IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

Estimation of the Parameters of the Binary Logistic Regression Model Using the Bootstrap Method and the Employment of the Genetic Algorithm for Epilepsy Patients

Asmaa Ghalib Jaber[1], Fahad Hussein Enad[2] and Zainab Nihad Mohammed[3]

[1]University of Baghdad/ College of Administration & Economics
[2]University of Dhi Qar / Department of Studies and Planning
[3]University of Baghdad / Computer Center

ARTICLE INFO

Article history:
Received 15/11/2024
Revised 15/11/2024
Accepted 13/1/2025,
Available online 15/5/2025

Keywords:
logistic regression
Bootstrap
genetic algorithm
mean square error

ABSTRACT

This research aims to conduct a systematic comparison between the efficiency of the bootstrap method and the application of the genetic algorithm in estimating the parameters of the binary logistic regression model, which is considered one of the pivotal statistical models in the analysis of binary medical data. The study problem lies in the need to develop accurate and effective estimation methods when dealing with real medical data that often exhibit variability and heterogeneity among independent variables.The two methods were applied to actual data from epilepsy patients, using a random sample of 142 patients, with the aim of evaluating the performance of each method in estimating the statistical model. The evaluation was based on several statistical indicators, the most prominent of which are: Mean Squared Error (MSE), model accuracy, and Maximum Likelihood Estimation for parameter estimation.
The results revealed a significant superiority of the genetic algorithm method compared to the bootstrap method, as the genetic algorithm recorded the lowest mean squared error (0.646), compared to (2.446) for the bootstrap method, reflecting its efficiency in improving estimation accuracy and reducing variance. The results also showed that the residence variable (X2) had the most statistically significant impact on determining the length of hospital stay for patients, while the other variables did not show significant relevance within the model.Based on the above, the study highlights the effectiveness of genetic algorithms as a smart and promising tool in analyzing medical data and estimating predictive models, especially in environments that require more flexible and accurate alternatives than traditional methods. It also emphasizes the importance of the logistic regression model in supporting medical decisions and improving the quality of healthcare through reliance on advanced statistical analysis.

1. Introduction

The study of economic, social, and medical phenomena is one of the vital topics that many researchers are currently interested in. In this context, the binary logistic regression model is used to study the relationship between the dependent variable (response variable) and the independent variables related to the patients. In this research, the binary logistic regression model is applied to analyze the data of epilepsy patients, where the dependent variable is divided into two categories: a five-day stay and a stay of more than five days. The research aims to compare the effectiveness of two statistical methods, namely the bootstrap method and the genetic algorithm method, in estimating the most suitable model for epilepsy patients, in addition to evaluating the accuracy of these models using the criterion of the precise classification rate of observations. The problem lies in the need to select the most suitable model for estimating the parameters of binary logistic regression that provides

* Corresponding author.E-**mail** address: Drasmaa.ghalib@coadec.uobaghdad.edu.iq
https://doi.org/10.62933/pt8nf008

accurate estimates, as both the bootstrap method and the genetic algorithm method are considered advanced statistical techniques that can contribute to improving the accuracy of estimates and reducing errors. The importance of this research increases given the limited studies that compare these two methods in the context of epilepsy patients. Therefore, this research aims to compare two methods for estimating the parameters of the binary logistic regression model: the bootstrap method and the genetic algorithm method, to determine the most suitable for epilepsy patients. These models are evaluated using the accuracy rate criterion to ensure the precision of the estimates. In turn, this study contributes to enhancing knowledge related to parameter estimation and providing more accurate solutions for analyzing patient data, thereby improving the effectiveness of medical and therapeutic decision-making, especially in cases like epilepsy. In a similar context, many previous studies have addressed similar advanced statistical techniques. For instance, Davidson et al. (1996) compared three methods for estimating the parameters of random models for histopathological image data using genetic algorithms, logistic regression, and the Newton method. The results showed a good match with industrial data, enhancing the effectiveness of these methods in random tissue modeling applications. While Nakamichi et al. (2004) combined logistic regression with genetic algorithms to study the interactions of binary pathological traits with SNPs and environmental factors, demonstrating effective applications in identifying types of diabetes using data from diabetic patients. In the study by Liu and Zhang (2014), a multinomial logistic model was developed to classify the technical efficiency of public projects using Data Envelopment Analysis (DEA) and supported statistical inferences using bootstrap, which showed a significant impact of the environmental factor on performance classification in public projects. Meanwhile, in the study by Stripling et al. (2015), a profit-maximizing logistic regression model was presented to predict customer churn in customer management campaigns in the telecommunications sector, where the results showed a significant improvement in profits compared to the traditional model. On the other hand, the study by Jain et al. (2017) addressed the issue of class imbalance in medical diagnosis and proposed a new sampling method using the genetic algorithm, which showed superiority over traditional methods such as oversampling and SMOTE in improving classification accuracy in imbalanced medical data. In the study by Li et al. (2018), two non-parametric graphical methods for estimating the parameters of the binary logistic regression model were compared, and the results showed the effectiveness of the non-parametric Bayesian bootstrap in estimating the parameters with higher accuracy when the samples were small. In Schmid and Desmarais (2017), two methods for estimating the parameters of the Exponential Random Graph Model (ERGM) in large networks were compared, and the study showed that combining MPLE with parametric bootstrap provides accurate confidence intervals and high computational efficiency. The study by Shaheen and Ja Sm. (2019) also compared the bootstrap method for estimating the binary logistic regression model using criteria such as Mean Square Error (MSE) and Mean Absolute Error (MAE), where the results showed the effectiveness of the bootstrap method in improving the accuracy of estimates compared to the traditional model. Finally, the study by Hao et al. (2023) reviewed the progress in classifying electroencephalogram (EEG) signals using the Support Vector Machine (SVM) algorithm, where the study demonstrated the strength of this algorithm in classifying EEG signals, especially with the development of various optimization methods that support medical applications such as epilepsy.

