Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Performance of Liu-Type Estimator in Inverse Gaussian Regression Model

Muhammed Jumaa Azzawi Al-Sahan
Department of Mathematics, College of Education for pure Sciences, Tikrit University, Tikrit, Iraq.
Email: muhammed.g.azzawi@st.tu.edu.iq

Zakariya Yahya Algamal *
Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
E-mail: zakariya.algamal@uomosul.edu.iq
ORCID: 0000-0002-0229-7958

Qasim Nasir Husain
Department of Mathematics, College of Education for pure Sciences, Tikrit University, Tikrit, Iraq.
Email: qasim11@tu.edu.iq

***Corresponding Author**
Zakariya Yahya Algamal
E-mail: zakariya.algamal@uomosul.edu.iq
Telephone number: +964 7701640834

Article Info
Page Number: 5009 - 5022
Publication Issue:
Vol 71 No. 4 (2022)

Abstract
The ridge regression model has been shown to be an effective shrinking strategy for reducing the impacts of multicollinearity on a number of occasions. When the response variable is positively skewed, the inverse Gaussian regression model (IGR) is a popular model to use. Multicollinearity, on the other hand, is known to reduce the variance of the maximum likelihood estimator of inverse Gaussian regression coefficients. A novel estimator is proposed in this paper by presenting a generalization of the Liu-type estimator using inverse Gaussian regression (NIGLTE). The performance of NIGLTE is fully depending on the shrinkage parameter, k. In this paper, three selection methods of the shrinkage parameter are explored and investigated. In addition, their predictive

Article History
***Article Received**: 25 March 2022*
***Revised**: 30 April 2022*
***Accepted**: 15 June 2022*
***Publication**: 19 August 2022*

performances are considered. Our Monte Carlo simulation and real application results suggest that some estimators can bring significant improvement relative to others, in terms of mean squared error.

Keywords: Multicollinearity; ridge estimator; inverse Gaussian regression model; Liu-type estimator; Monte Carlo simulation.

1. Introduction

Inverse Gaussian regressionis widely applied model for studying several real data problems, such as automobile insurance claims, healthcare economics, and medical science[1-3]. "Specifically, inverse Gaussian regression model is used when the response variable under the study is not distributed as normal distribution or the response variable is positively skewed. Consequently, the inverse Gaussian regression assumes that the response variable has a inverse Gaussian distribution [4, 5].

In dealing with the inverse Gaussian regression model, it is assumed that there is no correlation among the regressors. In practice, however, this assumption often not holds, which leads to the problem of multicollinearity. In the presence of multicollinearity, when estimating the regression coefficients for inverse Gaussian regression model using the maximum likelihood (ML) method, the estimated coefficients are usually become unstable with a high variance, and therefore low statistical significance[6, 7]. Numerous remedial methods have been proposed to overcome the problem of multicollinearity[8-28]. The ridge regression method [29]has been consistently demonstrated to be an attractive and alternative to the ML estimation method".

In classical linear regression models the following relationship is usually adopted

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations of the response variable, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is an $n \times p$ known design matrix of explanatory variables, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is a $p \times 1$ vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors with mean 0 and variance σ^2.

Ridge regression is a shrinkage method that shrinks all regression coefficients toward zero to reduce the large variance [6, 30]. "This done by adding a positive amount to the diagonal of $\mathbf{X}^T\mathbf{X}$. As a result, the ridge estimator is biased, but it guaranties a smaller mean squared error than the ML estimator.

In linear regression, the ridge estimator is defined as

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^T\mathbf{X} + k\,\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2}$$

where $\mathbf{I}$ is the identity matrix with dimension $p \times p$ and $k \geq 0$ represents the ridge parameter (shrinkage parameter). The ridge parameter, k, controls the shrinkage of $\boldsymbol{\beta}$ toward zero. For larger value of k, the $\hat{\boldsymbol{\beta}}_{Ridge}$ estimator yields greater shrinkage approaching zero[29].

2. Statistical methodology

2.1. Inverse Gaussian ridge regression model

Positively skewed data often arise in epidemiology, social, and economic studies. This type of data consists of nonnegative values. Inverse Gaussian distribution is a well-known distribution that fits to such type of data. Inverse Gaussian regression model (IGR) is used to model the relationship between the positively skewed response variable and potentially regressors [31].

