Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

The Effect of Divisive Analysis Clustering Technique on Goodness-of-Fit Test for Multinomial Logistic Regression

Hamzah Abdul Hamid[1,*], Yap Bee Wah[2], Khatijahhusna Abdul Rani[1], Xian Jin Xie[3]

[1] Institute of Engineering Mathematics, Universiti Malaysia Perlis, Pauh Putra Main Campus, 02600 Arau, Perlis, Malaysia
[2] School of Information Technology, UNITAR International University, Jalan SS6/3, 47301 Petaling Jaya, Selangor, Malaysia
[3] Division of Biostatistics and Computational Biology, Iowa Institute for Oral Health Research, University of Iowa College of Dentistry, Iowa City, Iowa, USA

ARTICLE INFO

Article history:
Received 4 November 2023
Received in revised form 4 March 2024
Accepted 15 June 2024
Available online 15 July 2024

Keywords:
Simulation; multinomial logistic regression; goodness-of-fit test

ABSTRACT

The relationship between a categorical dependent variable and independent variable(s) is usually modelled using the logistic regression method. There are three types of logistic regression: binary, multinomial, and ordinal. When there are two categories of dependent variable, binary logistic regression is used while when there are more than two nominal categories of dependent variable, multinomial logistic regression is employed. Ordinal logistic regression is used when the dependent variable contains more than two ordinal categories. All regression models should be checked after being fitted to the data to see whether it matches the data or not. For multinomial logistic regression, there are a number of goodness-of-fit tests proposed that can be used to evaluate the fit of the model. One of the proposed tests is based on a clustering partitioning strategy. However, the proposed test only considered agglomerative nesting (AGNES) hierarchical clustering technique, which is Ward's to group the data. The performance of the test using the divisive analysis (DIANA) hierarchical clustering technique remains unknown. Thus, this study attempts to examine the power of the test using the divisive analysis clustering technique. A simulation technique was used to evaluate the performance of the test. The results showed that the test using DIANA clustering technique has controlled type I error and the mean are close to hypothesized values. It also has almost equivalent power with the test using Ward's clustering technique in detecting the omission of a quadratic term. However, the test using Ward's clustering technique shows noticeably higher power in detecting omission of an interaction term.

1. Introduction

Regression models are usually used to describe the relationship between independent and dependent variables. When the dependent variable is categorical, the logistic regression method is utilized. The logistic regression can be divided into three types, which are binary, multinomial, and ordinal. If there are only two categories of dependent variable, the application of binary logistic

* *Corresponding author.*
E-mail address: hamzahhamid@unimap.edu.my

https://doi.org/10.37934/araset.48.2.3948

regression is appropriate. When there are more than two categories of dependent variable, the multinomial logistic regression should be applied. However, when there are more than two ordinal categories of dependent variable, the ordinal logistic regression is more suitable to be applied [1].

The logistic regression is applicable in various study areas such as epidemiology [2-5], medical [6-7], psychology [8-10], finance [11-13], safety science [14], computer security [15-16], social [17-18] and hydrology [19-21]. The unneeded assumptions of normality, linearity, or homoscedasticity of covariates make this type of regression popular among researchers. Discriminant function analysis is an alternative to logistic regression. However, due to the assumptions of multivariate normality, this method is not widely employed [22].

Once the regression model is obtained, the model fit is normally tested by using a goodness-of-fit test. There are several tests available to assess the fit of the model for multinomial logistic regression. Usually, the deviance and Pearson chi-square test are employed. However, since an asymptotic chi-square distribution of the test statistic does not hold when single counts are dominating for the contingency table, neither the Pearson chi-square test nor the deviance test are relevant if the model contains one or more continuous covariate [22]. Hamid *et al.*, [1] proposed a test adapted from Xie *et al.*, [22] which is based on clustering partitioning strategy in the covariate space. There are a lot of clustering techniques available. Since non-hierarchical clustering approaches are so sensitive to initial partitions [22], Xie and Bian, [23] suggested hierarchical clustering techniques to be used for the test. Hamid et al., only considered AGNES (Agglomerative Nesting) clustering technique, which is Ward's hierarchical clustering technique in their study. Agglomerative clustering builds a hierarchy by merging small clusters into larger ones, while divisive clustering builds a hierarchy by dividing a large cluster into smaller ones. However, the performance of the test using divisive clustering is still unknown. Thus, the main objective of this study is to investigate the power of the goodness-of-fit test using divisive clustering technique, which is known as Divisive Analysis (DIANA).