These studies demonstrate the importance of advanced statistical models such as bootstrap and genetic algorithms in improving the accuracy of model estimates and analyzing medical data, thereby enhancing the effectiveness of medical and therapeutic applications in various fields, including epilepsy.

2. Methods and Techniques

2.1. Logistic Regression [1] ,[5],[6], [15]

A statistical model for categorical data analysis, logistic regression is regarded as a specific instance of general linear regression models. The logistic model or logit model are other names for it. (Model Logit). This model examines the connection between the response variable, which accepts categorical values, and independent (explanatory) factors. Numerous disciplines, including the arts, social sciences, and medicine, as well as the engineering sciences, use logistic regression

2.1.1. Model of binary logistic regression

One of the nonlinear regression models is the logistic regression model, in which there is a nonlinear relationship between the explanatory factors (x_1, x_2, x_3, …, x_k) and the dependent variable (y, the response variable). The basic premise of logistic regression is that the dependent variable (y) is binary, accepting only one of two values (0 or 1). representing either success or failure, where success occurs with a probability of pi and failure occurs with a probability of (1- pi) Therefore, the dependent variable yi follows a Bernoulli distribution [5],[15].
i.e

$$\boldsymbol{Y_i \sim Ber(p_i)} \qquad \boldsymbol{i=1,2,\ldots..n}$$

Additionally, the probability density function looks like this:

$$\boldsymbol{P_r(Y=Y_i)=p_i^{Yi}\,(1-p_i)^{1-Yi}} \quad \textbf{(1)} \quad \boldsymbol{Y_i=0,1}$$

Since: -
Y_i : is a binary response dependent variable
P_i: is the likelihood that the reaction will take place at Yi=1
1-pi: is the likelihood that the reaction won't happen when Yi=0
Consequently, the likelihood that the response will occur (p_i) is represented by the expectation of the variable Y_i.

$$\boldsymbol{E(Y)=P_r(Y=1)=p_i} \qquad \textbf{(2)}$$

Regarding the variable Y_i variance based on the Bernoulli distribution

$$\boldsymbol{V(Y_i)=p_i(1-p_i)} \qquad \textbf{(3)}$$

Let $(\boldsymbol{x_1.x_2.x_3.\ldots\ldots x_k})$ is a set of explanatory variables, representing the observations of these variables which form the following matrix: -

$$\boldsymbol{X=(x_{ij})n*k} \qquad (4)$$

Since: -
X: is the matrix of the independent variables
$\boldsymbol{i=1.2.3.\ldots\ldots n}$ **n** is the number of observations (sample size)
$\boldsymbol{j=1.2.3.\ldots\ldots k}$ k is the number of explanatory variables
if it was$\boldsymbol{y_i}$=$[\boldsymbol{y_1.y_2.y_3\ldots\ldots y_n}]$ The two-response variable $y_i \in \{0.1\}$ is represented by a random sample.**.**
As a result, the logistic regression model may be written as follows:

$$\boldsymbol{y_i=p+\varepsilon i} \qquad \textbf{(5)}$$

and that: -
$\boldsymbol{\mu_i}$ is the logistic regression function (the logistic response function)

$$\boldsymbol{\mu_i=p(y=1)=\frac{e^{xi'\beta}}{1+e^{xi'\beta}}} \qquad \textbf{(6)}$$

