Regularized right-censored zero-inflated Poisson regression for correlated count data: Applications to social contact and environmental studies

Essoham Ali[1,2*†], Adewale F. Lukman[3†] and Solym Mawaki Manou-Abi[4]

[1]Univ Bretagne Sud, CNRS UMR 6205, LMBA, Vannes, France.
[2]Institut de Mathématiques Appliquées, UCO, 49000, Angers, France.
[3]Department of Mathematics and Statistics, University of North Dakota, Grand Forks, North Dakota, 58202 USA.
[4]Laboratoire de Mathématiques et Applications, Université de Poitiers, Poitiers, UMR CNRS 7348, France.

*Corresponding author(s). E-mail(s): essoham.ali@univ-ubs.fr ;
Contributing authors: adewale.lukman@und.edu; solym.Manou.abi@univ-poitiers.fr;
†These authors contributed equally to this work.

Abstract

In regression models for count data, the presence of both multicollinearity among covariates and excess zeros often leads to unstable and biased parameter estimates, especially when data are subject to censoring. To address this, we propose a regularized estimation framework for the Right-Censored Zero-Inflated Poisson (RCZIP) regression model, integrating Ridge, Liu, and a Modified Ridge-Type (MRT) estimator specifically adapted to this context. We develop the theoretical properties of these estimators, providing matrix-based and eigenvalue-based criteria for comparing their mean squared errors. An extensive Monte Carlo simulation study evaluates the performance of the proposed methods under varying degrees of multicollinearity and censoring. The results show that the MRT estimator consistently outperforms traditional Maximum Likelihood and standard regularized estimators, offering greater robustness and accuracy. The practical utility of our approach is demonstrated through two real-world applications: a social contact survey in Mayotte and an environmental dataset, both characterized by zero inflation, censoring, and high multicollinearity. The findings confirm the relevance of regularization, particularly MRT, in improving model stability and predictive performance for censored count data models.

Keywords: Zero-Inflated Poisson Regression; Right-Censoring; Multicollinearity; Regularization; Monte Carlo simulation

1 Introduction

Count data frequently arise in various fields such as social sciences, biostatistics, economics, and environmental studies, where they typically represent the number of occurrences of discrete events within a specific time or space interval (Cox and Oakes, 1984; Cameron and Trivedi [8]). These data are often modeled using Poisson or Negative Binomial distributions. However, in practice, count data may display excess zeros (zero inflation) and be subject to right-censoring due to measurement limitations or detection thresholds (Lambert [19]; Kalbfleisch and Prentice [1]). Failure to account for censoring can bias parameter estimates and compromise inferential validity, especially when the censoring mechanism is related to observable covariates (Mehta and Patel [33]).

Another major issue in count data analysis is multicollinearity among covariates, which occurs when explanatory variables are highly correlated. This situation inflates the variance of the parameter estimates, making them unstable and difficult to interpret ([17]; Menard [29]). In Poisson regression models, these problems can be exacerbated due to the discrete and often skewed nature of count data. Regularization methods, such as Ridge regression (Hoerl et al. [13]; Tikhonov [37]), have been widely used to mitigate multicollinearity in linear and generalized linear models. However, their application to zero-inflated and censored models remains limited, despite the practical importance of these cases (Fitzmaurice et al. [12]).

The aim of this paper is to propose a regularized estimation framework for the Right-Censored Zero-Inflated Poisson (RCZIP) regression model, integrating Ridge, Liu (Liu, [21]), and a Modified Ridge-Type (MRT) estimator (Lukman et al.[25]) specifically adapted to this complex context. We derive the theoretical properties of these estimators and compare them using matrix-based and eigenvalue-based mean squared error criteria. Extensive Monte Carlo simulations are conducted to evaluate the performance of these methods under varying levels of multicollinearity and censoring. Finally, the practical relevance of our approach is illustrated through two real-world applications: a social contact survey from Mayotte and an environmental dataset characterized by zero inflation and high multicollinearity.

The remainder of this paper is organized as follows. Section 2 presents the RCZIP model and discusses the limitations of the Maximum Likelihood Estimator (MLE) in the presence of multicollinearity, and presents several regularized estimators, including Ridge, Liu, and the proposed MRT estimator. Section 3 provides a theoretical comparison of these estimators based on their mean squared error properties. Section 4 reports the results of extensive Monte Carlo simulation studies. Section 5 applies the proposed methods to two real datasets. Finally, Section 6 concludes the paper by summarizing the findings and outlining future research directions.

2 The right-censored zero-inflated Poisson model

Let the response variable be denoted by Y_i, for $i = 1, \dots, n$, and let π_i be the probability that Y_i is a structural zero, where $0 \le \pi_i \le 1$. The random variable Y_i follows a Zero-Inflated Poisson (ZIP) distribution if:

$$P(Y_i = y) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, & \text{if } y = 0, \\ (1-\pi_i)\frac{e^{-\lambda_i}\lambda_i^y}{y!}, & \text{if } y > 0. \end{cases} \tag{1}$$

with expectation $\mathbb{E}[Y_i] = (1-\pi_i)\lambda_i$ and variance $\mathrm{Var}(Y_i) = (1-\pi_i)\lambda_i(1+\pi_i\lambda_i)$, for $i = 1, \dots, n$. In this model, zeros in the outcome variable Y_i can originate from two sources: (i) structural zeros with probability π_i, or (ii) Poisson-distributed zeros with probability $(1-\pi_i)$.

The parameters λ_i and π_i are modeled using link functions to establish linear predictors as follows:

$$\mathrm{logit}(\pi_i) = \boldsymbol{\gamma}^\top \mathbf{G}_i = \gamma_0 + \gamma_1 G_{1,i} + \cdots + \gamma_q G_{q,i}, \quad i = 1, \dots, n, \tag{2}$$

where $\mathbf{G}_i = (1, G_{1,i}, \dots, G_{q,i})^\top$ is a vector of $(q+1)$ covariates, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)^\top$ is a vector of parameters.

For the count component:

$$-\log(\lambda_i) = \boldsymbol{\beta}^\top \mathbf{X}_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i}, \quad i = 1, \dots, n, \tag{3}$$

where $\mathbf{X}_i = (1, X_{1,i}, \dots, X_{p,i})^\top$ is a vector of $(p+1)$ covariates, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ is the corresponding parameter vector. We consider the case where the response variable Y_i in the zero-inflated Poisson (ZIP) model is subject to right censoring. In this framework, we observe the tuple $(Y_i^*, \nu_i, \mathbf{X}_i, \mathbf{G}_i)$ for each individual $i = 1, 2, \dots, n$, where $Y_i^* = \min(Y_i, C_i)$, with C_i denoting the censoring threshold, and $\nu_i = I(Y_i < C_i)$ indicating whether the observation is censored. If $Y_i = C_i$, then $Y_i^* = C_i$ and $\nu_i = 0$. We further define $J_i = I(Y_i^* = 0)$ to distinguish structural zeros.

Given these definitions, the log-likelihood function of the right-censored ZIP (RCZIP) model, with parameter vector $\boldsymbol{\theta} := (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, is given by [35]:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \nu_i \left[J_i \ln \left(e^{\boldsymbol{\gamma}^\top \mathbf{G}_i} + e^{-\exp(\boldsymbol{\beta}^\top \mathbf{X}_i)} \right) + (1 - J_i) \left(Y_i^* \boldsymbol{\beta}^\top \mathbf{X}_i - \exp(\boldsymbol{\beta}^\top \mathbf{X}_i) - \ln(Y_i^*!) \right) \right] \right.$$
$$\left. + (1-\nu_i)(1-J_i) \ln \left(1 - \sum_{t=0}^{Y_i^*-1} \frac{e^{-\exp(\boldsymbol{\beta}^\top \mathbf{X}_i)} \exp(\boldsymbol{\beta}^\top \mathbf{X}_i)^t}{t!} \right) - \ln(1 + e^{\boldsymbol{\gamma}^\top \mathbf{G}_i}) \right\}. \tag{4}$$

The maximum likelihood estimator (MLE), denoted by $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$, is obtained by solving the score equation $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = 0$ using nonlinear optimization. In practice, the MLE is computed numerically using the Iterative Reweighted Least Squares (IRLS) algorithm. The parameter update at each iteration is given by:

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + I(\boldsymbol{\theta}^{(m)})^{-1} S(\boldsymbol{\theta}^{(m)}), \tag{4}$$

where $S(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the score function, and $I(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$ is the Fisher information matrix.

