REVIEW ARTICLE
pISSN 0976 3325 | eISSN 2229 6816
Open Access Article (CC BY-SA)
www.njcmindia.org
DOI: 10.55489/njcm.134202222

A Review on Logistic Regression in Medical Research

Nihar Ranjan Panda[1], Jitendra Kumar Pati[2], Jatindra Nath Mohanty[3], Ruchi Bhuyan[4]

[1]IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, Odisha
[2]CV Raman global university, Bhubaneswar
[3]IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, Odisha
[4]IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, Odisha

ABSTRACT

In today's scenarios many healthcare decisions are being taken by predictive modelling and machine learning techniques. With this review, we focused on logistic regression model, a kind of predictive modelling used in machine learning, and how healthcare researchers take decisions by the help of predictive modelling. For a better data analysis in healthcare, we need to understand the concept of logistic regression as well as others terms, which are linked with it. so that we can clearly understand the concept behind it and implement in medical research. In this review we worked on an example and illustrated how to perform logistic regression using R programming language. The aim of this paper is to understand logistic regression in healthcare and implement it for decision making.

Key words: Logistic regression, Odds ratio, R programming

INTRODUCTION

Logistic regression (LR) modelling is now an popular statistical tool in healthcare analysis and medical research, from last three decades.[1] its origin was established in 19th century[2], It is the most common statistical method to predict the dichotomous dependent variable using one or more than one independent variables.[3,4,5] The French mathematician Pierre François Verhulst invented logistic function in 19th century for the description of growth of human populations.[6] In between 1838 to 1847 Verhulst published his suggestions which were edited by Quetelet.[7] Pearl and Reed discovered a new logistic function in 1920 in USA for a study of the population growth.[8] Logistic regression (LR) is used when the dependent variable in the study is dichotomous and contains factor like decision making (yes or no), disease state (diseased or healthy).Some complex form of logistic regression can be solved using multinomial regression analysis, where predict variable takes more than two categories.[9,10] In order to fit a model ,certain assumptions are made. but in case of logistic regression model, it does not assume a kind of linear relationship between the dependent variable and independent variables.[10] In most of the research analysis logistic regression can be used to deduce how the independent variables are affecting the dependent variable in a particular study.[11,14,15] for example we may consider in a healthcare the patient will die or survive after an intervention.[12,13] whether a new born baby will be overweight or underweight. So, for this kind of prediction purpose the LR model can be used.

MATERIALS AND METHODS

2.1. The Logistic Regression Model

The logistic equation is given by

$$(Y)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1)}} = \frac{e^{(\beta_0+\beta_1 x_1)}}{1+e^{(\beta_0+\beta_1 x_1)}} \text{ ....(1)}$$

When there are several predictors in the data, the equation becomes:

How to cite this article: Panda NR, Kumar PJ, Mohanty JN, Bhuyan R. A Review on Logistic Regression in Medical Research. Natl J Community Med 2022;13(4):265-270. DOI: 10.55489/njcm.134202222

Financial Support: None Declared

Conflict of Interest: None declared

Date of Submission: 04-02-2022
Date of Acceptance: 08-03-2022
Date of Publication: 30-04-2022

Correspondence: Nihar Ranjan Panda (Email: niharpanda1994@gmail.com)

$$P(y) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots\beta_i x_i)}} \text{……………………}(2)$$

$$=\frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots\beta_i x_i)}} \text{……………………………}(3)$$

Where

$\beta_0, \beta_1, \beta_2, \beta_i$ *are the coefficient of regression equation.* And x_1, x_2, x_3, x_i are the independent variables in the given equation.

a "link function" that links the Dependent variable and independent variable is

$$\ln\left[\frac{\pi_i}{1-\pi_i}\right] = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_i x_i)}} \text{……………………..}(4)$$

Now we can write the above equation as

$$\left[\frac{\pi_i}{1-\pi_i}\right] = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_i x_i)} \text{………………………..}(5)$$

Now solving equation (5) we can find out our required equation i.e equation (3)

2.2 Introduction to Logistic Curve

In logistic regression model the outcome variable takes the value 0 and 1 and the predicted values falls within the range 0 and 1, as we know the total probability is always 1. logistic regression uses the logistic curve to find out relationship between the independent and outcome variable. The probability follows 0, At very low values of the independent variable, never reaches 0.and if the independent variable increases the logistic regression curve approaches towards 1. But never equal to 1.