β is a parameter vector whose dimensions are (P×1)
$\boldsymbol{x_i=\{x_{i1}.x_{i2}\ldots x_{ip}\}}$ An array vector of independent variables whose dimensions are (1×p)
$\boldsymbol{\varepsilon_i}$ represents the random error.

$$\boldsymbol{\varepsilon_i=y_i-p} \qquad \textbf{(7)}$$

Additionally, the error term's variance is equal to the dependent variable's variance, and its mean is equal to zero, according to my agencies:

$$\boldsymbol{E(\varepsilon_i)=E(y_i)-E(p)} \quad = \quad \boldsymbol{p_i-p_i} \quad \boldsymbol{=0}$$

$$\boldsymbol{V(\varepsilon_i)=V(y_i)=p_i(1-p_i)} \qquad \textbf{(8)}$$

The following mathematical formula converts this model into a linear form that is represented by a linear relationship through the linear predictor (X_i) of the explanatory variables with the logit of the probability [logit $p(X_i)$].:

$$\boldsymbol{Z=logit\ p(\ )=ln\left[\frac{p(x_i)}{1-p(x_i)}\right]=\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}} \qquad \textbf{(9)}$$

to estimate the parameters $\beta_0.\beta_1.\beta_2.\ldots.\beta_p$ We employ the Maximum Likelihood Estimation approach, one of the estimation technique. [7].

For every pair (X_i, Y_i), we have a sample from a set of independent observations of size (n) such that i=(1,2,……,n).So that:

Y_i: represents the rank of the binary response variable for item i.

X_i: represents the value of the independent variable i.

We can express the greatest possibility function in the following form.

$$\boldsymbol{L(B) = \prod_{i=1}^{n} f(X)^{Yi}\,[1 - f(X)]^{1-Yi}} \quad \textbf{(10)}$$

The following equation is obtained by taking the logarithms of both sides:

$$\boldsymbol{L(B) = \ln L(B) = \sum_{i=1}^{n}\{Yi \ln[f(X)] + (1 - Yi)\ \ln[1 - f(X)]\}} \quad \textbf{(11)}$$

When the aforementioned equation is calculated for the parameters to be estimated (Bi) and set to zero, a series of equations are produced that can only be resolved using the iterative weighted least squares approach. [20].

2.2. The Bootstrap Method

The bootstrap method is considered one of the most widely used techniques for determining statistical estimates, and it belongs to the non-parametric category, meaning it does not require assumptions about the normal distribution of the data. This method was first developed by Efron in 1979, and since then, numerous applied research studies have demonstrated its effectiveness in reducing bias and variance in estimates. In 1993, Efron and Tibshirani further developed this method. The bootstrap method operates on fewer assumptions regarding computational processes, especially with the advancement of computing technologies that have contributed to speeding up these processes. The bootstrap method is used due to its proven ability to handle practical data with high accuracy and efficiency, without the need to know its distribution.The basic idea of the bootstrap method is that in the absence of information about the distribution, the available sample contains all the information available about the underlying distribution. Therefore, the process of resampling is considered the best method to predict what can be obtained from sample frequencies. This method provides more accurate estimates of true values on average compared to single-variable methods. However, one of the drawbacks of bootstrap is that it may take longer in calculations, but this problem can be overcome using computers. This method relies on generating unbiased estimates from a set of biased estimates by randomly extracting a large number of samples from the original data with replacement, and of the same size as the original sample[21]. In his quest to solve the problem of small sample size, Efron used this method of resampling to generate a large number of samples. After conducting B repetitions, B estimates are obtained, which are known as parameter estimates. The average of these estimates is considered the bootstrap estimate.And the average of these estimates is called an estimate.(Bootstrap). The sample is repeated at least (1000) times, i.e., (B>1000). To calculate the bootstrap method, the following steps can be followed:

1- Building the original sample: The original sample, which contains the dependent variable and the explanatory variables, is prepared in the following form:

$$z = (y.x_1.x_2 \ldots\ldots.x_k) \quad (12)$$

2-Bootstrap sampling: Random samples (with replacement) are drawn from the original sample, each of a certain size denoted by D each Bootstrap sample expressed in the form $[z^* = (y^*.x_1{}^*.x_2{}^* \ldots\ldots.x_k{}^*)]$. (13)

3-Model estimation for each Bootstrap sample: For each Bootstrap sample, the parameters of the logistic regression model are estimated using an appropriate method

(such as Maximum Likelihood or an advanced numerical algorithm).

4-Repetition of the process: Steps (2) and (3) are repeated many times $B > 1000$, preferably to achieve accurate results.