2. Methodology

2.1 Multinomial Logistic Regression

One of the logistic regression types is multinomial logistic regression (MLR). In order to develop MLR, let Y be an outcome variable with possible value of $m(0,1,\ldots,m-1)$ and the reference category is Y=0. Let the independent variable $x=(x_1,x_2,\ldots,x_k)$, the conditional probabilities of each outcome category can be defined as [24]:

$$P(Y=0\mid x)=\frac{1}{1+\mathrm{e}^{\mathrm{g}_1(x)+\ldots+\mathrm{g}_{\mathrm{c-1}}(x)}}$$
$$P(Y=1\mid x)=\frac{\mathrm{e}^{\mathrm{g}_1(x)}}{1+\mathrm{e}^{\mathrm{g}_1(x)+\ldots+\mathrm{g}_{\mathrm{c-1}}(x)}} \tag{1}$$
$$P(Y=\mathrm{c}-1\mid x)=\frac{\mathrm{e}^{\mathrm{g}_{\mathrm{c-1}}(x)}}{1+\mathrm{e}^{\mathrm{g}_1(x)+\ldots+\mathrm{g}_{\mathrm{c-1}}(x)}}$$

Thus, the logit function of category j versus the baseline category can be expressed as:

$$g_j(x)=\ln\ \frac{\mathrm{P}(\mathrm{Y}=\mathrm{j}\mid x)}{\mathrm{P}(\mathrm{Y}=0\mid x)}\ =\beta_{\mathrm{j0}}+\beta_{\mathrm{j1}}x_1+\ldots+\beta_{\mathrm{jp}}x_p \quad \text{For } j=1,2,\ldots,c-1 \tag{2}$$

Then, the parameter estimates, $\hat{\beta}$ is obtained using the maximum likelihood method.

2.2 Goodness-of-Fit Test

In this study, the power of the test based on the clustering partitioning strategy proposed by Xie *et al*., [22] is investigated using the DIANA hierarchical clustering technique.

$$\chi^2_{p*G} = \sum_{g=1}^{G}\sum_{j=1}^{J}\frac{\left(O_{gj} \quad E_{gj}\right)^2}{E_{gj}} \tag{3}$$

where g is the number of clusters while j refers to the response categories, O_{gj} is the number of responses with Y=j in the g-th cluster while E_{gj} is the model expected count for response j in the g-th cluster. The degrees of freedom are calculated by (*G*-2) (*J*-1) and the number of clusters used in this study is *G*=10.

2.3 Simulation Design

This study applied simulation technique in order to evaluate the power of the test. The simulation is carried out using R. The design and steps for simulation study follow the work by Fagerland *et al*., [25]:

(i) The covariate (x) is generated.
(ii) The multinomial logit equation is fitted to the simulated data.
(iii) The probability is calculated.
(iv) The random data, u is generated from a uniform distribution, U(0,1).
(v) The outcomes for multinomial logistic regression is generated using the rule; (i) y=2 if $u > \pi_0 + \pi_1$, (ii) y=1 if $u < \pi_0 + \pi_1$ and $u > \pi_0$, and (iii) y=0 otherwise.
(vi) The multinomial logistic regression model is fitted.
(vii) The test statistic is calculated.
(viii) The rejection rates for each model is computed.

The factors that are set to vary are the type of model departure from the underlying true model (omission of a quadratic term, omission of interaction term), the relative importance of the missing term (magnitude of β), the distribution of the omitted predictor, and sample size. In order to represent small and large sample sizes, there are two sample sizes used which are 100 and 400. The number of replication is 10,000. The distributions for covariate used are N(0,3), N(0,1), U(-6,6), U(-3,3), U(-1,1) and $\chi^2(4)$.

In the simulation study design for the true model, we examine the test statistics using the goodness-of-fit test when the model is true. For this model, we only simulate one continuous covariate with a dependent variable with 3 categories. The coefficients for each of the five models are provided in Table 1. For each model, six different covariate distributions were used as summarized in Table 2. Each model was used to reflect a variety of situations that might arise in practice.

