A Simulation Study on Group Variable Selection Methods for Binary Response

Hussein A. Hashem

Department of Mathematics, College of Science, University of Duhok, Kurdistan Region, IRAQ.

*Corresponding author E-mail: hussein.hashem@uod.ac

https://doi.org/10.29072/basjs.20240214

ARTICLE INFO

Keywords

Variable selection, Binary group regression, Logistic regression.

ABSTRACT

Binary data, denoting data having two alternative outcomes, is frequently observed across several research domains including finance, social sciences, psychology, and health. The logistic regression model is extensively employed for the analysis of binary data. It is essential to meticulously examine the detection and management of influential outliers to guarantee the suitability of the fitted binary logistic models. This article offers an extensive evaluation of various collective binary logistic techniques employed in regression models, emphasizing a comparison of the efficacy of four distinct logistic regression approaches. The methods encompass group Lasso binary estimates, group mcp binary estimates, group scad binary estimates, and binary regularization paths for generalized linear models by coordinate descent (glmnet) estimates. The comparisons derive from a simulation research aimed at determining which of these approaches exhibits superior performance across all regression scenarios. This review and comparison enable researchers and practitioners to discern the most effective methodologies for evaluating binary data via logistic regression.

Received 29 Nov 2024; Received in revised form 12 Dec 2024; Accepted 18 Dec 2024, Published 31 Dec 2024

1. Introduction

Logistic regression is a statistical technique widely employed for modelling binary dependent variables. It involves establishing a mathematical relationship between independent variables and a binary dependent variable, typically representing two categories as 0 or 1. The independent variables in this statistical model can take various forms, including continuous, discrete, binary, or combinations thereof. To handle atypical observations in data, researchers have developed diverse statistical models. For instance, Gelman suggested a model that modifies the chance of success in order to handle outlier problems in logistic regression [1]. . In Gelman's model, users are required to specify predetermined trimming values in advance. In a comprehensive study conducted by Wang et al. in 2004 [2], the researchers compared the performance of the Bayesian model averaging method with the stepwise selection method. Their in-depth analysis revealed that Bayesian Model Averaging exhibited superior performance compared to the stepwise selection method. Furthermore, Saker et al. (2009) [3] undertook a study involving both stepwise selection and best subset selection methods for variable selection in model fitting. Their findings indicated that both stepwise selection and best subset selection methods produced similar results. Wang in 2024 [4] used ordinal logistic regression for analysis the effective teaching practices. In 2024 Graham [5] studied the sparse network asymptotic for logistic regression under possible misspecification. Hou and Song [6] delve into the logistic regression transfer learning problem supported by differential privacy. Lewis and Battey [7] studied the inference in high-dimensional logistic regression models with separated data. Shareef et al. [8] used multinomial logistic regression for determining the factors Influencing blood pressure. Balboa et al. [9] evaluated several algorithms to predict evacuation decisions and found that being with a close family member is the most influential factor in responding to a fire alarm . Yuniarsih et al. [10] employed binary logistic regression model in the adoption of technological innovation of urban farming. This manuscript aims to provide a comprehensive review of the logistic regression analysis as a powerful method for defining the relationship between binary result variables and independent variables. The focus of the review lies in the applicability of various logistic regression methods in a simulation study, offering insights into their effectiveness and limitations.

2. Material and Method

2.1 General Form of Binary Logistic Regressions

Binary logistic regression is a powerful statistical method used to analyze the relationship between a binary dependent variable and one or more independent variables. In this type of regression, the dependent variable is binary, meaning it can have only two possible outcomes, often coded as 0 or 1. The primary aim of binary logistic regression is to model the probability of the binary outcome as a function of the independent variables. In contrast to linear regression, which is used for continuous dependent variables, binary logistic regression models the log odds of the dependent variable belonging to a particular category. This makes it well-suited for predicting the likelihood of a binary event occurring based on one or more predictor variables. One of the key advantages of binary logistic regression is that it does not assume a linear relationship between the independent variables and the log odds of the dependent variable. Instead, it employs the logistic function, also known as the sigmoid function, to model the relationship. The logistic function ensures that the predicted probabilities fall within the range of 0 to 1, making it particularly suitable for modeling binary outcomes. This capability makes binary logistic regression a valuable tool for a wide range of applications, including in fields such as healthcare, marketing, and social sciences. Overall, binary logistic regression is an important statistical method used for predicting the probability of a binary outcome based on one or more predictor variables. It is widely employed in fields such as medicine, social sciences, and business for modelling and predicting binary outcomes. This versatile tool allows researchers and analysts to examine the relationships between input variables and the likelihood of a particular event occurring. Its applications range from predicting the likelihood of a patient developing a specific disease to forecasting the success of a marketing campaign. To model and predict binary outcomes based on relevant factors [11] : $p = P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$.

The function described above is a non-linear S-shaped curve, commonly known as the logistic regression function. In this function, β represents the coefficient of the predictor or input variable x in a regression equation. Although this function can handle multiple input variables in a simplified form, it is essentially linear. It is considered superior to the logistic response function, $p = P(Y = 1|X_1 = x_1,\ldots,X_p = x_p) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_{2+}\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_{2+}\ldots+\beta_p x_p}}$.