5-Final estimates calculation: After repeating the estimates across all Bootstrap samples, the arithmetic mean of the parameter estimates across the samples is calculated to obtain stable estimates of the model parameters.

$$\hat{B}_j = \frac{1}{B}\sum_{b=1}^{B} \hat{B}_j^p \quad j = 1,2,\ldots,k \qquad (14)$$

6-Statistical results analysis: Bootstrap results can be used to calculate confidence intervals for parameters, or to assess the stability and statistical accuracy of the model.[12],[13],[4],[21].

2.3. Genetic Algorithm

It is a technique from artificial intelligence methods for finding the best solutions to problems under study in a strong and fast manner. It is considered one of the random search methods that address a problem with the aim of achieving the best possible results. It revolves around the evolution technique, which states that the fittest survive by mimicking nature's work through retaining good traits from the parent generation and passing them on to the offspring generation, with the goal of obtaining descendants that possess the best traits inherited from the parents, at the very least[7].

To find the parameters of the binary logistic regression using the genetic algorithm, we follow the following steps[19],[25].

1-The beginning: The formation of chromosomes through the βp values that constitute the genes of the chromosomes, where (P=0,1,...,p) within the real numbers.

2- Initialization: Creating an initial generation by finding initial gene values with random values for the other constraints.Initialization: Creating an initial generation by finding initial values for the genes with random values for the other set of constraints.

3- In the objective function, the chromosome is evaluated in terms of efficiency until reaching the optimal solution by determining the values of β_p.

4- Conducting the selection process for the chromosome that has a small objective function value by choosing a high probability for it and finding its evaluation function from the following fitness function equation:

fitness function=1/(1+objective function)

Since the objective function: represents the objective function, and through the evaluation function formula, the probability of this function can be found according to the formula[9].

$$\boldsymbol{C_i = \frac{f_{(i)}}{\sum_{i=1}^{N} f_{(i)}}} \qquad \textbf{(15)}$$

Where:

C_i: Represents the probability of individual i :$f_{(i)}$: represents the evaluation function for individual i.

The roulette wheel criterion can be used to generate a random number R(c) confined within the range [0,1].If R(c)<C(1),the first chromosome (the best) will be selected. Alternatively, the selection can be made such that the probability is confined within R(c) < C(p) < C(p-1) or the random number is confined withinC(p-1) < R(c) < C(p) Each time, one chromosome is determined for the new population based on the evaluation function.

5- After completing the testing process, the next step is the hybridization of the chromosomes with good traits through the pairing of each chromosome. One of its standards is organized hybridization based on the hybridization probability (Pc). This value is compared with the gene values of the parent chromosomes to form the new generation (offspring). Exchange occurs when the gene value is greater than or equal to the probability value[17].

6- The last step that chromosomes can undergo is the mutation process, which also depends on a probabilistic measure (Pm) for the parameters by replacing randomly selected genes with a new value that we also obtained

randomly[8]. The following equation can be applied:

Total genes = number of genes in the chromosome * population size

Thus, the parameters of logistic regression can be estimated by employing the genetic algorithm in the maximum likelihood method.

2-4- criterion of comparison between methods

Mean Square Error (MSE) was adopted and calculated according to the following relationship:

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i=1}}^{\mathbf{N}} (\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}})^{\mathbf{2}} \qquad \textbf{(16)}$$

2-5-Analysis of logistic regression's explanatory power

Use one of his statistics, R^2 (Cox & Snel), or R^2 (Nagelkerke).Since the two statistics Identifying the quality of the estimated regression equation in evaluating the relationship between the dependent variable and the explanatory factors is the aim of logistic regression, which shares the same objective as the coefficient of determination R2 used in multiple linear regression. with the exception that In contrast to R^2Nagelkerke, it cannot be adjusted to the proper one. Restrictions can R^2Nagelkerke Because it spans from zero to the integer one, it is more dependable than R^2(Cox&Snel). Its value is typically more than R^2(cox&snel). Using the formula below, we may compute a statistic, R^2 (cox&snel) [27].

$$\mathbf{R^2_{cox\&snel}} = \mathbf{1} - [\frac{\mathbf{L_0}}{\mathbf{L_1}}]^{(\frac{\mathbf{2}}{\mathbf{n}})} \qquad \textbf{(17)}$$

So that

L0: the maximum possibility function in the case of the fixed term model.

