International Journal of Sustainable Applied Sciences (IJSAS)
Vol. 4, No. 4, 2026: 261-270

Application of The Inverse Gaussian Hybrid Estimator (Igh) To Address Multicollinearity in The Number of Tuberculosis Cases

Gracia Trifena Sintauli[1], Netti Herawati[2], Misgiyati[3*], Nusyirwan[4]
University of Lampung
Corresponding Author: Misgiyati misgiyati@fmipa.unila.ac.id

A R T I C L E I N F O

Keywords: Inverse Gaussian Regression, IGML, IGH, Multicolinearitas, Tuberculosis, Mean Square Error.

Received : 21, February
Revised : 22, March
Accepted: 28, April

A B S T R A C T

The Inverse Gaussian Regression (IGR) model is one approach within the Generalized Linear Model (GLM) framework for modeling data with a positively skewed distribution. Parameter estimation is typically performed using the Inverse Gaussian Maximum Likelihood (IGML) method. However, under conditions of high multicollinearity, IGML becomes unstable due to increased coefficient variance, which leads to a higher MSE. This study compares IGML with the Inverse Gaussian Hybrid Estimator (IGH) in addressing multicollinearity in Tuberculosis cases across 28 districts/cities in West Java Province from 2022-2024. The analysis results indicate the presence of multicollinearity, characterized by high correlation values and large VIF values. The IGH method produces coefficient shrinkage, making the model more stable and superior to IGML.

DOI: https://doi.org/10.59890/ijsas.v4i4.411
ISSN-E: 3025-5597
https://dmimultitechpublisher.my.id/index.php/ijsas

INTRODUCTION

Inverse Gaussian Regression (IGR) is a model for positively skewed response variables. Inverse Gaussian Regression is an appropriate approach for modeling continuous data that is not normally distributed and exhibits a high degree of skewness. Parameter estimation in Inverse Gaussian regression generally uses the Maximum Likelihood (ML) method, known as Inverse Gaussian Maximsum Likelihood (IGML). However, when there is high multicollinearity among variables, the ML estimator becomes unstable, the coefficient variance increases, and the Mean Squared Error (MSE) value grows.

Several previous studies have addressed this issue. Multicollinearity can lead to instability in parameter estimation due to a design matrix approaching singularity, thereby increasing the variance of the coefficients (Gujarati & Porter, 2009). In the context of Inverse Gaussian Regression, Amin et al. (2020) demonstrated that multicollinearity can increase the MSE of the Maximum Likelihood estimator. To address this issue, shrinkage-based methods such as Ridge Regression, introduced by Hoerl and Kennard (1970), are used to reduce the variance of the estimator. By shrinking the effect of small eigenvalues in the design matrix, IGH effectively stabilizes coefficient estimation and mitigates the inflation of variance caused by multicollinearity.

The Inverse Gaussian Hybrid Estimator (IGH) is an estimator that modifies the matrix using two shrinkage parameters to reduce estimator variance and improve model stability under conditions of high multicollinearity. This approach combines the advantages of maximum likelihood and shrinkage methods, resulting in a more stable estimator with lower variance (Kibria, 2003).

This study examines the number of Tuberculosis cases, which is influenced by demographic and socioeconomic factors. These variables are intercorrelated and have the potential to cause multicollinearity. This study applies Inverse Gaussian Regression (IGR) with IGML estimation and compares it with the Inverse Gaussian Hybrid Estimator (IGH) in modeling the number of Tuberculosis cases to evaluate the stability of the estimator under conditions of high multicollinearity.

LITERATURE REVIEW

Generalized Linear Models (GLMs) are used to assess and measure the relationship between response variables and explanatory variables. According to Jong & Heller (2008), Generalized Linear Models are an extension of linear modeling that allows for the estimation of a model from data where the random variables do not necessarily have a normal distribution, provided that the distribution belongs to the exponential family. Thus, GLM can be used not only for normally distributed response variables but also for response variables with other distributions and non-constant variances (McCullagh & Nelder, 1989). The general GLM model can be expressed as:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

The Inverse Gaussian distribution is a continuous distribution characterized by greater skewness and sharp kurtosis (Jong & Heller, 2008). This distribution belongs to the exponential family, with the following mean and variance:

$$E(y) = \mu \text{ and } Var(Y) = \phi\mu^3$$

Parameter estimation in Inverse Gaussian regression generally uses the Maximum Likelihood method introduced by Fisher (1922). This method determines the parameter values that maximize the likelihood of the observed sample. This estimator is known as the Inverse Gaussian Maximum Likelihood (IGML) and is consistent, asymptotically unbiased, and efficient if the model assumptions are met (Greene, 2018). In matrix form, the IGML estimator can be written as:

$$\hat{\beta}_{IGML} = (\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W z}$$

