**
100	0.85	36.433	4.000	4.000	4.000	4.000	4.000	4.000	**3.999**	**3.999**
	0.90	41.120	4.000	4.000	**3.999**	**3.999**	**3.999**	**3.999**	**3.999**	**3.999**
	0.95	48.418	3.999	3.999	3.998	3.998	3.998	3.998	**3.997**	**3.997**
	0.95	107.540	3.962	3.967	3.939	3.939	3.941	3.948	**3.902**	**3.902**
150	0.85	7.665	4.000	4.000	4.000	4.000	4.000	4.000	**3.999**	**3.999**
	0.90	9.255	3.999	3.999	3.999	3.999	3.999	3.999	**3.998**	**3.998**
	0.95	13.784	3.982	3.983	3.979	3.979	3.974	3.975	**3.969**	**3.969**
	0.95	51.558	3.800	3.816	3.772	3.772	3.763	3.775	**3.732**	**3.732**
200	0.85	5.477	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**
	0.90	6.100	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**	**4.000**
	0.95	7.947	4.000	4.000	4.000	4.000	4.000	**3.999**	**3.999**	**3.999**
	0.95	22.929	3.983	3.984	3.972	3.972	3.971	3.973	**3.953**	**3.953**

The best values are in bold font.

We can conclude the following points from Tables 1-4

(1) Increasing the multicollinearity level, ρ, with fixed values of n, p, has a negative impact on the MLE estimator and in some cases of the AUZIPRE and ZIPRE. This is because the values of the MSE increase as the level of the multicollinearity, ρ, increases. However, increasing the value of ρ with $p = 4$ and $q_i = 1$ has a positive impact on AUZIPRE and ZIPRE as the MSE values become smaller.

(2) The values of the MSE of the estimators, AUZIPRE, ZIPRE, and MLE, increase when the number of explanatory variables, p, increased with fixed values of ρ and n.

(3) All of the AUZIPRE estimators, k_1, k_2, k_3 and k_4 are better than the corresponded ones of the ZIPRE estimators in that they have smaller values of the MSE. In

contrast, the MLE has, in general, the worse performance in that it has the highest values of the MSE.

(4) Among the AUZIPRE estimators, the k_3 and k_4 estimators outperform the k_1 and k_2 estimators as they have smaller values of the MSE.

It can be concluded from the simulation study that the MSE of AUZIPRE is always smaller than those of the ZIPRE and the MLE. All the selection methods of k are superior to the MLE in terms of MSE. Moreover, the AUZIPRE with the k_3 and k_4 improved the AUZIPRE performance compared with the ZIPRE and the MLE in most of the cases. Furthermore, k_4 and k_4 are the optimal estimation methods for k of the AUZIPRE. On the contrast, the MLE estimator values are the poorest compared with the other estimators.

6. Real data application

In this section, we consider the dataset of bioChemists, by [20]. The bioChemists dataset consists of $n = 915$ observations. The Articles is the dependent variable that represents articles number published during the Ph.D study in the last 3 years. The dependent variable depends on five explanatory variables as described in Table 7.

Table 7. The description of the explanatory variables of the bioChemists data.

Variable names	Description
Female	the student gender, 0 if male 1 and if female.
MentorArts	the articles number published during the last 3 Ph.D. years.
Prestige	the Ph.D. student prestige.
Married	the marital status, 0 if single and 1 if married.
Children	the children number of aged 5 or younger

The ZIP regression model was fitted to the bioChemists data using equation (2). Then, the AUZIPRE, ZIPRE and MLE were calculated. Table 8 presents the estimated values of the MSE and the estimated values of the coefficient parameters of the ZIP model for the bioChemists dataset for different estimators, AUZIPRE, ZIPRE, and MLE. We can notice that the AUZIPRE has the smallest value of the MSE in comparison with the ZIPRE and the MLE. In addition, the k_3 and k_4 estimators of the ridge parameter have the best performance among the other ridge parameter estimators for the AUZIPRE as they have the smallest values of the MSE compared with the values of the k_1 and k_2 estimators.

Table 8. The estimated coefficient parameters and the estimated MSE for the AUZIPRE, ZIPRE and MLE.

	MLE	ZIPRE				AUZIPRE			
		k_1	k_2	k_3	k_4	k_1	k_2	k_3	k_4
Intercept	0.756	0.001	0.759	0.001	0.001	0.001	0.756	0.001	0.001
Female	-0.518	0.001	-0.518	0.001	0.001	0.001	-0.518	0.001	0.001
MentorArts	0.398	0.001	0.398	0.001	0.001	0.001	0.398	0.001	0.001
Prestige	-0.465	0.001	-0.465	0.001	0.001	0.001	-0.465	0.001	0.001
Married	0.382	0.001	0.382	0.002	0.002	0.002	0.382	0.003	0.003
Children	0.038	0.001	0.038	0.002	0.002	0.002	0.038	0.004	0.004
MSE	170.6	4.183	170.6	4.153	4.153	4.158	170.6	4.098	4.098

The best value of the MSE is in bold font.