L1: the maximum possibility function in the case of the model including all variables.

while counting it($\mathbf{R^2_{Nagelkerke}}$)Calculated through the following relationship.

$$\mathbf{R^2_{Nagelkerke}} = \frac{\mathbf{R^2_{cox\&snel}}}{\mathbf{1-[L_0]^{(\frac{2}{n})}}} \qquad \textbf{(18)}$$

3. The side that is applied

The practical aspect of this research involves estimating the parameters of the binary logistic regression model using two advanced statistical methods: the bootstrap method and the genetic algorithm, with the aim of comparing their accuracy and efficiency in analyzing real medical data. The performance of the two methods was evaluated using the Mean Squared Error (MSE) criterion to measure the accuracy of the estimates. The study was based on real data from epilepsy patients collected from the hospital, due to the significant importance of this topic in improving healthcare and making medical decisions based on accurate statistical foundations.The study sample included 142 epilepsy patients in 2023, who were classified into two groups based on the length of hospital stay: the first group included 76 patients whose stay was five days, while the second group included 65 patients whose stay exceeded five days. Four explanatory variables were used in the model.The statistical analysis was conducted using both MATLAB 2018 to apply the bootstrap method and the genetic algorithm, in addition to SPSS version 26 to perform some ready-made analyses. This experimental design contributed to classifying the sample into two distinct types based on the duration of stay, which enabled a precise study of the relationship between the explanatory variables and the dependent variable.

3.1. Descriptive statistics

Table 1: shows the descriptive statistics

Variation coefficient	**Standard deviation**	**mean**	**Percentage**	**Replication**	Variable
0.420	**0.463**	**1.2**	**54.6**	**75**	***Stay for 5 days***
			44.5	**67**	***Stay for more than 5 days***
0.431	**0.502**	**1.3**	**54.6**	**81**	***male***
			43.3	**61**	***female***
0.292	**0.602**	**1.4**	**63.4**	**88**	***civilian***

			38.7	**54**	***Country***
0.201	**0.553**	**1.5**	**40.3**	**50**	***injured***
			64.3	**92**	***Not injury***

Table (1) shows a statistical analysis of several variables related to epilepsy patients, by studying the sample characteristics based on frequency, percentage, mean, standard deviation, and coefficient of variation. Regarding the length of stay variable, the results indicated that 54.6% of the patients stayed for five days, with a mean of 1.2 and a standard deviation of 0.463, and a coefficient of variation of 0.420, indicating relative homogeneity in this category. As for the patients whose length of stay exceeded five days, they constituted 44.5% of the sample, with no additional data available to measure the variation.
Regarding the gender variable, males constituted 54.6% of the sample, with a mean of 1.3, a standard deviation of 0.502, and a coefficient of variation of 0.431, indicating moderate dispersion in this category. Meanwhile, females accounted for 43.3%, with no supporting data to determine the level of dispersion. As for the place of residence variable, it showed that 63.4% of the patients were civilians, with a mean of 1.4, a standard deviation of 0.602, and the lowest coefficient of variation at 0.292, indicating relatively high homogeneity in this group, compared to rural residents who made up 38.7% without supplementary statistics.As for the health status in terms of infection, the percentage of infected individuals was 40.3%, with a mean of 1.5 and a standard deviation of 0.553, and a low coefficient of variation of 0.201, indicating a high degree of homogeneity. Meanwhile, the non-infected individuals constituted 64.3% of the sample, with no detailed statistical distribution data available for them.Through the analysis of these results, it is observed that the "civilians" and "injured" categories exhibited the greatest degree of homogeneity compared to the other categories, reflecting the importance of these two variables in explaining the behavior of the dependent variable within the logistic regression model, and enhancing the accuracy of future predictions based on these variables.

3.2. Applying the classical logistic regression model

Table 2: shows some indicators of the logistic regression model

Nagelkerke R Square	**Cox & Snell R Square**	**Hosmer and Lemeshow (sig)**	**Chi-square (sig)**
0.0866	**0.072**	**0.692**	**0.034**

The table(2) shows the results of several important statistical indicators to evaluate the quality and suitability of the binary logistic regression model used in the study. The Chi-square test value was significant at (0.034), which is lower than the usual significance level (0.05), indicating that the model as a whole is appropriate and has good explanatory power compared to a model without independent variables.As for the Hosmer and Lemeshow test, it recorded a significance value of (0.692), which is higher than 0.05, indicating that the differences between the expected and observed values are not statistically significant. Therefore, the model has a good goodness of fit with the actual data.
Regarding the determination coefficients, the results showed that the Cox & Snell R Square value was (0.072), while the Nagelkerke R Square value was slightly higher at (0.0866). These values indicate that the model explains approximately 7.2% to 8.7% of the variance in the dependent variable, which is relatively low, reflecting that there may be other variables with greater impact that were not included in the current model.In general, these results indicate that the model used is acceptable in terms of statistical fit, but its interpretive power is limited, necessitating the consideration of introducing additional variables in the future to improve the model's predictive performance.