The inverse Gaussian distribution is a continuous distribution with two positive parameters: location parameter, μ, and scale parameter, τ, denoted as $IG(\mu,\tau)$. Its probability density function is defined as

$$f(y,\mu,\tau) = \frac{1}{\sqrt{2\pi y^3 \tau}} \exp\left[-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sqrt{\tau}}\right)^2\right], \quad y > 0. \tag{3}$$

The mean and variance of this distribution are, respectively, $E(y) = \mu$ and $\operatorname{var}(y) = \tau\mu^3$.

Inverse Gaussian regression model is considered a member of the generalized linear models (GLM) family, extending the ideas of linear regression to the situation where the response variable is following the inverse Gaussian distribution. Following the GLM methodology, Eq. (1) can re-write in terms of exponential family function as

$$f(y,\mu,\tau) = \frac{1}{\tau}\left\{-\frac{y}{2\mu^2} + \frac{1}{\mu}\right\} + \left\{-\frac{1}{2}\ln(2\pi y^3) - \frac{1}{2}\ln(\tau)\right\}, \tag{4}$$

where $C(y,\tau) = -(1/2)\ln(2\pi y^3) - (1/2)\ln(\tau)$ and $\frac{y\theta - a(\theta)}{\phi} = \frac{1}{\tau}\left\{-\frac{y}{2\mu^2} + \frac{1}{\mu}\right\}$. Here, τ represents the dispersion parameter and $1/\mu^2$ represents the canonical link function.

In GLM, a monotonic and differentiable link function connects the mean of the response variable with the linear predictor $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$ and $\boldsymbol{\beta}$ is a $(p+1)\times 1$ vector of unknown regression coefficients. Because η_i depends on $\boldsymbol{\beta}$ and the mean of the response variable is a function of η_i, then $E(y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$. Related to the IGR, the $\mu = 1/\sqrt{\mathbf{x}_i^T \boldsymbol{\beta}}$. Another possible link function for the IGR is log link function, $\mu = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$.

The model estimation of the IGR is based on the maximum likelihood method (ML). The log likelihood function of the IGR under the canonical link function is defined as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{\frac{1}{\tau}\left[\frac{y_i \mathbf{x}_i^T \boldsymbol{\beta}}{2} - \sqrt{\mathbf{x}_i^T \boldsymbol{\beta}}\right] - \frac{1}{2\tau y_i} - \frac{\ln\tau}{2} - \ln(2\pi y_i^3)\right\}. \quad (5)$$

The ML estimator is then obtained by computing the first derivative of the Eq. (3) and setting it equal to zero, as

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{1}{2\tau}\left[y_i - \frac{1}{\sqrt{\mathbf{x}_i^T \boldsymbol{\beta}}}\right]\mathbf{x}_i = 0. \quad (6)$$

Unfortunately, the first derivative cannot be solved analytically because Eq. (4) is nonlinear in $\boldsymbol{\beta}$. The iteratively weighted least squares (IWLS) algorithm or Fisher-scoring algorithm can be used to obtain the ML estimators of the IGR parameters. In each iteration, the parameters are updated by

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + I^{-1}(\boldsymbol{\beta}^{(r)})S(\boldsymbol{\beta}^{(r)}), \quad (7)$$

where $S(\boldsymbol{\beta}^{(r)})$ and $I^{-1}(\boldsymbol{\beta}^{(r)})$ are $S(\boldsymbol{\beta}) = \partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $I^{-1}(\boldsymbol{\beta}) = \left(-E\left(\partial^2\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\right)\right)^{-1}$ evaluated at $\boldsymbol{\beta}^{(r)}$, respectively. The final step of the estimated coefficients is defined as

$$\hat{\boldsymbol{\beta}}_{IGR} = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{\mathbf{u}}, \quad (8)$$

where $\hat{\mathbf{W}} = \text{diag}(\hat{\mu}_i^3)$, $\hat{\mathbf{u}}$ is a vector where ith element equals to $\hat{u}_i = (1/\hat{\mu}_i^2) + ((y_i - \hat{\mu}_i)/\hat{\mu}_i^3)$, and $\hat{\mu} = 1/\sqrt{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}$. The covariance matrix of $\hat{\boldsymbol{\beta}}_{IGR}$ equals

$$\text{cov}(\hat{\boldsymbol{\beta}}_{IGR}) = \left[-E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) \right]^{-1} = \tau (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}, \tag{9}$$

