Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 2093-2104
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.6039

Jackknifing K-L estimator in generalized linear models

Abed Ali Hamad[a], Zakariya Yahya Algamal[b,*]

[a] *Department Of Economics, College of Administration and Economics, University of Anbar, Anbar, Iraq*
[b] *Department of Statistics and Informatics, University of Mosul, Mosul, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

It is a challenge in the real application when modelling the relationship between the response variable and several explanatory variables when the existence of collinearity. Traditionally, in order to avoid this issue, several shrinkage estimators are proposed. Among them is the Kibria and Lukman estimator (K-L). In this study, a jackknifed version of the K-L estimator is proposed in the generalized linear model that combines the Jackknife procedure with the K-L estimator to reduce the biasedness. Our Monte Carlo simulation results and the real data application related to the inverse Gaussian regression model suggest that the proposed estimator can bring significant improvement relative to other competitor estimators, in terms of absolute bias and mean squared error.

Keywords: Collinearity, K-L estimator, Inverse Gaussian regression model, Jackknife estimator, Monte Carlo simulation

1. Introduction

Statistical modeling is essential in many scientific research areas because it explains the relationship between the response variable of interest and a number of explanatory variables. In linear regression model, there is assumption that the response variable must have normal distribution. In many real applications, however, this assumption may not hold. In medical sciences, for instance, the response variable can be positively skewed. Therefore, using linear regression model may not be suitable. Generalized linear model (GLM) is a broad class of regression models which it is gaining popularity as a statistical modeling method for continuous and discrete response variable [4, 5, 6, 7, 8].

*Corresponding author

Email addresses: `abidh1965@uoanbar.edu.iq` (Abed Ali Hamad), `zakariya.algamal@uomosul.edu.iq` (Zakariya Yahya Algamal)

Received: October 2021 *Accepted:* December 2021

In real applications, the design data matrix $\boldsymbol{X}$ has multicollinearity between explanatory variables, and, therefore, $\boldsymbol{X^T X}$ is singular or can be inflating the variance of the maximum likelihood estimator (MLE). Therefore, the traditional estimation methods, such as MLE, tend to perform poorly. The ridge, Liu, Liu-type, and others estimator that given by several authors is an alternative to MLE to overcome the multicollinearity in linear regression model [20, 27]. These estimators have been extended to the GLMs [31, 37, 40, 22, 32, 24, 2, 26, 42].

Although the powerful of these shrinkage estimators, but they have a smaller bias. It is possible to reduce bias by applying a jackknife procedure to these estimators. This procedure enables processing of experimental data to get statistical estimator for unknown parameters. The advantage of the jackknife procedure is that it presents an estimator that has a small bias while still providing beneficial properties of large samples [25, 41, 3].

The main objective given in this paper is to use Jackknife approach with the new ridge-type estimator (K-L estimator) of [23]. Our proposed estimator will efficiently help to decrease the biasness of K-L estimator in GLM. The superiority of our proposed estimator in different simulated examples and a real data application is proved.

2. K-L estimator in GLM

The nonlinear relationship between the response variable and the predictors can be transformed to linear relationship in GLM by linking them with a differentiable and monotonic link function [33]. "In GLM, the response variable belongs to the exponential family that includes normal, inverse Gaussian, and gamma distributions.

Assume that $(y_i, \boldsymbol{x_i}), i = 1, 2, \ldots, n$ is independent observed data with the predictor vector $x_i \in R^{p+1}$ and the response variable $y_i \in R$ which follows a distribution that belongs to the exponential family. Then, the density function of y_i can be expressed as

$$f(y_i; \theta_i, \phi) = exp\{\frac{y_i\theta_i - a_1(\theta_i)}{a_2(\phi)} + c(y_i, \phi)\} \tag{1}$$

where $a_1(.), a_2(.)$, and $c(.)$ are specific functions corresponding to the related distribution of Eq. (1). The parameter ϕ_i represents the natural (canonical) parameter, and the parameter $\phi > 0$ is representing the dispersion parameter. The mean and the variance of Eq. (1) are, respectively, defined as $E(y_i) = \mu_i = \partial a_1(\theta_i)/\partial(\theta_i)$ and $V(y_i) = a_2(\phi)\partial^2 a_1(\theta_i)/\partial\theta_i^2$. In GLM, the mean of the response variable, $\mu_i = E(y_i)$, is conditionally related to a linear function of predictors through a link function. The linear function is stated as $\eta_i = \beta_0 + \sum_{j=1}^{P} x_{ij}\beta_j = x_i^T \boldsymbol{\beta}$ with $x_i^T = (1, x_{i2}, x_{i3}, \ldots, x_{ip})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. The link function is providing the relation of the mean and the natural parameter as $\mu_i = g^{-1}(\eta_i) = g^{-1}(x_i^T \boldsymbol{\beta})$.

The parameter estimation in the GLM is achieved through using the MLE based on the iteratively reweighted least-squares algorithm. The log-likelihood of Eq. (1) is defined

$$l(\boldsymbol{\beta}, \phi) = S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - a_1(\theta_i)}{a_2(\phi)} + c(y_i, \phi) \right\} \tag{2}$$

Then, the MLE is derived by equaling the first derivative of Eq. (2) to zero as:

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \frac{1}{a_2(\phi)} \sum_{i=1}^{n} \left\{ y_i - \frac{\partial a_1(\theta_i)}{\partial(\theta_i)} den \right\} \boldsymbol{x_i} = 0 \tag{3}$$

Equation (3) cannot be solved analytically because it is nonlinear in $\boldsymbol{\beta}$. Fisher-scoring algorithm can be used to obtain the MLE where in each iteration, the parameter is updated by

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + I^{-1}(\boldsymbol{\beta}^{(r)})S(\boldsymbol{\beta}^{(r)}) \tag{4}$$

where $I^{-1}(\boldsymbol{\beta}) = (-E(\partial^2 l(\boldsymbol{\beta}, \phi)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T))^{-1}$. After that, the estimated coefficients are defined as is updated by

$$\hat{\boldsymbol{\beta}}_{MILE} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{u}} \tag{5}$$

where $\hat{\boldsymbol{W}} = diag[(\partial\mu_i/\partial\eta_i)^2/V(y_i)]$ and $\hat{\mathbf{u}}$ is a vector where i^{th} element equals to $\hat{u}_i = \hat{\mu}_i + [(y_i - \hat{\mu}_i)(\partial\mu_i/\partial\eta_i)]$. The MLE is distributed asymptotically normal with a covariance matrix as

$$cov(\hat{\boldsymbol{\beta}}_{MILE}) = \left[-E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)\right]^{-1} = [a_{2(\phi)}]^2(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1} \tag{6}$$

In the presence of multicollinearity, the $(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}) \leq rank(\mathbf{X})$, and, therefore, the near singularity of $(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})$ makes the estimation unstable and enlarges the variance [28]. The ridge estimator (RE) [20], Liu estimator [27] and their extensions have been consistently demonstrated to be an attractive and alternative to the MLE, when the multicollinearity exists”.