Table 3: shows the observed and expected values for the duration of stay

Total	*Stay for more than 5 days*		*Stay for 5 days*		steps
	Expected	Observed	Expected	Observed	
12	4.174	2	8.654	10	1
8	1.801	2	4.232	6	2
15	5.215	5	11.659	10	3
16	4.571	5	11.197	11	4
19	11.329	12	9.432	7	5
14	6.720	9	5.652	5	6
21	11.722	10	9.259	11	7
13	7.540	6	5.741	7	8
8	4.428	5	3.503	3	9
14	8.162	8	5.897	6	10

Table (3) shows the results of the Hosmer and Lemeshow test, which is used to assess the goodness of fit of the logistic regression model with the actual data by comparing the observed values with the expected values for the number of patients who stayed in the hospital for five days or more across ten groups (steps). In general, we observe that the observed and expected values are close in most steps, reflecting an acceptable agreement between the model and the data. For example, in the first step, the actual number of patients who stayed for five days was (10), while the expected value was (8.654). In the same step, the actual number of those who stayed for more than five days was (2) compared to an expectation of (4.174), indicating the accuracy of the prediction. As shown in the fifth step, there is a good convergence between the observed and expected values in the category of stays longer than five days (12 observed versus 11.329 expected), reflecting the model's accuracy. Although there are some minor differences in some steps such as (6), (7), and (10), these differences remain within acceptable limits, indicating that the model is capable of reasonably predicting the number of cases. These results show that the model has a good level of goodness of fit, as there are no significant or systematic differences between the expected and observed values. Therefore, these results enhance the model's reliability in predicting the length of hospital stay for epilepsy patients based on the input variables, thereby supporting the model's effectiveness in analyzing medical data related to epilepsy.

Table 4: shows the results of the traditional logistic regression

Upper	Lower	Exp(B)	Sig.	Df	Wald	S.E.	B	Variables
1.801	0.452	0.801	0.710	1	0.132	0.231	-0.145	X_1
0.602	0.213	0.201	0.013	1	6.341	0.202	-1.342	X_2
2.103	0.541	0.891	0.812	1	0.123	0.401	-0.302	X_3
2.107	0.501	0.841	0.701	1	0.010	0.310	-0.102	X_4
		4.322	0.221	1	2.313	0.802	1.301	Constant

Table (4) shows the results of the binary logistic regression analysis for the independent variables. The results showed that variable X2 had a significant effect on the length of hospital stay, with a p-value of Sig. = 0.013, which is less than 0.05, indicating a statistically significant relationship between the stay and the length of hospital stay. The Exp(B) value for X2 was 0.201, indicating that patients who stay longer in the hospital are less likely to stay for shorter periods.As for the other variables, none of them showed a significant effect on the dependent variable. For example, X1 was not statistically significant as Sig. = 0.710, which means there is no significant correlation between gender and length of stay. Similarly, X3 and X4 did not have a significant effect on the dependent variable, as their Sig. values were higher than 0.05, indicating that neither variable has a statistically significant impact on predicting the length of hospital stay. Regarding the constant value, it was also not statistically significant, as Sig. = 0.221, indicating that it has no significant effect on the model.Based on these results, it can be concluded that (X2) is the only variable that significantly contributes to determining the length of hospital stay, while the other variables did not show a statistically significant effect.

3.3. Applying bootstrap logistic regression method

When the bootstrap logistic regression model was applied to the data, the same results were obtained as in the traditional logistic regression model regarding the indicators(Chi-square (sig), Cox & Snell R Square, and Nagelkerke R Square). The difference in results was according to Table (5), which includes the regression coefficients.

Table 5: the results of bootstrap logistic regression method

Upper	Lower	Sig. (2-tailed)	Std. Error	Bias	B	Variables
0.683	-0.762	0.764	0.42	0.02	-0.14	X_1
-0.491	-1.543	0.010	0.30	-0.03	-1.14	X_2
0.801	-0.430	0.775	0.40	-0.10	-0.17	X_3
0.891	-0.863	0.892	0.74	-0.00	-0.10	X_4
3.439	-0.553	0.430	1.01	0.04	1.51	Constant

The results of the table (5) indicate the analysis of the logistic regression model coefficient estimates after applying the bootstrap method to examine the stability of the estimates and assess their accuracy.The results showed that the variable X2 (residence) was the only variable that demonstrated a statistically significant significance, with a p-value (Sig.) of 0.010, which is less than the significance level of 0.05, indicating a significant effect of residence on the dependent variable (length of hospital stay). The value of B = -1.14 indicates that residing in certain environments reduces the likelihood of a short hospital stay, which is consistent with previous sestimates.In contrast, the other variables did not show significant significance, as their p-values exceeded 0.05 (respectively: 0.764, 0.775, and 0.892), indicating that their effect is not significant in the model. Similarly, the confidence intervals (Lower – Upper) for all these variables contain zero, which supports the absence of a significant effect.