The equation provided calculates the likelihood of the response variable being 1, taking into account multiple predictor variables. The model is inherently non-linear, but it is transformed into linearity through the use of the logit response function. The logistic response function equation is then utilized to achieve this transformation: $\frac{p}{1-p} = e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots\beta_px_p}$.

The term $\frac{p}{1-p}$ the above equation is called the odds ratio of the event. Taking the natural logarithm on both sides, $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2+\ldots\beta_px_p$

Where $p = P(Y = 1) = 1 - P(Y = 0)$

$P(Y = 1), P(Y = 0)$ is the probability of success and failure of an observation respectively

$\beta_0 = \log - \text{odds when all } x_j \text{are } 0\ \beta_j = \text{increase in log} -$
$\text{odds when } x_j \text{ is increased by one unit}, j = 1, \ldots, p e^{\beta_j} =$
$\text{increase in odds when is increased by one unit}, j = 1, \ldots, p$

The equation provided describes a linear relationship between variables and can be utilized to analyze and understand the connections between the variables of interest [11].

2.2 Logistic Curve

When the outcome or dependent variable takes on only two possible values, such as 0 and 1, and the predictor or independent variable is numerical, a logistic regression model is employed to analyze the relationship between the two variables. This model fits a logistic curve to the data, which is characterized by a distinct "S" shape known as a sigmoid curve [12]. The logistic curve depicts the probability of the outcome variable as a function of the predictor variable. In logistic regression, the logistic function is commonly used and can be defined by the following formula

$y = \frac{e^x}{1+e^x}$ .This equation can be extended to the form $y = \frac{e^{\beta_0+\beta_1x}}{1+e^{\beta_0+\beta_1x}}\ \frac{1}{1+e^{-(\beta_0+\beta_1x)}}$, which is graphed in Figure 1.

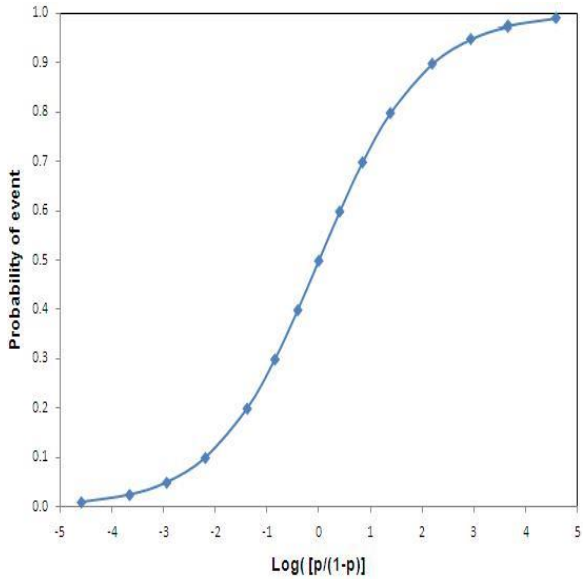

Figure 1**:** The logistic function

2.3 Assumptions of Binary Logistic Regression

Binary logistic regression is distinct from general linear models in that it does not rely on certain key assumptions. First, it does not necessitate a linear relationship between the dependent and independent variables, nor does it require the error distribution to be normal or the errors to be homoscedastic. Additionally, it is not contingent on the measurement level of the independent variables .Nevertheless, binary logistic regression does have other assumptions that need to be considered [13]:

1. The dependent variable must be binary.

2. The dependent variable must be coded to represent the probability of an event occurring ($P(Y = 1)$).

3. The model must be well-fitted, including all meaningful variables while avoiding unnecessary ones to prevent over-fitting.

4. Each observation in the data should be independent, and the independent variables should exhibit minimal or no multicollinearity.

5. There should be linearity between the independent variables and the log odds, but not necessarily between the dependent and independent variables.

6. Large sample sizes are required, as small samples can lead to an overestimation of the effect measure. Including more independent variables in the model also necessitates a larger sample size.

2.4 Maximum Likelihood Estimation

The logistic regression model may look similar to a simple linear regression model, but there are important differences in the underlying distribution. In logistic regression, the dependent variable follows a binomial distribution, requiring a different approach to estimating the parameters. The α and β parameters in logistic regression cannot be estimated in the same way as in simple linear regression. Instead, the coefficients are typically estimated using the Maximum Likelihood Estimation (MLE) method. MLE involves finding the values of the parameters that maximize the likelihood function. The likelihood function represents the probability of obtaining the observed values of the dependent variable given the observed values of the independent variables. Similar to other probabilities, the likelihood ranges from 0 to 1, and the goal of MLE is to find the parameter values that make the observed data most probable under the assumed statistical model [13].

This information is based on the work of Torosyan (2017) [6].$P(Y = y_i) = P_i^{1-y_i}(1 - P_i)^{y_i}$, where P_i is the probability of the ith observation, y_i is the value of random variable Y that takes value 0 or 1. Assuming that our n observations are independent the likelihood of the data is equal to $L = \prod_{i=1}^{n} P_i^{1-y_i}(1 - P_i)^{y_i}$.The maximum Likelihood method will provide values for β_0 and β which maximize L function.