At convergence, the MLE can be expressed as:

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \Lambda^{-1} \mathbf{X}^\top \mathbf{W} \hat{\mathbf{z}}, \tag{5}$$

where $\Lambda = \mathbf{X}^\top \mathbf{W} \mathbf{X}$, $\mathbf{W} = \text{diag}\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\theta}} \frac{\partial \pi_i}{\partial \boldsymbol{\theta}}\right)$ is the weight matrix, and

$$\hat{z}_i = \log(\hat{\lambda}_i) + \frac{(y_i^* - \hat{\lambda}_i)}{\sqrt{\mathbf{W}\mathbf{V}}}, \quad \text{with} \quad \mathbf{V} = \text{diag}\left((1-\pi_i)\lambda_i(1+\pi_i\lambda_i)\right)$$

being the variance matrix of the pseudo-responses. The Mean Squared Error Matrix (MSEM) and Scalar Mean Squared Error (SMSE) are:

$$\text{MSEM}(\widehat{\boldsymbol{\theta}}_{MLE}) = \Lambda^{-1}, \tag{6}$$

$$\text{SMSE}(\widehat{\boldsymbol{\theta}}_{MLE}) = \sum_{j=1}^{p+q} \frac{1}{\eta_j}, \tag{7}$$

where η_j are the eigenvalues of $\mathbf{X}^\top \mathbf{W} \mathbf{X}$.

The asymptotic properties of the MLE $\hat{\boldsymbol{\theta}}$ under the censored ZIP model have been rigorously studied in [35]. However, in practice, this estimator may become unstable when covariates are highly correlated, leading to unreliable inference.

2.1 Proposed estimators

To address the challenges posed by multicollinearity and potential instability of the MLE in the RCZIP model, we develop alternative estimation strategies. Building on the formulation and likelihood-based inference introduced earlier, we propose several penalized estimation approaches that aim to enhance robustness, particularly in the presence of correlated covariates or small sample sizes.

2.1.1 RCZIP Ridge Estimator

In this section, we propose an extension of the Ridge estimator for the Right-Censored Zero-Inflated Poisson Regression (RCZIP) model. The main objective of this estimator is to address multicollinearity issues while improving the stability of parameter estimation. Ridge regression was introduced by Hoerl and Kennard to address multicollinearity in the linear regression model. The standard ridge estimator is given by:

$$\widehat{\beta}^k = (\mathbf{X}^\top \mathbf{X} + kI)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{8}$$

where k is the penalization parameter. In the context of the Right-Censored Zero-Inflated Poisson (RCZIP) model, we propose an extension of this estimator using the weighted information from the model. The RCZIP Ridge Estimator is denoted by $\widehat{\boldsymbol{\theta}}_R$, and is defined as:

$$\widehat{\boldsymbol{\theta}}_{Ridge} = (\Lambda + kI)^{-1}\Lambda\widehat{\theta}_{MLE}. \tag{9}$$

The scalar mean squared error (SMSE) is given by:

$$\text{SMSE}(\hat{\theta}_{Ridge}) = \sum_{j=1}^{p+q}\left(\frac{\eta_j}{(\eta_j + k)^2}\right) + k^2\sum_{j=1}^{p+q}\left(\frac{\hat{\alpha}_j^2}{(\eta_j + k)^2}\right) \tag{10}$$

where η_j is the eigenvalues of $\mathbf{X}^\top\mathbf{W}\mathbf{X}$ and $\hat{\alpha}_j$ are the components of $Q^\top\widehat{\boldsymbol{\theta}}_{MLE}$. The matrix mean squared error (MSEM) of the ridge estimator is given by:

$$\text{MSEM}(\widehat{\theta}_{Ridge}) = Q\Lambda_k\Lambda\Lambda_k Q^\top + k^2\Lambda_k\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\Lambda_k, \tag{11}$$

where $\Lambda = \mathbf{X}^\top\mathbf{W}\mathbf{X}$, $\Lambda_k = (\mathbf{X}^\top\mathbf{W}\mathbf{X} + kI_p)^{-1}$ and Q is the matrix of eigenvectors of $\mathbf{X}^\top\mathbf{W}\mathbf{X}$.

Hoerl and Kennard suggested the following estimator for k in a linear regression model:

$$\hat{k}_{HM} = \frac{p\sigma^2}{\sum_{j=1}^{p+q}(\hat{\alpha}_j)^2}$$

Schaefer and Kibria proposed another approach based on the maximum of the components of $\hat{\alpha}_j$:

$$\hat{k}_1 = \frac{1}{\max((\hat{\alpha}_j)^2)}$$

We will use these approaches to estimate k in the RCZIP model.

2.1.2 RCZIP Liu Regression Estimator

Mayer and Wilke [32] introduced the contraction estimator, which is defined as:

$$\hat{\beta}_\rho = (1 + \rho)^{-1}\hat{\beta}_{\text{MLE}}. \tag{12}$$

Building upon this, Liu [21] proposed an alternative method called the Liu estimator that integrates both the ridge and contraction estimators. This approach has demonstrated competitive performance against the traditional ridge estimator and garnered considerable regularization interest. The Liu estimator is defined as:

$$\hat{\theta}_{Liu} = \Lambda_a\Lambda_d\hat{\theta}_{\text{MLE}}, \quad 0 \leq d < 1, \tag{13}$$

where $\Lambda_a = (X'WX + I)^{-1}$, $\Lambda_d = X'WX + dI$, and d represents the penalization parameter and is obtained through the following equation in this study.

$$\hat{d} = \min\left\{\frac{\hat{\alpha}_i^2}{\frac{\hat{\sigma}^2}{\lambda_i} + \hat{\alpha}_i^2}\right\}$$

The properties of the estimator are as follows:

The bias of $\widehat{\theta}_{Liu}$ is given by:

$$\text{bias}(\hat{\theta}_{Liu}) = [\Lambda_a\Lambda_d - I]\,\theta$$

The variance-covariance matrix is:

$$\text{cov}(\hat{\theta}_{Liu}) = Q\Lambda_a\Lambda_d\Lambda^{-1}\Lambda_d\Lambda_a Q^T.$$

Hence, the MSEM and SMSE are given as:

$$MSEM(\hat{\theta}_{Liu}) = Q\Lambda_a\Lambda_d\Lambda^{-1}\Lambda_d\Lambda_a Q^T + (d-1)^2\Lambda_a\alpha'\alpha\Lambda_a \tag{14}$$

$$SMSE(\hat{\theta}_{Liu}) = \sum_{j=1}^{q}\frac{(n_j + d)^2}{(n_{j+1})^2 n_j} + (d-1)^2\sum_{j=1}^{q}\frac{\alpha_j^2}{(n_{j+1})^2} \tag{15}$$

2.1.3 RCZIP-MRT

Lukman et al. (2019) developed the modified ridge-type (MRT) estimator to account for multicollinearity (linear dependency among the predictors) in the linear regression model by incorporating an additional shrinkage parameter to improve the performance of the ridge estimator. Their results indicate that the MRT estimator behaves similarly to the ridge estimator in parameter estimation while demonstrating superior predictive accuracy. It has attracted attention from researchers, leading to its extension across different generalized linear models. Recent studies have applied the MRT estimator to models such as the inverse Gaussian regression model (Akram et al.[2]), Poisson regression model (Lukman et al. [25]), Bell regression model (Bulut et al.[7]), and Beta regression model (Akram et al.[3]), among others. Hence, the MRT is defined as follows:

$$\widehat{\beta}^{kd} = (\mathbf{X}^\top\mathbf{X} + k(1+d)I)^{-1}\mathbf{X}^\top\mathbf{y} \tag{16}$$

Consequently, we propose an extension of this estimator to RCZIP and define RCZIP-MRT as follows:

$$\widehat{\boldsymbol{\theta}}_{MRT} = (\Lambda + k(1+d)I)^{-1}\Lambda\widehat{\theta}_{MLE} \tag{17}$$

where k and d are the penalization parameters.

The MSEM of the MRT estimator is given by:

$$\mathrm{MSEM}(\widehat{\theta}_{\mathrm{MRT}}) = Q\Lambda_{kd}\Lambda\Lambda_{kd}Q^\top + (k(1+d))^2\Lambda_{kd}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\Lambda_{kd} \tag{18}$$

where $\Lambda_{kd} = (\mathbf{X}^\top\mathbf{W}\mathbf{X} + k(1+d)I_p)^{-1}$.