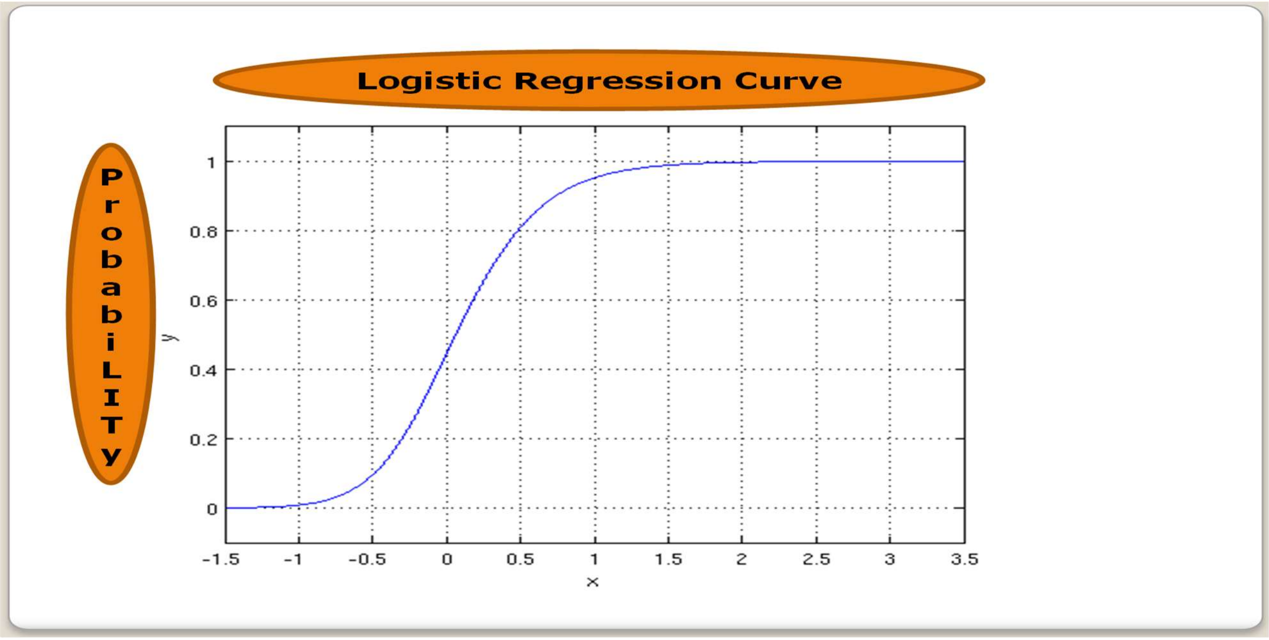


Fig 1: The curve of logistic regression

2.3 Converting a Probability into Odds and Logit Values

The estimated values, which are given by logistic regression model, do not fall outside the range of 0 and 1. we can obtain this by two step methods. firstly, we can restate the probability as odds, which can be defined as the ratio of probability of occurring with probability of non-occurring. for example, for example, if a doctor has a probability of 0.7 to succeed in a certain operation, then the odds of success are 0.7/(1-0.7).