However, the stability of the estimator is highly dependent on the nature of the matrix $`X^T WX `$. If this matrix approaches singularity due to multicollinearity, the inverse becomes very large and the variance of the estimator increases significantly (Gujarati & Porter, 2009). The covariance of the estimator in GLM is given by:

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$

Multicollinearity is a condition in which two or more predictor variables are strongly correlated, leading to problems in estimating regression coefficients. According to Stock & Watson (2018), multicollinearity can be detected using the Variance Inflation Factor (VIF), with a VIF value greater than 10 indicating the presence of multicollinearity. Mathematically, multicollinearity causes the design matrix to approach singularity, resulting in very large estimator variance (Montgomery et al., 2021). VIF is formulated as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

In the context of Inverse Gaussian Regression, multicollinearity causes an increase in the Mean Square Error (MSE) of the Maximum Likelihood estimator (Amin et al., 2020), thus requiring an alternative shrinkage-based sssssapproach such as ridge regression. Ridge regression was introduced by Hoerl and Kennard (1970) as an estimation method to address multicollinearity by adding a penalty term to the diagonal of the matrix

$$\hat{\beta}_{ridge} = (\boldsymbol{X^T X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X^T Y}$$

The Inverse Gaussian Hybrid Estimator (IGH) is an extension of the Maximum Likelihood estimator for the Inverse Gaussian regression model, combined with a shrinkage approach similar to that used in ridge regression

(Hoerl & Kennard, 1970). The hybrid approach introduces a shrinkage parameter and a bias adjustment parameter to control the effects of multicollinearity. The IGH estimator is defined as:

$$\hat{\beta}_{IGH} = (\boldsymbol{X^T W X} + \boldsymbol{kI})^{-1}(\boldsymbol{X^T W X} - \boldsymbol{kI})(\boldsymbol{X^T W X} + \boldsymbol{k(1+d)I})^{-1}\boldsymbol{X^T W X}\hat{\beta}_{IGML}$$

The addition of the shrinkage parameter aims to improve the numerical conditions of the matrix so that the variance of the estimator can be reduced, while the bias adjustment parameter is used to control the level of bias so that it is not too large (Kibria, 2003). IGH produces an estimator that is biased but has smaller variance compared to IGML under conditions of high multicollinearity. Thus, IGH often yields a smaller Mean Squared Error (MSE) compared to the pure Maximum Likelihood estimator.

METHODOLOGY

The data used in this study are secondary data obtained from official publications of the West Java Provincial Central Statistics Agency (https://jabar.bps.go.id/id). The data cover 28 regencies/cities in West Java Province for the years 2022–2024 and include the following variables: number of Tuberculosis cases (Y), population density (X_1), total population (X_2), area (X_3), number of health facilities (X_4), percentage of the poor population (X_5), and population growth rate (X_6). All these variables were selected because they have epidemiological relevance in describing the sociodemographic factors that influence the spread of Tuberculosis in a region.

This research method analyzes and applies the Inverse Gaussian Maximum Likelihood Estimator (IGML) and the Inverse Gaussian Hybrid Estimator (IGH) to address the issue of multicollinearity in modeling the number of Tuberculosis cases in West Java Province from 2022 to 2024. The research process began with a descriptive analysis to characterize the data using histograms and boxplots. Next, a distribution test was conducted using the Kolmogorov–Smirnov test to ensure the data's conformity with the Inverse Gaussian distribution. Parameter estimation was then performed using the IGML method by calculating the regression parameters and dispersion parameters. Following this, multicollinearity was tested using the correlation matrix and the Variance Inflation Factor (VIF).

To address multicollinearity, the IGH method was applied by determining the shrinkage parameter k and the adjustment parameter d based on the parameter values and the matrix's eigenvalues. Model parameter estimates were then calculated using the IGH approach to obtain more stable coefficients. Next, a performance evaluation and comparison of the IGML and IGH models was conducted using the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values. The final stage of the study involved testing the model parameters, which included simultaneous testing using the Likelihood Ratio test and partial testing using the Wald test, to determine the significance of the independent variables' effects on the number of Tuberculosis cases.

