TWMS J. App. and Eng. Math. V.16, N.8, 2026, pp. 1035-1041

REGULARIZED LOGISTIC REGRESSION WITH THE ATAN IN HIGH DIMENSIONAL DATA

ALI HAMEED YOUSIF[1*], ZAINAB KADHUM MEZHER[1], NAJLAA ALI DHUMAD[1], §

Abstract. Logistic regression models play an important role in analyzing binary classification problems in medical and biological data. One common method for estimating the parameters of a logistic regression model is the maximum likelihood method. However, this method does not perform well in high-dimensional settings or in the presence of multicollinearity. To overcome these problems, a penalty term is added to the objective function. In this paper, we propose a method for parameter estimation and variable selection in logistic regression models using an L_0 -like arctangent (Atan) regularization approach. The Atan penalty, which is based on the arctangent function, enjoys oracle properties. The performance of the regularized logistic regression model with the Atan penalty is compared with that of the fused lasso and the SELO penalty. Monte Carlo simulation studies are conducted under different sample sizes and different standard deviation settings. In addition, a real data set is used to evaluate the performance of the proposed method. The results show that the proposed estimator outperforms the competing methods (fused lasso and SELO) in terms of both estimation accuracy and variable selection.

Keywords: logistic regression, fused lasso, SELO, BIC, squares, tuning parameter selection.
AMS Subject Classification: 62J12, 62J07, 62H30, 62P10.

1. introduction

Logistic regression models play an important role in analyzing medical and biological data involving binary classification problems. In some situations, the dependent variable is categorical; therefore, traditional linear regression models are not suitable for handling this type of data. An alternative approach is logistic regression, which is designed to accommodate such data [1]. In logistic regression models, the dependent variable takes values in $\{0, 1\}$. Many studies have addressed logistic regression models. For example, [2] demonstrated the possibility of measuring disease progression over time using logistic

[1] College of Administration and Economic, Wasit University, Iraq
e-mail: ahameed@uowasit.edu.iq ORCID: 0000-0003-3604-2491.
e-mail: zmezher@uowasit.edu.iq; ORCID: 0009-0006-5487-1057.
e-mail: Nalaboody@uowasit.edu.iq; ORCID: 0009-0007-3166-3172
[*] Corresponding author.
§ Manuscript received: October 05, 2025; accepted: January 12, 2026.

regression. In practice, maximum likelihood estimation (MLE) is commonly used to estimate the parameters of logistic regression models in low-dimensional settings. However, MLE becomes unstable in high-dimensional data or in the presence of strong correlations among explanatory variables. To overcome these challenges, regularized regression methods based on penalty functions have been proposed. Several penalty functions have been introduced in the literature, including the Bridge penalty [3], lasso [4], SCAD [5], Elastic net [6], fused lasso [7], MCP [8], SELO [9], and the Atan penalty [10]. Many researchers have studied regularized logistic regression models (see, for example, [11, 20]). Other related approaches have also been proposed, such as LAD–Atan regression [12], and the Atan penalty for quantile regression [13].

On the other hand, some regularization methods for logistic regression perform well in both parameter estimation and variable selection. For instance, ridge regression for logistic models was studied in [11]. Iterated Lasso and adaptive Lasso approaches for logistic regression were proposed in [14], along with efficient computational algorithms.

The remainder of this paper is organized as follows. Section 2 introduces the logistic regression model. Section 3 presents regularized logistic regression using the Fused Lasso and SELO penalties. Section 4 explores the Atan penalty in logistic regression. Section 5 describes the method used for selecting the penalty parameter. Section 6 analyzes the simulation results, while Section 7 presents the analysis of the real data. Finally, Section 8 concludes the paper.

2. Logistic Regression

There are two types of logistic regression: binary logistic regression and multinomial logistic regression. In binary logistic regression, the dependent variable takes two values, 0 and 1; that is, $Y_i \in \{0, 1\}$. In contrast, in multinomial logistic regression, the dependent variable takes three or more categories. Binary logistic regression is appropriate for binary classification problems. In this case, the dependent variable follows a Bernoulli distribution with probability:

$$p\left(Y_i = 1|x_i\right) = \pi_i \tag{1}$$

$$p\left(Y_i = 0|x_i\right) = 1 - \pi_i \tag{2}$$

and

$$p\left(y_i\right) = \pi_i^{y_i}\left(1 - \pi_i\right)^{1-y_i} \ ; \ y_i = 0, 1. \tag{3}$$