and the mean squared error (MSE) equals

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}_{IGR}) &= E(\hat{\boldsymbol{\beta}}_{IGR} - \hat{\boldsymbol{\beta}})^T (\hat{\boldsymbol{\beta}}_{IGR} - \hat{\boldsymbol{\beta}}) \\ &= \tau\, tr[(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}] \\ &= \tau \sum_{j=1}^{p} \frac{1}{\lambda_j}, \end{aligned} \tag{10}$$

where λ_j is the eigenvalue of the $\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$ matrix and the dispersion parameter, τ, is estimated by 31

$$\hat{\tau} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3}. \tag{11}$$

In the presence of multicollinearity, the matrix $\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$ becomes ill-conditioned leading to high variance and instability of the ML estimator of the inverse Gaussian regression parameters. As a remedy, the inverse Gaussian ridge regression estimator (IGRR) can be defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{IGRR}} &= (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \mathbf{X} \hat{\boldsymbol{\beta}}_{\text{IGR}} \\ &= (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{\mathbf{u}}, \end{aligned} \tag{12}$$

where $k \geq 0$. The ML estimator can be considered as a special estimator from Eq. (11) with $k = 0$ ".

2.2.The proposed estimator

K. Liu (1993) proposed an estimator, which is called Liu estimator, combining the Stein estimator with the ridge estimator. "Comparing with ridge estimator, the Liu estimator is a linear function of the shrinkage parameter, therefore it is easy to choose the shrinkage parameter than to choose ridge parameter.

Consequently, the Liu estimator in GRM, the inverse Gaussian Liu estimator (IGLE), is defined as

$$\hat{\boldsymbol{\beta}}_{IGLE} = (\mathbf{X}^T \hat{\mathbf{W}}\mathbf{X} + \mathbf{I})^{-1}(\mathbf{X}^T \hat{\mathbf{W}}\mathbf{X} + d\,\mathbf{I})\hat{\boldsymbol{\beta}}_{IGR} \quad (13)$$

where d is a shrinkage parameter, $0 < d < 1$. For $d = 1$, $\hat{\boldsymbol{\beta}}_{IGLE} = \hat{\boldsymbol{\beta}}_{IGR}$ and for $0 < d < 1$, $\left\|\hat{\boldsymbol{\beta}}_{IGLE}\right\| < \left\|\hat{\boldsymbol{\beta}}_{IGR}\right\|$.

Liu estimator is upgraded by proposing Liu-type estimator to overcome the problem of sever multicollinearity, Liu-type estimator is defined as follows [32]

$$\hat{\boldsymbol{\beta}}_{IGLTE} = (\mathbf{X}^T \hat{\mathbf{W}}\mathbf{X} + k\,\mathbf{I})^{-1}(\mathbf{X}^T \hat{\mathbf{W}}\mathbf{X} - d\,\mathbf{I})\hat{\boldsymbol{\beta}}_{IGR} \quad (14)$$

where $-\infty < d < \infty$ and $k \geq 0$. Liu-type estimator has superior over ridge estimator [33]. The MSE of $\hat{\boldsymbol{\beta}}_{IGLTE}$ is

$$MSE(\hat{\boldsymbol{\beta}}_{IGLTE}) = \nu^{-1}\sum_{j=1}^{J}\frac{(\lambda_j - d)^2}{\lambda_j(\lambda_j + k)^2} + (d + k)^2\sum_{j=1}^{J}\frac{\alpha_j^2}{(\lambda_j + k)^2} \quad (15)$$

In the context of the linear regression model,Kurnaz and Akay [34], proposed the a new generalized Liu-type estimator to alleviate the problem of multicollinearity in linear regression model.The theoretical of the generalized Liu-type estimator have been studied by Kurnaz and Akay [34] and Ertan and Akay [35].

According to Kurnaz and Akay [34] and Ertan and Akay [35], our proposed new Liu-type estimator for inverse Gaussian regression model (NIGLTE) is defined as:

$$\hat{\boldsymbol{\beta}}_{NIGLTE} = (\mathbf{X}^T \hat{\mathbf{W}}\mathbf{X} + k\,\mathbf{I})^{-1}(\mathbf{X}^T \hat{\mathbf{W}}\mathbf{X} + f(k)\mathbf{I})\hat{\boldsymbol{\beta}}_{IGR} \quad (16)$$

where $f(k)$ is a continuous function of the k. Usually, $f(k)$ is selected as a linear function of the biasing parameter such as $f(k) = \mathrm{a}k + b$, where $\mathrm{a,b} \in R$. As a result, the NGLTE becomes a general estimator which includes the other biased estimators[34, 35].