The scalar mean squared error (SMSE) is given by:

$$\mathrm{SMSE}(\hat{\theta}_{MRT}) = \sum_{j=1}^{p+q}\left(\frac{\eta_j}{(\eta_j + k(1+d))^2}\right) + (k(1+d))^2\sum_{j=1}^{p+q}\left(\frac{\hat{\alpha}_j^2}{(\eta_j + k(1+d))^2}\right) \tag{19}$$

Having introduced the proposed penalized estimators for the RCZIP model, we now proceed to a theoretical investigation of their statistical properties. In particular, we assess and compare their bias, variance, and overall estimation accuracy using matrix and scalar mean squared error criteria.

3 Main Results

In this section, we present a theoretical comparison of the proposed estimators with the maximum likelihood estimator (MLE) in the context of the RCZIP model. By analyzing their respective mean squared error matrices (MSEM) and scalar mean squared errors (SMSE), we derive conditions under which each regularized estimator outperforms the MLE. The results are established through key matrix inequalities and existing lemmas in the literature, providing insight into the trade-offs between bias and variance inherent to each method.

Lemma 1. (Farebrother [16]). *Let T be a positive definite matrix, namely $T > 0$, and let ν be some vector, then*

$$T - \nu\nu' \geq 0 \text{ if and only if } \nu' T^{-1}\nu \leq 1.$$

Lemma 2. (Trenkler and Toutenburg [36]). *Let $\hat{\theta}_j = C_j y$, $j = 1, 2$ be two competing estimators for θ. Suppose that $V = V(\hat{\theta}_1) - V(\hat{\theta}_2) > 0$, where $V(\hat{\theta}_j)$, $j = 1, 2$ denotes the variance-covariance matrix of $\hat{\theta}_j$. Then*

$$\Delta(\hat{\theta}_1, \hat{\theta}_2) = MSEM(\hat{\theta}_1) - MSEM(\hat{\theta}_2) \geq 0$$

if and only if

$$m_2'(V + m_1 m_1')^{-1} m_2 \leq 1,$$

where $MSEM(\hat{\theta}_j)$ and m_j denote the mean squared error matrix and bias vector of $\hat{\theta}_j$, respectively.

3.1 Comparison of $\hat{\boldsymbol{\theta}}_{MLE}$ Versus $\hat{\boldsymbol{\theta}}_{R}$

We examined the difference between the MSEM of the MLE method and the RCZIP Ridge Regression Estimator as follows:

$$MSEM(\hat{\theta}_{MLE}) - \mathrm{MSEM}(\hat{\theta}_{Ridge}) = (\Lambda^{-1} - \Lambda_k\Lambda\Lambda_k) - k^2\Lambda_k\theta_{MLE}\theta_{MLE}^{'}\Lambda_k. \tag{20}$$

Given that $k > 0$, we have the following theorem.

Theorem 1. *Given two linear estimators $\hat{\theta}_{MLE}$ and $\hat{\theta}_{RCZIPRE}$. If $k > 0$, $\hat{\theta}_{RCZIPRE}$ dominates $\hat{\theta}_{MLE}$ given that MSEM($\hat{\theta}_{MLE}$) − MSEM($\hat{\theta}_{Ridge}$) > 0 if and only if*

$$k^2\Lambda_k\theta'_{MLE}[\Lambda^{-1} - \Lambda_k\Lambda\Lambda_k]^{-1}\theta_{MLE}\Lambda_k \leq 1.$$

Proof. Using the SMSE for both estimators in (7) and (19), we obtain

$$V(\hat{\theta}_{MLE}) - V(\hat{\theta}_{Ridge}) = (\Lambda^{-1} - \Lambda_k\Lambda\Lambda_k)$$

$$= \text{diag}\left\{\frac{1}{\eta_j} - \frac{\eta_j}{(\eta_j + k)^2}\right\}_{j=1}^{p+q}.$$

The variance-covariance difference $\Lambda^{-1} - \Lambda_k^{-1}\Lambda\Lambda_k^{-1}$ will be positive definite (pd) if and only if

$$(\eta_j + k)^2 - \eta_j^2 > 0 \quad \text{or} \quad (\eta_j + k) - \eta_j > 0.$$

For $k > 0$, we can certainly see that $(\eta_j + k) - \eta_j > 0$. Therefore, $\Lambda^{-1} - \Lambda_k^{-1}\Lambda\Lambda_k^{-1}$ is pd. The proof is completed by Lemma 2.

3.2 Comparison of $\hat{\theta}_{Ridge}$ Versus $\hat{\theta}_{MRT}$

We examined the difference between the MSEM of the RCZIP Ridge Estimator and the RCZIP Modified ridge-type Estimator as follows:

$$\begin{aligned} MSEM(\hat{\theta}_{Ridge}) - \text{MSEM}(\hat{\theta}_{MRT}) =& (\Lambda_k\Lambda\Lambda_k - \Lambda_{kd}\Lambda\Lambda_{kd}) + k^2\Lambda_k\theta_{MLE}\theta'_{MLE}\Lambda_k \\ & - [k(1+d)]^2\Lambda_{kd}\theta_{MLE}\theta'_{MLE}\Lambda_{kd}). \end{aligned} \tag{21}$$

Given that $k, d > 0$, we have the following theorem.

Theorem 2. *Given two linear estimators $\hat{\theta}_{Ridge}$ and $\hat{\theta}_{MRT}$. If $k, d > 0$, $\hat{\theta}_{MRT}$ dominates $\hat{\theta}$ giveRidgen that MSEM($\hat{\theta}_{Ridge}$) − MSEM($\hat{\theta}_{MRT}$) > 0 if and only if*

$$[k(1+d)]^2\Lambda_{kd}\theta'_{MLE}[\Lambda_k\Lambda\Lambda_k - \Lambda_{kd}\Lambda\Lambda_{kd} + k^2\Lambda_k\theta_{MLE}\theta'_{MLE}\Lambda_k]^{-1}\theta_{MLE}\Lambda_{kd} \leq 1.$$

Proof. Using the SMSE for both estimators in (10) and (19), we obtain

$$V(\hat{\theta}_{Ridge}) - V(\hat{\theta}_{MRT}) = (\Lambda_k\Lambda\Lambda_k - \Lambda_{kd}\Lambda\Lambda_{kd})$$

$$= \text{diag}\left\{\frac{\eta_j}{(\eta_j + k)^2} - \frac{\eta_j}{(\eta_j + k(1+d))^2}\right\}_{j=1}^{p+q}.$$

The variance-covariance difference $\Lambda_k\Lambda\Lambda_k - \Lambda_{kd}\Lambda\Lambda_{kd}$ will be positive definite (pd) if and only if

$$(\eta_j + k(1+d)) - (\eta_j + k) > 0$$

For $k > 0$, we can certainly see that $(\eta_j + k(1+d)) - (\eta_j + k) > 0$. Therefore, $\Lambda_k\Lambda\Lambda_k - \Lambda_{kd}\Lambda\Lambda_{kd}$ is pd. The proof is completed by Lemma 2.

3.3 Comparison of $\hat{\theta}_{Lui}$ Versus $\hat{\theta}_{MRT}$

We examined the difference between the MSEM of the RCZIP-LE and RCZIP-MRT as follows:

$$\begin{aligned} MSEM(\hat{\theta}_{Liu}) - \text{MSEM}(\hat{\theta}_{MRT}) =& (\Lambda_d\Lambda_a\Lambda^{-1}\Lambda_a\Lambda_d - \Lambda_{kd}\Lambda\Lambda_{kd}) + (d-1)^2\Lambda_a\theta_{MLE}\theta'_{MLE}\Lambda_a \\ & - [k(1+d)]^2\Lambda_{kd}\theta_{MLE}\theta'_{MLE}\Lambda_{kd}). \end{aligned} \tag{22}$$

Given that $k, d > 0$, we have the following theorem.