3. Jackknifing K-L estimator in GLM

In 2020, Kibria and Lukman proposed a new ridge-type estimator for the linear regression model. This proposed estimator is called as Kibria-Lukman (KL) estimator, which is defined as [23]:

$$\hat{\boldsymbol{\beta}}_{KL} = (I + k(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(I - k(\boldsymbol{X}^T\boldsymbol{X})^{-1})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y \tag{7}$$

where $k > 0$ is the shrinkage parameter. The estimator $\hat{\boldsymbol{\beta}}_{KL}$ is biased but more stable and has less mean square error than the ordinary least square estimator. For the GLM, Eq. (7), $\hat{\boldsymbol{\beta}}_{KL-GLM}$, can be defined as [29, 30].

$$\hat{\boldsymbol{\beta}}_{KL-GLM} = (I + k(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1})^{-1}(I - k(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1})(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{u}} \tag{8}$$

The bias and variance of Eq. (8) are defined as, respectively,

$$Bias(\hat{\boldsymbol{\beta}}_{KL-GLM}) = -2k\boldsymbol{Q}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} + k\boldsymbol{I})^{-1}\alpha \tag{9}$$

$$\begin{aligned} Variance(\hat{\boldsymbol{\beta}}_{KL-GLM}) =& \phi\boldsymbol{Q}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} - k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1} \\ & (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} - k\boldsymbol{I})^{-1}\boldsymbol{Q}^T \end{aligned} \tag{10}$$

where ϕ is the dispersion parameter, $\boldsymbol{Q} = (q_1, q_2, \ldots, q_p)$ represents the matrix of eigenvectors of the $\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ matrix, and $\alpha = \boldsymbol{Q}^T\boldsymbol{\beta}$. In simple way, the mean square error (MSE) of Eq. (8) can be written as

$$MSE(\hat{\boldsymbol{\beta}}_{KL-GLM}) = \phi\sum_{j=1}^{P}\frac{(\lambda_j - k)^2}{\lambda_j(\lambda_j + k)^2} + 4k^2\sum_{j=1}^{P}\frac{\alpha_j^2}{(\lambda_j + k)^2} \tag{11}$$

Shrinkage estimators are biased estimators. In linear regression model, Singh et al. [43] proposed the Jackknife procedure to alleviate the problem of bias in generalized ridge estimator. The theoretical and application of the jackknife estimator have been studied by several authors [36, 18, 1, 25, 44, 45, 10, 11, 41, 42, 38, 15, 35, 39, 34, 29, 30, 9, 3, 4, 5, 38, 7, 12, 13, 14].

The proposed estimator, Jackknifed K-L estimator (JKL-GLM), in GLM can be expressed and derived. Let $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ is the matrix of eigenvalues of the $\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ matrix, such that$\boldsymbol{Q}^T\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} = \boldsymbol{\Lambda}$, where $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Q}$. Consequently, the MLE estimator of Eq. (5) can be re-written as

$$\hat{\boldsymbol{\beta}}_{MILE} = \boldsymbol{Q}\hat{\boldsymbol{\theta}}_{MILE} \tag{12}$$

where $\hat{\boldsymbol{\theta}}_{MILE} = \boldsymbol{\Lambda}^{-1}\boldsymbol{M}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{u}}$. As a result, the KL-GLM estimator of Eq. (8) is re-written as

$$\hat{\boldsymbol{\theta}}_{KL-GLM} = (\boldsymbol{\Lambda} + k\boldsymbol{I})^{-1}(\boldsymbol{\Lambda} - k\boldsymbol{I})^{-1}\boldsymbol{M}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{u}} \tag{13}$$

Following the idea of Jackknife approach [19], let $\boldsymbol{u}_{(-i)}$, $\boldsymbol{m}_{(-i)}$, and $\boldsymbol{W}_{(-i)}$, respectively, are the $\boldsymbol{i}^{th}$ row deleted from the vector $\boldsymbol{u}$, the $\boldsymbol{i}^{th}$ row deleted from the matrix $\boldsymbol{M}$, and the $\boldsymbol{i}^{th}$ row and column deleted from the matrix $\boldsymbol{W}$. Let $\hat{\boldsymbol{\theta}}_{KL-GLM}$ be given by Eq. (12) with replacing $\boldsymbol{M}$, $\boldsymbol{W}$, and $\boldsymbol{u}$ by $\boldsymbol{M}_{(-i)}$, $\boldsymbol{W}_{(-i)}$, and $\boldsymbol{u}_{(-i)}$, thus,

$$\hat{\boldsymbol{\theta}}_{KL-GLM(-i)} = (\boldsymbol{M}^T_{(-i)}\hat{\boldsymbol{W}}_{(-i)}\boldsymbol{M}_{(-i)} + k\boldsymbol{I})^{-1}(\boldsymbol{M}^T_{(-i)}\hat{\boldsymbol{W}}_{(-i)}\boldsymbol{M}_{(-i)} - k\boldsymbol{I})^{-1}\boldsymbol{M}^T_{(-i)}\hat{\boldsymbol{W}}_{(-i)}\hat{\boldsymbol{u}}_{(-i)} \tag{14}$$

where $(\boldsymbol{M}^T_{(-i)}\hat{\boldsymbol{W}}_{(-i)}\boldsymbol{M}_{(-i)} \mp k\boldsymbol{I})^{-1}$ is calculated according to Sherman-Morrison Woodbury theorem. Consequently, Eq. (13) can be expressed as

$$\hat{\boldsymbol{\theta}}_{KL-GLM(-i)} = \hat{\boldsymbol{\theta}}_{KL-GLM} - \frac{(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} - k\boldsymbol{I})^{-1}\boldsymbol{m}_i^T(\hat{\boldsymbol{u}}_i - \boldsymbol{m}_i^T\hat{\boldsymbol{\theta}}_{KL-GLM})}{1 - \boldsymbol{m}_i^T(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} - k\boldsymbol{I})^{-1}\boldsymbol{m}_i} \tag{15}$$

Using the weighted pseudo values [19], which are calculated as

$$T_i = \hat{\boldsymbol{\theta}}_{KL-GLM} + n(1 - \boldsymbol{m}_i^T(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} - k\boldsymbol{I})^{-1}\boldsymbol{m}_i)(\hat{\boldsymbol{\theta}}_{KL-GLM} - \hat{\boldsymbol{\theta}}_{KL-GLM}(-i)) \tag{16}$$

Then, our proposed estimator, JKL-GLM, is defined as

$$\hat{\boldsymbol{\theta}}_{JKL-GLM} = \hat{\boldsymbol{\theta}}_{KL-GLM} + (\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} - k\boldsymbol{I})^{-1}\sum_{i=1}^{n}\boldsymbol{m}_i^T(\hat{\boldsymbol{u}}_i - \boldsymbol{m}_i^T\hat{\boldsymbol{\theta}}_{KL-GLM}) \tag{17}$$