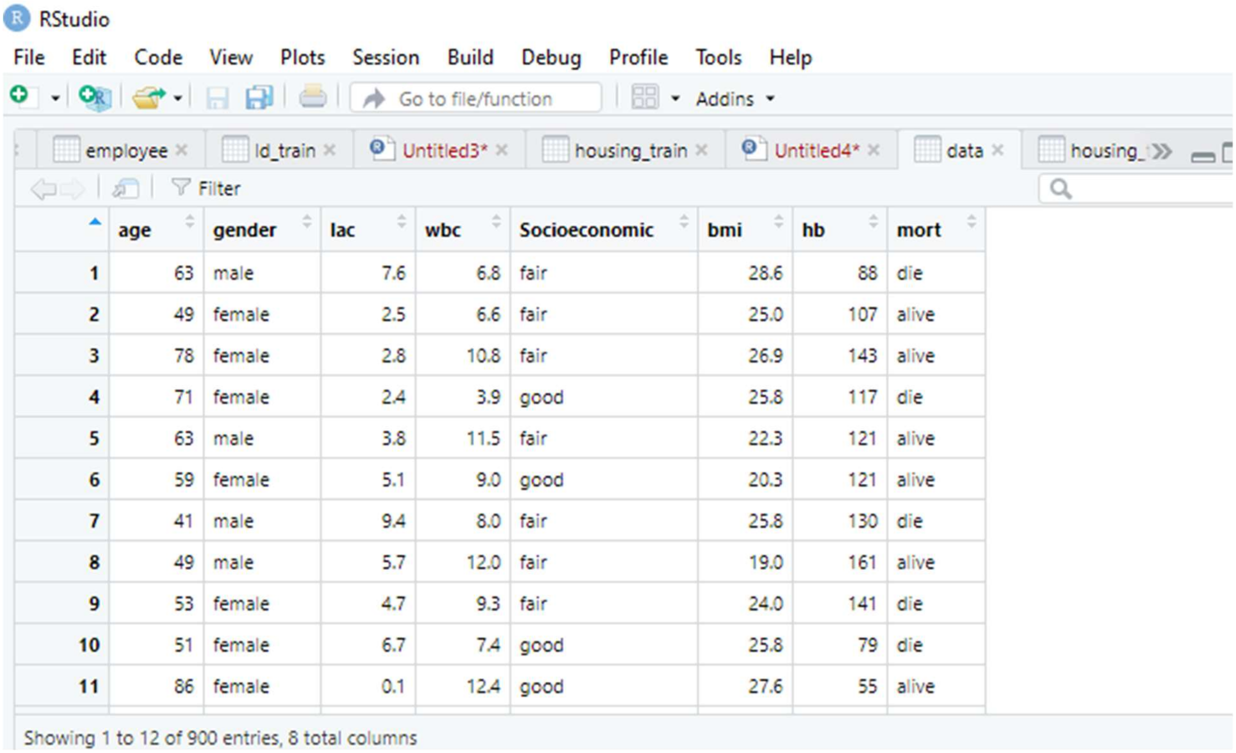
3. Introduction to R programming Language

According to the survey in 2021, R programming language is a widely used and most preferable statistical language. Now a day's R programming language is an emerging programming language. Many health care researchers are using this for data management, data cleaning, data preparation and data analysis. It has many statistical functions, which are very easy to use. it contains a large library, which enables the researchers to prefer this language. Many complex graphics, bivariate plots, 3D plots can be drawn using this language. so here in this paper we used the R programming language to implement a working example of logistic regression.

3.1 Working example of Logistic Regression using R Studio

Here in this working example, to show the application of logistic regression, we have created some variables using R studio. Our objective was to show the factors affected mortality using modelling. in this example, we have considered 7 independent variables to predict the mortality of the patient.

```
set.seed (999)
#age of the patients
age<-abs (round (rnorm(n=900, mean=67,sd=14)))
#socioeconomic status of the patients
Socioeconomic<-factor (rbinom(n=900, size=1,prob=0.6),labels=c("good","fair"))
#bmi of the patients
bmi<-abs (round (rnorm(n=900, mean=25,sd=2.5),1))
#lac of the patients
lac<-abs(round(rnorm(n=900,mean=5,sd=3),1))
# gender of the patients
gender<-factor(rbinom(n=900,size=1,prob=0.6),labels=c("male","female"))
# wbc scores of the patients
wbc<-abs(round(rnorm(n=900,mean=10,sd=3),1))
# hb of the patients
hb<-abs(round(rnorm(n=900,mean=120,sd=40)))
# Calculating z value
z<-0.1*age-0.02*hb+lac-10
pr = 1/(1+exp(-z))
y = rbinom(900,1,pr)
#mortality of the patients
mort<-factor(rbinom(900,1,pr),labels=c("alive","die"))
#creating data frame
data<-data.frame(age,gender,lac,wbc,Socioeconomic,bmi,hb,mort)
```



	age	gender	lac	wbc	Socioeconomic	bmi	hb	mort
1	63	male	7.6	6.8	fair	28.6	88	die
2	49	female	2.5	6.6	fair	25.0	107	alive
3	78	female	2.8	10.8	fair	26.9	143	alive
4	71	female	2.4	3.9	good	25.8	117	die
5	63	male	3.8	11.5	fair	22.3	121	alive
6	59	female	5.1	9.0	good	20.3	121	alive
7	41	male	9.4	8.0	fair	25.8	130	die
8	49	male	5.7	12.0	fair	19.0	161	alive
9	53	female	4.7	9.3	fair	24.0	141	die
10	51	female	6.7	7.4	good	25.8	79	die
11	86	female	0.1	12.4	good	27.6	55	alive