The MSE of $\hat{\boldsymbol{\beta}}_{NIGLTE}$ is

$$MSE(\hat{\boldsymbol{\beta}}_{NIGLTE}) = \nu^{-1}\sum_{j=1}^{p+1}\frac{(\lambda_j + f(k))^2}{\lambda_j(\lambda_j + k)^2} + \sum_{j=1}^{p+1}\frac{(f(k) - k)^2\alpha_j^2}{(\lambda_j + k)^2} \quad (17)$$

2.3.Selecting methods of k

Since the performances of the biased estimators depend on the choice of biasing parameters, it is a significant problem to find reasonable estimates of the biasing parameters. Three well-known methods are used. They defined as:

$$k_1 = \nu^{-1} \frac{p+1}{\hat{\alpha}'_{GR}\hat{\alpha}_{GR}} \tag{18}$$

$$k_2 = \max_j \left(\frac{1}{m_j}\right) \tag{19}$$

$$k_3 = \text{median}\left(\frac{1}{q_j}\right) \tag{20}$$

where $m_j = \sqrt{\frac{\nu^{-1}}{\hat{\alpha}^2}}$ and $q_j = \frac{\lambda_{\max}}{(n-p-1)\nu^{-1} + \lambda_{\max}\hat{\alpha}_j^2}$, .

3. Simulation study

In this section, a Monte Carlo simulation experiment is used to examine the performance of our proposed estimatorunder different degrees of multicollinearity.

3.1. Simulation design

The response variable of n observations from inverse Gaussian regression model is generated as $y_i \sim IG(\theta_i, \nu)$, where $\nu \in \{0.50, 1.5\}$ and $\theta_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ with $\sum_{j=1}^{p} \beta_j^2 = 1$ and $\beta_1 = \beta_2 = ... = \beta_p$ [36].The explanatory variables $\mathbf{x}_i^T = (x_{i1}, x_{i2}, ..., x_{in})$ have been generated from thefollowing formula

$$x_{ij} = (1-\rho^2)^{1/2} w_{ij} + \rho w_{ip}, \; i = 1, 2, ..., n, \quad j = 1, 2, ..., p, \tag{21}$$

where ρ represents the correlation between the explanatory variables and w_{ij} 's are independent standard normal pseudo-random numbers. Because the sample size has direct impact on the prediction accuracy, three representative values of the sample size are considered: 50, 100, and 150. In addition, the number of the explanatory variables is considered as $p = 4$ and $p = 8$ because increasing the number of explanatory variables can lead to increase the MSE. Further, because we are interested in the effect of multicollinearity, in which the degrees of correlation considered more important, three values of the pairwise correlation are considered with $\rho = \{0.90, 0.95, 0.99\}$.For a combination of these different values of n, ν, p, and ρ the generated data is repeated 1000 times and the average absolute bias and average MSE are determined as

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{22}$$

3.2. Simulation results

The averaged MSE all the combination of n, ν, p, and ρ, are respectively summarized in Tables 1 and 2.The best value of the averaged bias and MSE is highlighted in bold. As Table 1 shows, the proposed method, k3, gives lowbias comparing with IGR, k1, and k2. On other hand, k3 performances better than IGR. It is noted from Table 2 that k3 ranks first with respect to MSE. In the second rank, k2 estimator performs better than both IGR and k1 estimators. Additionally, IGR estimator has the worst performance among k1, k2, and k3which is significantly impacted by the multicollinearity.