The bias, variance and MSE of $\hat{\boldsymbol{\theta}}_{JKL-GLM}$ is respectively defined as

$$Bias(\hat{\boldsymbol{\theta}}_{JKL-GLM}) = \begin{bmatrix} [\boldsymbol{I} - 2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}]^2 \\ \\ [\boldsymbol{I} + 2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1-1}] - \boldsymbol{I} \end{bmatrix}\boldsymbol{\theta} \tag{18}$$

$$\begin{aligned} Variance(\hat{\boldsymbol{\theta}}_{JKL-GLM}) =& \phi[\boldsymbol{I} - (2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1})^2]^2(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1} \\ & [\boldsymbol{I} - (2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1})]^2, \end{aligned} \tag{19}$$

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}_{JKL-GLM}) =& \phi\sum_{j=1}^{P}\frac{\left((\lambda_j + k)^2 - 4k^2\right)^2(\lambda_j - k)^2}{\lambda_j(\lambda_j + k)^6} \\ & \sum_{j=1}^{P}\frac{\left((\lambda_j - k)^2(\lambda_j + 3k) - (\lambda_j + k)^3\right)^2\alpha_j^2}{(\lambda_j + k)^6} \end{aligned} \tag{20}$$

4. Theoretical comparison between $\hat{\boldsymbol{\theta}}_{JKL-GLM}$ and $\hat{\boldsymbol{\theta}}_{KL-GLM}$

With availability of different estimators for a parameter in the regression model, it is of interest to compare their performances in terms of MSE. For two given estimators $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ of $\hat{\boldsymbol{\beta}}$, the estimator $\hat{\boldsymbol{\beta}}_B$ is said to be superior to $\hat{\boldsymbol{\beta}}_B$ under the MSE criterion if and only if $\Delta = MSE(\hat{\boldsymbol{\beta}}_A) - MSE(\hat{\boldsymbol{\beta}}_B) \geq 0$.

Lemma 4.1. *[17] Let $\boldsymbol{G}$ is a $p \times p$ positive definite matrix, $\boldsymbol{b}$ is a $p \times 1$ vector, and c is a positive constant. Then $c^{\boldsymbol{G}-\boldsymbol{b}\boldsymbol{b}^{\boldsymbol{T}}}$ is a nonnegative definite if and only if $\mathbf{b}^T \boldsymbol{G}^{-1}\mathbf{b} \leq c$ is hold*

Theorem 4.2. *The proposed estimator $\hat{\boldsymbol{\theta}}_{JKL-GLM}$ is superior to estimator $\hat{\boldsymbol{\theta}}_{KL-GLM}$ if and only if*

$$\theta^T[\boldsymbol{I} - 2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}]^2[\boldsymbol{I} + 2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}] - \boldsymbol{I}]$$
$$[\phi(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} - k\boldsymbol{I})^2(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M})^{-1}(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-2} - \phi\left[\boldsymbol{I} - (2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1})^2\right]^2$$
$$(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M})\left[\boldsymbol{I} - (2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1})\right]^2 + 4k^2(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-2}\boldsymbol{\theta}^T\boldsymbol{\theta}]$$
$$\left[\boldsymbol{I} - 2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}\right]^2[\boldsymbol{I} + 2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1}] - \boldsymbol{I}]0 < 1 \tag{21}$$

Proof . The difference between $MSE(\hat{\boldsymbol{\theta}}_{JKL-GLM})$ and $MSE(\hat{\boldsymbol{\theta}}_{KL-GLM})$ is

$$\phi diag\left\{\frac{1}{\lambda_j}\left(\frac{\lambda_j - k}{\lambda_i + k}\right)^2 - \frac{((\lambda_j + k)^2 - 4k^2)^2(\lambda_j - k)^2}{\lambda_j(\lambda_j + k)^6}\right\}_{j=1}^{p} \tag{22}$$

Consequently,

$$\left[\phi(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^2(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M})^{-1}(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-2} - \phi\left[\boldsymbol{I} - (2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1})^2\right]^2\right]$$
$$(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M})\left[\boldsymbol{I} - (2k(\boldsymbol{M}^T\hat{\boldsymbol{W}}\boldsymbol{M} + k\boldsymbol{I})^{-1})\right]^2 \tag{23}$$

is positive definite provided

$$(\lambda_j - k)^2(\lambda_j + k)^4 > \left((\lambda_j + k)^2 - 4k^2\right)^2(\lambda_j - k)^2 \tag{24}$$

the proof is completed. □

5. Simulation results

In this section, a Monte Carlo simulation experiment is used to examine the performance of JKL-GLM with different degrees of multicollinearity for the inverse Gaussian regression model (IGRM). “The IGRM has been widely used in industrial engineering, life testing, reliability, marketing, and social sciences [4, 5, 6, 7, 8, 10, 11, 12, 13]. It considered as an GLM. Specifically, IGRM is used when the response variable under the study is positively skewed.

The response variable is drawn from inverse Gaussian distribution $y_i \sim IG(\mu_i, \phi)$ with sample sizes $n = 50, 100$ and 200 , respectively, where $\phi \in \{0.5, 3\}$. The explanatory variables $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{in})$ have been generated from the following formula

$$x_{ij} = (1 - \rho^2)^{1/2}W_{ij} + \rho W_{ip}, \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, p \tag{25}$$

where ρ represents the correlation between the explanatory variables and W_{ij} 's are independent pseudo-random numbers. The number of p is 4, 8, and 12, the values of ρ are 0.90, 0.95, and 0.99 are considered. The log link function is investigated, which is defined as

$$\mu_i = exp(\boldsymbol{x}_i^T \boldsymbol{\beta}), \quad i = 1, 2, \ldots, n \tag{26}$$

Here, the vector $\boldsymbol{\beta}$ is chosen as the normalized eigenvector corresponding to the largest eigenvalue of the $\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}$ matrix subject to $\boldsymbol{\beta}^T \boldsymbol{\beta} = 1$ [22]. In addition, the in Eq. (25) are generated from normal distribution (0,1). The estimated average MSE and the average absolute bias are calculated as

$$MSE(\hat{\boldsymbol{\beta}}_{IGLE}) = \frac{1}{R} \sum_{i=1}^{R} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{27}$$

$$Bias(\hat{\boldsymbol{\beta}}) = \frac{1}{1000} \sum_{i=1}^{1000} |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}| \tag{28}$$

where R equals 1000 corresponding to the number of replicates used in our simulation. All the calculations are computed by R program. The optimum value of k can be obtained by using Hoerl et al. [21] formula as

