Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Variables Selection in Inverse Gaussian Regression Model using Modified Heuristic Search Algorithm

Sura Mohamed Jamalalden Hussein*

College of Computer Science and Mathematics ,Department of Statistics and Informatics ,University of Mosul ,Mosul ,Iraq.

Article information

Article history:
Received September 26 ,2024
Revised: December 25,2024
Accepted :April 24,2025
Available online: June 1,2026

Keywords:
Variable Selection, Invasive Weed Algorithm ,Simulation, Inverse Gaussian Regression Model

Correspondence:
Sura Mohamed Jamalalden Hussein*
sura.alalwlia@uomosul.edu.iq

Abstract

The inverse Gaussian regression model is one of the most widely recognized models, frequently used across various applications. It is part of the generalized linear model families and serves as a foundational model. Like other regression models, it may include numerous independent variables, which can negatively impact both the model's accuracy and the simplicity of interpreting its results. This study aims to apply the modified invasive weed optimization algorithm and compare it with other methods for variable selection in the inverse Gaussian regression model, using both simulations and real-world data. The Monte Carlo simulation approach was employed, setting sample sizes n to four different values—30, 50, 100, and 150—to facilitate comparisons across sample sizes (small, medium, large). Results indicated that the proposed method reduces the mean square error of the model and outperforms previously used techniques. The AIC method emerged as the least effective in variable selection, as it yielded the highest prediction error (PE) and tended to select irrelevant explanatory variables.

DOI: 10.33899/iqjoss.v23i1.62118 ,

1.Introduction

One of the most important foundations of scientific research is the study of problems or phenomena in various fields such as economics, sociology, and medicine . The main goal of studying it is to determine the main equation that represents that phenomenon accurately, by collecting data related to it from various available sources. Then, this data is analyzed using statistical techniques and mathematical analysis to determine the relationships between the different variables and design statistical models that describe these relationships, which constitutes the basic entry for understanding more deeply and identifying its main features. This process is referred to statistics as modeling phenomena[Ross,2020].

Of all the generalized linear regression models

1- the inverse Gaussian regression model is one of the most popular, being widely used in many applications. The inverse Gaussian model is listed in the tables of the generalized linear model families as of a basic model, rarely, it discussed adequately despite its importance. Even (McCullagh, P.) and (J. Nelder) [McCullagh et al ,1989] give only a passing comment on its existence[peter et al,2018].

2- Widely the inverse Gaussian regression method is used in many fields of industrial engineering, life testing, reliability, marketing, and social sciences and this method is most useful in cases where the responding variable is skewed in the positive direction [yonis et al,2022].

3- Most of the data in real-world application contain problems such as the problem of the large number of independent variables studied, which is a well-known problem among statistical researchers, and negatively affects the estimation process. This problem can lead to ignoring some important explanatory variables in some cases.

Traditional methods for selecting subsets such as forward selection, backward elimination, and stepwise selection have become less efficient in their function and more expensive to compute. In addition, information criteria for selecting variables such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) have become impractical in selecting explanatory variables due to their computational complexity, which grows exponentially with the increase in the number of explanatory variables [Alkhateeb et al,2021].

By highlighting a number of factors that may affect the quality of these methods and the necessity of using them under certain conditions rather than other methods. this research aims to employ the modified Raven algorithm and compare it with other methods for selecting explanatory variables in the inverse Gaussian regression model using simulation and real data.

2. Inverse Gaussian Regression Model (IGRM)

The inverse Gaussian distribution, which has two positive parameters, the location parameter (μ) and the dispersion parameter(τ), is used as a continuous distribution when the response variable follows y_i a skewed pattern positively. This distribution is denoted by the symbol $IG\sim(\mu,\tau)$, and the probability density function for this distribution is defined as follows:

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

Equation (1) can be converted to the exponential family form because the inverse Gaussian regression model (IGRM) belongs to the family of generalized linear models (GLM), by rewriting it as follows [Akram et al,2020 ,Yaha,2019]:

$$f(y,\mu,\tau)=\left\{\frac{y\,\theta-b(\theta)}{\phi}+C(y,\phi)\right\} \qquad (2)$$

Where:

θ: The canonical parameter or link function

b(θ) : The cumulate function

ϕ : The dispersion parameter

C(y, ϕ): The normalization term: It is a normalization function that ensures that equation (2) is a probability function. So $c(y,\phi)$، is a function with significance $\phi$and $y$guarantees that $\int f_y(y;\theta,\phi)dy=1$ whether the variable is y continuous or $\sum_y f_y(y;\theta,\phi)=1$ if y is discrete[Olsson,2002 ,peter et al ,2018].