2.5 Classification table

The classification table, also known as a confusion matrix, is a valuable tool for evaluating the predictive performance of a logistic regression model. It is used to compare the observed values for the dependent outcome with the predicted values generated by the model. The table cross-classifies these values to provide insights into the model's accuracy [11].

To illustrate, let's consider a scenario where a cut-off value of 0.5 is used. Any predicted values above 0.5 are designated as predicting an event, while any predicted values below 0.5 are considered as not predicting the event. This allows for clear categorization of model predictions, aiding in the assessment of its accuracy in predicting specific outcomes [14].

Table 1. Sample Classification Table

Observed	Predicted	
	1	0
1	a(True Positive)	b(False Negative)
0	c(False Positive)	d(True Negative)

where a and d are the number of cases that are predicted correctly, and b and c are the numbers of cases that are not predicted correctly. "In a predictive model, 'a' represents the number of true positive cases, 'b' represents the number of false negative cases, 'c' represents the number of false positive cases, and 'd' represents the number of true negative cases." When assessing the accuracy of a test that predicts binary outcomes, we can consider two key indicators: sensitivity and specificity. Sensitivity refers to the proportion of true positives ($Y = 1$), while specificity refers to the proportion of true negatives ($Y = 0$).

Sensitivity is calculated using the formula $d\ /\ (c + d)$, and specificity is calculated using the formula $a\ /\ (a + b)$. It's important to note that the values of sensitivity and specificity are influenced by the chosen cut-off value. For instance, if we increase the cut-off point, fewer observations will be predicted as positive. This leads to fewer $Y = 1$ observations being predicted as positive, causing a decrease in sensitivity. Conversely, more $Y = 0$ observations will be predicted as negative, leading to an increase in specificity. In an ideal scenario, a model would exhibit 100% sensitivity and 100% specificity, but in practical terms, achieving such results is rare. Therefore, understanding how sensitivity and specificity are impacted by the chosen cut-off value is crucial for accurately evaluating the performance of the model.

2.6 ROC curve (Receiver Operating Characteristics)

The Receiver Operating Characteristic (ROC) curve is a valuable graphical tool used to evaluate the performance of diagnostic tests. Unlike the traditional two-by-two table, the ROC curve considers a wide range of cutoff values from 0 to 1. For each cutoff value, a corresponding two-by-two table is constructed, allowing for a more nuanced analysis. The ROC curve visually represents the relationship between sensitivity (True Positive rate) and one minus the specificity (False Positive rate) as the cutoff value increases from 0 to 1. This visualization provides a comprehensive understanding of the test's performance across various cutoff values, enabling insights into its predictive power that go beyond what's obtainable from a standard classification table [14].

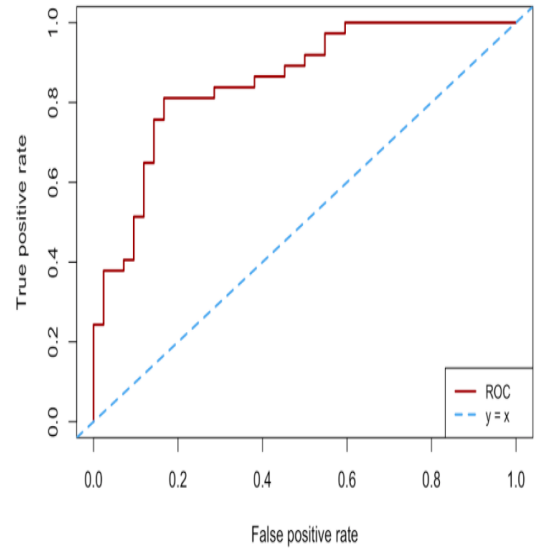

Figure2. An Example of the ROC Curve

The ROC curve, short for receiver operating characteristic curve, is a graphical plot that shows the performance of a binary classifier model across different discrimination thresholds. The curve illustrates the trade-off between the classifier's sensitivity and specificity. Sensitivity (true positive rate) measures the proportion of actual positive cases that are correctly identified, while specificity (true negative rate) measures the proportion of actual negative cases that are correctly identified. The area under the ROC curve (AUC) quantifies the overall performance of the model. It provides a single value to represent how well the model can distinguish between the two classes. A perfect model has an AUC of 1, indicating that it achieves perfect discrimination between the classes. On the other hand, an AUC of 0.5 suggests that the model's performance is no better than random guessing. In practice, an AUC above 0.7 is generally considered indicative of a very good model, showing strong discriminatory ability. Models with higher AUC values are more effective at distinguishing between the classes, making them valuable for many applications in machine learning and predictive analytics.

2.7 Regularization Paths for Generalized Linear Models via Coordinate Descent (glmnet)

In 2010, Friedman and Hastie [15] made significant contributions by introducing rapid algorithms for estimating generalized linear models with convex penalties. These algorithms were a breakthrough in the field as they could effectively handle a wide range of applications, including but not limited to linear regression, two-class logistic regression, and multinomial regression problems. The penalties involved in these models encompassed the lasso, ridge regression, and the hybrid

elastic net, making them highly versatile. The algorithms developed by Friedman and Hastie were based on cyclical coordinate descent computed along a regularization path. This unique approach allowed the algorithms to efficiently handle large-scale problems and effectively manage sparse features, which was a crucial advancement in the field of machine learning and statistical modeling.