$$\hat{k} = \frac{\hat{\phi} p}{\hat{\alpha}^T \hat{\alpha}} \tag{29}$$

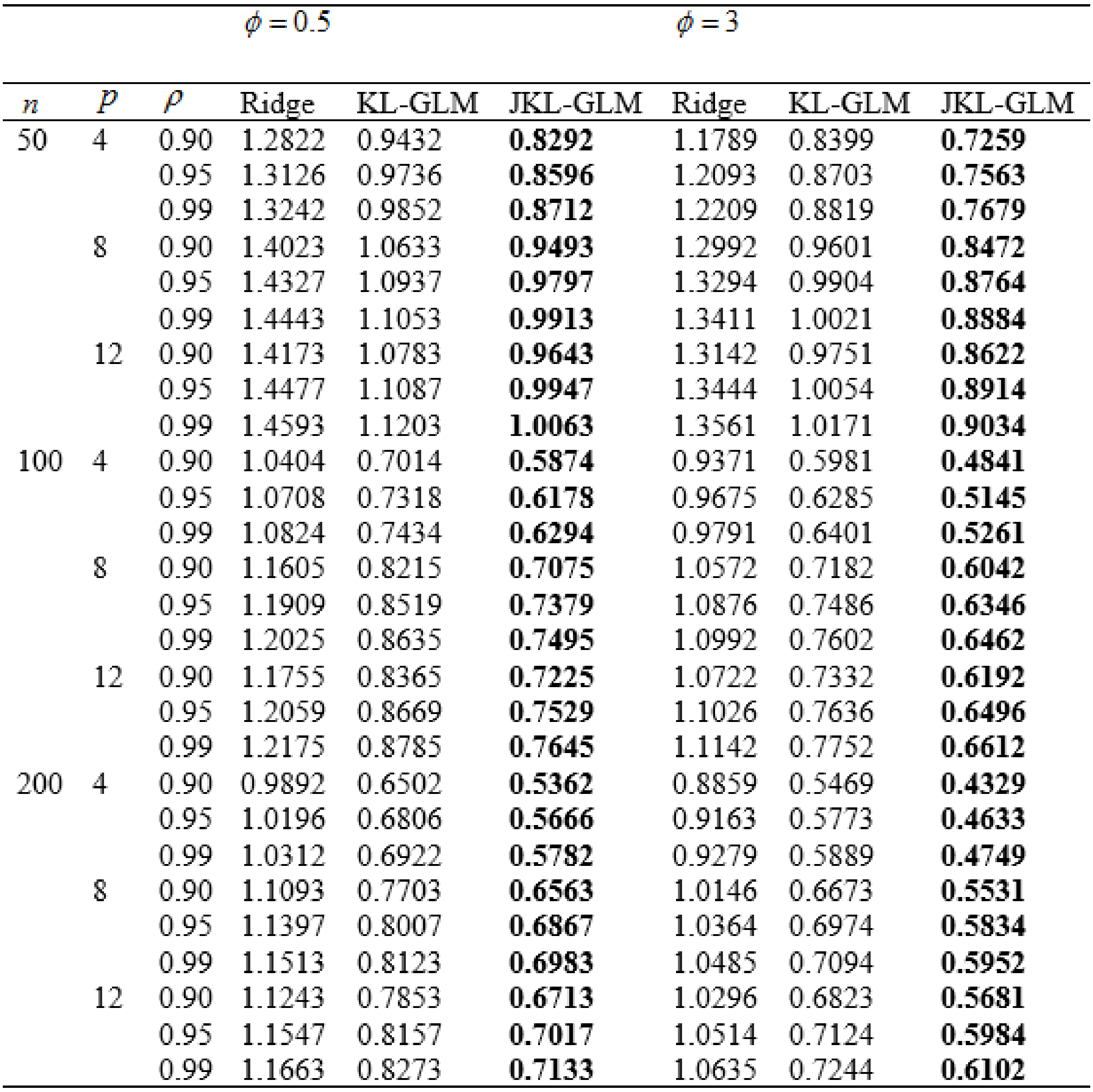
where $\hat{\phi}$ is the estimated dispersion parameter which is calculated by

$$\hat{\phi} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3} \tag{30}$$

The averaged bias and MSE all the combination of n, ϕ, p and ρ, are respectively summarized in Tables 1–3. The best value of the averaged bias and MSE is highlighted in bold. As Table 1 shows, the proposed estimator, JKL-GLM, gives low bias comparing with Ridge estimator and KL-GLM estimator. On other hand, KL-GLM performances better than Ridge estimator. This finding indicates that the Jackknifed estimator is significantly decreasing the bias. Meanwhile, JKL-GLM estimator performs well not only in terms of bias but also in terms of MSE (Table 2). It is noted from Table 2 that JKL-GLM estimator ranks first with respect to MSE. In the second rank, KL-GLM estimator performs better than both Ridge and MLE estimators. Additionally, MLE estimator has the worst performance among Ridge, KL-GLM, and JKL-GLM which is significantly impacted by the multicollinearity.

Furthermore, with respect to ρ , there is increasing in the bias and MSE values when the correlation degree increases regardless the value of n, ϕ and ρ . Regarding the number of explanatory variables, it is easily seen that there is a negative impact on both bias and MSE, where there are increasing in their values when the p increasing from four variables, eight variables, to 12 variables. In Addition, in terms of the sample size n , the bias and the MSE values decrease when n increases, regardless the value of ρ and p ". Clearly, in terms of the dispersion parameter ϕ , both bias and MSE values are decreasing when ϕ increasing.