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

Table 1: Averaged MSE values for the four estimators when $\nu = 0.5$

n	p	ρ	IGR	k1	k2	k3
50	4	0.90	6.031	4.737	4.398	**4.284**
		0.95	6.075	4.787	4.448	**4.334**
		0.99	6.341	5.053	4.714	**4.6**
	8	0.90	6.145	4.857	4.518	**4.404**
		0.95	6.195	4.907	4.568	**4.454**
		0.99	6.461	5.173	4.834	**4.72**

100	4	0.90	5.783	4.495	4.156	**4.042**
		0.95	5.833	4.545	4.206	**4.092**
		0.99	6.099	4.811	4.472	**4.358**
	8	0.90	5.909	4.615	4.276	**4.162**
		0.95	5.953	4.665	4.326	**4.212**
		0.99	6.219	4.931	4.592	**4.478**
150	4	0.90	5.732	4.444	4.105	**3.991**
		0.95	5.782	4.494	4.155	**4.042**
		0.99	6.048	4.76	4.421	**4.307**
	8	0.90	5.852	4.564	4.225	**4.112**
		0.95	5.902	4.614	4.275	**4.161**
		0.99	6.168	4.88	4.541	**4.427**

Table 2: Averaged MSE values for the four estimators when $\nu = 1.5$

n	p	ρ	IGR	k1	k2	k3
50	4	0.90	5.922	4.634	4.295	**4.181**
		0.95	5.971	4.683	4.344	**4.23**
		0.99	6.238	4.95	4.611	**4.497**
	8	0.90	6.042	4.754	4.415	**4.301**
		0.95	6.091	4.803	4.464	**4.349**
		0.99	6.358	5.07	4.731	**4.617**
100	4	0.90	5.68	4.392	4.053	**3.939**
		0.95	5.73	4.441	4.102	**3.988**
		0.99	5.996	4.708	4.369	**4.255**
	8	0.90	5.8	4.512	4.173	**4.059**
		0.95	5.85	4.562	4.223	**4.109**
		0.99	6.116	4.828	4.489	**4.375**
150	4	0.90	5.629	4.341	4.002	**3.888**
		0.95	5.678	4.39	4.052	**3.937**
		0.99	5.945	4.657	4.318	**4.204**
	8	0.90	5.749	4.461	4.122	**4.008**

0.95	5.798	4.51	4.172	**4.057**
0.99	6.065	4.777	4.438	**4.324**

4. Real Data Application

To demonstrate the usefulness of the NIGLTE in real application, we present here a chemistry dataset with $(n, p) = (65,15)$, where n represents the number of imidazo[4,5-b]pyridine derivatives, which are used as anticancer compounds. While p denotes the number of molecular descriptors, which are treated as explanatory variables [37]. The response of interest is the biological activities (IC_{50}). Quantitative structure-activity relationship (QSAR) study has become a great deal of importance in chemometrics. The principle of QSAR is to model several biological activities over a collection of chemical compounds in terms of their structural properties [38]. Consequently, using of regression model is one of the most important tools for constructing the QSAR model. A description of the used explanatory variables is provided in Table 3. All the variables are numerical.

First, to check whether the response variable belongs to the inverse Gaussian distribution, a Chi-square test is used. The result of the test equals to 5.2762 with p-value equals to 0.2601. It is indicated form this result that the inverse Gaussian distribution fits very well to this response variable. That is, the following model is set

$$\hat{y}_{IC_{50}} = \exp(\sum_{j=1}^{15} \mathbf{x}_j \hat{\beta}_j). \tag{23}$$

Second, to check whether there is a relationship among the explanatory variables or not, Figure 1 displays the correlation matrix among the 15 explanatory variables. It is obviously seen that there are correlations greater than 0.90 among MW, SpMaxA_D, and ATS8v ($r = 0.96$), between SpMax3_Bh(s) and ATS8v ($r = 0.92$), and between Mor21v with Mor21e ($r = 0.93$).

Third, to test the existence of multicollinearity after fitting the inverse Gaussian regression model using log link function and the estimated dispersion parameter is 0.00103, the eigenvalues of the matrix $\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$ are obtained as 1.884×10^9, 3.445×10^6, 2.163×10^5, 2.388×10^4, 1.290×10^3, 9.120×10^2, 4.431×10^2, 1.839×10^2, 1.056×10^2, 5525, 3231, 2631, 1654, 1008, and 1.115. The determined condition number $\text{CN} = \sqrt{\lambda_{\max} / \lambda_{\min}}$ of the data is 40383.035 indicating that the severe multicollinearity issue is exist.

The estimated inverse Gaussian regression coefficients and MSE values for the IGR, K1, K2, and K3 estimators are listed in Table 4. According to Table 4, it is clearly seen that the NGLTE shrinkages the value of the estimated coefficients efficiently. Additionally, in terms of the MSE, there is an important reduction in favor of the NGLTE. Specifically, it can be seen that the MSE of the NGLTE estimator was about 44.52%, 36.61%, and 22.11% lower than that of IGR, K1, and K2estimators, respectively".