In addition to their groundbreaking research findings, Friedman and Hastie generously released their R package to the public. This act of openness and generosity greatly contributed to wider access and application of their work, enabling researchers and practitioners to benefit from their innovative algorithms and models.

2. Results and Discussions

In our current research section, we aim to perform a thorough comparative analysis of different logistic regression methods using a simulation study. This analysis is crucial due to the widespread use of these methods in real-world applications. To guarantee the robustness and precision of our results, we will generate simulated data from diverse sources and meticulously assess the efficacy of each logistic regression method. This rigorous evaluation process will enable us to make informed conclusions regarding the most suitable logistic regression method for our specific research context.

$y_i^\star = x_i^T\beta + \varepsilon_i, \quad i = 1,. \ . \ ., n$ and $y_i = h(y_i^\star)$,

where ε_i are the errors and h is a link function. For binary response data, the link function is given by $h(y^\star) = I(y^\star > 0)$, with I the indicator function. In real applications, y is the observed binary response and the interest is to predict y from knowledge of x. $y^\star$ is unobserved and used mainly for modelling purposes. We are considering error distributions similar to those in the studies by Yu et al. (2013) [16] and Li et al. (2010) [17]:

■Normal: $N(0;\ 1)$

■Normal: $N(0;\ 9)$

■ A mixture of two normal distributions: $0.1N(0, 10000) + 0.9N(0, 1)$

■A t distribution with 1 degree of freedom : t_1

■ A t distribution with 3 degrees of freedom: t_3

■ Laplace distribution with location 0 and scale 10: $Laplace(0, 10)$

■ A mixture of two Laplace distributions: $0.1Laplace\ (0, 1) + \ 0.9Laplace(0, 5)$

■Skewed (skew): $\frac{1}{5}N\left(-\frac{22}{25}, 1\right) + \frac{1}{5}N(-\frac{49}{125}\ , (\frac{3}{2})^2) + \frac{3}{5}N(\frac{49}{250}, (\frac{5}{9})^2)$

■Kurtotic: (kur): $\frac{2}{3}N(0,1)+\frac{1}{3}N(0,(\frac{1}{10})^2)$

■Bimodal (bim): $\frac{1}{2}N(-1,(\frac{2}{3})^2)+\frac{1}{2}N(1,(\frac{2}{3})^2)$

■Bimodal, with separate modes (bim-sep): $\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2)+\frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$

■Skewed Bimodal (skew-bim): $\frac{3}{4}N\left(-\frac{43}{100},1\right)+\frac{1}{4}N(\frac{107}{100},(\frac{1}{3})^2)$

■Trimodal (tri): $\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2)+\frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2)+\frac{1}{10}N(0,(\frac{1}{4})^2)$.

These distributions were selected to have a median close to or equal to zero. We used a sample size of 50 for the simulation.

For the β vector, we consider the case of a large number of predictors, i.e. $p \gg n$. Similar to Li et al. (2010) [13], we draw the independent variables x from a multivariate normal distribution, $N(0,\Sigma_x)$. The pairwise covariance between x_i and x_j is set to be $(\Sigma_x)_{ij}=r^{|i-j|}$. For the correlation r, we experiment both with $r=0.5$ and $r=0.95$. For the β values we consider three cases:

(1) The values for the first three groups are given by
$\beta_j=$
$(0.5,1,1.5,2,2.5,2,2,2,2,2),(2,2,1,1,1,1,3,3,3,3),(1,1,1,2,2,2,3,3,3,3)$, The coefficients are set to zero for all groups except the specified ones, reflecting a sparse scenario with structured grouping in the predictor variables.

(2) $\beta_j=(3,1.5,0,0,2,0,\ldots,0)$, The coefficients are set to zero for all groups except the specified ones. This means that the model has a very sparse structure with specific group patterns present in the predictors.

(3) $\beta_j=0.85$ for all j, which corresponds to a dense case.

(4) In this section, we will compare the following binary group logistic regression methods:

■"grp. lasso": binary group Lasso penalty (Yuan and Lin, 2006) [18].

■"grp. scad": binary group smoothly clipped absolute deviation (Xiong et al., 2016) [19].

■"grp. mcp": binary group minimax concave penalty (Xiong et al., 2016) [19].

■" glmnet": Regularization Paths for Generalized Linear Models via Coordinate Descent (Friedman and Hastie, 2010) [15].

For the grp. lasso, grp. scad and grp. mcp methods we use the R package grpreg and for the generalized linear models via coordinate descent, we use the R package glmnet. In the analysis, we evaluated various methods and error distributions, calculating the AUC values over 500 iterations on a test set of the same size as the training set. Our investigation involved different scenarios for the β values, with r set at 0.5 and 0.95 for Tables 2 and 3, and r set at the same values for Tables 4 and 5. Additionally, Tables 6 and 7 explored the case of all βs equal to 0.85, with r at 0.5 and 0.95. The results, presented in Tables 2, 3, 4, 5, 6, and 7, revealed no significant differences between three prediction approaches for grMCP, grSCAD, and the regularization paths for generalized linear models via coordinate descent (glmnet R package). Interestingly, the analysis indicated that the grLasso R package performed as the best-performing method in most cases.