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

Equation (3) can be written in a simpler form as follows:

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

Therefore:

$$x_i^T\beta=\frac{1}{\mu},\sqrt{\eta}=\frac{1}{\mu},\ \mu=\frac{1}{\sqrt{\eta}}$$

Through compare equation (4) with equation (1), it obtaining:

$$\theta = \frac{1}{2\mu^2},\ \ b(\theta) = \frac{1}{\mu} = \frac{1}{\sqrt{-2\theta}} = -\sqrt{-2\theta}, \phi = \frac{1}{\tau}$$

The link function can be used to obtain the mean and variance of equation (4) as follows:

$$E[y] = b'(\theta) = \frac{\partial b}{\partial \theta} = \frac{\partial b}{\partial \mu}\frac{\partial \mu}{\partial \theta}\ , b = \frac{1}{u}, \frac{\partial b}{\partial \mu} = \frac{-1}{\mu^2}, \theta = \frac{1}{2\mu^2}\ , \frac{\partial \theta}{\partial u} = \frac{-2\mu^{-3}}{2}\ \ ,\ \frac{-1}{-\mu^3} = -\mu^3 \tag{5}$$

$$v[y] = \emptyset\, b''(\theta) = \frac{\sigma 2}{(-2\theta)^{3/2}} = \sigma 2\, \mu^3 \tag{6}$$

Where $\phi = \frac{1}{\tau}$

The logarithmic likelihood function of the inverse Gaussian regression model can be defined using the maximum likelihood method to estimate its parameters. This function takes the following form [Yahya,2019]:

$$\ell(\beta) = \sum_{i=1}^{n}\left(\tau\left(\frac{-y_i}{2\mu_i^2} + \frac{1}{\mu_i}\right) + (-\frac{1}{2})(ln\left(\frac{2\pi y_i^3}{\tau}\right) + \frac{\tau}{y_i})\right) \tag{7}$$

$$\ell(\beta) = \sum_{i=1}^{n}\left(\tau\left(\frac{-y_i x_i^T \beta}{2} - \sqrt{x_i^T \beta}\right) + (-\frac{1}{2})(ln\left(\frac{2\pi y_i^3}{\tau}\right) + \frac{\tau}{y_i})\right) \tag{8}$$

The first partial derivative of the parameters β of equation (8) is calculated and set equal to zero. By this procedure, the maximum likelihood value of (IGRM) is obtained according to the above formula.

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n}\frac{1}{2\phi}\left[y_i - \frac{1}{\sqrt{x_i^T \beta}}\right]x_i^T = 0 \tag{9}$$

The first derivative cannot be computed due to the non-linearity of equation (9) with respect to the parameter β. To overcome this problem, numerical techniques as mentioned in the study by[Salh et al ,2021], such as the Newton Raphson iterative method[Mawlood,2021] or the iterative regular least squares (IRLS) algorithm, can be used to obtain the parameters of the inverse Gaussian regression (IGRM), where the parameters are updated at each iteration using the following formula:

$$\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + \left(X^T \widehat{W}^r X\right)^{-1} X^T(y - \hat{\mu}^r) \tag{10}$$