Showing 1 to 12 of 900 entries, 8 total columns

Fig 2: Preparation of data frame using R studio

3.2 Step one: Univariate data analysis using R

To find out the unadjusted link between outcome and independent variables we perform an univariate data analysis of our data. We separately included each of the independent variable to our model.

```
# univariate analysis for age .........
univariate.age<-glm(mort~age, family = binomial)
 summary(univariate.age)
# univariate analysis forSocioeconomic status...........
 univariate.Socioeconomic<-glm(mort~Socioeconomic, family = binomial)
 summary (univariate. Socioeconomic)
```

Table 1 Results of Univariate data analysis

Variable	Coefficient	Standard error	Z value	P value
Age	0.042012	0.005505	7.631	<0.000
Socioeconomic	-0.1807	0.1381	-1.308	0.191
bmi	0.01263	0.02730	0.463	<0.000
lac	0.81361	0.05343	15.23	<0.000
gender	-0.23938	0.13882	-1.724	0.0846
wbc	0.007481	0.022465	0.333	0.739
hb	-0.009103	0.001721	-5.289	<0.000

Logistic Regression is generalized linear model in R [16]. The result of the univariate regression can be access through the summary () function. if a p value is less than 0.25, we can include those variables for further analysis for clinical relevance. [17, 18]. Now from table 1 we can observe that four variables age, bmi, lac, hb are statistically significant (p<0.05) in univariate analysis. But for clinical relevance we shall include all the variables in further analysis.

3.3 Step two: multivariate model comparisons using R

Here in this step, we fitted two logistic regression model i.e. model 1 and model2 to analyze our data set. From the first model it is observed that the variables socioeconomic status, gender,wbc are insignificant(p>0.05).whereas the factors age ,bmi ,lac,hb are statistically significant(p<0.05).

Hence to create more accurate model, we exclude the insignificant variables like socioeconomic status,gender,wbc from our model and created a new model i.e. model2.Now by observing the result we can verify that all the factors in model 2 are statistically significant.

```
#Multivariate data analysis
 model1<- glm(mort~lac+hb+wbc+age+bmi+Socioeconomic+gender, family =  binomial)
 summary(model1)

Call:
glm(formula = mort ~ lac + hb + wbc + age + bmi + Socioeconomic +  gender, family = binomial)
    Min        1Q      Median      3Q        Max
-2.64350  -0.40850  -0.09155   0.39316   3.10143
```

Table 2: Results of Model 1

Coefficients:	Estimate	Std. Error	Z value	Pr(>|z|)
Intercept	-10.228369	1.382006	-7.401	1.35e-13***
lac	1.037040	0.071166	14.572	<2e-16***
hb	-0.019164	0.002885	-6.643	3.08e-11***
wbc	-0.008139	0.035393	-0.230	0.818
age	0.094568	0.009808	9.642	<2e-16***
bmi	0.037173	0.042619	0.872	<2e-16***
Socioeco	-0.229991	0.219703	1.047	0.295
gender	-0.340773	0.224199	-1.520	0.129

```
Null deviance: 1234.19  on 899  df
Residual deviance:  564.42  on 892  df
AIC: 580.42
```

Now from the above analysis we can see that the factors wbc, bmi, Socioeconomic status and gender are not statistically significant (p>0.05). Hence, we shall create another model by removing these three variables and create model 2.

```
Call:
glm(formula = mort ~ lac + hb + age + bmi + gender, family = binomial)

Deviance Residuals:
    Min        1Q       Median      3Q         Max
-2.66375  -0.41459  -0.09164   0.38986   3.05810
```

Table 3: Results of Model2

Coefficients:	Estimate	Std. Error	Z value	Pr(>|z|)
Intercept	-10.349741	1.352668	-7.651	1.99e-14***
lac	1.036494	0.071090	14.580	<2e-16***
hb	-0.019219	0.002882	-6.668	3.08e-11***
age	0.093628	0.009736	9.617	<2e-16***
bmi	0.035833	0.042603	0.841	<2e-16***

```
Null deviance: 1234.19  on 899  df
Residual deviance:  565.56  on 894  df
AIC: 577.56
Number of Fisher Scoring iterations: 6
```

Now we found a better result in model 2.all the variables associated with mortality are statistically significant (p<0.05). Also we can extract the estimated coefficients from fitted model using the function coef().Now we compared model1 with model2 by using partial likelihood ratio test.

```
#initiate library lmtest
 library(lmtest)
lrtest(model1,model2)
Model 1:  mort ~ lac + hb + wbc + age + bmi + Socioeconomic + gender
Model 2:  mort ~ lac + hb + age + bmi
  #Df  LogLik Df Chisq Pr(>Chisq)
1   8 -282.21
2   6 -282.78 -2 1.142      0.565
```

Now from the above result the chi-square value (0.565) indicates that there is no difference between model1 and model2 according to fit of the data. we can say model2 is a good model as all the independent variables are significant(p<0.05) .