Theorem 3. *Given two linear estimators $\hat{\theta}_{Liu}$ and $\hat{\theta}_{MRT}$. If $k, d > 0$, $\hat{\theta}_{RCZIP-MRT}$ dominates $\hat{\theta}_{Liu}$ given that $MSEM(\hat{\theta}_{Liu}) - MSEM(\hat{\theta}_{MRT}) > 0$ if and only if*

$$[k(1+d)]^2 \Lambda_{kd} \theta'_{MLE} [\Lambda_k \Lambda \Lambda_k - \Lambda_{kd} \Lambda \Lambda_{kd} + (d-1)^2 \Lambda_a \theta_{MLE} \theta'_{MLE} \Lambda_a]^{-1} \theta_{MLE} \Lambda_{kd} \leq 1.$$

Proof. Using the SMSE for both estimators in (15) and (19), we obtain

$$V(\hat{\theta}_{Liu}) - V(\hat{\theta}_{MRT}) = (\Lambda_a \Lambda_d \Lambda^{-1} \Lambda_d \Lambda_a - \Lambda_{kd} \Lambda \Lambda_{kd})$$

$$= \operatorname{diag} \left\{ \frac{(\eta_j + d)^2}{\eta_j(\eta_j+1)^2} - \frac{\eta_j}{(\eta_j + k(1+d))^2} \right\}_{j=1}^{p+q}.$$

The variance-covariance difference $\Lambda_a \Lambda_d \Lambda^{-1} \Lambda_d \Lambda_a - \Lambda_{kd} \Lambda \Lambda_{kd}$ will be positive definite (pd) if and only if

$$(\eta_j + k(1+d))(\eta_j + d) - (\eta_j + k)(\eta_j + 1) > 0$$

For $k, d > 0$, we can certainly see that $(\eta_j + k(1+d))(\eta_j + d) - (\eta_j + k)(\eta_j + 1) > 0$. Therefore, $\Lambda_a \Lambda_d \Lambda^{-1} \Lambda_d \Lambda_a - \Lambda_{kd} \Lambda \Lambda_{kd}$ is pd. The proof is completed by Lemma 2.

4 The Monte Carlo simulations

4.1 The design of the experiment

In this section, we analyze the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Squared Difference Error (MSDE) of the ML, Ridge, Lui, and Modified Ridge Type estimators for the Zero-Inflated Poisson Regression Model with right-censoring. We assess these estimators based on various approaches, including the classical Maximum Likelihood (ML) method and proposed estimators. This evaluation is carried out using a Monte Carlo simulation scheme. To evaluate the performance of the different estimators, we primarily focus on the following criteria:

- MSE, using the following equation:

$$\text{MSE} = \sum_{i=1}^{N} \frac{(\hat{\beta}_i - \beta)^\top (\hat{\beta}_i - \beta)}{N},$$

- MAE, using the following equation:

$$\text{MAE} = \sum_{i=1}^{N} \frac{|\hat{\beta}_i - \beta|}{N},$$

- MSDE, using the following equation:

$$\text{MSDE} = \sum_{i=1}^{N} \frac{(\hat{\beta}_i - \beta)^2}{N},$$

where $\hat{\beta}$ is the estimator of β obtained from the Maximum Likelihood (ML) or other proposed estimators. N is equal to 1000, corresponding to the number of replications used in the Monte Carlo simulation with various sample sizes: $n = 100, 300, 500, 1000$.

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

$$x_{ij} = \sqrt{1-\rho^2} w_{ij} + \rho w_{ip}, = 1, 2, \ldots, n; j = 1, 2, \ldots, p \tag{23}$$

Here, w_{ij} represents pseudo-random numbers drawn from the standard normal distribution, and ρ indicates the correlation between the explanatory variables, with values of $\rho = 0.95, 0.99, 0.999$. We also simulate the count data Y_i fromto ZIP regression model (1)-(2)-(3) The number of regressors p equals 4 and 8. The average percentage of zero-inflation in the simulated data sets is 15%. The censoring values are simulated from a zero-truncated Poisson model with the parameter λ, where λ is chosen to yield various average censoring proportions C in the simulated samples, namely $C = 15\%$ and $C = 40\%$. For comparison purposes, we also provide results that would be obtained if there were no censoring (i.e., when $C = 0\%$), as these results will serve as a benchmark for assessing the performance of the proposed estimators.

4.2 The discussion of results

To interpret the results presented in Tables 1 and 2, which summarize the performance of various estimators (MLE, Ridge, Liu, and MRT) under different levels of correlation (ρ) and sample sizes (n), it is crucial to analyze the impact of both multicollinearity and censoring on their robustness. The performance of the estimators varies with censoring levels (0%, 15%, and 40%), with increasing censoring leading to a significant degradation, particularly for MLE. At 0% censoring, all estimators perform well, especially for large sample sizes, with MRT showing slight superiority. At 15% censoring, the errors increase considerably, with MLE being the most affected, while MRT and LIU maintain relatively stable performance. At 40% censoring, traditional estimators like MLE struggle, whereas MRT and LIU remain more robust. Similarly, multicollinearity, measured by ρ, affects the estimators differently. At low multicollinearity ($\rho = 0.95$), all estimators perform adequately, but at high multicollinearity ($\rho = 0.999$), MLE exhibits instability, with a significant rise in error, whereas MRT and LIU provide more stable estimates. Overall, MRT emerges as the most robust estimator under both censoring and multicollinearity, followed closely by LIU. RIDGE benefits from regularization but does not match the performance of MRT and LIU in extreme conditions. Conversely, MLE is the most vulnerable estimator, suffering considerable performance degradation as censoring and multicollinearity intensify. These findings indicate that MRT and LIU are the most reliable choices for handling challenging datasets characterized by high correlation and censoring.

5 Applications to real datasets

5.1 Example 1: Mayotte social contact survey dataset

The information on social contacts was collected through cross-sectional surveys conducted by the Regional Health Agency (ARS) of Mayotte. These surveys were conducted between October and December 2021 with the oral informed consent of the participants. Participants were randomly assigned a day of the week to record each person they had come into contact with. In summary, only one person per household was invited to participate in the study. Paper diaries were handed out face-to-face to participants, who were trained on how to fill them out. They were required to note each person contacted only once in the diary. A contact was defined as either a physical contact (e.g., a handshake or kiss) or a two-way conversation involving at least three words in the physical presence of another person, without physical contact (a non-physical contact). Participants were also required to provide information on the age and sex of each person contacted. If the exact age was not known, participants were asked to estimate the corresponding age group. For each contact, participants had to indicate the location (home or outside the home) and the average number of usual contacts with that person.

The sample for the survey in Mayotte was constructed through quota sampling based on age, sex, and municipality, randomly selected from population registers, excluding individuals under 1 year of age. Participants received a written invitation for a face-to-face interview. Ifnecessary, respondents or their parents were visited at home and approached in a language other than French. During the interviews, participants stated their age and were asked about the number of people living in their household, excluding themselves, as well as the number of different people they conversed with during a typical week, excluding household members. A total of 4670 participants responded and completed the questionnaire.

The survey journal classified participants into five age groups: [0-18], [19-24], [25-49], [50-64], and [65+]. The journals also recorded basic socio-demographic information about the participants, including employment status, education level, household composition, age, gender, etc. It should be noted that during the interviews, responses from participants with high values (> 900) corresponded to special cases. In this study, the training data concerns only values where the number of contacts is less than or equal to 100. Although we carefully reviewed the listed variables and removed some clearly unusual values, the data now consists of 4,015 observations. In this case, 50% of individuals had no contact. This suggests a problem related to an excess of zeros.

In addition to the number of contacts, which is the dependent variable and has an excess of zero values, the following information was available in the data and used as follows. The goal of the analysis

Table 1: Performance metrics (MSE, MAE, MSDE) for different estimators when $p = 4$