The mean and the variance of dependent variable are:

$$E\left(y_i\right) = \pi_i$$

$$var\left(y_i\right) = \pi_i(1 - \pi_i)$$

Also

$$E\left[Y_i|x_i\right] = p(Y_i = 1|x_i)$$

And let $(t) = \frac{e^{-t}}{1+e^{-t}}$; then in logistic regression assume that:

$$p\left(Y_i = 1|x_i\right) = \pi\left(x_i'\beta\right) = \frac{\exp\left(x_i'\beta\right)}{1 + \exp\left(x_i'\beta\right)} \tag{4}$$

where x is explanatory variables. The log likelihood function:

$$\log L(\beta) = -\sum_{i=1}^{n} \left\{ Y_i \log \pi \left(x_i'\beta\right) + (1 - Y_i) \log \left[1 - \pi \left(x_i'\beta\right)\right] \right\} \tag{5}$$

The parameters vector β were estimated by maximum likelihood method [15].

3. Regularized Logistic Regression

Standard logistic regression cannot handle high-dimensional data, nor can it effectively deal with situations in which explanatory variables are highly correlated. A suitable alternative for addressing these issues is regularized logistic regression, which incorporates a penalty term into the objective function. The objective function of regularized logistic regression is given as follows:

$$L(\beta; \lambda) = \log L(\beta) + \lambda J(\beta) \tag{6}$$

Here, λ is the tuning parameter, $J(\cdot)$ denotes the penalty function, and $\lambda J(\beta)$ represents the penalty term. The tuning parameter λ controls the trade-off between bias and variance. The estimator of the regularized logistic regression model can be expressed as:

$$\widehat{\beta}_{RLR} = \arg\min_{\beta} \left\{ L(\beta; \lambda) \right\}. \tag{7}$$

Various penalty functions $J(\cdot)$ have been proposed in the literature. A ridge-type penalty for logistic regression was introduced by [11] where $J(\beta_j) = \sum_{j=1}^{p} {\beta_j}^2$. The fused lasso penalty, proposed by [7] is defined as follows:

$$J(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \tag{8}$$

The fused lasso for logistic regression had been proposed by [16]. The estimator of regularized logistic regression with the fused lasso is found by minimizing:

$$L(\beta; \lambda_1, \lambda_2) = \log L(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \tag{9}$$

Also CD algorithm was proposed to compute the fused lasso for logistic regression.

The SELO penalty in generalized linear models had been proposed by [17]. [9] proposed a new version of penalty called SELO penalty that is taking the following formula

$$p_{SELO}(\beta_j) = \frac{\lambda}{\log(2)} \log \left(\frac{|\beta_j|}{|\beta_j| + \epsilon} + 1 \right). \tag{10}$$

In practice the tuning parameter ϵ is equal to (0.01). Regularized logistic regression with the SELO penalty was proposed by [19]. The estimator of the regularized logistic regression model with the SELO penalty is obtained by minimizing the following objective function:

$$L(\beta; \lambda) = \log L(\beta) + \frac{\lambda}{\log(2)} \log \left(\frac{|\beta_j|}{|\beta_j| + \epsilon} + 1 \right). \tag{11}$$

4. Regularized Logistic Regression with Atan penalty

Recently, a new penalty function named Atan was proposed by [10]. This penalty resembles the L_o norm and possesses oracle properties. The Atan penalty is defined as follows:

$$P_{\lambda,\alpha}\left(|\beta|\right) = \lambda\left(\alpha + \frac{2}{\pi}\right)\arctan\left(\frac{|\beta|}{\alpha}\right). \tag{12}$$

Where $\alpha > 0$. The estimator of regularized logistic regression with Atan penalty is found by minimizing:

$$L(\beta;\lambda) = \log L + P_{\lambda,\alpha}\left(|\beta|\right) \tag{13}$$

5. Selection of penalty parameter

The selection of the penalty parameter is a crucial issue in regularized regression, as it controls the amount of shrinkage applied to the regression coefficients. One of the most commonly used criteria for selecting the penalty parameter is the Bayesian Information Criterion (BIC). The BIC formula is given by

$$BIC(\lambda) = \log\left(\widehat{\sigma}\right) + df(\lambda)\frac{\log(n)}{n}. \tag{14}$$

Where $(\widehat{\sigma})$ represent to the residual and $df(\lambda)$ represent the degree of freedom [18].

6. Simulation Study

In the simulation study, we investigate the performance of the proposed estimator and compare it with the following estimators: the fused lasso estimator (implemented using the genlasso package in R by [21]) and the SELO estimator (based on the algorithm described by [17]). The data are generated according to a logistic regression model, and the details are described as follows. The dependent variable is generated from a binomial distribution with:

$$p\left(Y_i = 1|x_i\right) = \pi\left(x_i'\beta\right) = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}. \tag{15}$$

The explanatory variables were generated from a multivariate normal distribution, with correlations defined as $Cov\left(X_i, X_j\right) = 0.6^{|i-j|}$. The true parameter vector was set as $\beta = \{3, 1.5, 0, 0, 2, 0, 0, -1\}$, and the error term was assumed to follow a standard normal distribution with mean zero and standard deviation σ, $i.e\ \varepsilon \sim N(0,\sigma)$; where $\sigma = 0.1, 0.3$. The sample sizes considered were 30, 60, and 100, with 1000 replications.