Table 1: Averaged bias values for used estimators

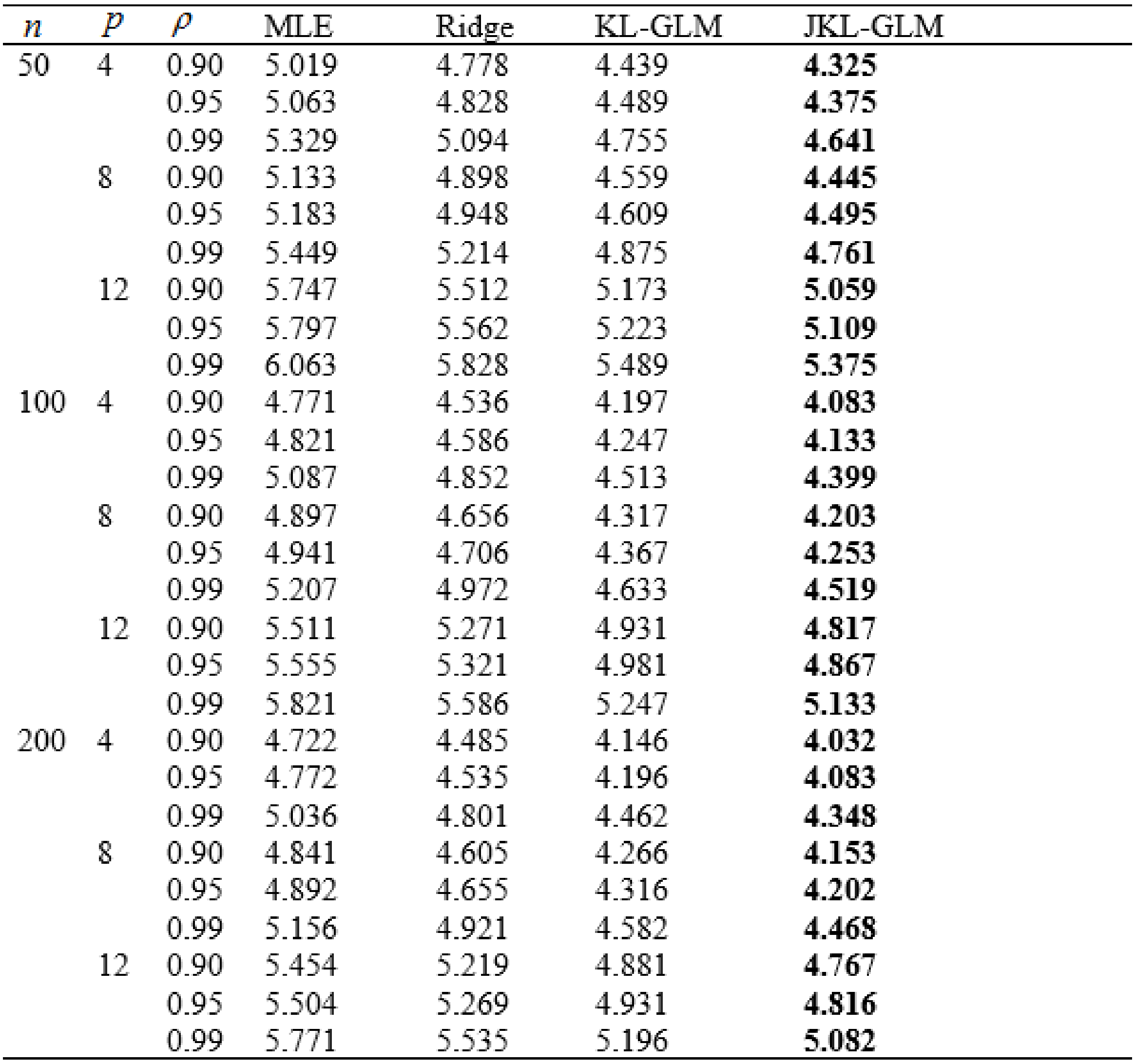


			$\phi = 0.5$			$\phi = 3$		
n	p	ρ	Ridge	KL-GLM	JKL-GLM	Ridge	KL-GLM	JKL-GLM
50	4	0.90	1.2822	0.9432	**0.8292**	1.1789	0.8399	**0.7259**
		0.95	1.3126	0.9736	**0.8596**	1.2093	0.8703	**0.7563**
		0.99	1.3242	0.9852	**0.8712**	1.2209	0.8819	**0.7679**
	8	0.90	1.4023	1.0633	**0.9493**	1.2992	0.9601	**0.8472**
		0.95	1.4327	1.0937	**0.9797**	1.3294	0.9904	**0.8764**
		0.99	1.4443	1.1053	**0.9913**	1.3411	1.0021	**0.8884**
	12	0.90	1.4173	1.0783	**0.9643**	1.3142	0.9751	**0.8622**
		0.95	1.4477	1.1087	**0.9947**	1.3444	1.0054	**0.8914**
		0.99	1.4593	1.1203	**1.0063**	1.3561	1.0171	**0.9034**
100	4	0.90	1.0404	0.7014	**0.5874**	0.9371	0.5981	**0.4841**
		0.95	1.0708	0.7318	**0.6178**	0.9675	0.6285	**0.5145**
		0.99	1.0824	0.7434	**0.6294**	0.9791	0.6401	**0.5261**
	8	0.90	1.1605	0.8215	**0.7075**	1.0572	0.7182	**0.6042**
		0.95	1.1909	0.8519	**0.7379**	1.0876	0.7486	**0.6346**
		0.99	1.2025	0.8635	**0.7495**	1.0992	0.7602	**0.6462**
	12	0.90	1.1755	0.8365	**0.7225**	1.0722	0.7332	**0.6192**
		0.95	1.2059	0.8669	**0.7529**	1.1026	0.7636	**0.6496**
		0.99	1.2175	0.8785	**0.7645**	1.1142	0.7752	**0.6612**
200	4	0.90	0.9892	0.6502	**0.5362**	0.8859	0.5469	**0.4329**
		0.95	1.0196	0.6806	**0.5666**	0.9163	0.5773	**0.4633**
		0.99	1.0312	0.6922	**0.5782**	0.9279	0.5889	**0.4749**
	8	0.90	1.1093	0.7703	**0.6563**	1.0146	0.6673	**0.5531**
		0.95	1.1397	0.8007	**0.6867**	1.0364	0.6974	**0.5834**
		0.99	1.1513	0.8123	**0.6983**	1.0485	0.7094	**0.5952**
	12	0.90	1.1243	0.7853	**0.6713**	1.0296	0.6823	**0.5681**
		0.95	1.1547	0.8157	**0.7017**	1.0514	0.7124	**0.5984**
		0.99	1.1663	0.8273	**0.7133**	1.0635	0.7244	**0.6102**

6. Real-life application

The chemical dataset adopted in this study was employed in the study of Correa-Basurto et al. [21]. This dataset consists of 88 compounds from N-aryl derivatives which is used as inhibitors of acetylcholinesterase (AChE) and butyrylcholinesterase (BuChE). The inhibitory activity of N-aryl derivatives is reported as the inhibition constant Ki (nM) and is transformed into the negative logarithmic scale pKi (M) which is used as a response variable for QSAR analysis. However, the regression modeling is employed when the response variable is skewed. In this study, the variables of interest are described in Table 4.