The maximum likelihood estimate (MLE)

$$\hat{\beta}_{IGRM} = D^{-1} X^T \widehat{W} \hat{z} \tag{11}$$

Where $D = \left(X^T \widehat{W} X\right)$,

$\widehat{W} = \text{diag}(\hat{\mu}_i^3)$

$\hat{Z}$ represents the response rate variable and calculated its value as follows:

$$\hat{z}_i = \left(\frac{1}{\hat{\mu}_i^2}\right) + \left(\frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i^3}\right)$$

$$\hat{\mu} = \frac{1}{\sqrt{x_i^T \hat{\beta}}}$$

3. Invasive Weed Algorithm

Invasive Weed Optimization Algorithm (IWO) is numerical random improvement algorithm biologically inspired by weeds, which was first proposed by Mehrabian and Lucas in (2006). Simply, this algorithm simulates the natural behavior of weeds in colonizing and finding a suitable place to grow and reproduce. To simulate the colonial behavior of weeds, some basic properties of this process must be taken into account:

1. A limited number of seeds are spread over the search area (population initialization).
2. All seeds grow into flowering plants and produce seeds based on the fitness function (reproduction).
3. The produced seeds are randomly spread over the research area to grow into new plants (spatial dispersal).
4. This process continues until the maximum plants is reached.

Only plants with the highest fitness function can survive and produce seeds, but the others are eliminated (competitive elimination). The process continues until the maximum iterations is reached, in hope that the plant with the best fitness function will be the closest to the optimal solution.

The IWO algorithm includes a number of basic steps. These steps are interconnected with each other and this algorithm cannot be applied to any problem unless all of these steps are applied and otherwise, the IWO algorithm will lose its value and benefit in finding and improving the solution. The steps of the algorithm can be explained as follows:

The first step: Initialize A Population

An initial population of solutions is generated and propagated over d dimensions of the problem space with random locations and the value of the fitness function for this population is calculated.

The second step: Reproduction

A plant in a plant community is allowed to produce seeds (reproduce) depending on the value of its fitness function as well as the maximum and minimum limits of the fitness function in the colony. The number of seeds produced by the plant increases linearly from the minimum possible limit for seed production to the maximum possible limit. In other words, the plant produces seeds depending on the value of its fitness function, the minimum fitness function of the colony, and the maximum fitness function of the colony to ensure that the increase is linear.

Based on this technique, to complete and achieve the research aim each element (crow) in the set will have d locations that represent the number of explanatory variables in the inverse Gaussian regression model. Accordingly, the employment of the crow algorithm is according to the following steps:

The first step: Determine the size of the group

(number of herbs) which is 20, since each herb will have a vector of the number of independent variables, in addition to determining the number of iterations within the algorithm, as the results stabilized at 300 iterations.

The second step: the initial values

that the algorithm will represents the assumptive initial values needs generated from a continuous uniform distribution according to the interval [0,1].

The third step: For choosing the optimal values

the Fitness Function was relied upon according to the following formula:

$$Fitness\ Function = \min\left[\frac{\sum_{i=1}^{n}\left(y_i - \widehat{m}(x)\right)^2}{n}\right] \quad (12)$$

The fourth step: Depending on the lowest value obtained by any herb according to equation (12), the locations of the remaining herbs are updated.

The fifth step: continuing the solving until reach the maximum iteration of the algorithm, which was determined in the first step and which will represent the optimal solution.

X_1	X_2	…………..	x_{p-1}	X_P
1	0	…………..	1	0

Figure 1: Variables selection mechanism according to the weed algorithm

3- Criteria for Evaluating Methods for Selecting Variables

3-1 Prediction accuracy evaluation criterion

First: Prediction Error

the difference between the actual value and the expected value in the Prediction model and it measures how accurately the Prediction method predicts the outcome.

$$PE = (y - \hat{y})^T (y - \hat{y}) \quad (13)$$

Based on this criterion, the best method is determined, which gives the lowest value compared to other methods.

Second: Criteria for evaluating the accuracy of variable selection

Since the proposed methods generally work on variables selection, it is important to evaluate and measure the ability and quality of these methods in how to select important variables. Therefore, two criteria were relied upon in our study for this purpose, as follows:

1- Evaluation criteria "C"

Symbolized by (C) to evaluation criterion which is defined as the number of true coefficients with zero values that were correctly estimated as having zero values.