Table 1
Coefficients for the five models for simulation study

Model	β_{10}	β_{11}	β_{20}	β_{21}
1	-2.10	-0.35	-1.90	-0.21
2	1.00	0.20	1.80	0.30
3	0.00	1.00	2.00	0.50
4	1.40	0.25	0.20	0.30
5	0.50	-0.20	-0.20	0.10

Table 2
Covariate distributions

Setting	Covariate distribution
1	U(-6,6)
2	U(-3,3)
3	U(-1,1)
4	N(0,3)
5	N(0,1)
6	χ^2 (4)

Next is the design for omission of a quadratic term. The coefficient for the model used is as in Model 1 in Table 1. There are four different distributions of covariate used as summarized in Table 3. In this design, the response variable was generated using the logit function:

$$g_j(x) = \beta_{j0} + \beta_{j1}x + \beta_{j2}x^2 \tag{4}$$

However, when we fit the model, we only include the first two terms ($\beta_{j2} = 0$). If the test is able to reject the model, it means the test is good at detecting the omission of the quadratic term. The coefficient of β_{j2} were 0.01, 0.05, 0.10, 0.20, 0.30, 0.40 and 0.50.

The final one is the design for omission of an interaction term. The power of the test is examined in detecting the departure of the true model by considering the omission of the interaction term. The coefficient used is as in Model 1 shown in Table 1. In this design, the response variable was generated using the logit function:

$$g_j(x) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \beta_{j3}x_1x_2 \tag{5}$$

However, when the model is fitted, we only include the first three terms $(\beta_{j3} = 0)$. The power of the test is evaluated by observing the ability of the test to reject the model. There are three combinations of β_{j2} values (0.2, 0.6, 1.0) and four β_{j3} values (0.2, 0.4, 0.6, 1.0) ~~were use~~d to investigate the effect of increasing levels of interaction of the continuous term.

3. Results

The simulation result for the rejection rates of the true model is summarized in Table 3. The result shows that the rejection rates for both clustering techniques are close to the nominal 5 percent. As expected, the result is better when the sample size is increased. The results for both tests are better for sample size of 400 compared to 100. The rejection rates for both tests are noticeably affected for skewed distribution when the sample size is smaller (*n*=100).

Table 3
Rejection rates at 5 percent nominal level

	n=100		*n*=400	
	A	B	A	B
U(-6,6)				
Model 1	4.91	5.27	4.45	4.97
Model 2	4.87	4.93	5.07	4.76
Model 3	6.87	6.91	7.67	7.08
Model 4	4.33	4.93	4.84	5.06
Model 5	4.45	4.55	4.64	4.95
U(-3,3)				
Model 1	4.34	4.62	4.11	4.44
Model 2	4.60	5.06	4.92	4.74
Model 3	4.73	5.00	5.16	5.11
Model 4	4.40	4.63	4.94	5.13
Model 5	4.54	4.69	4.62	4.67
U(-1,1)				
Model 1	4.51	4.69	4.34	4.51
Model 2	4.68	4.90	5.05	4.69
Model 3	4.64	4.90	5.29	4.85
Model 4	4.38	4.88	4.82	5.01
Model 5	4.55	4.54	5.07	5.07
N(0,3)				
Model 1	5.56	6.35	6.03	6.67
Model 2	5.16	5.98	5.84	6.56
Model 3	5.57	5.43	6.61	6.61
Model 4	5.82	6.52	5.52	6.61
Model 5	5.86	6.43	5.66	6.65
N(0,1)				
Model 1	5.28	6.07	4.91	6.07
Model 2	5.32	5.87	5.17	5.76
Model 3	5.72	6.71	5.94	7.10
Model 4	5.22	5.61	5.00	5.51
Model 5	5.15	5.13	5.25	4.81
$\chi^2(4)$				
Model 1	1.50	1.66	4.32	4.61
Model 2	2.89	3.18	4.79	5.01
Model 3	1.12	1.37	3.56	3.43
Model 4	3.15	3.38	4.75	5.08
Model 5	5.26	5.96	5.68	6.52

Note: A = Test using Ward's Clustering Technique
B = Test using Diana Clustering Technique

The adherence of sampling distribution to the hypothesized null distribution is studied by examining the sample moments. The mean values of the test statistic using different clustering techniques are summarized in Table 4. Based on the results, the mean values are close to the hypothesized value, (G-2)*(J-1), which is 16, for both clustering techniques. However, the values for the test using DIANA clustering technique are more affected for the models with skewed covariate.