Table 2: Averaged MSE values for used estimators when $\phi = 0.5$



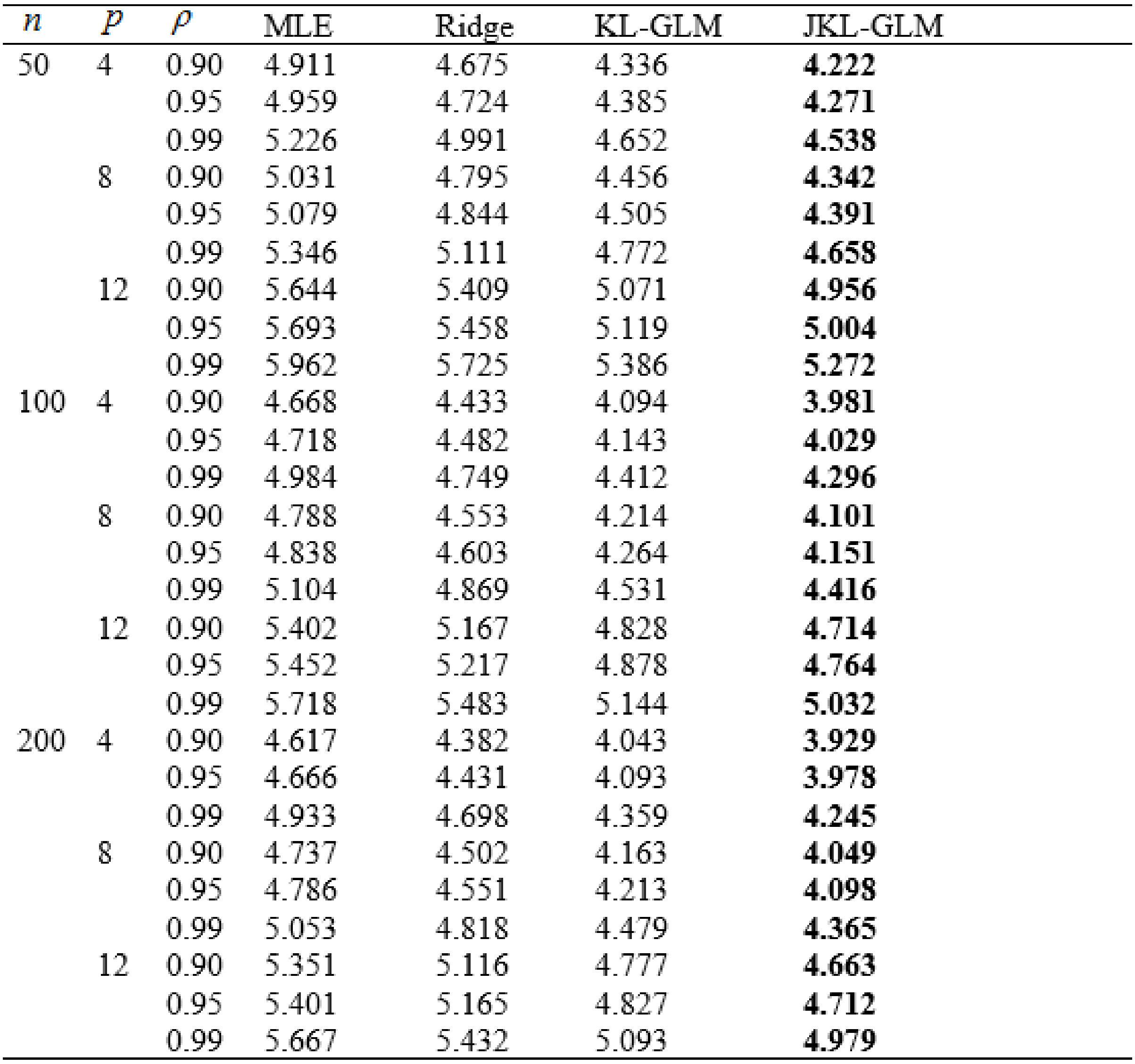
n	p	ρ	MLE	Ridge	KL-GLM	JKL-GLM
50	4	0.90	5.019	4.778	4.439	**4.325**
		0.95	5.063	4.828	4.489	**4.375**
		0.99	5.329	5.094	4.755	**4.641**
	8	0.90	5.133	4.898	4.559	**4.445**
		0.95	5.183	4.948	4.609	**4.495**
		0.99	5.449	5.214	4.875	**4.761**
	12	0.90	5.747	5.512	5.173	**5.059**
		0.95	5.797	5.562	5.223	**5.109**
		0.99	6.063	5.828	5.489	**5.375**
100	4	0.90	4.771	4.536	4.197	**4.083**
		0.95	4.821	4.586	4.247	**4.133**
		0.99	5.087	4.852	4.513	**4.399**
	8	0.90	4.897	4.656	4.317	**4.203**
		0.95	4.941	4.706	4.367	**4.253**
		0.99	5.207	4.972	4.633	**4.519**
	12	0.90	5.511	5.271	4.931	**4.817**
		0.95	5.555	5.321	4.981	**4.867**
		0.99	5.821	5.586	5.247	**5.133**
200	4	0.90	4.722	4.485	4.146	**4.032**
		0.95	4.772	4.535	4.196	**4.083**
		0.99	5.036	4.801	4.462	**4.348**
	8	0.90	4.841	4.605	4.266	**4.153**
		0.95	4.892	4.655	4.316	**4.202**
		0.99	5.156	4.921	4.582	**4.468**
	12	0.90	5.454	5.219	4.881	**4.767**
		0.95	5.504	5.269	4.931	**4.816**
		0.99	5.771	5.535	5.196	**5.082**

To check whether the pKi variable belongs to the inverse Gaussian distribution, a Chi-square test is used. The result of the test equals to 10.211 with p-value equals to 0.538. It is indicated form this result that the inverse Gaussian distribution fits very well to this pKi variable. That is, the following model is set

$$\hat{y}_{pKi} = exp(\sum_{j=1}^{8} \boldsymbol{x}_j \hat{\beta}_j) \qquad (31)$$

In order to check whether there is a relationship among the eight explanatory variables or not, Table 5 displays the correlation matrix among the 8 explanatory variables. “It is obviously seen that there are correlations greater than 0.90 among several variables. To test the existence of collinearity the eigenvalues of the matrix $\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ are obtained as 2.547×10^8, 1.368×10^7, 4.251×10^5, 3.659×10^3, 2.227×10^3 , 5.320×102, 3184, 10.528. The determined condition number $CN = \sqrt{\lambda_{max}/\lambda_{min}}$ of the data is 4918.6 indicating that the severe multicollinearity issue is exist. Further, estimated dispersion parameter is 0.00103 after fitting the inverse Gaussian regression model with log link function.

Table 3: Averaged MSE values for used estimators when $\phi = 3$



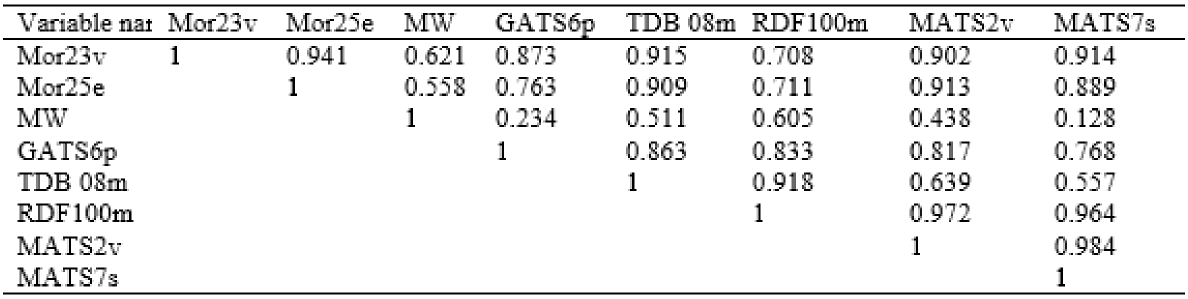
n	p	ρ	MLE	Ridge	KL-GLM	JKL-GLM
50	4	0.90	4.911	4.675	4.336	**4.222**
		0.95	4.959	4.724	4.385	**4.271**
		0.99	5.226	4.991	4.652	**4.538**
	8	0.90	5.031	4.795	4.456	**4.342**
		0.95	5.079	4.844	4.505	**4.391**
		0.99	5.346	5.111	4.772	**4.658**
	12	0.90	5.644	5.409	5.071	**4.956**
		0.95	5.693	5.458	5.119	**5.004**
		0.99	5.962	5.725	5.386	**5.272**
100	4	0.90	4.668	4.433	4.094	**3.981**
		0.95	4.718	4.482	4.143	**4.029**
		0.99	4.984	4.749	4.412	**4.296**
	8	0.90	4.788	4.553	4.214	**4.101**
		0.95	4.838	4.603	4.264	**4.151**
		0.99	5.104	4.869	4.531	**4.416**
	12	0.90	5.402	5.167	4.828	**4.714**
		0.95	5.452	5.217	4.878	**4.764**
		0.99	5.718	5.483	5.144	**5.032**
200	4	0.90	4.617	4.382	4.043	**3.929**
		0.95	4.666	4.431	4.093	**3.978**
		0.99	4.933	4.698	4.359	**4.245**
	8	0.90	4.737	4.502	4.163	**4.049**
		0.95	4.786	4.551	4.213	**4.098**
		0.99	5.053	4.818	4.479	**4.365**
	12	0.90	5.351	5.116	4.777	**4.663**
		0.95	5.401	5.165	4.827	**4.712**
		0.99	5.667	5.432	5.093	**4.979**