2- Evaluation riteria "I"

symbolized by (I) to evaluation criterion which is defined as the number of true coefficients with non-zero values that were incorrectly estimated as having zero values. The quality of penalty methods in terms of the criteria for evaluating the accuracy of variable selection depends on who gives maximum value for (C) and the lowest value for (I).

5. Simulation Results

The Monte Carlo method was used in the simulation, to designed An experiment and simulated by using the programming language (R), where the variable (y_i) was generated in the inverse Gaussian regression model, then the values of the sample size (n) were set, where four sample sizes were used, which are (30, 50, 100, 150), in order to study the comparison according to the samples of different types (small, medium, large). The comparison will be made with both the Bayesian criterion and the Akaky criterion.

First: The data of the variable y were generated following an inverse Gaussian distribution model and for a value of the dispersion parameter equal to $\tau = 0.5,1.5,3$ and as follows:

$$y_i \sim \text{inverse Gaussain}(\mu_i, \tau)$$
$$\mu_i = exp(x_i^T \beta)$$

Second: An explanatory variables matrix X were generated with dimensions ($n \times p$) that follows the multivariate normal distribution:

$$X \sim \text{MN}(\mu, \text{M})$$

M is a common variance matrix, where $M_{ij} = r^{|i-j|}$, when $(i, j = 1,2, \dots, p)$ if the explanatory variables are correlated.

Third: The experiment was repeated (100) times in order to reduce bias in Monte Carlo experiments.

Fourth: The Poisson regression model data were generated according to the values of the regression parameter vector β whose dimensions are ($1 \times p$). The values of the regression parameter vector β were as follows:

$(1.5,2,0.8,-3.5,5,0,\dots,0)^{\text{T}}$, where the number of non-zero parameters is $q = 5$, and the zero parameters are equal to $p - q$. Pearson regression is used to answer questions such as what factors can predict the frequency of an event.

The results of the simulation experiment will be analyzed and interpreted according to the criteria of prediction accuracy and the criteria of variables selection accuracy. (1), (2), (3) and (4) tables show the values of the criteria of (PE, C, I) for the BIC and AIC methods and the proposed IWO method, and it can be concluded the following:

1- It is clear that the (IWO) method gave the minimum (PE) values, when the value of the dispersion parameter changes. The amount of improvement in prediction based on the (PE) criterion reached 35.04% and 29.66% at (n=50) $\tau = 1.5$ compared to (AIC) and (BIC) respectively, regardless of the sample size value.

2- The (IWO) method gave the best results,. when the sample size changes and regardless of the value of the dispersion parameter.

3- Depending on the criteria for variables selection, the (IWO) method had the maximum values of (C), which is the number of real coefficients with zero values that were correctly estimated as having zero values, and gave the minimum values of (I), which is known as the number of real coefficients with non-zero values that were incorrectly estimated as having zero values.

4- The AIC method appeared as the worst method in variables selection because it gives the maximum values of (PE) and also as the worst method in variables selection because it tends to unimportant explanatory variables selection.

Table (1): Average of evaluating selection methods criteria when n=30

τ	Method	PE	C	I
0.5	AIC	23.206	1	0
	BIC	21.658	3	0
	IWO	16.425	5	0
1.5	AIC	21.582	3	0
	BIC	20.034	4	0
	IWO	14.801	5	0
3	AIC	20.815	4	1
	BIC	19.267	4	0
	IWO	14.034	5	0

Table (2): Average of evaluating selection methods criteria when n=50

τ	Method	PE	C	I
0.5	AIC	22.168	1	0
	BIC	20.62	2	0
	IWO	15.387	5	0
1.5	AIC	20.544	3	0
	BIC	18.996	4	0
	IWO	13.763	5	0
3	AIC	19.777	3	1
	BIC	18.229	4	0
	IWO	12.996	5	0

Table (3): Average of evaluating selection methods criteria when n=100

τ	Method	PE	C	I
0.5	AIC	20.39	1	0
	BIC	18.842	**3**	0
	IWO	13.609	**5**	0
1.5	AIC	18.766	3	0
	BIC	17.218	4	0
	IWO	11.985	5	0
3	AIC	17.999	**3**	1
	BIC	16.451	4	0
	IWO	11.218	5	0