As for the constant, although the value B = 1.51 indicates a positive trend, the p-value of 0.430 confirms the absence of a statistically significant effect, as the confidence interval ranges from -0.553 to 3.439 and includes zero.

Therefore, it can be concluded that the bootstrap-estimated logistic regression model reinforces the previous results regarding the significance of the residence variable (X2) only, while the other variables show a weakness in their statistical impact, highlighting the need to focus on factors with significant effects when analyzing the factors associated with the length of hospital stay for epilepsy patients.

$$\widehat{Y} = 1.510 - 0.146X_1 - 1.143X_2 - 0.175X_3 - 0.103X_4$$

3.4. Employment of the genetic algorithm

To determine the effect of the explanatory variables on the dependent variable, the genetic algorithm method was used to estimate the coefficients of the logistic regression model, and the results are presented in Table (6).

Table 6: shows the results of the Employment of the genetic algorithm method

Wald	Sig. (2-tailed)	Std. Error	B	Variables
2.195	0.865	0.225	1.452	X_1
0.376	0.013	0.103	-0.103	X_2
12.35	0.964	0.023	-1.350	X_3
0.964	0.702	0.342	0.205	X_4
0.013	0.332	0.4598723	0.018	Constant

The results of the table (6) indicate the use of the genetic algorithm to estimate the parameters of the binary logistic regression model, where the effect of the explanatory variables on the dependent variable was analyzed. This estimation resulted in the following model equation:

$$\widehat{Y} = 0.018 + 1.452X_1 - 0.103X_2 - 1.350X_3 + 0.205X_4$$

This equation illustrates the trends represented by the estimated coefficients for each of the independent variables, and indicates that variable X2 (residence) is the only one that showed significant statistical significance, with a probability value (Sig.) of 0.013, which is less than the significance level of 0.05. This suggests that it has an effective impact on determining the likelihood of a patient staying longer in the hospital. This result is supported by a Wald statistic value of 0.376, reflecting the relative importance of this variable within the model.

In contrast, the other variables (X1, X3, X4) did not show statistical significance, as their Sig. values exceeded the significance level. The significance level (respectively: 0.865, 0.964, 0.702), indicating their lack of meaningful impact on the dependent variable in this model. Although the Wald value for variable X3 was relatively high (12.35), its lack of statistical significance reduces the possibility of interpreting it as an influential variable.

As for the constant, it was 0.018 with a standard error of 0.4598 and a significance value of 0.332, indicating its lack of statistical significance in the absence of explanatory variables.

Based on these results, it can be concluded that the genetic algorithm is an effective tool in improving the accuracy of predictive models through its ability to distinguish variables with real impact. Additionally, identifying the residence variable as a key influencer supports directing efforts towards improving the living environment as a factor in reducing the length of stay, which adds value in enhancing the quality of healthcare and making more precise treatment decisions.

3.5. Comparison between the Bootstrap method and the Employment of the genetic algorithm method

The mean squared error (MSE) was used as a criterion to evaluate the performance of both the bootstrap method and the employment of the genetic algorithm in estimating the parameters of the binary logistic regression model. The results showed that the MSE value when using the bootstrap method was 2.446, while it decreased to 0.646 when using the genetic algorithm. Based on these results, it can be concluded that the genetic algorithm method

provides more accurate estimates, given the lower mean squared error, making it more efficient in model estimation compared to the bootstrap method..

4. Conclusions and Recommendations

The results showed that the data used align well with the binary logistic regression model, based on the statistical tests applied. It was found that using the genetic algorithm achieves better performance compared to the bootstrap method, as it recorded a lower mean squared error (MSE) of (0.646) compared to (2.446) for the bootstrap method, reflecting higher efficiency in estimating the model parameters. The analysis results also indicated that the variable X2 (housing) was the only variable with a significant impact on the dependent variable, while the other variables did not show any noteworthy statistical significance. Based on these results, it is recommended to adopt the genetic algorithm as an effective and accurate tool for estimating the parameters of the logistic regression model, especially in medical studies dealing with small samples or unbalanced data. This model also stands out as an important means of predicting the length of hospital stay for epilepsy patients, which can support the quality of healthcare and improve the management of medical resources. The results highlight the importance of focusing on influential variables when designing therapeutic and predictive programs. Additionally, it is recommended to expand the scope of data collection to include all governorates, ensuring the comprehensiveness and accuracy of the results, along with integrating artificial intelligence techniques with traditional statistical methods, which contributes to enhancing predictive capability and improving the efficiency of medical data analysis, and supports the development of decision support systems in the healthcare sector more effectively.