The estimated inverse Gaussian regression coefficients and the estimated theoretical MSE values for the MLE, and the used estimators are listed in Table 6. According to Table 8, it is clearly seen that the Ridge, KL-GLM, and JKL-GLM estimators have MSE values less than the MSE of the MLE, in general. Moreover, it is clearly seen that the JKL-GLM shrinkages the value of the estimated coefficients efficiently. Additionally, in terms of the MSE, there is an important reduction in favor of the JKL-GLM in comparison with KL-GLM". Specifically, it can be seen that the MSE of the JKL-GLM estimator was about 63.71%, 44.90%, and 17.77% lower than that of MLE, Ridge, and JKL-GLM estimators, respectively. These findings come in agreement with the results of simulation.

Table 4: Description of the used explanatory variable

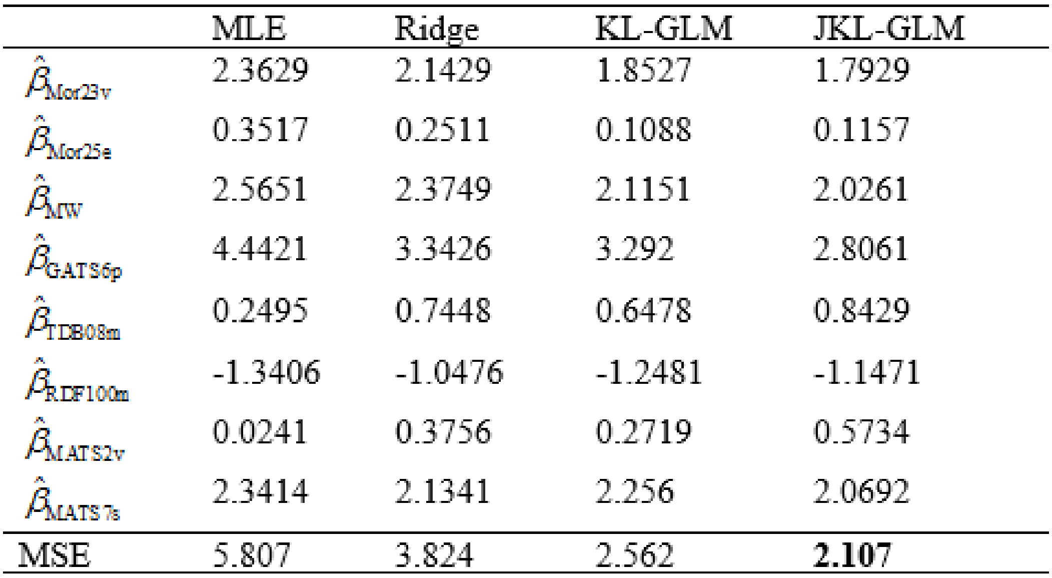
Variable names	Description
Mor23v	Signal 23/ weighted by van der Waals volume
Mor25e	Signal 25/weighted by Sanderson electronegativity
MW	molecular weight
GATS6p	Geary autocorrelation of lag 6 weighted by polarizability
TDB 08m	3D Topological distance based descriptors -lag 8 weighted by mass
RDF100m	Radial Distribution Function-100/ weighted by mass
MATS2v	Moran autocorrelation of lag 2 weighted by van der Waals volume
MATS7s	Moran autocorrelation of lag 7 weighted by l-state

Table 5: The correlations among the 8 variables



Variable nar	Mor23v	Mor25e	MW	GATS6p	TDB 08m	RDF100m	MATS2v	MATS7s
Mor23v	1	0.941	0.621	0.873	0.915	0.708	0.902	0.914
Mor25e		1	0.558	0.763	0.909	0.711	0.913	0.889
MW			1	0.234	0.511	0.605	0.438	0.128
GATS6p				1	0.863	0.833	0.817	0.768
TDB 08m					1	0.918	0.639	0.557
RDF100m						1	0.972	0.964
MATS2v							1	0.984
MATS7s								1

Table 6: The estimated coefficients and MSE values for the four used estimators.



	MLE	Ridge	KL-GLM	JKL-GLM
$\hat{\beta}_{Mor23v}$	2.3629	2.1429	1.8527	1.7929
$\hat{\beta}_{Mor25e}$	0.3517	0.2511	0.1088	0.1157
$\hat{\beta}_{MW}$	2.5651	2.3749	2.1151	2.0261
$\hat{\beta}_{GATS6p}$	4.4421	3.3426	3.292	2.8061
$\hat{\beta}_{TDB08m}$	0.2495	0.7448	0.6478	0.8429
$\hat{\beta}_{RDF100m}$	-1.3406	-1.0476	-1.2481	-1.1471
$\hat{\beta}_{MATS2v}$	0.0241	0.3756	0.2719	0.5734
$\hat{\beta}_{MATS7s}$	2.3414	2.1341	2.256	2.0692
MSE	5.807	3.824	2.562	**2.107**

7. Conclusions

We have presented a new proposed estimator of K-L estimator for GLM in the presence of collinearity. The proposed estimator combines Jackknife procedure with K-L estimator to reduce the biasedness. Our experimental results with both simulated and real application, which is related to the inverse Gaussian regression model, demonstrated that the proposed estimator could successfully deal with collinearity. Moreover, compared with MLE, Ridge, and KL-GLM, the proposed estimator can efficiently reduce the MSE.

References

[1] E. Akdeniz Duran and F. Akdeniz, *Efficiency of the modified jackknifed Liu-type estimator*, Statist. Papers 53(2) (2012) 265–280.

[2] M.N. Akram, M. Amin and M. Amanullah, *Two-parameter estimator for the inverse Gaussian regression model*, Commun. Statist. Simul. Comput. 2020 (2020) 1–19.

[3] A. Alkhateeb and Z. Algamal, *Jackknifed liu-type estimator in poisson regression model*, J. Iran. Statist. Soc. 19(1) (2020) 21–37.

[4] Z.Y. Algamal, *Developing a ridge estimator for the gamma regression model*, J. Chemometrics 32(10) (2018).

[5] Z.Y. Algamal, *A new method for choosing the biasing parameter in ridge estimator for generalized linear model*, Chemometrics Intell. Lab. Syst. 183 (2018) 96–101.

[6] Z.Y. Algamal, *Performance of ridge estimator in inverse Gaussian regression model*, Commun. Statist. Theory Methods 48(15) (2018) 3836–3849.