Table2: Average AUC values over 500 iterations (with standard deviations in brackets) for n = 50, p = 100, $r = 0.5$, and β values as in case (1).The best mean is indicated in bold.

	grLasso	grMCP	grSCAD	glmnet
N(0,1)	**0.787** (0.102)	0.701 (0.123)	0.775 (0.104)	0.735 (0.123)
N(0,3)	**0.772** (0.107)	0.685 (0.121)	0.758 (0.112)	0.714 (0.136)
Normal M.	**0.666** (0.116)	0.613 (0.114)	0.66 (0.115)	0.629 (0.122)
t_1	**0.736** (0.118)	0.661 (0.124)	0.728 (0.118)	0.703 (0.132)
t_3	**0.776** (0.111)	0.688 (0.126)	0.76 (0.116)	0.727 (0.133)
Laplace	**0.62** (0.114)	0.58 (0.101)	0.618 (0.112)	0.585 (0.106)
Laplace M.	**0.728** (0.122)	0.643 (0.124)	0.718 (0.122)	0.689 (0.138)

Skew	**0.786** (0.109)	0.698 (0.125)	0.774 (0.108)	0.741 (0.13)
Kur	**0.773** (0.113)	0.693 (0.126)	0.763 (0.118)	0.732 (0.134)
Bim	**0.788** (0.099)	0.693 (0.123)	0.769 (0.11)	0.736 (0.131)
bim − sep	**0.776** (0.11)	0.689 (0.122)	0.765 (0.107)	0.731 (0.127)
skew − bim	**0.771** (0.109)	0.684 (0.124)	0.763 (0.109)	0.716 (0.134)
Tri	**0.774** (0.111)	0.687 (0.127)	0.76 (0.112)	0.732 (0.128)

Table3: Average AUC values over 500 iterations (with standard deviations in brackets) for $n = 50, p = 100, r = 0.95$, and β values as in case (1). The best mean is indicated in bold.

	grLasso	grMCP	grSCAD	glmnet
N(0,1)	0.778 (0.107)	0.69 (0.122)	0.763 (0.113)	**0.955** (0.036)
N(0,3)	0.778 (0.107)	0.698 (0.119)	0.763 (0.107)	**0.949** (0.044)
Normal M.	0.73 (0.119)	0.664 (0.12)	0.721 (0.124)	**0.911** (0.06)
t_1	0.755 (0.117)	0.677 (0.126)	0.744 (0.119)	**0.931** (0.052)
t_3	0.774 (0.108)	0.69 (0.124)	0.767 (0.107)	**0.954** (0.039)
Laplace	0.696 (0.122)	0.645 (0.119)	0.698 (0.115)	**0.877** (0.08)

Laplace M.	0.761 (0.114)	0.679 (0.126)	0.75 (0.117)	**0.939** (0.046)
Skew	0.773 (0.109)	0.69 (0.121)	0.768 (0.103)	**0.956** (0.041)
Kur	0.784 (0.107)	0.698 (0.128)	0.771 (0.112)	**0.956** (0.038)
Bim	0.781 (0.111)	0.688 (0.13)	0.771 (0.111)	**0.956** (0.041)
bim − sep	0.772 (0.112)	0.686 (0.121)	0.763 (0.11)	**0.956** (0.039)
skew − bim	0.776 (0.111)	0.686 (0.125)	0.759 (0.119)	**0.957** (0.033)
Tri	0.775 (0.108)	0.696 (0.123)	0.768 (0.106)	**0.956** (0.04)

Table4: Average AUC values over 500 iterations (with standard deviations in brackets) for $n = 50, p = 100, r = 0.5$, and β values as in case (2). The best mean is indicated in bold.

	grLasso	grMCP	grSCAD	glmnet
N(0,1)	0.927 (0.046)	**0.949** (0.033)	0.948 (0.034)	0.949 (0.037)
N(0,3)	0.817 (0.082)	**0.839** (0.088)	0.832 (0.082)	0.823 (0.1)
Normal M.	**0.527** (0.067)	0.521 (0.057)	0.525 (0.07)	0.525 (0.061)
t_1	0.825 (0.085)	**0.845** (0.092)	0.843 (0.077)	0.822 (0.118)
t_3	0.908 (0.054)	**0.931** (0.04)	0.928 (0.042)	0.927 (0.048)