		MLE			RIDGE			LIU			MRT		
ρ	n	MSE	MAE	MSDE	MSE	MAE	MSDE	MSE	MAE	MSDE	MSE	MAE	MSDE
						Level of censorship = 0%							
0.95	100	0.0650	0.4012	0.0287	0.0591	0.3835	0.0259	0.0641	0.3991	0.0284	0.0553	0.3709	0.0242
	300	0.0181	0.2132	0.0084	0.0175	0.2102	0.0082	0.0180	0.2131	0.0084	0.0171	0.2075	0.0079
	500	0.0154	0.1957	0.0069	0.0151	0.1936	0.0067	0.0154	0.1957	0.0068	0.0147	0.1916	0.0066
	1000	0.0067	0.1303	0.0031	0.0066	0.1297	0.0031	0.0067	0.1303	0.0031	0.0066	0.1291	0.0030
0.99	100	0.2716	0.8162	0.1191	0.1953	0.6956	0.0872	0.2342	0.7650	0.1082	0.1795	0.6636	0.0766
	300	0.0868	0.4644	0.0370	0.0768	0.4373	0.0332	0.0851	0.4605	0.0365	0.0709	0.4195	0.0304
	500	0.0652	0.4033	0.0290	0.0595	0.3856	0.0264	0.0645	0.4014	0.0288	0.0557	0.3726	0.0246
	1000	0.0294	0.2722	0.0136	0.0282	0.2667	0.0130	0.0293	0.2720	0.0136	0.0271	0.2618	0.0126
0.999	100	2.5301	2.4637	0.9952	0.7684	1.3125	0.2592	0.9647	1.5059	0.3700	0.6701	1.2348	0.2372
	300	0.8663	1.4744	0.3800	0.4295	1.0201	0.1683	0.5633	1.1831	0.2392	0.3873	0.9700	0.1515
	500	0.6176	1.2452	0.2819	0.3493	0.9289	0.1475	0.4452	1.0579	0.1989	0.3206	0.8897	0.1347
	1000	0.3074	0.8784	0.1321	0.2252	0.7489	0.0950	0.2688	0.8242	0.1178	0.2076	0.7146	0.0847
						Level of censorship = 15%							
0.95	100	0.6613	1.3769	0.4764	0.5762	1.3392	0.4538	0.6296	1.3583	0.4664	0.5690	1.3370	0.4532
	300	0.4925	1.2125	0.4152	0.4614	1.1988	0.3989	0.4835	1.2076	0.4098	0.4573	1.1974	0.3985
	500	0.3642	1.1169	0.3171	0.3559	1.1165	0.3152	0.3626	1.1166	0.3165	0.3536	1.1168	0.3155
	1000	0.4655	1.3217	0.4440	0.4625	1.3224	0.4434	0.4650	1.3218	0.4441	0.4618	1.3227	0.4432
0.99	100	1.5639	2.0018	0.8210	0.8704	1.5463	0.5340	1.0787	1.6817	0.6182	0.8320	1.5221	0.5271
	300	1.1738	1.7845	0.8461	0.8233	1.4987	0.6034	0.9822	1.6293	0.7207	0.7893	1.4721	0.5907
	500	0.6050	1.2851	0.3936	0.4892	1.2051	0.3572	0.5583	1.2479	0.3749	0.4772	1.1986	0.3575
	1000	0.5905	1.3703	0.4870	0.5436	1.3530	0.4693	0.5764	1.3631	0.4822	0.5390	1.3521	0.4690
0.999	100	11.5511	5.4551	5.5422	1.7462	2.0759	0.7679	1.8050	2.0623	0.7508	1.5437	1.9674	0.7234
	300	8.8280	4.9548	5.6973	2.4813	2.5210	1.4872	2.8713	2.6976	1.6420	2.0826	2.3189	1.3039
	500	3.2137	2.8316	1.5517	0.9803	1.5745	0.4896	1.2433	1.7470	0.6065	0.8847	1.5125	0.4734
	1000	1.9271	2.2690	1.0864	0.8873	1.5913	0.5683	1.0981	1.7338	0.6923	0.8478	1.5650	0.5593
						Level of censorship = 40%							
0.95	100	0.5442	1.2271	0.3713	0.4691	1.1677	0.3431	0.5152	1.2015	0.3607	0.4618	1.1634	0.3421
	300	0.3899	1.0464	0.3351	0.3657	1.0229	0.3177	0.3824	1.0385	0.3297	0.3619	1.0198	0.3140
	500	0.2985	0.9855	0.2569	0.2911	0.9837	0.2550	0.2972	0.9850	0.2566	0.2881	0.9836	0.2541
	1000	0.2692	0.9503	0.2503	0.2665	0.9504	0.2489	0.2689	0.9503	0.2501	0.2650	0.9506	0.2476
0.99	100	1.6164	2.0819	0.9385	0.9146	1.5664	0.5320	1.1178	1.7321	0.6522	0.8659	1.5316	0.5211
	300	1.2494	1.8878	0.9763	0.9443	1.6243	0.7479	1.0885	1.7532	0.8520	0.8967	1.5846	0.7226
	500	0.5481	1.1967	0.3694	0.4424	1.0920	0.3111	0.5061	1.1538	0.3454	0.4262	1.0785	0.3057
	1000	0.4286	1.0977	0.3411	0.3858	1.0576	0.3131	0.4157	1.0842	0.3348	0.3769	1.0514	0.3073
0.999	100	15.3952	6.4222	8.9188	3.4197	2.9066	1.5197	3.1659	2.7285	1.2337	2.9070	2.6947	1.3180
	300	11.7192	5.7750	9.2963	4.9183	3.5960	3.8443	5.4707	3.8024	3.9129	4.0627	3.2550	3.2075
	500	3.8268	3.1398	2.0982	1.3501	1.8215	0.6666	1.7649	2.0901	0.8980	1.1522	1.6954	0.5993
	1000	2.2843	2.4382	1.5306	1.1378	1.7003	0.7046	1.3605	1.8843	0.9061	1.0586	1.6436	0.6626

Table 2: Performance metrics (MSE, MAE, MSDE) for different estimators when $p = 8$

		MLE			RIDGE			LIU			MRT		
ρ	n	MSE	MAE	MSDE	MSE	MAE	MSDE	MSE	MAE	MSDE	MSE	MAE	MSDE
						Level of censorship = 0%							
0.95	100	0.1353	0.7850	0.0514	0.1164	0.7292	0.0446	0.1304	0.7722	0.0500	0.1102	0.7075	0.0419
	300	0.0303	0.3889	0.0133	0.0294	0.3834	0.0129	0.0303	0.3888	0.0132	0.0287	0.3786	0.0126
	500	0.0133	0.2573	0.0058	0.0132	0.2557	0.0058	0.0133	0.2573	0.0058	0.0130	0.2541	0.0057
	1000	0.0048	0.1561	0.0022	0.0048	0.1557	0.0022	0.0048	0.1561	0.0022	0.0047	0.1554	0.0022
0.99	100	0.6301	1.7001	0.2454	0.3905	1.3402	0.1536	0.4988	1.5189	0.2027	0.3759	1.3161	0.1474
	300	0.1189	0.7719	0.0522	0.1071	0.7330	0.0473	0.1166	0.7651	0.0515	0.1018	0.7133	0.0447
	500	0.0533	0.5157	0.0234	0.0506	0.5029	0.0223	0.0531	0.5149	0.0234	0.0486	0.4928	0.0214
	1000	0.0186	0.3039	0.0082	0.0182	0.3009	0.0080	0.0185	0.3038	0.0082	0.0178	0.2981	0.0079
0.999	100	6.2056	5.4796	2.6877	1.5684	2.6912	0.5687	1.8824	2.8838	0.6935	1.7125	2.6628	0.5473
	300	1.1455	2.3999	0.5190	0.6446	1.7942	0.2787	0.8338	2.0531	0.3739	0.6223	1.7650	0.2719
	500	0.4875	1.5593	0.2118	0.3427	1.3050	0.1475	0.4199	1.4489	0.1851	0.3307	1.2810	0.1413
	1000	0.1812	0.9520	0.0810	0.1534	0.8759	0.0686	0.1729	0.9316	0.0779	0.1461	0.8522	0.0642
						Level of censorship = 15%							
0.95	100	1.3329	2.2561	0.8503	1.0625	2.0020	0.6955	1.2091	2.1445	0.7903	1.0235	1.9647	0.6768
	300	0.8958	2.3168	0.7894	0.8260	2.2222	0.7269	0.8733	2.2868	0.7684	0.8136	2.2058	0.7171
	500	0.4346	1.4877	0.3364	0.4188	1.4612	0.3267	0.4313	1.4820	0.3344	0.4141	1.4540	0.3232
	1000	0.9167	2.0118	0.8779	0.9066	2.0001	0.8684	0.9154	2.0101	0.8771	0.9002	1.9927	0.8611
0.99	100	7.1717	5.4058	4.4607	4.5190	4.2251	2.8637	4.7141	4.3441	3.0840	4.3020	4.1177	2.7254
	300	5.3265	5.6969	4.8110	4.5800	5.2783	4.1004	4.8202	5.4208	4.3222	4.4423	5.1969	3.9640
	500	2.0724	3.1272	1.4626	1.8585	2.9533	1.3168	1.9775	3.0522	1.4042	1.8139	2.9177	1.2826
	1000	4.8834	5.0080	4.6041	4.7603	4.9421	4.4860	4.8295	4.9799	4.5528	4.7053	4.9126	4.4461
0.999	100	83.3491	19.5787	54.2293	28.5092	10.9846	18.3132	22.1213	7.6925	7.8680	31.3108	10.6733	17.0165
	300	60.1763	18.9098	54.7829	41.0686	15.5118	36.9737	33.3346	13.9541	28.8232	37.7220	14.8165	33.6204
	500	21.5569	10.2394	15.1108	15.4747	8.6049	10.6282	14.0143	8.2212	9.7123	14.7761	8.4037	10.1889
	1000	50.6605	16.6772	46.8296	46.8447	16.0305	43.2377	46.0552	15.8999	42.4951	45.2070	15.7480	41.7621
						Level of censorship = 40%							
0.95	100	1.3842	2.3285	0.9134	1.1340	2.0971	0.7594	1.2745	2.2307	0.8511	1.0897	2.0552	0.7329
	300	1.0052	2.4707	0.8960	0.9372	2.3851	0.8350	0.9825	2.4425	0.8757	0.9235	2.3681	0.8238
	500	0.4668	1.5303	0.3651	0.4515	1.5048	0.3545	0.4637	1.5250	0.3630	0.4460	1.4960	0.3505
	1000	0.9949	2.1034	0.9547	0.9849	2.0922	0.9454	0.9936	2.1019	0.9538	0.9783	2.0848	0.9388
0.99	100	7.7772	5.7320	5.1131	5.1532	4.6145	3.4848	5.3471	4.7231	3.6742	4.9080	4.4987	3.3440
	300	5.9326	5.9963	5.3872	5.2146	5.6143	4.7094	5.4436	5.7428	4.9247	5.0599	5.5277	4.5703
	500	2.3163	3.3209	1.6355	2.1087	3.1607	1.4924	2.2238	3.2513	1.5740	2.0586	3.1229	1.4555
	1000	5.3258	5.2629	5.0374	5.2054	5.2012	4.9197	5.2723	5.2361	4.9828	5.1499	5.1728	4.8734
0.999	100	88.4719	20.2027	59.5767	32.5264	11.9916	23.1023	15.9133	8.2637	10.2027	30.1277	11.5068	21.1220
	300	68.3415	20.0910	62.2485	49.5538	16.9932	44.8887	41.2513	15.4767	36.2129	45.6285	16.2497	40.8642
	500	23.3972	10.6773	16.7665	17.3701	9.1325	12.1614	15.7140	8.7139	11.0409	16.6177	8.9266	11.6438
	1000	55.2471	17.4257	51.2694	51.4742	16.8179	47.7805	50.4724	16.6612	46.5229	49.8412	16.5499	46.1929