Tables 1, 2 and 3 show that the Atan estimator was more efficient and outperformed the other two estimators (SELO and Fused Lasso) in terms of parameter estimation, achieving the lowest mean squared error (MSE). The SELO estimator ranked second, while the Fused Lasso estimator performed the least effectively. In terms of variable selection, the Atan estimator again performed the best, followed by SELO and Fused Lasso. Additionally, as the sample size increased, the mean squared errors of all estimators tended to converge.

7. Application

In this part, we applied regularized logistic regression to the Iraq Breast Cancer dataset. Breast cancer is one of the most common diseases among women in Iraq. According to recent statistics, 6,000 to 7,000 new cases are registered annually, with an incidence rate of approximately 22 cases per 100,000 females per year. The data were collected from Alzahra Hospital. The dependent variable indicates whether the patient has the disease or not. The predictor variables include: age, grade, surgery, chemotherapy, tumor volume, status, immunotherapy, lymph nodes, proliferation, radiotherapy, behaviors, and extent.

TABLE 1. Results for the first experiment, with $\sigma = 0.1$.

FNR	FPR	MAPE	estimators	n
0	0.24	0.6289	fused lasso	30
0	0.22	0.3465	SELO	
0	0.13	0.3042	Atan	
0	0.2	0.4851	fused lasso	60
0	0.18	0.3064	SELO	
0	0.03	0.2341	Atan	
0	0.17	0.2777	fused lasso	100
0	0.14	0.2654	SELO	
0	0.01	0.1832	Atan	

TABLE 2. Results for the second experiment, $\sigma = 0.3$.

FNR	FPR	MAPE	estimators	n
0	0.07	0.9575	fused lasso	30
0	0.05	0.9363	SELO	
0	0.01	0.9047	Atan	
0	0.03	0.8626	fused lasso	60
0	0.03	0.8604	SELO	
0	0.01	0.8786	Atan	
0	0.03	0.8596	fused lasso	100
0	0.03	0.8556	SELO	
0	0.06	0.8267	Atan	

Applying regularized logistic regression methods to this dataset, the results show differences in variable selection. The Fused Lasso method selected eight variables: age, BMI, glucose, insulin, HOMA, leptin, adiponectin, and MCP.1. The SELO method selected seven variables: age, BMI, glucose, insulin, HOMA, adiponectin, and MCP.1. The Atan method selected six variables: age, BMI, glucose, insulin, HOMA, and MCP.1.

TABLE 3. Coefficients for the state of patient.

Atan	SELO	Fused lasso	Estimators
0.3918	0.4209	0.4729	Age
-0.0889	-0.1397	-0.1924	BMI
-0.3788	-0.4268	-0.4706	Glucose
1.1265	1.1785	1.1994	Insulin
0.2757	0.3684	0.3425	HOMA
0	0	0.2302	Leptin
0	-0.0505	-0.1231	Adiponectin
0	0	0	Resistin
0.3737	0.4397	0.4999	MCP.1

8. CONCLUSIONS

In this paper, an L_o (Atan) regularized approach was applied to the regression parameters to improve prediction accuracy in sparse high-dimensional logistic regression models.

We introduced the sparse Atan estimator, which is used for both parameter estimation and variable selection. This estimator is capable of handling high-dimensional data as well as multicollinearity among explanatory variables. Simulation studies and a real data application demonstrated the superior performance of the Atan estimator in terms of parameter estimation and variable selection, compared to the Fused Lasso and SELO methods.

References

[1] Z. Y. Algamal and M. H. Lee, (2015), Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification, *Expert Syst. Appl.*, vol. 42, no. 23, pp. 9326–9332, 2015.

[2] R. L. Prentice and R. Pyke, (1979), Logistic disease incidence models and case-control studies, *Biometrika*, vol. 66, no. 3, pp. 403–411, 1979.

[3] Ll. E. Frank and J. H. Friedman, (1993), A statistical view of some chemometrics regression tools, *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[4] R. Tibshirani, (1996), Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] J. Fan and R. Li, (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[6] H. Zou, T. Hastie, (2005), Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B (statistical Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[7] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, (2005), Sparsity and smoothness via the fused lasso, *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, vol. 67, no. 1, pp. 91-108, 2005.