Laplace	0.528 (0.078)	**0.519** (0.071)	0.522 (0.078)	0.518 (0.062)
Laplace M.	0.655 (0.12)	0.652 (0.133)	**0.659** (0.122)	0.653 (0.129)
Skew	0.942 (0.043)	0.963 (0.03)	**0**.962 (0.031**)**	**0.987** (0.015)
Kur	0.938 (0.042)	**0.958** (0.03)	0.957 (0.03)	0.957 (0.042)
Bim	0.94 (0.043)	**0.96** (0.032)	0.959 (0.033)	0.96 (0.036)
bim − sep	0.94 (0.046)	0.961 (0.035)	0.96 (0.036)	**0.966** (0.033)
skew − bim	0.937 (0.042)	0.955 (0.033)	0.954 (0.034)	**0.956** (0.036)
Tri	0.942 (0.044)	0.963 (0.03)	0.962 (0.031)	**0.967** (0.033)

Table5: Average AUC values over 500 iterations (with standard deviations in brackets) for $n = 50, p = 100, r = 0.95$, and β values as in case (2). The best mean is indicated in bold.

	grLasso	grMCP	grSCAD	glmnet
N(0,1)	0.934 (0.046)	0.957 (0.031)	0.956 (0.033)	**0.98** (0.017)
N(0,3)	0.868 (0.063)	0.89 (0.061)	0.886 (0.054)	**0.927** (0.037)
Normal M.	0.56 (0.095)	0.552 (0.093)	0.56 (0.095)	**0.618** (0.124)
t_1	0.854 (0.076)	0.879 (0.069)	0.871 (0.075)	**0.918** (0.052)

t_3	0.919 (0.051)	0.943 (0.035)	0.941 (0.038)	**0.971** (0.02)
Laplace	0.549 (0.092)	0.547 (0.095)	0.55 (0.096)	**0.598** (0.121)
Laplace M.	0.742 (0.111)	0.746 (0.13)	0.753 (0.111)	**0.826** (0.091)
Skew	0.942 (0.043)	0.963 (0.03)	0.962 (0.031)	**0.987** (0.015)
Kur	0.939 (0.042)	0.961 (0.03)	0.96 (0.03)	**0.985** (0.017)
Bim	0.942 (0.046)	0.964 (0.029)	0.962 (0.031)	**0.988** (0.014)
$bim-sep$	0.946 (0.039)	0.965 (0.029)	0.964 (0.029)	**0.99** (0.011)
$skew-bim$	0.941 (0.039)	0.962 (0.029)	0.961 (0.029)	**0.985** (0.014)
Tri	0.941 (0.043)	0.963 (0.029)	0.962 (0.031)	**0.989** (0.013)

Table6: Average AUC values, over 500 iterations (with standard deviations in brackets) for $n = 50, p = 100, r = 0.5$, and β values as in case (3). The best mean is indicated in bold.

	grLasso	grMCP	grSCAD	glmnet
N(0,1)	**0.591** (0.096)	0.541 (0.074)	0.588 (0.093)	0.573 (0.097)
N(0,3)	**0.596** (0.098)	0.542 (0.075)	0.579 (0.091)	0.583 (0.103)
Normal M.	**0.546** (0.076)	0.52 (0.06)	0.541 (0.075)	0.539 (0.075)
t_1	**0.569**	0.536	0.563	0.563

	(0.089)	(0.073)	(0.087)	(0.091)
t_3	**0.596** (0.098)	0.542 (0.075)	0.579 (0.091)	0.583 (0.103)
Laplace	**0.541** (0.074)	0.523 (0.061)	0.538 (0.069)	0.525 (0.065)
Laplace M.	**0.568** (0.088)	0.533 (0.07)	0.567 (0.088)	0.561 (0.093)
Skew	**0.599** (0.103)	0.551 (0.081)	**0.595** **(0.098)**	0.58 (0.1)
Kur	**0.6** (0.097)	0.546 (0.078)	0.591 (0.093)	0.584 (0.1)
Bim	**0.6** (0.094)	0.548 (0.076)	0.588 (0.089)	0.577 (0.095)
$\text{bim} - \text{sep}$	**0.607** (0.099)	0.551 (0.08)	0.594 (0.094)	0.576 (0.1)
$\text{skew} - \text{bim}$	0.587 (0.096)	0.543 (0.075)	**0.587** (0.094)	0.57 (0.097)
Tri	**0.602** (0.097)	0.545 (0.077)	0.593 (0.095)	0.581 (0.102)

Table7**:** Average AUC values over 500 iterations (with standard deviations in brackets) for $n = 50, p = 100, r = 0.95$, and β values as in case (3). The best mean is indicated in bold.

	grLasso	grMCP	grSCAD	glmnet
N(0,1)	0.602 (0.103)	0.546 (0.078)	0.591 (0.096)	**0.811** (0.121)
N(0,3)	0.598 (0.105)	0.546 (0.081)	0.595 (0.1)	**0.79** (0.125)