RESULTS AND DISCUSSION

This study began with a descriptive analysis of the data to provide an overview of the characteristics of the variables used in modeling the number of Tuberculosis cases. This analysis aimed to identify patterns in data distribution and variations across regions as a basis for determining an appropriate modeling approach. Descriptive analysis can be seen in the table

Table 1. Descriptive Analysis

Variable	Min	Max	Mean	Median
Y	274	224,798	12,443.45	4,668
X_1	382	15,176	3,779.27	1,399
X_2	206	50,345	3,562.95	1,878
X_3	39.44	37,044.86	2,645.83	1,267.24
X_4	14	58,513	2,804.92	1,193
X_5	2.34	12.77	8.26	8.43
X_6	0.41	1.86	1.19	1.18

Based on Table 1, the variable for the number of Tuberculosis cases (Y) has a minimum value of 274 and a maximum value of 224,798, indicating a very high disparity in the number of cases across regions. The mean value, which is significantly larger than the median, indicates that the data distribution is asymmetric and skewed to the right. This suggests that most regions have relatively low to moderate case counts, but there are a few regions with very high case counts, which pull the mean value higher.

The variables population density (X_1), population size (X_2), and number of health facilities (X_4) also exhibit a fairly wide range. This indicates significant differences in regional characteristics, particularly between densely populated and sparsely populated areas. Meanwhile, the variables percentage of the poor population (X_5) and population growth rate (X_6) exhibit smaller variations, suggesting that socioeconomic factors are relatively more homogeneous compared to demographic factors. Given the data characteristics indicating an asymmetric distribution, the next step is to test the data distribution's suitability to determine the appropriate regression model.

Table 2. Kolmogorov–Smirnov Test

Kolmogorov-Smirnov	
D	0.15237
P-value	0.13618

Table 2, provides a p-value of 0.13618 was obtained, which is greater than the significance level of 0.05; therefore, the null hypothesis is not rejected. This indicates that the data on the number of Tuberculosis cases follow an inverse Gaussian distribution. Interpretatively, this result confirms that the right-skewed distribution pattern of the data aligns with the characteristics of the inverse Gaussian distribution, making the IGR model suitable for use in the analysis. Once it is established that the data follow an Inverse Gaussian

distribution, the next step is to estimate the parameters using the Inverse Gaussian Maximum Likelihood (IGML) method.

Table 3. IGML Model Estimation

Variable	IGML Coefficient
Intercept	0.098348
X_1	0.005775
X_2	0.009969
X_3	0.111595
X_4	0.208327
X_5	-0.003687
X_6	0.004480

Table 3, presents the coefficients estimated by the IGML method are relatively large, particularly for certain variables such as area and health facilities. This indicates that small changes in the independent variables can result in significant changes in the response variable. However, the magnitude of these coefficient values also indicates instability in parameter estimation. This condition arises due to strong relationships among the independent variables, making the IGML estimator sensitive to changes in the data and resulting in high variance. However, to ensure the stability of the resulting model, it is necessary to test for multicollinearity among the independent variables.

Table 4. Multicollinearity Test

Variable	VIF
X_1	2.41
X_2	97.76
X_3	96.38
X_4	2.97
X_5	1.90
X_6	1.09

Based on table 4, the VIF values for the population (X_2) and land area (X_3) variables are very high, at 97.76 and 96.38, respectively. These values far exceed the general threshold (VIF > 10), indicating the presence of very serious multicollinearity. This causes the estimates to become unstable and less reliable when using the IGML method. Based on the test results indicating severe multicollinearity, an alternative estimation method capable of addressing this issue is required, namely the Inverse Gaussian Hybrid Estimator (IGH).

Table 5. IGH Parameter Test

Parameter	Value
k	80.685.727.208
d	8.22×10^{-9}

Based on Table 5, the very large value of the shrinkage parameter k indicates that a strong penalty is needed to address multicollinearity. Meanwhile, the very small value of the parameter d indicates that the contribution of bias correction is relatively small and that the model suffers from severe multicollinearity, requiring significant shrinkage correction to stabilize the parameter estimates. After the shrinkage parameter is determined, the next step is to estimate the parameters using the IGH method.

Table 6. IGH model estimation

Variable	IGH Coefficients
Intercept	0.000396
X_1	0.000048
X_2	-0.000141
X_3	-0.000153
X_4	-0.000121
X_5	0.000019
X_6	0.000108

According to Table 6, all coefficients showed a very significant reduction compared to the IGML results. The coefficient values became very small and more controlled, indicating that the IGH method successfully reduced the variance of the estimators by mitigating the effects of multicollinearity, making them more stable and no longer overly sensitive to changes in the data. Next, a performance comparison was conducted between the IGML and IGH models to determine which method yields the best results.