Table 3: Description of the used explanatory variables

Variable name's	description
MW	molecular weight
IC3	Information Content index (neighborhood symmetry of 3-order)
SpMaxA_D	normalized leading eigenvalue from topological distance matrix
ATS8v	Broto-Moreau autocorrelation of lag 8 (log function) weighted by van der Waals volume
MATS7v	Moran autocorrelation of lag 7 weighted by van der Waals volume
MATS2s	Moran autocorrelation of lag 2 weighted by I-state
GATS4p	Geary autocorrelation of lag 4 weighted by polarizability
SpMax8_Bh(p)	largest eigenvalue n. 8 of Burden matrix weighted by polarizability
SpMax3_Bh(s)	largest eigenvalue n. 3 of Burden matrix weighted by I-state
P_VSA_e_3	P_VSA-like on Sanderson electronegativity, bin 3
TDB08m	3D Topological distance based descriptors - lag 8 weighted by mass
RDF100m	Radial Distribution Function - 100 / weighted by mass
Mor21v	signal 21 / weighted by van der Waals volume
Mor21e	signal 21 / weighted by Sanderson electronegativity
HATS6v	leverage-weighted autocorrelation of lag 6 / weighted by van der Waals volume

Table 4: The estimated coefficients and MSE values for the four used estimators.

	Estimators			
	GR	k1	k2	k3

$\hat{\beta}_{\text{MW}}$	1.002	0.3397	0.3157	0.2727
$\hat{\beta}_{\text{IC3}}$	1.237	0.0244	0.0004	-0.0426
$\hat{\beta}_{\text{SpMaxA\_D}}$	-1.102	0.6021	0.5781	0.5351
$\hat{\beta}_{\text{ATS8v}}$	-1.379	0.0244	0.0004	-0.0426
$\hat{\beta}_{\text{MATS7v}}$	-1.219	-0.8652	-0.8892	-0.9322
$\hat{\beta}_{\text{MATS2s}}$	-1.215	0.08084	0.05684	0.01384
$\hat{\beta}_{\text{GATS4p}}$	-1.237	-1.2411	-1.2651	-1.3081
$\hat{\beta}_{\text{SpMax8\_Bh(p)}}$	2.506	0.11684	0.09284	0.04984
$\hat{\beta}_{\text{SpMax3\_Bh(s)}}$	2.069	-1.0827	-1.1067	-1.1497
$\hat{\beta}_{\text{P\_VSA\_e\_3}}$	2.001	0.0931	0.0691	0.0261
$\hat{\beta}_{\text{TDB08m}}$	-2.103	-1.0968	-1.1208	-1.1638
$\hat{\beta}_{\text{RDF100m}}$	1.571	0.0593	0.0353	-0.0077
$\hat{\beta}_{\text{Mor21v}}$	-2.434	-1.1259	-1.1499	-1.1929
$\hat{\beta}_{\text{Mor21e}}$	-2.352	0.04284	0.01884	-0.02416
$\hat{\beta}_{\text{HATS6v}}$	2.211	0.8611	0.8371	0.7941
MSE	3.2951	1.9366	1.709	**1.655**

5. Conclusions

Numerous selection methods of the k parameter are explored and investigated of Liu-type inverse Gaussian regression model. In addition, their predictive performances are considered. According to Monte Carlo simulation studies, it has been seen that some estimator can bring significant improvement relative to others, in terms of MSE. The K_3 improved the performance of the inverse Gaussian Liu-type regression compared to GM estimator in all the cases without any domination but with superiority of K_3 in terms of MSE. In contrast, k_1 estimator showed poor results comparing with others in all cases. Moreover, in real data application, compared to GR estimators, the developed estimator, k1, k2, and k3 can efficiently reduce the MSE.