Table 4
Mean values for true models

	n=100		*n*=400	
	A (16)	B (16)	A (16)	B (16)
U(-6,6)				
Model 1	15.888	16.008	15.930	15.995
Model 2	16.161	16.212	16.139	16.090
Model 3	14.763	14.841	15.913	15.784
Model 4	16.246	16.321	16.097	16.121
Model 5	16.295	16.353	16.022	16.014
U(-3,3)				
Model 1	16.211	16.243	15.907	16.001
Model 2	16.311	16.385	16.140	16.087
Model 3	15.894	15.933	16.144	16.078
Model 4	16.312	16.374	16.123	16.133
Model 5	16.325	16.385	16.074	16.059
U(-1,1)				
Model 1	16.325	16.374	15.94	16.045
Model 2	16.399	16.481	16.203	16.057
Model 3	16.278	16.300	16.234	16.130
Model 4	16.352	16.398	16.085	16.086
Model 5	16.315	16.368	16.102	16.091
N(0,3)				
Model 1	15.525	15.436	16.003	15.820
Model 2	16.040	15.965	16.097	16.087
Model 3	15.168	14.982	15.365	15.288
Model 4	16.136	16.103	16.085	16.026
Model 5	16.330	16.324	16.012	15.929
N(0,1)				
Model 1	15.965	15.794	16.012	15.968
Model 2	16.212	16.265	16.058	16.059
Model 3	15.894	15.803	16.025	16.042
Model 4	16.269	16.292	16.123	16.079
Model 5	16.472	16.510	16.152	16.103
$\chi^2(4)$				
Model 1	11.383	10.332	13.230	11.898
Model 2	13.488	12.856	14.455	13.809
Model 3	10.588	9.828	11.643	10.472
Model 4	14.542	14.072	15.212	15.062
Model 5	15.921	15.884	15.901	15.740

Note: A = Test using Ward's Clustering Technique
B = Test using Diana Clustering Technique

The results for the power of the test in detecting the omission of quadratic terms using different clustering techniques are shown in Table 5. Only two distributions of covariate were considered in this simulation study to represent normal (N(0,1)) and non-normal ($\chi^2(4)$) distribution of the covariate. Based on the result, the power is almost the same for both clustering techniques. However, the test using DIANA clustering technique is slightly better for the model with normal distribution covariate, while the test using Ward's clustering technique has slightly higher power for the model with non-normal distribution of covariate.

Table 5
Rejection rates at 5 percent nominal level

	Omission of a quadratic term			
	n=100		*n*=400	
β_{j2}	A	B	A	B
N(0,3)				
0.01	8.73	10.01	11.97	13.68
0.05	40.26	43.70	87.82	90.81
0.10	80.47	82.14	100	100
0.20	98.10	98.08	100	100
0.30	99.64	99.57	100	100
0.40	99.82	99.77	100	100
0.50	99.88	99.82	100	100
χ^2(4)				
0.01	3.46	3.54	11.21	13.55
0.05	9.99	9.86	37.09	33.50
0.10	8.45	7.52	31.82	24.29
0.20	4.96	3.77	16.44	11.18
0.30	3.34	2.46	8.67	5.36
0.40	2.48	1.80	4.81	3.21
0.50	1.60	1.30	3.11	2.26

Note: A = Test using Ward's Clustering Technique
B = Test using Diana Clustering Technique

Table 6 summarizes the simulation results of the rejection rate for omission of a continuous interaction term. There are also only two distributions of covariate are considered in this simulation study to represent normal (N(0,1)) and non-normal (χ^2(4)) distribution of the covariate. The test using Ward's clustering technique has noticeably higher power in detecting omission of interaction term compared to the test using the DIANA clustering technique.

Table 6
Rejection rates at 5 percent nominal level

		Omission of a continuous interaction term			
		n=100		*n*=400	
β_{j2}	β_{j3}	A	B	A	B
N(0,3)					
0.2	0.2	9.48	9.31	23.19	12.79
	0.4	24.38	13.27	76.13	23.68
	0.6	42.74	16.22	95.71	31.17
	1.0	68.30	17.57	99.77	38.64
0.6	0.2	8.16	8.55	21.15	13.35
	0.4	22.56	12.31	74.44	21.4
	0.6	40.82	14.67	95.42	27.3
	1.0	66.99	16.62	99.72	33.92
1.0	0.2	6.76	7.39	15.96	11.74
	0.4	19.18	11.16	66.91	18.34
	0.6	37.50	13.37	93.14	23.46
	1.0	65.01	15.00	99.52	30.25
χ^2 (4)					
0.2	0.2	3.17	2.40	6.75	5.79
	0.4	6.36	1.70	15.42	2.15
	0.6	8.29	0.96	22.43	1.15
	1.0	9.49	0.63	26.49	0.64
0.6	0.2	3.51	2.12	5.25	3.89
	0.4	5.79	1.06	11.02	1.79
	0.6	6.90	0.26	15.57	0.99
	1.0	8.08	0.69	19.15	0.76
1.0	0.2	3.97	1.42	4.58	2.58
	0.4	5.24	0.98	8.38	1.50
	0.6	6.38	0.81	11.62	1.05
	1.0	7.53	0.62	14.15	0.71