7. Conclusions

In this article, we proposed an almost unbiased ridge estimator based on the ridge estimator for the zero-inflated Poisson regression model. The proposed estimator is able to solve the inflation problem of the maximum likelihood estimation method that is applied to estimate the ZIP model parameters. The performance of the proposed estimator was investigated by conducting a Monte Carlo simulation experiment and a real dataset using the MSE as a measure. Based on our results, the performance of the AUZIPRE is better than that of the MLE and ZIPRE as it has smaller MSE values than the other estimators for the ZIP model when multicollinearity exists in the data.

From the simulated results and the real dataset, we have seen that when multicollinearity is presented, the MLE becomes inflated. The performance of the k_3 and k_4 estimators for the AUZIPRE are much better than the MLE and those for the ZIPRE as they have smaller values of the MSE. Hence, we recommended the k_3 and k_4 estimators for estimating the ridge parameter for the ZIP regression model.

Acknowledgement. The authors are very grateful to the University of Mosul/ College of Education for Pure Science and College of Computers Sciences and Mathematics for their provieded facilities, which helped to improve the quality of this work.

References

[1] Månsson, K., (2012). On ridge estimators for the negative binomial regression model, Economic Modelling, 29(2), pp. 178-184.
[2] Kibria, G., Månsson, K., Shukur, G., (2015), A simulation study of some biasing parameters for the ridge type estimation of Poisson regression, Communications in Statistics-Simulation and Computation, 44(4), pp. 943-957.
[3] King, G., (1989), Event count models for international relations: Generalizations and applications, International Studies Quarterly, 33(2), pp. 123-147.
[4] Lambert, D., (1992), Zero-inflated Poisson regression, with an application to defects in manufacturing, Technometrics, 34(1), pp. 1-14.
[5] Greene, W., (1994), Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, Technometrics.
[6] Månsson, K. and Shukur, G., (2011), A Poisson ridge regression estimator, Economic Modelling, 28(4), pp. 1475-1481.
[7] Algamal, Z., (2018), Shrinkage estimators for gamma regression model, Electronic Journal of Applied Statistical Analysis, 11(1) pp. 253-268.
[8] Kibria, G., Månsson, K., Shukur, G., (2013), Some ridge regression estimators for the zero-inflated Poisson model, Journal of Applied Statistics, 40(4) pp. 721-735.

[9] Singh, B., Chaubey, Y. P., Dwivedi, T. D., (1986), An almost unbiased ridge estimator, Sankhyā: The Indian Journal of Statistics, Series B, pp. 342-346.
[10] Jansakul, N., Hinde, JP., (2002), Score tests for zero-inflated Poisson models, Computational statistics & data analysis, 40(1) pp. 75-96.
[11] Li, CS., (2011), A lack-of-fit test for parametric zero-inflated Poisson models, Journal of Statistical Computation and Simulation, 81(9), pp. 1081-1098.
[12] Numna, S., (2009), Analysis of extra zero counts using zero-inflated Poisson models, Doctoral dissertation, Prince of Songkla University.
[13] Mackinnon, M. J., Puterman, M. L., (1989), Collinearity in generalized linear models, Communications in statistics-theory and methods, 18(9), pp. 3463-3472.
[14] Liu, G., Piantadosi, S., (2017), Ridge estimation in generalized linear models and proportional hazards regressions, Communications in Statistics-Theory and Methods, 46(23), pp. 11466-11479.
[15] Hoerl, A., Kennard, R., (1970), Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12(1), pp. 55-67.
[16] Algamal, Z., (2018), Performance of ridge estimator in inverse Gaussian regression model, Communications in Statistics-Theory and Methods, 48(15), pp. 1-14.
[17] Asar, Y., Ahmed, S. E., Yüzbaşı, B., (2018), Efficient and Improved Estimation Strategy in Zero-Inflated Poisson Regression Models, International Conference on Management Science and Engineering Management, pp. 329-342.
[18] Kibria, G., (2003), Performance of some new ridge regression estimators, Communications in Statistics-Simulation and Computation, 32(2) pp. 419-435.
[19] Bhat, S., (2016), A comparative study on the performance of new ridge estimators, Pakistan Journal of Statistics and Operation Research, 12(2) pp. 317-325.
[20] Long, J., (1990), The origins of sex differences in science, Social forces, 68(4), pp. 1297-1316.
[21] Sakallioğlu, S., Kaçiranlar, S., Akdeniz, F., (2001), Mean squared error comparisons of some biased regression estimators, Communications in Statistics-Theory and Methods, 30(2), pp. 347-361.
[22] Akdeniz, F., Roozbeh, M., (2019), Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models, Statistical Papers, 60(5), pp. 1717-1739.
[23] Akdeniz, F., Roozbeh, M., (2017), Efficiency of the generalized-difference-based weighted mixed almost unbiased two-parameter estimator in partially linear model, Communications in Statistics-Theory and Methods, 46(24), pp. 12259-12280.

Younus Al-Taweel is a lecturer in the Department of Mathematics, College of Education for Pure Science, University of Mosul, Iraq. He got his Ph.D. in statistics from the University of Sheffield, UK in 2018. His research interests are in Bayesian statistics, machine learning, uncertainty quantification for computer models, regression models.

Zakariya Algamal graduated from the Department of Mathematical Sciences of Universiti Teknologi Malaysia in 2016. He is working as a professor of statistics in the Department of Statistics and Informatics at the University of Mosul-Iraq. His research areas include high dimensional data, sparse methods, genealized linear model, machine learning.