Table 7. Model Comparison

Model	MSE	RMSE
IGML	1 267 407 893	35 600.67
IGH	1 254 997 456	35 425.94

Table 7 demonstrates that the model comparison results, it was found that the Inverse Gaussian Hybrid Estimator (IGH) method performs better than the Inverse Gaussian Maximum Likelihood (IGML) method. This is demonstrated by the Mean Squared Error (MSE) value of 1,267,407,893 for the IGML model, whereas the MSE value for the IGH model is 1,254,997,456. It is evident that the IGH model has a smaller MSE value compared to the IGML. Additionally, based on the Root Mean Squared Error (RMSE) values, the IGML RMSE is 35,600.67, while the IGH RMSE is 35,425.94. The smaller RMSE value in the IGH further reinforces that this model provides more accurate predictions.

Thus, the Inverse Gaussian Hybrid (IGH) model outperforms the IGML model in modeling the number of Tuberculosis cases. This indicates that the IGH method is capable of addressing the issue of multicollinearity by producing more stable estimates. Therefore, the best model used in this study is the Inverse Gaussian Regression model with IGH estimation. The Inverse Gaussian regression model with the log link function obtained is as follows:

$$\log(\mu) = 0{,}000396 + 0{,}000048X_1 - 0{,}000141X_2 - 0{,}000153X_3 - 0{,}000121X_4 + 0{,}000019X_5 + 0{,}000108X_6$$

The model was used to describe the relationship between Population Density (X_1), Population Size (X_2), Area (X_3), Number of Health Facilities (X_4), Percentage of Poor Population (X_5), and Population Growth Rate (X_6) with the number of tuberculosis cases using a log-link function, so that the model can capture the right-skewed data pattern. The results indicate that these variables influence the number of tuberculosis cases, assuming all other variables remain constant; and after confirming that the model is simultaneously significant, partial tests were conducted to evaluate the effect of each independent variable.

Table 8. Simultaneous Parameter Test

Test Statistics	Value
Likelihood Ratio Value (G)	0.015574
Degrees of Freedom (df)	6
p-value	0.0000002

The very small p-value indicates that the independent variables collectively have a significant effect on the number of Tuberculosis cases and that the model built has a good ability to explain the variation in the total number of Tuberculosis cases. After determining that the model is significant simultaneously, a partial test was conducted to determine the effect of each independent variable.

Table 9. Partial Parameter Test

Variable	*p-value*	Decision
Intercept	< 2e-16	Significant
X_1	1.27e-08	Significant
X_2	4.72e-07	Significant
X_3	0.908	Not significant
X_4	1.26e-11	Significant
X_5	0.144	Not significant
X_6	0.333	Not significant

The results of the partial analysis indicate that the variables of population density, total population, and health facilities have a significant effect on the number of Tuberculosis cases, and that demographic factors and the availability of health facilities are the primary factors influencing the number of Tuberculosis cases. Meanwhile, the other variables were not significant, indicating that their influence on the number of Tuberculosis cases was not statistically significant.

CONCLUSIONS AND RECOMMENDATIONS

The analysis results indicate that the Inverse Gaussian Hybrid Estimator (IGH) method is the best estimator for addressing multicollinearity issues in the Inverse Gaussian Regression model. This is demonstrated by the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, which are smaller than those of the IGML method. Thus, the IGH method is capable of producing more stable and accurate estimates in modeling the number of Tuberculosis cases.

FURTHER STUDY

This study is subject to several limitations, highlighting the need for future studies to further refine the IGH estimator's application in modeling tuberculosis cases or other healthcare-related data with high multicollinearity.

REFERENCES

Amin, M., Lukman, A. F., Ayinde, K., & Ogundimu, E. O. (2020). On the performance of some biased estimators in inverse Gaussian regression model. Journal of Statistical Computation and Simulation, 90(12), 2141–2160.

Badan Pusat Statistik Provinsi Jawa Barat. (2024). Statistik kesehatan Jawa Barat 2024.

Fisher, R. A. (1992). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons.

Greene, W. H. (2018). Basic econometrics (5th ed.). McGraw-Hill International Edition.

Gujarati, D. N., & Porter, D. C. (2009). Basic econometrics (5th ed.). McGraw-Hill International Edition.

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

Jong, P. de, & Heller, G. Z. (2008). Generalized linear models for insurance data. Cambridge University Press.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. Communications in Statistics – Simulation and Computation, 32(2), 419–435.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2nd ed.). Chapman & Hall.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis (6th ed.). John Wiley & Sons.

Schrödinger, E. (1915). Zur Theorie der Fall- und Steigversuche an Teilchen mit

Brownscher Bewegung. Physikalische Zeitschrift, 16, 289–295.

World Health Organization. (2024). Global tuberculosis report 2024.