Note: A = Test using Ward's Clustering Technique
B = Test using Diana Clustering Technique

4. Conclusion

The power of the goodness-of-fit test based on clustering partitioning strategy using Ward's and DIANA clustering techniques was compared using simulation design as in Fagerland *et al*., [25]. The results show that the test using both clustering techniques has almost equivalent power for the true model and in detecting the omission of a quadratic term. The mean values also are close to the hypothesized values for both clustering techniques. However, the test using Ward's clustering technique shows noticeably higher power in detecting omission of an interaction term compared to the DIANA clustering technique. Thus, it can be concluded that the DIANA clustering technique is unable to improve the power of the goodness-of-fit test based on the clustering partitioning strategy proposed by Xie *et al*., [22]. Future studies can address the effect of distance measures on the test.

Acknowledgement

This research was not funded by any grant.

References

[1] Abdul Hamid, Hamzah, Yap Bee Wah, Xian-Jin Xie, and Ong Seng Huat. "Investigating the power of goodness-of-fit tests for multinomial logistic regression." *Communications in Statistics-Simulation and Computation* 47, no. 4 (2018): 1039-1055. https://doi.org/10.1080/03610918.2017.1303727

[2] Sakly, Houneida, Mourad Said, Rachid Jennane, and Moncef Tagina. "Epidemiological Study of Cardiopathies and Valvulopathies using Binary Logistic Regression." *Journal of Data, Information and Management* 5, no. 3 (2023): 151-163. https://doi.org/10.1007/s42488-023-00094-w

[3] Nusinovici, Simon, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. "Logistic regression was as good as machine learning for predicting major chronic diseases." *Journal of clinical epidemiology* 122 (2020): 56-69. https://doi.org/10.1016/j.jclinepi.2020.03.002

[4] Kim, Taekyoung, Sang D. Choi, and Shuping Xiong. "Epidemiology of fall and its socioeconomic risk factors in community-dwelling Korean elderly." *PLoS one* 15, no. 6 (2020): e0234787. https://doi.org/10.1371/journal.pone.0234787

[5] Muhammad, L. J., Ebrahem A. Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty, and Ibrahim Alh Mohammed. "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset." *SN computer science* 2 (2021): 1-13. https://doi.org/10.1007/s42979-020-00394-7

[6] Boateng, Ernest Yeboah, and Daniel A. Abaye. "A review of the logistic regression model with emphasis on medical research." *Journal of data analysis and information processing* 7, no. 4 (2019): 190-207. https://doi.org/10.4236/jdaip.2019.74012

[7] Zeqo, Lorena Margo, Eljona Milo Tasho, and Juliana Karanxha. "Bootstrapping the coefficients of multiple logistic regression model in medicine data." *Asian-European Journal of Mathematics* 15, no. 10 (2022): 2250248. https://doi.org/10.1142/S1793557122502485

[8] Stewart, Genéa, Akihito Kamata, Rona Miles, Evan Grandoit, Faigy Mandelbaum, Crystal Quinn, and Laura Rabin. "Predicting mental health help seeking orientations among diverse Undergraduates: An ordinal logistic regression analysis." *Journal of affective disorders* 257 (2019): 271-280. https://doi.org/10.1016/j.jad.2019.07.058

[9] Wang, Wei, Wenqin Song, Zhongyuan Xia, Yuhong He, Linghua Tang, Jiabao Hou, and Shaoqing Lei. "Sleep disturbance and psychological profiles of medical staff and non-medical staff during the early outbreak of COVID-19 in Hubei Province, China." *Frontiers in psychiatry* 11 (2020): 733. https://doi.org/10.3389/fpsyt.2020.00733