is to assess whether age (X1), the number of household members under 18 years of age (X2), the number of household members aged 19-24 years (X3), the number of household members aged 50-64 years (X4), the number of household members aged 65 years and older (X5), household size (X6), and weight (X7) all influence the dependent variable: (y) the number of contacts.

We calculated the condition index (CI) to check for potential multicollinearity within the data. The CI is given by: $CI = \sqrt{\max(\lambda_j)/\min(\lambda_j)} = 692.83$ The high value of the CI indicates the presence of strong multicollinearity among the independent variables. The correlation matrix of the regressions, presented in Table 3, shows the correlation coefficients between the independent variables.

In the initial data, the response variable y is not censored. We chose to use it and artificially censor y. This censorship involves truncating a proportion of the values of y according to a defined censorship rate, such as 15% or 40%. The censorship is performed by generating thresholds based on zero-truncated Poisson values, and all values exceeding these thresholds are considered censored. The average censorship proportions for the sample of y are set to 0.15 and 0.4, respectively. The Zero-Inflated Poisson (ZIP) regression model is used to estimate the number of contacts, taking into account the censored data. The model is adjusted on a training dataset using the `zeroinfl` function from the `pscl` library. The optimization of the penalty parameters, such as Ridge and Liu penalties, is done by cross-validation to minimize the mean squared error (MSE). Once the models are adjusted, predictions are generated using the estimated coefficients, and the performance of the different models is evaluated using error metrics, such as MSE, MAE, and MSDE. These results allow for the comparison of model performances and examination of the impact of artificial censorship. The adjusted coefficients and associated errors are then stored for further analysis, with a final display of results for each censorship rate tested. The estimated coefficients from the Maximum Likelihood (ML) estimator and shrinkage estimators are presented in Table 4.

The results of different estimation methods, such as Maximum Likelihood Estimation (MLE), Ridge, LIU, and MRT, were compared under various censoring conditions (none, 15%, and 40%). Three model validation criteria were used: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Squared Difference of Errors (MSDE). Regarding the estimated coefficients, the intercept remains relatively stable despite the methods and censoring conditions, indicating robustness. For the explanatory variables (X_1 to X_8), the coefficients vary slightly across the estimation methods, but the trends are generally consistent, particularly for X_1, which remains negative, and for X_5, which shows a positive correlation with the number of contacts, especially under 40% censoring. In terms of validation, the MSE indicates that MRT performs the best without censoring, followed by RIDGE, with similar values at 15% and 40% censoring. For MAE, RIDGE and MRT are at the top, with RIDGE slightly excelling at 40% censoring. Regarding MSDE, both MRT and RIDGE provide the best performances, with a slight advantage for RIDGE at 15% and 40% censoring. The introduction of censoring seems to improve the models' performance, particularly for RIDGE and MRT, which show notable reductions in errors. In conclusion, RIDGE and MRT methods are the most effective for handling censored data, providing better prediction accuracy and greater stability in the estimates, and the introduction of censoring appears to have a beneficial effect, especially for these two methods.

5.2 Example 2: Environmental data

As an illustration, we analyze the pollutant emissions data described in [11]. There are three predictors, namely: the average ozone concentration (O_3), the daily air quality index (AQI), and the daily average temperature ($Temp_{moy}$). The dataset exhibits a zero-inflation ratio of 0.53. We set $C = 5$ and $C = 6$ for this data, resulting in censoring rates of approximately 15% and 40%, respectively. The condition index for this dataset was calculated to be 254.3948, which strongly indicates the presence of severe multicollinearity among the explanatory variables. We randomly split the dataset into a training set of size 80% and a test set of size 30%, and compare the results of four methods: MLE, Ridge, Liu, and MRT. The cross-validation method was employed to select the optimal regularization parameters for Ridge and Liu methods. Table 5 presents the estimated coefficients as well as the error values for each of the four methods: MLE, Ridge, Liu, and MRT. The estimated coefficients for the different variables (intercept, average ozone concentration O_3, Air Quality Index (AQI), and average daily temperature ($Temp_{moy}$) vary depending on the censorship rate and the estimator used. Without censorship (0%), MLE provides more extreme coefficients than Ridge, LIU, and MRT, with the latter reducing the variance of the estimates. LIU falls between MLE and Ridge, offering a trade-off between bias and variance. With

15% censorship, the amplitude of the coefficients decreases, especially for MLE and LIU, while Ridge and MRT maintain more stable estimates. At 40% censorship, MLE is the most affected, with sign inversions for some coefficients, whereas Ridge and MRT remain more robust.

Regarding the evaluation of prediction errors using MSE, MAE, and MSDE, MRT and Ridge exhibit the most stable performances with the lowest errors, while MLE and LIU show greater sensitivity to data variations. When censorship reaches 40%, MRT maintains the best overall performance with the lowest errors (MSE = 1.5402, MAE = 1.0895, MSDE = 0.6828), with Ridge being close in terms of efficiency. Overall, MRT and Ridge appear to be the most effective estimators under censorship, while MLE becomes unstable and LIU provides an intermediate alternative. Therefore, for better robustness and accuracy, Ridge and MRT are recommended, especially in the presence of censored data.