Normal M.	0.578 (0.095)	0.541 (0.077)	0.574 (0.09)	**0.719** (0.133)
t_1	0.577 (0.094)	0.534 (0.071)	0.569 (0.092)	**0.767 (**0.134**)**
t_3	0.607 (0.101)	0.549 (0.08)	0.591 (0.094)	**0.809** (0.121)
Laplace	0.561 (0.089)	0.531 (0.071)	0.557 (0.085)	**0.688** (0.13)
Laplace M.	0.589 (0.098)	0.543 (0.074)	0.583 (0.093)	**0.771** (0.134)
Skew	0.595 (0.092)	0.537 (0.066)	0.585 (0.091)	**0.795** (0.129)
Kur	0.599 (0.094)	0.546 (0.076)	0.593 (0.093)	0.8 (0.123)
Bim	0.599 (0.099)	0.544 (0.08)	0.59 (0.098)	**0.809** (0.124)
bim − sep	0.599 (0.097)	0.55 (0.082)	0.593 (0.095)	**0.81** (0.115)
skew − bim	0.602 (0.099)	0.546 (0.076)	0.59 (0.093)	**0.806** (0.118)
Tri	0.599 (0.098)	0.547 (0.076)	0.588 (0.095)	**0.799** (0.124)

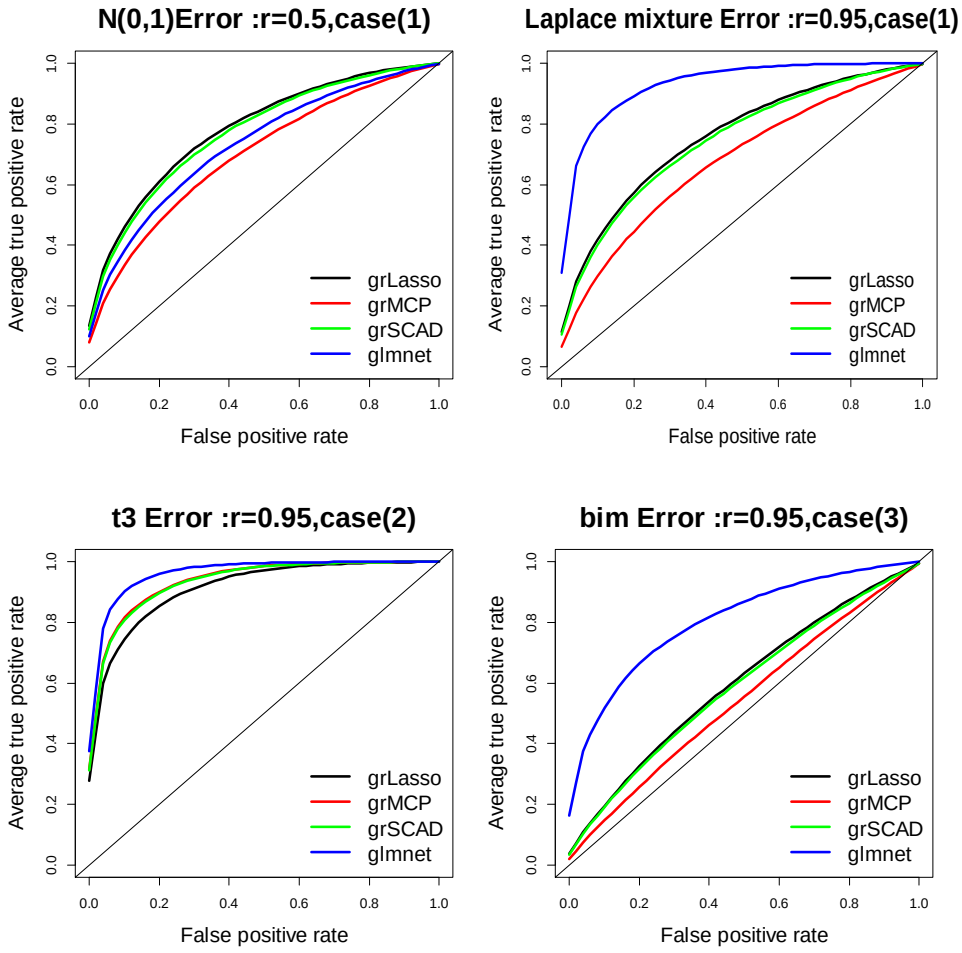

Figure2: Comparison of Average ROC Curves over 500 Iterations Under Various Error Distributions: Normal (0, 1), Laplace Mixture, t3, and bim.

4. Conclusions

In this paper, our objective is to conduct a thorough comparison of regression methods in scenarios where the response variable is binary, utilizing simulated data. Upon extensive simulation studies, we have found compelling evidence that the grLasso R package consistently outperforms other methods, especially in cases where there is a low correlation. This conclusion is based on a comprehensive analysis of the average ROC curve. Moreover, our results indicate that the glmnet R package emerges as the top-performing method in situations characterized by high correlation. This comprehensive comparison provides valuable insights into the performance of various regression methods under different conditions.

References

[1] A. Gelman, Parameterization and Bayesian modeling, J. Am. Stat. Assoc., 99(2004)537- 545, https://doi.org/10.1198/016214504000000458.

[2] D. Wang, W. Zhang, A. Bakhai. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression, Stat. Med., 23(2004) 3451–3467, https://doi.org/10.1002/sim.1930.

[3] S. K. Sarkar, H. Midi, Optimization Techniques for Variable Selection in Binary Logistic Regression Model Applied to Desire for Children Data, J. Math. Stat., 5(2009) 387-394 https://doi.org/10.3844/jmssp.2009.387.394.

[4] H. Wang .Ordinal Logistic Regression Analysis in Effective Teaching Practices, J Math Statis, 20(2024)13-17, https://doi.org/10.3844/jmssp.2024.13.17.