[7] Z.Y. Algamal, *Shrinkage estimators for gamma regression model*, Electron. J. Appl. Statist. Anal. 11(1) (2018) 253–268.

[8] Z.Y. Algamal, *Shrinkage parameter selection via modified cross-validation approach for ridge regression model*, Commun. Statist. Simul. Comput. 49(7) (2018) 1922–1930.

[9] Z.Y. Algamal and M.R. Abonazel, *Developing a Liu-type estimator in beta regression model*, Concur. Comput. Pract. Exper. 2021 (2021).

[10] Z.Y. Algamal, *Performance of ridge estimator in inverse Gaussian regression model*, Commun. Statist. Theory Methods 48(15) (2018) 3836–3849.

[11] Z.Y. Algamal and M.M. Alanaz, *Proposed methods in estimating the ridge regression parameter in Poisson regression model*, Electron. J. Appl. Statist. Anal. 11(2) (2018) 506–15.

[12] Z.Y. Algamal and Y. Asar, *Liu-type estimator for the gamma regression model*, Commun. Statist. Simul. Comput. 49(8) (2018) 2035–2048.

[13] Z. Algamal, *Shrinkage estimators for gamma regression model*, Electron. J. Appl. Statist. Anal. 11(1) (2018) 253–268.

[14] Z. Algamal, *Generalized ridge estimator shrinkage estimation based on particle swarm optimization algorithm*, Iraqi J. Statist. Sci. 17(32) (2020) 37–52.

[15] Y. Al-Taweel, Z. Algamal and N. Sciences, *Some almost unbiased ridge regression estimators for the zero-inflated negative binomial regression model*, Period. Engin. Natural Sci. 8(1) (2020) 248–255.

[16] J. Correa-Basurto, C. Flores-Sandoval, J. Marín-Cruz, A. Rojo-Domínguez, L.M. Espinoza-Fonseca and J.G.J. Trujillo-Ferrara, *Docking and quantum mechanic studies on cholinesterases and their inhibitors*, Eur. J. Medicinal Chem. 42(1) (2007) 10–19.

[17] R. Farebrother, *Further results on the mean square error of ridge regression*, J. Royal Statist. Soc. Ser. B 38(3) (1976) 248–250.

[18] M.H. Gruber, *The efficiency of jack-knifed and usual ridge type estimators: A comparison*, Statist. Probab. Lett. 11(1) (1991) 49–51.

[19] D.V. Hinkley, *Jackknifing in unbalanced situations*, Technometrics 19(3) (1977) 285–292.

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

[21] A.E. Hoerl, R.W. Kannard and K.F. Baldwin, *Ridge regression: Some simulations*, Comm. Statist. Theory Methods 4(2) (1975) 105–123.

[22] B.M.G. Kibria, *Performance of some new ridge regression estimators*, Commun. Statist. Simul. Comput. 32(2) (2003) 419–435.

[23] B.M.G. Kibria and A.F. Lukman, *A new ridge-type estimator for the linear regression model: simulations and applications*, Scientifica 2020 (2020).

[24] G. Kibria, K. Månsson and G. Shukur, *Performance of some logistic ridge regression estimators*, Comput. Econ. 40(4) (2012) 401–414.

[25] M. Khurana, Y.P. Chaubey and S. Chandra, *Jackknifing the ridge regression estimator: A revisit*, Comm. Statist. Theory Methods 43(24) (2014) 5249–5262.

[26] F. Kurtoğlu and M.R. Özkale, *Liu estimation in generalized linear models: application on gamma distributed response variable*, Statist. Papers 57(4) (2016) 911–928.

[27] K. Liu, *A new class of biased estimate in linear regression*, Comm. Statist. Theory Methods 22 (1993) 393–402.

[28] G. Liu and S. Piantadosi, *Ridge estimation in generalized linear models and proportional hazards regressions*, Comm. Statist. Theory Methods. 46(23) (2016) 11466–11479.

[29] A.F. Lukman, Z.Y. Algamal, B.M.G. Kibria and K. Ayinde, *The KL estimator for the inverse Gaussian regression model*, Concurr. Comput. Pract. Exper. 33(13) (2021).

[30] A.F. Lukman, I. Dawoud, B.M.G. Kibria, Z.Y. Algamal and B. Aladeitan, *A new ridge-type estimator for the gamma regression model*, Scientifica 2021 (2021).

[31] M.J. Mackinnon and M.L. Puterman, *Collinearity in generalized linear models*, Commun. Statist. Theory Methods. 18(9) (1989) 3463–3472.

[32] K. Månsson and G. Shukur, *A Poisson ridge regression estimator*, Economic Modell. 28(4) (2011) 1475–1481.

[33] P. McCullagh, J.A. Nelder and P. McCullagh, *Generalized Linear Models*, 2nd ed. New York: Chapman and Hall London, 1989.

[34] H.S. Mohammed and Z.Y. Algamal, *Shrinkage estimators for semiparametric regression model*, J. Phys. Conf. Ser. 1897(1) (2021).

[35] F. Noeel and Z.Y. Algamal, *Almost unbiased ridge estimator in the count data regression models*, Electron. J. Appl. Statist. Anal. 14(1) (2021) 44–57.

[36] H. Nyquist, *Applications of the jackknife procedure in ridge regression*, Comput. Statist. Data Anal. 6(2) (1988) 177–183.

[37] H. Nyquist, *Restricted estimation of generalized linear models*, J. Royal Statist. Soc. Ser. C 40(1) (1991) 133–141.

[38] N.K. Rashad and Z.Y. Algamal, *A new ridge estimator for the poisson regression model*, Iran. J. Sci. Technol. Trans. A: Sci. 43(6) (2019) 2921–2928.

[39] N.K. Rashad, N.M. Hammood and Z.Y. Algamal, *Generalized ridge estimator in negative binomial regression model*, J. Phys. Conf. Ser. 1897(1) (2021).

[40] B. Segerstedt, *On ordinary ridge regression in generalized linear models*, Commun. Statist. Theory Methods. 21(8) (1992) 2227–2246.

[41] M.R. Özkale and E. Arıcan, *A first-order approximated jackknifed ridge estimator in binary logistic regression*, Comput. Statist. 34(2) (2018) 683–712.

[42] R. Shamany, N.N. Alobaidi and Z.Y. Algamal, *A new two-parameter estimator for the inverse Gaussian regression model with application in chemometrics*, Electron. J. Appl. Statist. Anal. 12(2) (2019) 453–464.

[43] B. Singh, Y. Chaubey and T. Dwivedi, *An almost unbiased ridge estimator*, Indian J. Statist. Ser. B. 13 (1986) 342–346.

[44] S. Türkan and G. Özel, *A new modified Jackknifed estimator for the Poisson regression model*, J. Appl. Statist. 43(10) (2015) 1892–1905.

[45] N. Yıldız, *On the performance of the Jackknifed Liu-type estimator in linear regression model*, Commun. Statist. Theory Methods 47(9) (2018) 2278–2290.