3.4 Step three: Test of linearity using R

Here we considered the four independent variable which are statistically significant for the outcome variable. To test the linearity of these variable to the logit of the outcome, we used scatter. Smooth function.

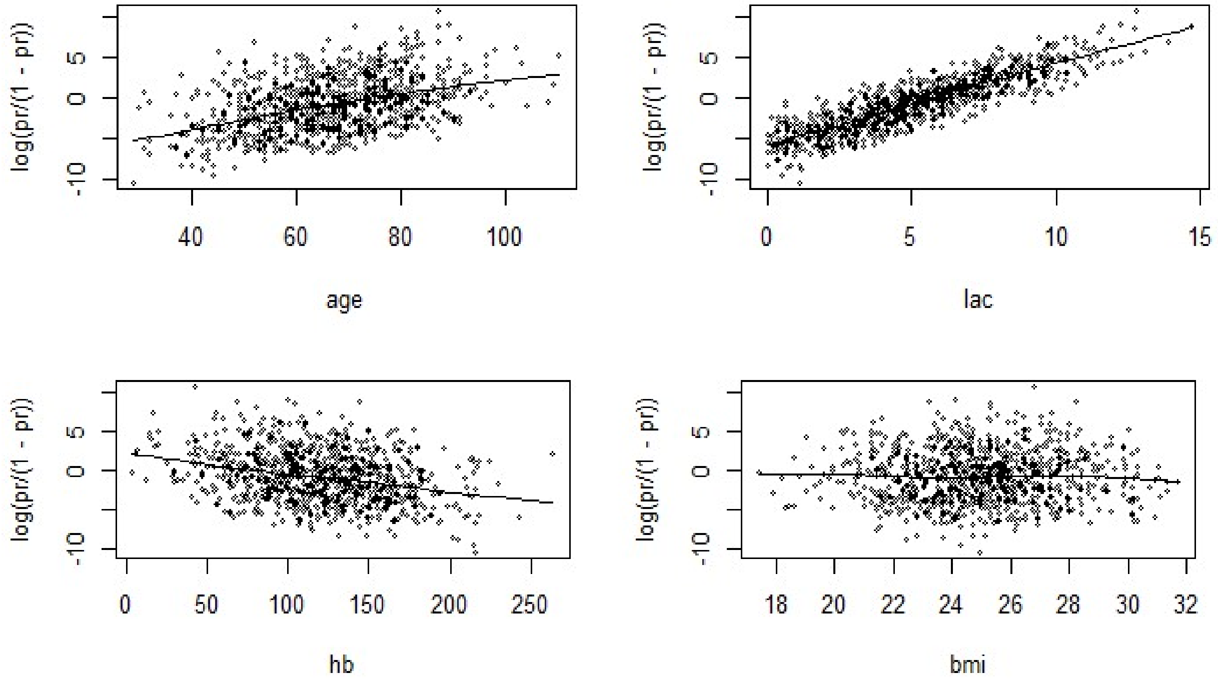


Fig 3: Test of linearity

Now from the above graphics we can see that the significant variables in model2 linearly associated with mortality outcome in logit scale (Figure 3).in case the smooth scatter plot shows nonlinearity, we can use alternative method to build our model.

3.5 Step four models fitting using R

Finally, we checkd the fit of our model. We examined is there any significant difference between obtained data and fitted data. We can check the model fit in two ways. Firstly, the model fit can be checked by using goodness of fit (GOF). Secondly by using regression diagnostics plot. we shall verify the fit of the model. The most common method for model fit is Hosmer-Lemeshow test. We performed the Hosmer-Lemeshow test in R and the results are shown below. Here from the table, we can observe that the p value is 0.069, which indicates there is no significant difference between observed and predicted values.