References

[1] Abdulqader, Q. M. (2015). Comparison of discriminant analysis and logistic regression analysis: An application on caesarean births and natural births data. Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 20(1-2), 34-46.

[2] Cramer, J. S. (2002). The origins of logistic regression (No. 02-119/4). Tinbergen Institute discussion paper.

[3] Davidson, J. L., Hua, X., & Ashlock, D. (1996, April). A comparison of genetic algorithm, regression, and Newton's method for parameter estimation of texture models. In Proceeding of Southwest Symposium on Image Analysis and Interpretation (pp. 201-206). IEEE.

[4] Davison, A. C., & Kuonen, D. (2002). An introduction to the bootstrap with applications in R. Statistical computing & Statistical graphics newsletter, 13(1), 6-11.

[5] Demaris, A. (1992). Logit modeling: Practical applications (Vol. 86). Sage.

[6] Enad, F. H. (2022). The use of logistic regression method in data classification with practical application of Covid-19 patients in Nasiriya General Hospital. University of Thi-Qar Journal, 17(2).

[7] Gayou, O., Das, S. K., Zhou, S. M., Marks, L. B., Parda, D. S., & Miften, M. (2008). A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes. Medical physics, 35(12), 5426-5433.

[8] Goldberg, D. E., & Deb, K. (1991). A comparative analysis of selection schemes used in genetic algorithms. In Foundations of genetic algorithms (Vol. 1, pp. 69-93). Elsevier.

[9] Hadji, S., Gaubert, J. P., & Krim, F. (2015). Theoretical and experimental analysis of genetic algorithms based MPPT for PV systems. Energy Procedia, 74, 772-787.

[10] Hamza, I. S. N., & Jas, M. N. (2019). Using the bootstrap method in estimating the logistic regression model via the maximum likelihood method: An applied study. Journal of Administration & Economics, (121).

[11] Hao, C., Chen, C., Wen, Y., Meng, F., Chi, Z., Zhao, R., ... & Cheng, L. (2023, October). Research progress of electroencephalogram (EEG) classification method and its application based on support vector machine (SVM). In 2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-4). IEEE.

[12] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[13] Jaber, A. G., Enad, F. H., & Alrawi, Z. N. (2023). Estimating the Multiple Regression Model Using the Bootstrap Method to Study the Effect of Environmental Wastes on the Waters of the Euphrates River. Al-Rafidain University College For Sciences, (54).

[14] Jain, A., Ratnoo, S., & Kumar, D. (2017, August). Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach. In 2017 international conference on information, communication, instrumentation and control (ICICIC) (pp. 1-8). IEEE.
[15] Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.
[16] Li, R., Zhou, J., & Wang, L. (2018). Estimation of the binary logistic regression model parameter using bootstrap re-sampling. Latin American Applied Research, 48, 199-204.
[17] Liu, H. H., & Ong, C. S. (2008). Variable selection in clustering for marketing segmentation using genetic algorithms. Expert systems with applications, 34(1), 502-510.
[18] Liu, Z., & Zhang, Q. (2014, June). A multinomial logistic model for ranking technical efficiency of public project. In 2014 11th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-4). IEEE.
[19] Mahdavi, I., Paydar, M. M., Solimanpur, M., & Heidarzade, A. (2009). Genetic algorithm approach for solving a cell formation problem in cellular manufacturing. Expert Systems with Applications, 36(3), 6598-6604.
[20] McCullagh, P. (2019). Generalized linear models. Routledge.
[21] Mostajeran, A., Iranpanah, N., & Noorossana, R. (2016). A new bootstrap based algorithm for Hotelling's T2 multivariate control chart. Journal of Sciences, Islamic Republic of Iran, 27(3), 269-278.
[22] Nakamichi, R., Imoto, S., & Miyano, S. (2004, May). Case-control study of binary disease trait considering interactions between SNPs and environmental effects using logistic regression. In Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering (pp. 73-78). IEEE.
[23] Rahamneh, A., & Hawamdeh, O. (2017). The Factors Affecting Eye Patients (Cataract) In Jordan by Using the Logistic Regression Model. Modern Applied Science, 11(8), 1-38.
[24] Schmid, C. S., & Desmarais, B. A. (2017, December). Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In 2017 IEEE international conference on big data (Big Data) (pp. 116-121). IEEE.
[25] Sefiane, S., & Benbouziane, M. (2012). Portfolio selection using genetic algorithm.
[26] Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2015, October). Profit maximizing logistic regression modeling for customer churn prediction. In 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). IEEE.
[27] Wuensch, K. L. (2014). Binary logistic regression with SPSS. Retrieved March, 18, 2015.