Table (4): Average of evaluating selection methods criteria when n=150

τ	Method	PE	C	I
0.5	AIC	19.352	1	0
	BIC	17.804	3	0
	IWO	12.571	5	0
1.5	AIC	17.728	4	0
	BIC	16.18	4	0
	IWO	10.947	5	0
3	AIC	16.961	3	1
	BIC	15.413	4	0
	IWO	10.18	5	0

5- The Practical Side

In this side, a comparison is made between the performance of the proposed estimator IGDK and other estimators by using actual data. The performance of the estimators is evaluated using the MSE criterion. To verify the performance of the proposed IGDK method using actual data, specific chemical data were used $(n, \text{p}) = (65,15)$, n representing the number of imidazole[4,5-b]pyridine derivatives which are anticancer compounds, while the symbol p represents the explanatory variables which symbolize the molecular properties [Yahya,2019]. This section discusses the role of the IC50 variable as a response variable in evaluating the biological activities of anticancer compounds, and highlights the importance of studying the quantitative relationship between chemical structures and biological activity using QSAR modeling. QSAR is defined as a model of biological activities based on the structural properties of a group of chemical compounds [Algamal et al,2017].

The chi-square test of fit was used to determine the appropriateness of the inverse Gaussian distribution for the response variable (IC50), where the results showed a value equal 5.2762 and p-value equal 0.2601. Based on these results, the inverse Gaussian distribution can be considered appropriate for the dependent response variable [Yahya,2019]. molecular descriptors were included as explanatory variables (independent) regarding the study.

The inverse Gaussian regression model was evaluated using variables selection methods, which indicated by calculating the mean square error values as well as the number of independent variables that were selected. The results summarized in table 5 show that the proposed method IWO outperformed the other methods, achieving the minimum value to MSE and the minimum number of independent variables that were selected.

Table 5 :The results of practical side

Method	MSE	# variables
AIC	43.681	11
BIC	41.368	9
IWO	33.174	7

6- Conclusions

Through the simulation and the real data, the results indicate that using the IWO method leads to excellent results when using the MSE and PE criteria, making it reliable for users in predicting outcomes and evaluating statistical models. In addition, the sample size n has a great effect on the PE values, he larger the sample size, the better the accuracy of the model improves and the PE value decreases,, that which means the accuracy increases. On the other hand, noticing a decrease in the PE value when the dispersion parameter value increases. It is worth noting that using the MSE criterion leads to better results in predicting outcomes and evaluating statistical models. Moreover, the proposed method showed its strength by choosing the minimum number of independent variables.

Acknowledgments
The authors are very grateful to the University of Mosul, and College of Computer Science and Mathematics, which helped improve this work's quality.

Conflict of Interest
The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Ethical Approval
Ethical approval was not required for this study as it did not involve human participants, personal data.