Table 3: Correlation matrix for Mayotte social contact survey dataset

Regressors	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.0000							
X_2	-0.6786							
X_3	-0.6493	0.9525						
X_4	-0.6534	0.9489	0.9614					
X_5	-0.6553	0.9333	0.9592	0.9560				
X_6	-0.6292	0.9207	0.9541	0.9448	0.9633			
X_7	-0.6227	0.9155	0.9534	0.9480	0.9637	0.9903		
X_8	-0.5684	0.3648	0.3488	0.3475	0.3483	0.3390	0.3377	1.0000

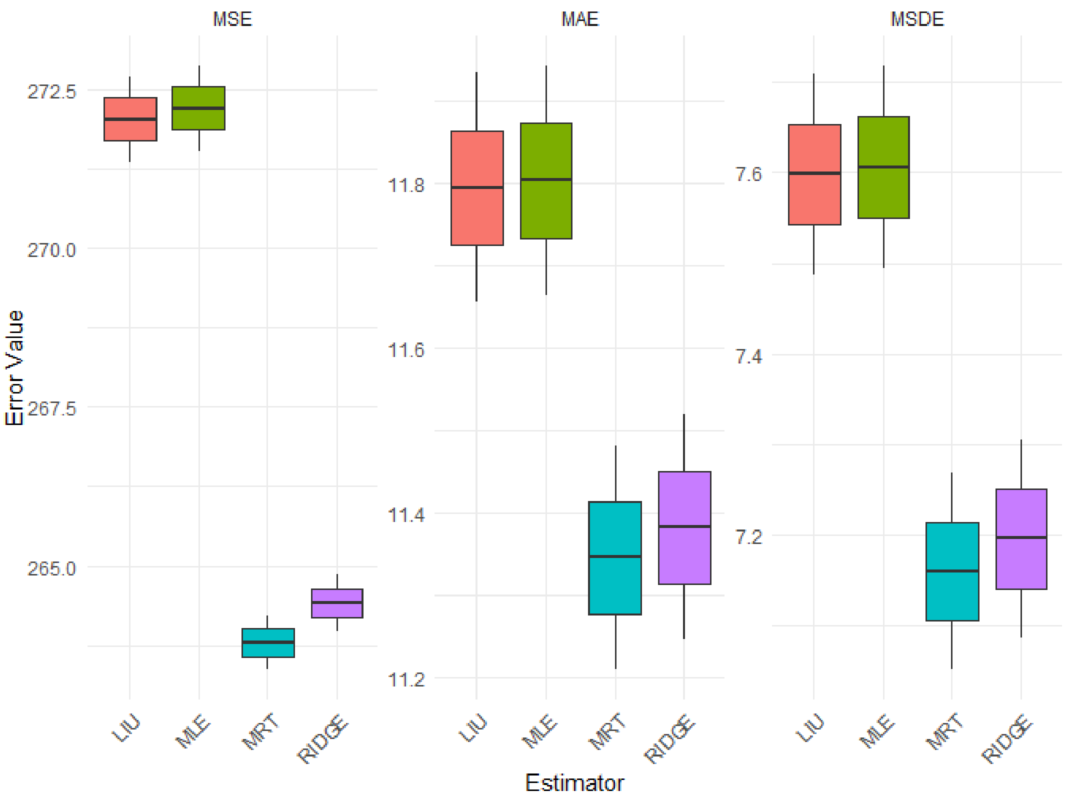

Figure 1: Boxplot of the MLE, Ridge, Liu, and MRT estimators in the presence of censorship for the error metrics provided in Table 4

Table 4: Estimated results for MLE, RIDGE, LIU, and MRT, along with model validation using MSE, MAE, and MSDE on Mayotte social contact data

Variables	**No Censoring**				**Censoring = 15%**				**Censoring = 40%**			
	MLE	**RIDGE**	**LIU**	**MRT**	**MLE**	**RIDGE**	**LIU**	**MRT**	**MLE**	**RIDGE**	**LIU**	**MRT**
Intercept	2.8621	2.8330	2.8615	2.8302	2.3010	2.2568	2.3001	2.2526	2.3540	2.3112	2.3532	2.3072
X_1	-0.0373	-0.0333	-0.0373	-0.0329	-0.0370	-0.0316	-0.0369	-0.0311	-0.0384	-0.0332	-0.0383	-0.0327
X_2	0.0325	-0.0026	0.0318	-0.0059	0.0683	0.0180	0.0673	0.0135	0.0593	0.0106	0.0583	0.0061
X_3	-0.0565	-0.0554	-0.0565	-0.0553	0.0277	0.0273	0.0278	0.0271	0.0271	0.0268	0.0271	0.0266
X_4	0.0309	0.0236	0.0308	0.0229	-0.0238	-0.0379	-0.0241	-0.0393	-0.0304	-0.0436	-0.0307	-0.0448
X_5	0.1384	0.1248	0.1381	0.1236	-0.0129	-0.0311	-0.0132	-0.0327	-0.0129	-0.0305	-0.0133	-0.0321
X_6	-0.2378	-0.2133	-0.2373	-0.2112	-0.1789	-0.1471	-0.1782	-0.1446	-0.1941	-0.1608	-0.1933	-0.1582
X_7	0.1885	0.1778	0.1883	0.1769	0.1130	0.1003	0.1127	0.0994	0.1363	0.1209	0.1359	0.1199
X_8	-0.0110	-0.0103	-0.0109	-0.0102	0.0108	0.0119	0.0109	0.0120	0.0100	0.0110	0.0100	0.0111
MSE	379.4876	356.8009	379.0252	354.8338	271.5270	263.9562	271.3596	263.3606	272.8792	264.8575	272.7036	264.2193
MAE	16.8128	16.2356	16.8016	16.1824	11.6638	11.245	11.6553	11.2093	11.9429	11.5185	11.9343	11.4815
MSDE	12.0195	11.4523	12.0085	11.4002	7.4947	7.0858	7.4864	7.0502	7.7184	7.3048	7.7100	7.2686

Table 5: Estimated results of MLE, Ridge, LIU, and MRT, along with model validation using MSE, MAE, and MSDE on environmental data.

Variables	**No Censoring**				**Censoring = 15%**				**Censoring = 40%**			
	MLE	**RIDGE**	**LIU**	**MRT**	**MLE**	**RIDGE**	**LIU**	**MRT**	**MLE**	**RIDGE**	**LIU**	**MRT**
Intercept	0.3500	0.1729	0.3581	0.1641	0.3090	0.1429	0.3091	0.1356	-0.0145	0.0126	-0.0106	0.0133
O_3	-0.7383	0.0046	-0.4551	0.0066	-0.5251	0.0188	-0.3106	0.0200	0.1341	0.0460	0.0936	0.0452
AQI	0.7183	0.0532	0.4305	0.0515	0.4888	0.0522	0.2833	0.0509	-0.0820	0.0386	-0.0357	0.0388
$Temp_{moy}$	0.0531	0.0386	0.0795	0.0375	0.1034	0.0475	0.1195	0.0459	0.1784	0.0680	0.1613	0.0653
MSE	2.0982	1.5669	1.9529	1.5608	1.9326	1.5628	1.8439	1.5576	1.5912	1.5424	1.5908	1.5402
MAE	1.2694	1.1233	1.2425	1.1199	1.2338	1.1166	1.2147	1.1137	1.1024	1.0903	1.1034	1.0895
MSDE	0.9223	0.7395	0.8901	0.7349	0.8711	0.7296	0.8490	0.7258	0.6874	0.6833	0.6889	0.6828

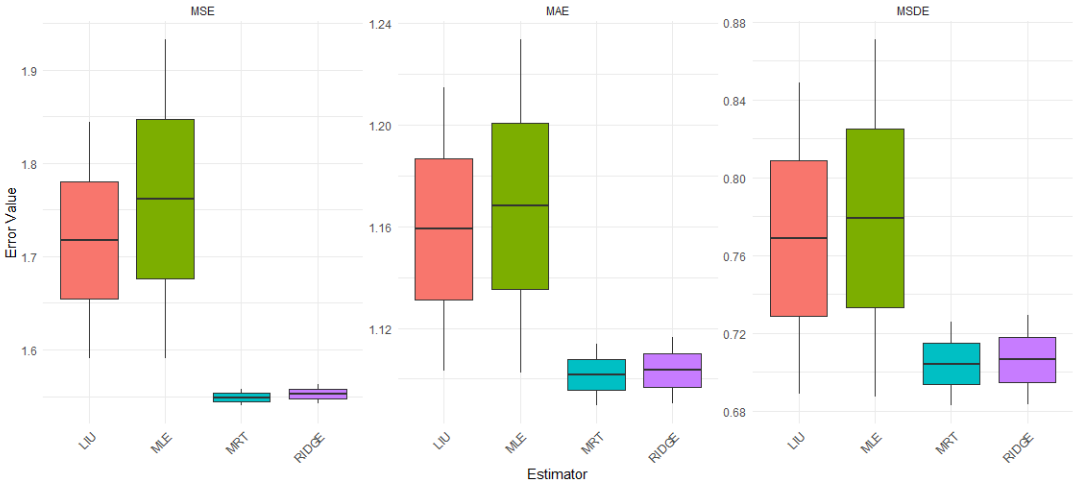

Figure 2: Boxplot of the MLE, Ridge, Liu, and MRT estimators in the presence of censorship for the error metrics provided in Table 5

6 Conclusions and future research directions

Zero-inflated (ZI) models, originally introduced by [19], address datasets with an excess of zeros, a limitation of traditional count models such as the Poisson and Negative Binomial distributions. In many practical applications, data are further complicated by censoring—either due to observational limits or detection thresholds—which, if ignored, can result in biased parameter estimates [19]. This has motivated the development of the Right-Censored Zero-Inflated Poisson (RCZIP) Regression Model, which offers a distinct advantage over standard zero-inflated approaches.