REFERENCES

1. P. De Jong and G. Z. Heller, *Generalized linear models for insurance data*. (Cambridge University Press Cambridge, 2008).
2. E. Dunder, S. Gumustekin and M. A. Cengiz, Journal of Applied Statistics, 1-9 (2016).
3. A. S. Malehi, F. Pourmotahari and K. A. Angali, Health Economics Review **5**, 1-11 (2015).
4. A. M. Al-Abood and D. H. Young, Communications in Statistics - Theory and Methods **15** (6), 1865-1874 (1986).
5. W. M. Hattab, Journal of Statistical Computation and Simulation **86** (17), 3512-3526 (2016).
6. Y. Asar and A. Genç, Communications in Statistics - Simulation and Computation **45** (3), 1094-1103 (2015).
7. F. Kurtoğlu and M. R. Özkale, Statistical Papers **57** (4), 911-928 (2016).
8. F. Noeel and Z. Y. Algamal, Electronic Journal of Applied Statistical Analysis **14** (1), 44-57 (2021).
9. Z. Y. Algamal and M. R. Abonazel, Concurrency and Computation: Practice and Experience (2021).
10. Z. Y. Algamal, Electronic Journal of Applied Statistical Analysis **5** (2), 178-186 (2012).
11. A. Alkhateeb and Z. Algamal, Journal of the Iranian Statistical Society **19** (1), 21-37 (2020).
12. A. A. Hamad, Z. Y. J. I. J. o. N. A. Algamal and Applications, **12** (Special Issue), 2093-2104 (2021).
13. A. F. Lukman, Z. Y. Algamal, B. M. G. Kibria and K. Ayinde, Concurrency and Computation: Practice and Experience **33** (13) (2021).
14. Z. Y. Algamal, Chemometrics and Intelligent Laboratory Systems **183**, 96-101 (2018).
15. N. N. Alobaidi, R. E. Shamany and Z. Y. J. T. S. Algamal, **19** (1), 116-125 (2021).
16. N. K. Rashad and Z. Y. Algamal, Iranian Journal of Science and Technology, Transactions A: Science **43** (6), 2921-2928 (2019).
17. A. F. Lukman, I. Dawoud, B. M. G. Kibria, Z. Y. Algamal and B. Aladeitan, Scientifica (Cairo) **2021**, 5545356 (2021).
18. R. Shamany, N. N. Alobaidi and Z. Y. J. E. J. o. A. S. A. Algamal, **12** (2), 453-464 (2019).
19. Z. Y. Algamal, Journal of Basrah Researches **37** (4), 201-211 (2011).

20. Z. Yahya Algamal, Communications in Statistics - Theory and Methods **48** (15), 3836-3849 (2018).
21. Z. Y. Algamal and M. M. J. E. J. o. A. S. A. Alanaz, **11** (2), 506-515 (2018).
22. S. W. Mahmood and Z. Y. J. T. S. Algamal, **19** (2), 308-316 (2021).
23. Z. Y. J. E. J. o. A. S. A. Algamal, **11** (1), 253-268 (2018).
24. H. S. Mohammed and Z. Y. Algamal, Journal of Physics: Conference Series **1897** (1) (2021).
25. Z. Y. Algamal, Communications in Statistics - Simulation and Computation **49** (7), 1922-1930 (2018).
26. Y. Al-Taweel, Z. J. P. o. E. Algamal and N. Sciences, **8** (1), 248-255 (2020).
27. N. A. Al-Thanoon, O. S. Qasim and Z. Y. Algamal, Journal of Physics: Conference Series **1591** (1) (2020).
28. A. N. Alkhateeb and Z. Y. J. E. J. o. A. S. A. Algamal, **14** (1), 266-276 (2021).
29. A. E. Hoerl and R. W. Kennard, Technometrics **12** (1), 55-67 (1970).
30. F. S. M. Batah, T. V. Ramanathan and S. D. Gore, Surveys in Mathematics and its Applications **3**, 111 – 122 (2008).
31. E. Uusipaikka, *Confidence intervals in generalized regression models*. (Chapman & Hall/CRC Press, NW, 2009).
32. Z. Y. Algamal and Y. Asar, Communications in Statistics - Simulation and Computation **49** (8), 2035-2048 (2020).
33. K. Liu, Communications in Statistics-Theory and Methods **32** (5), 1009-1020 (2003).
34. F. S. Kurnaz and K. U. Akay, Statistical Papers **56** (2), 495-517 (2014).
35. E. Ertan and K. U. Akay, Communications in Statistics - Theory and Methods, 1-25 (2020).
36. B. M. G. Kibria, Communications in Statistics - Simulation and Computation **32** (2), 419-435 (2003).
37. Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih and M. Aziz, Journal of Chemometrics **29** (10), 547-556 (2015).
38. Z. Y. Algamal and M. H. Lee, Journal of Chemometrics **31** (10), e2915 (2017).