References

1. Ross, S.M., Introduction to probability and statistics for engineers and scientists. 2020: Academic press.
2. McCullagh, P. and J. Nelder, Generalized Linear Models. 1989, London: Chapman and Hall.
   DOI: https://doi.org/10.1201/9780203753736
3. Peter K. D., G.K.S., Generalized Linear Models With Examples in R. 2018, New York: Springer.
4. yonis, F.a. and R.A. Othma, Shrinkage estimators in inverse Gaussian regression model: Subject review. IRAQI JOURNAL OF STATISTICAL SCIENCES, 2022. 19(35): p. 72-82. DOI: 10.33899/IQJOSS.2022.174331
5. Alkhateeb, A.N. and Z.Y.J.E.J.o.A.S.A. Algamal, Variable selection in gamma regression model using chaotic firefly algorithm with application in chemometrics. 2021. 14(1): p. 266-276. DOI: 10.1285/i20705948v14n1p266
6. Akram, M.N., et al., A new Liu-type estimator for the inverse Gaussian regression model. 2020. 90(7): p.1153-1172.DOI: 10.1080/00949655.2020.1718150
7. Yahya Algamal, Z., Performance of ridge estimator in inverse Gaussian regression model. Communications in Statistics-Theory ,Methods, 2019. 48(15): p. 3836-3849.DOI: 10.1080/03610926.2018.1481977
8. Olsson, U., Generalized linear models. An applied approach. Studentlitteratur, Lund, 2002. 18.ID: 118193531
9. Salh, A.P.D.S.M., et al., Using Multinomial Logistic Regression model to study factors that affect chest pain. 2021. 17(53 part 2).DOI:10.25130/tjaes.17.53.2.31
10. Mawlood, K.I., Estimating Hazard Function and Survival Analysis of Tuberculosis Patients in Erbil city. Tikrit Journal of Administration ,Economics Sciences, 2021. 17(54 part 3). DOI:https://doi.org/10.25130/tjaes.17.54.3.30
11. Askarzadeh, A., A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Computers & Structures, 2016. 169: p. 1-12.DOI: 10.1016/j.compstruc.2016.03.001
12. Sayed, G.I., A.E. Hassanien, and A.T. Azar, Feature selection via a novel chaotic crow search algorithm. Neural Computing and Applications, 2017.https://doi.org/10.1007/S00521-017-2988-6
13. Islam, M.J., X. Li, and Y. Mei, A time-varying transfer function for balancing the exploration and exploitation ability of a binary PSO. Applied Soft Computing, 2017. 59: p. 182-196.DOI: 10.1016/j.asoc.2017.04.050

14. Mafarja, M., et al., Binary dragonfly optimization for feature selection using time-varying transfer functions. Knowledge-Based Systems, 2018. 161: p. 185-204. DOI: 10.1016/j.knosys.2018.08.003
15. Algamal, Z.Y., & Lee, M. H. , A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression. Journal of Chemometrics, 2017. 31(10).https://doi.org/10.1002/cem.2915

اختيار المتغيرات في نموذج انحدار كاوس المعكوس باستخدام الخوارزمية الحدسية المعدلة

سرى محمد جمال الدين حسين

كلية علوم الحاسبوب والرياضيات، قسم الإحصاء والمعلوماتية، جامعة الموصل، الموصل، العراق

الخلاصة: أن نموذج انحدار كاوس المعكوس هو أحد أشهر هذه النماذج، حيث يتم استخدامه بشكل واسع في العديد من التطبيقات. يتم وضع نموذج كاوس المعكوس في جداول عائلات النماذج الخطية المعممة كونه من النماذج الأساسية. وكغيره من سائر نماذج الانحدار, قد يحتوي النموذج على متغيرات مستقلة كثيرة ما يؤثر سلباً على دقة النموذج وبساطته في تفسير النتائج. تهدف هذه الدراسة إلى استخدام خوارزمية أمثلة الأعشاب الضارة المعدلة ومقارنتها مع طرائق اخرى في إختيار المتغيرات في نموذج انحدار كاوس المعكوس باستخدام المحاكاة والبيانات الحقيقة . حيث تم استخدام اسلوب مونت كارلو (Mont Carlo) في المحاكاة حيث تم تعيين قيم حجم العينات (n) حيث تم استخدام اربع احجام من العينات وهي (30,50,100,150) وذلك لأجل دراسة المقارنة وفق العينات باختلاف أنواعها (صغيرة، متوسطة، كبيرة). حيثُ بيَّنت النتائج أنَّ الاسلوب المقترح يُساهم في تقليل متوسط مربعات الخطأ للنموذج ويحقق أداءً أفضل مُقارنةً بالطرائق الأخرى المستخدمة سابقاً. ظهرت طريقة AIC كاسوأ طريقة في اختيار المتغيرات لأنها تعطي أعلى قيم لـ (PE) وكذلك كاسوأ طريقة في اختيار المتغيرات كونها تميل الى اختيار متغيرات توضيحية غير مهمة.

الكلمات المفتاحية: اختيار المتغيرات، خوارزمية الأعشاب الضارة ، المحاكاة ، انموذج انحدار كاوس المعكوس