Despite the widespread use of maximum likelihood estimation (MLE) in RCZIP models, its performance is notably compromised in the presence of correlated regressors. To address this challenge, we have developed and investigated regularization methods, including both the conventional ridge regression and a novel modified ridge-type (MRT) estimator. The integration of these regularization techniques into the RCZIP model framework serves to stabilize parameter estimates by mitigating the adverse effects of multicollinearity, while preserving the predictive utility of correlated predictors.

Our extensive numerical simulations and empirical analyses provide compelling evidence of the superior performance of the MRT estimator over both the MLE and other regularization approaches. We evaluated the model under varying degrees of censoring (0%, 15%, and40%) and multicollinearity (moderate and high), and validated the methods using two real-world datasets. In these applications, data were partitioned into 80% training and 20% test sets, with regularization parameters optimally selected via cross-validation. Performance was assessed using mean squared error, mean absolute error, and median squared prediction error, with results consistently demonstrating reduced estimator variance, enhanced model fit, and robust predictive capabilities.

The promising outcomes of this study suggest that regularization, particularly through the proposed MRT estimator, offers a reliable and effective solution for managing multicollinearity in RCZIP regression models. Future research should consider extending this framework by incorporating adaptive ridge and L1-norm regularization techniques to further enhance model scalability and efficiency in high-dimensional settings.

Conflict of interest

The authors have no conflicts of interest to declare relevant to this article's content.

Funding

This research is part of a contract research work from ARS of Mayotte.

Data Availability Statement

The data will be made available upon request from the corresponding author

References

[1] Albfleisch, J. D., Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data (2nd ed.) *Wiley-Interscience.*

[2] Akram, M. N., Amin, M., Qasim, M. A new Liu-type estimator for the Inverse Gaussian Regression Model. Journal of Statistical Computation and Simulation, 90(7), 1153–1172,2020.

[3] Akram, M. N., Amin, M., Qasim, M. A new modified ridge-type estimator for the beta regression model: Simulation and application. *AIMS Mathematics* 7(1):1035-1057,2021

[4] Alreshidi, N. A., Alrasheedi, M. A., Lukman, A. F., Alrweili, H., & Farghali, R. A. Modified Kibria–Lukman Estimator for the Conway–Maxwell–Poisson Regression Model: Simulation and Application. *Mathematics*, 13(5), 794, 2025.

[5] Algamal, Z. Y., Lukman, A. F., Abonazel, M. R., & Awwad, F. A. Performance of the Ridge and Liu Estimators in the zero-inflated Bell Regression Model. *Journal of Mathematics*, 2022(1), 9503460, 2022.

[6] Amin, M., Akram, M. N., & Majid, A. On the estimation of Bell regression model using ridge estimator. *Communications in Statistics - Simulation and Computation*, 52(3), 854-867,2023.

[7] Bulut, Y. M., Lukman, A. F., Işılar, M., Adewuyi, E. T., & Algamal, Z. Y. Modified ridge estimator in the Bell regression model. *Journal of Inverse and Ill-posed Problems*, 32(6), 1081-1091, 2024.

[8] Cameron, A. C., Trivedi, P. K., 2013. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

[9] Cessie, S. L., and Houwelingen, J. V. "Ridge estimators in logistic regression." *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[10] El-Bassiouni, M. Y., Sayed, T., El-Basyouny, K. Modeling pedestrian injury severity in traffic collisions using a decision tree approach. Accident Analysis Prevention, 95, 122–130, 2016.

[11] Essoham Ali, Adewale F Lukman. Ridge-Penalized Zero-Inflated Probit Bell model for multicollinearity in count data. 2024. Available from: https://uco.hal.science/hal-04810240v1/document

[12] Fitzmaurice, G. M., Laird, N. M., Ware, J. H. (2011) Applied Longitudinal Analysis (2nd Edition). Wiley.

[13] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67, 1970.

[14] Hoerl, E., Kennard, R.W. and Baldwin, K.F. Ridge Regression: Some Simulations. *Communications in Statistics*, 4, 105-123, 1975.

[15] Firth, D. Bias reduction of maximum likelihood estimates. Biometrika, 80(1), 27–38, 1993.

[16] Farebrother, R. W. Further results on the mean square error of ridge regression. *J. Roy. Statist. Soc.* B 38:248–250, 1976.

[17] Kennedy, P. (2008) A Guide to Econometrics (6th Edition). Blackwell Publishing.

[18] Kibria, B. G., Månsson, K., & Shukur, G. Some ridge regression estimators for the zero-inflated Poisson model. *Journal of Applied Statistics*, 40(4), 721-735, 2013.

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

[20] Le Cessie, S., & Van Houwelingen, J. C. Ridge estimators in logistic regression. *Applied Statistics*, 41(1), 191-201, 1992.

[21] Liu, K. A New Class of Biased Estimators in Linear Regression. *Communications in Statistics - Theory and Methods*, 22(2), 393-402, 1993.

[22] Liu, Y., Wu, Y., & Zhang, H. H. Regularization methods for high-dimensional censored regression models: *A review. Statistical Modelling*, 20(2), 101–130, 2020.

[23] Lukman, A. F., Adewuyi, E., Månsson, K., & Kibria, B. M. G. A new estimator for the multicollinear Poisson regression model: simulation and application. *Scientific Reports*, 11(1), 3732, 2021.

[24] Lukman, A. F., Aladeitan, B., Ayinde, K., & Abonazel, M. R. Modified ridge-type for the Poisson regression model: simulation and application. *Journal of Applied Statistics*, 49(8), 2124–2136, 2022.

[25] Lukman, Adewale F and Ayinde, Kayode and Binuomote, Samuel and Clement, Onate A. Modified ridge-type estimator to combat multicollinearity: Application to chemical data, *Journal of Chemometrics*,33(5),e3125,2019.

[26] Månsson, K., & Shukur, G. A Poisson ridge regression estimator. *Economic Modelling*, 28(4), 1475-1481, 2011.

[27] Månsson, K. On ridge estimators for the negative binomial regression model. *Economic Modelling*, 29(2), 178-184, 2012.

[28] Mayer, L. S., Willke, T. A. On biased estimation in linear models. *Technometrics* 15:497–508, 1973.

[29] Menard, S. (2002) Applied Logistic Regression Analysis (2nd Edition). Sage Publications.

[30] McDonald, G.C. & Galarneau, D.I. A Monte Carlo Evaluation of Some Ridge-Type Estimators. *Journal of the American Statistical Association*, 70, 407-416, 1975.

[31] Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis. *Wiley*, 2012.

[32] Mayer, L. S., Willke, T. A. On biased estimation in linear models. *Technometrics* 15:497–508, 1973.

[33] Mehta, C. R., & Patel, N. R. (1995). Exact Inference for Categorical Data. In: Armitage, P., & Colton, T. (Eds.), *Encyclopedia of Biostatistics. Wiley.*

[34] Newhouse, J. P., Oman, S. D. An evaluation of ridge estimators. *Rand Corporation*, R-716-PR, 1–28, 1971.

[35] Nguyen, V. T., & Dupuy, J. F. Asymptotic results in censored zero-inflated Poisson regression. *Communications in Statistics - Theory and Methods*, 50(12), 2759–2779, 2019.

[36] Trenkler, G., Toutenburg, H. (1990). Mean squared error matrix comparisons between biased estimators: an overview of recent results. *Statst. Pap.* 31:165–179.

[37] Tikhonov, A. N. (1963) "Solution of incorrectly formulated problems and the regularization method." Soviet Mathematics Doklady, 4, 1035–1038.