[8] C.-H. Zhang (2010), Nearly unbiased variable selection under minimax concave penalty, *Ann. Stat.*, vol. 38, no. 2, pp. 894–942, 2010.

[9] L. Dicker, B. Huang, X. Lin, (2013), Variable selection and estimation with the seamless-L0 penalty, *Stat. Sin.*, vol. 23, no. 2, pp. 929–962, 2013, doi: 10.5705/ss.2011.074.

[10] Y. Wang and L. Zhu, (2016), Variable selection and parameter estimation with the Atan regularization method, *J. Probab. Stat.*, vol. 2016, 2016.

[11] A. H. Lee and M. J. Silvapulle, (1988) Ridge estimation in logistic regression, *Commun. Stat. Comput.*, vol. 17, no. 4, pp. 1231–1257, 1988.

[12] A. H. Yousif and O. A. Ali, ,(2020), Proposing Robust LAD-Atan Penalty of Regression Model Estimation for High Dimensional Data, Baghdad Sci. J., vol. 17, no. 2, pp. 550–555, 2020

[13] A. H. Yousif, W. J. Housain, (2021), Atan Regularized in Quantile Regression for High Dimensional Data, in *Journal of Physics: Conference Series*, 2021, vol. 1818, no. 1, p. 12098.

[14] J. Huang, S. Ma, C.-H. Zhang, (2008), The iterated lasso for high-dimensional logistic regression, *Univ. Iowa, Dep. Stat. Actuar. Sci.*, pp. 1–20, 2008.

[15] P. McCullagh, J. A. Nelder, (1989), Generalized linear models (Vol. 37) CRC press," 1989.

[16] S. Brightwood, (2024), Regularization Techniques in Logistic Regression, *Ladoke Akintola Univ. Technol. March*, 2024.

[17] K. A. Verchand, A. Montanari, (2024), High-dimensional logistic regression with missing data: Imputation, regularization, and universality, *arXiv Prepr. arXiv2410.01093*, 2024.

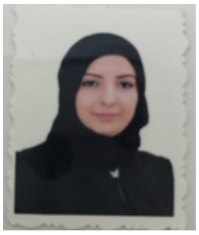
[18] S. H. Lee, D. Yu, A. H. Bachman, J. Lim, B. A. Ardekani, (2014), Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease, *J. Neurosci. Methods*, vol. 221, pp. 78–84, 2014.

[19] Z. Li, S. Wang, and X. Lin, (2012), Variable selection and estimation in generalized linear models with the seamless L0 penalty," *Can. J. Stat. Rev. Can. Stat.*, vol. 40, no. 4, p. 745, 2012.

[20] Z. Zhang, O. B. Morgan, D. L. Gillen, A. D. N. Initiative, (2025), Reweighted penalized regression for convenience samples," *Can. J. Stat.*, p. e70005, 2025.

[21] T. B. Arnold, R. J. Tibshirani, M. T. Arnold, T. ByteCompile, (2020), Package 'genlasso,' *Statistics (Ber).*, vol. 39, no. 3, pp. 1335–1371, 2020.

Dr. Ali Hameed Yousif is an Assistant Professor in the Department of Statistics at the College of Administration and Economics, Wasit University, Iraq. completed both his Master's and PhD in Statistics at University of Baghdad. His research focuses on Bayesian estimation, reliability analysis, regression models, and statistical simulation methods. Dr. Yousif has published numerous scientific studies in statistical modeling and estimation techniques and authored the book Theory of Statistical Decisions and its Applications. He teaches several undergraduate and postgraduate courses including Mathematical Statistics, Decision Theory, Operations Research, and Design and Analysis of Experiments.



Dr. Zainab Kazem Mezher is an Lecturer in the Department of Statistics at the College of Administration and Economics, Wasit University, Iraq. She specializes in Applied Statistics and earned her PhD in Statistics from University of Karbala in 2023. Her research focuses on statistical modeling, reliability analysis, and regression methods. She has published research in peer-reviewed academic journals and actively participates in international scientific conferences. Dr. Mezher has taught several university courses including Operations Research, Time Series Analysis, Financial Mathematics, and Sampling Methods.

Ms. Najlaa Ali Dhumad is an Assistant Lecturer in the Department of Statistics at the College of Administration and Economics, Wasit University, Iraq. She specializes in Applied Statistics and obtained her Bachelor's and Master's degrees in Statistics from Wasit University. Her research interests include statistical estimation methods and probability distributions, and she has several research papers presented at scientific conferences. Ms. Dhumad currently works as the Coordinator of the Statistics Department and previously served in the Studies and Planning Department at the university presidency.