[10] Saddik, Basema, Amal Hussein, Ammar Albanna, Iffat Elbarazi, Arwa Al-Shujairi, Mohamad-Hani Temsah, Fatemeh Saheb Sharif-Askari, Emmanuel Stip, Qutayba Hamid, and Rabih Halwani. "The psychological impact of the COVID-19 pandemic on adults and children in the United Arab Emirates: a nationwide cross-sectional study." *BMC psychiatry* 21, no. 1 (2021): 1-18. https://doi.org/10.1186/s12888-021-03213-2

[11] De Bortoli, Daiane, Newton da Costa Jr, Marco Goulart, and Jéssica Campara. "Personality traits and investor profile analysis: A behavioral finance study." *PloS one* 14, no. 3 (2019): e0214062. https://doi.org/10.1371/journal.pone.0214062

[12] Finishtya, Florentina Cindy. "The role of cash flow of operational, profitability, and financial leverage in predicting financial distress on manufacturing company in Indonesia." *Jurnal Aplikasi Manajemen* 17, no. 1 (2019): 110-117. https://doi.org/10.21776/ub.jam.2019.017.01.12

[13] Nguyen, Huan Huu, Vu Minh Ngo, and Anh Nguyen Tram Tran. "Financial performances, entrepreneurial factors and coping strategy to survive in the COVID-19 pandemic: case of Vietnam." *Research in International Business and Finance* 56 (2021): 101380. https://doi.org/10.1016/j.ribaf.2021.101380

[14] Bhattacharjee, Pushparenu, Vidyut Dey, and U. K. Mandal. "Risk assessment by failure mode and effects analysis (FMEA) using an interval number based logistic regression model." *Safety Science* 132 (2020): 104967. https://doi.org/10.1016/j.ssci.2020.104967

[15] Chiramdasu, Rupa, Gautam Srivastava, Sweta Bhattacharya, Praveen Kumar Reddy, and Thippa Reddy Gadekallu. "Malicious url detection using logistic regression." In *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, pp. 1-6. IEEE, 2021. https://doi.org/10.1109/COINS51742.2021.9524269

[16] Fan, Yongkai, Jianrong Bai, Xia Lei, Yuqing Zhang, Bin Zhang, Kuan-Ching Li, and Gang Tan. "Privacy preserving based logistic regression on big data." *Journal of network and computer applications* 171 (2020): 102769. https://doi.org/10.1016/j.jnca.2020.102769

[17] Liu, Bin, Tianshu Zhao, and Ranxia Zhang. "Prediction of Private Car Ownership in China Based on the Improved PCA-Logistic Model." In *5th International Conference on Social Sciences and Economic Development (ICSSED 2020)*, pp. 353-358. Atlantis Press, 2020. https://doi.org/10.2991/assehr.k.200331.074

[18] Chandler, Jesse, Cheskie Rosenzweig, Aaron J. Moss, Jonathan Robinson, and Leib Litman. "Online panels in social science research: Expanding sampling methods beyond Mechanical Turk." *Behavior research methods* 51 (2019): 2022-2038. https://doi.org/10.3758/s13428-019-01273-7
[19] Lee, Jae-Yeong, and Ji-Sung Kim. "Detecting areas vulnerable to flooding using hydrological-topographic factors and logistic regression." *Applied Sciences* 11, no. 12 (2021): 5652. https://doi.org/10.3390/app11125652
[20] Sujatha, Evangelin Ramani, and Venkataramana Sridhar. "Landslide susceptibility analysis: A logistic regression model case study in Coonoor, India." *Hydrology* 8, no. 1 (2021): 41. https://doi.org/10.3390/hydrology8010041
[21] Oñate-Valdivieso, Fernando, Arianna Oñate-Paladines, and Milton Collaguazo. "Spatiotemporal Dynamics of Soil Impermeability and Its Impact on the Hydrology of An Urban Basin." *Land* 11, no. 2 (2022): 250. https://doi.org/10.3390/land11020250
[22] Xie, X. J., Pendergast, J., & Clarke, W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Computational Statistics & Data Analysis* 52(5), 2703-2713. https://doi.org/10.1016/j.csda.2007.09.027
[23] Xie, Xian-Jin, and Aihua Bian. "Increasing the power: goodness-of-fit tests for ordinal response regression models with continuous covariates." JP Journal of Biostatistics 3, no. 3 (2009): 225-246.
[24] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013. https://doi.org/10.1002/9781118548387
[25] Fagerland, Morten W., David W. Hosmer, and Anna M. Bofin. "Multinomial goodness-of-fit tests for logistic regression models." *Statistics in medicine* 27, no. 21 (2008): 4238-4253. https://doi.org/10.1002/sim.3202