Table 4: Test of goodness of fit

Step	Chi-square	sig
1	14.23	.069

DISCUSSION

We have seen how to model a continuous numeric response with linear regression technique. But in many healthcare scenarios our target is binary. in medical research most of the classification problems deals by LR model. In this paper, we considered a working example to predict mortality of the patient. we created a data frame by using R programming language and finally we ended with performing a suitable logistic regression model. Now this data will be considered as “train” data, as we used this data to create our model. Now the performance of the model can be test by using a “test” data, which has not delivered in this paper. Most of the study does not focus on diagnosis plot, fit of the model and validation of the model [16] [17], while building a LR model, we have to take care of these things to create an accurate model.

CONCLUSION

In this paper we discussed the role of logistic regression in healthcare research. As it is a machine learning technique many researchers are using this as a predictive modelling algorithm. Apart from this R programming language is an emerging programming language, which is used in healthcare research for predictive modelling. This study put a light on how to build logistic regression model using R programming language. And finally, we checked linearity of the data, fit of the model, which are very essential for model building.

REFERENCE

1. Oommen, T., Baise, L.G. and Vogel, R.M. (2011) Sampling Bias and Class Imbalance in Maximum-Likelihood Logistic Regression. Mathematical Geosciences, 43, 99-120. https://doi.org/10.1007/s11004-010-9311-8
2. Cramer, J.S. (2002) The Origins of Logistic Regression. Tinbergen Institute Working Paper.
3. Tu, J.V. (1996) Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. Journal of Clinical Epidemiology, 49, 1225-1231. https://doi.org/10.1016/S0895-4356(96)00002-9
4. Hosmer D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Edition, Wiley, New York. https://doi.org/10.1002/0471722146 https://onlinelibrary.wiley.com/doi/book/10.1002/0471722146
5. King, G. and Zeng, L. (2001) Logistic Regression in Rare Events Data. Political Analysis, 9, 137-163. https://doi.org/10.1093/oxfordjournals.pan.a004868
6. Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. John Wiley & Sons, New York.
7. Bacaër, N. (2011) Verhulst and the Logistic Equation. In: A Short History of Mathematical Population Dynamics, Springer, London, 35-39. https://doi.org/10.1007/978-0-85729-115-8_6
8. Pearl, R. and Reed, L.J. (1920) On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. Proceedings of the National Academy of Sciences of the United States of America, 6, 275-288. https://doi.org/10.1073/pnas.6.6.275
9. Boateng, E.Y. and Oduro, F.T. (2018) Predicting Microfinance Credit Default: A Study of Nsoatreman Rural Bank Ghana. Journal of Advances in Mathematics and Computer Science, 26, 1-9. https://doi.org/10.9734/JAMCS/2018/33569
10. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (1989) The Multiple Logistic Regression Model. Applied Logistic Regression, 1, 25-37.
11. Srivastava, N. (2005) A Logistic Regression Model for Predicting the Occurrence of Intense Geomagnetic Storms. Annales Geophysicae, 23, 2969-2974. https://doi.org/10.5194/angeo-23-2969-2005
12. Khan, K.S., Chien, P.F. and Dwarakanath, L.S. (1999) Logistic Regression Models in Obstetrics and Gynecology Literature. Obstetrics & Gynecology, 93, 1014-1020. https://doi.org/10.1097/00006250-199906000-00024 https://www.ncbi.nlm.nih.gov/pubmed/10362173
13. Kim, Y., Kwon, S. and Song, S.H. (2006) Multiclass Sparse Logistic Regression for Classification of Multiple Cancer Types Using Gene Expression Data. Computational Statistics & Data Analysis, 51, 1643-1655. https://doi.org/10.1016/j.csda.2006.06.007
14. Jones, S.R. and McEwen, M.K. (2000) A Conceptual Model of Multiple Dimensions of Identity. Journal of College Student Development, 41, 405-414. https://www.researchgate.net/publication/292759031_A_conceptual_model_of_mul tiple_dimensions_of_identity
15. Vollmer, R.T. (1996) Multivariate Statistical Analysis for Pathologists: Part I, The Logistic Model. American Journal of Clinical Pathology, 105, 115-126.
16. Kabacoff R. R in action. Cherry Hill: Manning Publications Co; 2011.
17. Bendal RB, Afifi AA. Comparison of stopping rules in forward regression. Journal of the American Statistical Association 1977;72:46-53.
18. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol 1989;129:125-37