[5] B. S. Graham, Sparse network asymptotics for logistic regression under possible misspecification, Econometrica, 92 (2024)1837–1868, https://doi.org/10.3982/ECTA19051.

[6] Y. Hou, Y. Song, Transfer Learning for Logistic Regression with Differential Privacy, Axioms 13(2024)1-14, https://doi.org/10.3390/axioms13080517 .

[7] R.M. Lewis, H.S. Battey, On inference in high-dimensional logistic regression models with separated data, Biometrika 111(2024) 989–1011, https://doi.org/10.1093/biomet/asad065.

[8]A. A. Shareef, S. M. Ajeel, H.A. Hashem. Utilizing multinomial logistic regression for determining the factors influencing blood pressure, Sci. J University of Zakho , 12(2024) 367-374, https://doi.org/10.25271/sjuoz.2024.12.3.1322

[9] A. Balboa, A. Cuesta, J. González-Villa, G. Ortiz, D. Alvear, Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations, Saf. Sci. 174 (2024) 1-18, https://doi.org/10.1016/j.ssci.2024.106485.

[10] E.T. Yuniarsih, M. Salam, M.H. Jamil, A. Nixia Tenriawaru, Determinants determining the adoption of technological innovation of urban farming: employing binary logistic regression model in examining Rogers' framework, J. Open Innov. Technol. Mark. Complex. 10 (2024) 1-12, https://doi.org/10.1016/j.joitmc.2024.100307.

[11] Q. M. Abdulqader, Applying the Binary Logistic Regression Analysis on The Medical Data, Sci J University of Zakho , 5(2017) 330-334, https://doi.org/10.25271/2017.5.4.388.

[12] D. Kucharavy, R. D. Guioa, Application of Logistic Growth Curve, Procedia Eng., 131 (2015) 280 – 290, https://doi.org/10.1016/j.proeng.2015.12.390.

[13] S.Domínguez-Almendros, N.Benítez-Parejo, A.R.Gonzalez-Ramirezc, Logistic regression models, Allergologia et Immunopathologia, 39(2011)295-305, https://doi.org/10.1016/j.aller.2011.05.002.

[14] F. S. Nahm, Receiver operating characteristic curve: overview and practical use for clinicians, Korean J Anesthesiology, 75(2022)25-36, https://doi.org/10.4097/kja.21209.

[15] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw.,33(2010)1-22, https:// doi: 10.18637/jss.v033.i01.

[16] K. Yu, C. Cathy, C. Reed, D. Dunson, Bayesian variable selection in quantile regression, Statistics and Its Interface 6(2013) 261–274, https://dx.doi.org/10.4310/SII.2013.v6.n2.a9.

[17] Q. Li, R. Xi, N. Lin, Bayesian regularized quantile regression, Bayesian Analysis 5(2010) 533-556, https:// doi: 10.1214/10-BA521.

[18] M. Yuan, Y. Lin, Model Selection and Estimation in Regression with Grouped Variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2006)49–67, https://doi.org/10.1111/j.1467-9868.2005.00532.x.

[19] S.Xiong,B.Dai, J. Huling, P.Z.G. Qian, Orthogonalizing EM: A Design-Based Least SquaresAlgorithm,Technometrics,58(2016)285–293, https://doi.org/10.1080/00401706.2015.1054436.

دراسة محاكاة لطرق اختيار المتغيرات الجماعية للاستجابة الثنائية

حسين عبد الرحمن هاشم

قسم الرياضيات، كلية اللوم ، جامعة دهوك، دهوك، العراق

المستخلص

البيانات الثنائية، التي تشير إلى البيانات التي تحتوي على نتيجتين محتملتين فقط، شائعة الاستخدام في مجالات بحثية مختلفة مثل التمويل والعلوم الاجتماعية وعلم النفس والطب. يُستخدم نموذج الانحدار اللوجستي على نطاق واسع لتحليل البيانات الثنائية. ومع ذلك، من المهم التحقيق بشكل شامل في تحديد القيم المتطرفة المؤثرة والتعامل معها لضمان ملاءمة النماذج اللوجستية الثنائية الملائمة. تقدم هذه المقالة مراجعة شاملة للعديد من الأساليب اللوجستية الثنائية الجماعية المستخدمة في نموذج الانحدار وتركز على مقارنة أداء أربع طرق انحدار لوجستي محددة. تتضمن هذه الأساليب تقديرات مجموعة Lasso الثنائية، وتقديرات مجموعة mcp الثنائية، وتقديرات مجموعة scad الثنائية، ومسارات التنظيم الثنائي للنماذج الخطية المعممة عبر تقديرات الانحدار الإحداثي (glmnet). تستند المقارنات إلى دراسة محاكاة مصممة لتحديد أي من هذه الأساليب تعمل بشكل أفضل عبر جميع سيناريوهات الانحدار. من خلال هذه المراجعة والمقارنة، يمكن للباحثين اكتساب رؤى حول أكثر الأساليب فعالية لتحليل البيانات الثنائية باستخدام الانحدار اللوجستي.