Modifications of log-likelihood to measure floor and ceiling effects

I. Kulikovskikh[a], S. Prokhorov[a]

[a] *Samara National Research University, 443086, 34 Moskovskoye shosse, Samara, Russia*

Abstract

The problem of complete separation between classes may produce serious difficulties with the successful implementation of logistic regression due to the presence of floor and ceiling effects. To address this problem, the present study proposes two modifications of ordinary log-likelihood. To reveal the benefits of these modifications, we provided a theoretical and experimental basis for comparison with the mostly reported way of penalizing of log-likelihood – the regularization method. From these comparisons, we concluded that the proposed modifications produced less biased estimates and reached higher accuracy on prediction compared to the regularized log-likelihood.

Keywords: logistic regression; log-likelihood; regularization; floor effect; ceiling effect

1. Introduction

Despite a number of advantages such as lower bias and the higher level of simplicity and interpretability over other linear classifiers [1-4], the successful implementation of logistic regression (LR) seems to crucially depend on the accurate identification of complete separation between classes [1-3]. Although the problem of separation primarily arises in small datasets with several unbalanced, highly predictive features and results in a log-likelihood's (LL) failure to converge [5-7], it may also occur with small or medium-sized datasets when at least one LR estimate is infinite even if the likelihood converges. Moreover, the problem of separation may arise if the underlying model parameters are low in an absolute value. Consequently, creating a proper measure to handle perfect separability is of high importance.

A comprehensive review of the literature on this problem suggested a good deal of solutions [1,2,5-11], but the most reported of them is the regularization method [1,2,12,13]. This method implies penalizing of LL to make the estimates finite. Unfortunately, adopting regularization may lead to not asymptotically normal and highly biased estimates even if the regularized LL tends to produce lower prediction errors. Thus, the present research is an attempt to fill this gap by proposing the modifications of LL that make less biased estimates and ensure higher accuracy on prediction.

2. Problem statement

Let $\{x_i, y_i\}_{i=1}^m$ denote independent and identically distributed observations with binary responses $y_i \in \{0,1\}$. The matrix $X \in \mathbf{R}^{m \times n}$ can be viewed either as $X = [x_1, \ldots, x_n]^T$, with vectors of predictors $x_i \in \mathbf{R}^n$, or as $X = [x^1, \ldots, x^m]$, with vectors of features $x^j \in \mathbf{R}^m$. Let $y = [y_1, \ldots, y_n]^T$ be the response vector. Then, for any vector of regression coefficients $\theta \in \mathbf{R}^n$ LR models the class conditional probabilities $p(x_i, \theta) = P(y_i = 1 | x_i, \theta)$ by

$$\ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right) = \theta^T x_i .$$

Having defined the main parameters, let us now move on to pose the LR problem.

Problem 1. *Let g be the link (logit) function that defines the relationship between the class conditional expectation of the response variable and the underlying linear model $g(E[y_i \mid x_i]) = \theta^T x_i$. Taking into account that $E[y_i \mid x_i] = p(x_i, \theta)$ for LR,*

$$g(p(x_i,\theta)) = \ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right). \quad (1)$$

Under the model $p(x_i, \theta)$, the negative log-likelihood (LL) expressed as

$$\ln L(\theta) = -\sum_i y_i \ln(p(x_i,\theta)) + (1-y_i)\ln(1-p(x_i,\theta)) \quad (2)$$

or, what is the same,

$$\ln L(\theta) = -\sum_i y_i \theta^T x_i - \ln(1+\exp(\theta^T x_i)) \quad (3)$$

needs to be minimized to solve the following problem

$$\theta = \arg\min_{\theta \in \mathbf{R}^n} \ln L(\theta). \quad (4)$$

As a result of complete separation or no separation between the classes, the logit function (1) may go to $-\infty$ for 0 successes (floor effect) or ∞ for 0 failures (ceiling effect) [1,7], respectively. But, this means that (4) fails to converge. The traditional approach to deal with floor and ceiling effects is to penalize LL (4) for very large estimates and, thus, to shrink these estimates toward 0. In particular, a widely-used way [1,2] to do this adds an extra term to (2)

$$\theta^* = \arg\min_{\theta \in \mathbf{R}^n} \left\{ \ln L(\theta) + \lambda P(\theta) \right\}, \quad (5)$$

where λ is a regularization parameter, $P(\theta)$ is a function that penalizes coefficients θ as they get further away from zero.

Considering some shortcomings of (5) that were pointed out in Section 1, let us now propose two new modifications of LL to address the problem of floor and ceiling effects primarily stepped from complete separation of classes.

3. Modifications of log-likelihood

Before explaining these modifications, it is necessary to present the extension of *Problem 1*. Thus, *Problem 2* involves an extra parameter $c \in [0,1]$ to describe floor and ceiling effects.

Problem 2. *Let the logit function* $g\left(p(x_i,\theta)\right)$ *be extended to* $g\left(p(x_i,\theta),c\right)$, *where* $c \in [0,1]$, as follows

$$g\left(p(x_i,\theta),c\right) = \begin{cases} \ln\left(\dfrac{p(x_i,\theta)-c}{1-p(x_i,\theta)}\right) & \textit{if floor;} \\ \ln\left(\dfrac{p(x_i,\theta)}{1-p(x_i,\theta)-c}\right) & \textit{if ceiling.} \end{cases} \quad (6)$$

Then, minimizing the negative LL $\ln L(\theta,c)$ *solves the two-parameter problem*

$$(\theta,c) = \arg\min_{\substack{\theta \in \mathbf{R}^n \\ c \in [0,1]}} \ln L(\theta,c). \quad (7)$$

The present study considers two modifications of LL $\ln L(\theta,c)$ with regard to (6) and the way of passing c into (2).

3.1. 1-form modification

This form of modification implies introducing the parameter c in the LL loss function (2) as

$$\ln L(\theta,c) = \begin{cases} -\sum_i y_i \ln\left(g^{-1}\left(p(x_i,\theta),c\right)-c\right) + (1-y_i)\ln\left(1-g^{-1}\left(p(x_i,\theta),c\right)\right) & \textit{if floor;} \\ -\sum_i y_i \ln\left(g^{-1}\left(p(x_i,\theta),c\right)\right) + (1-y_i)\ln\left(1-g^{-1}\left(p(x_i,\theta),c\right)-c\right) & \textit{if ceiling.} \end{cases} \quad (8)$$

Based on the definition (8), let us state the following lemma.

Lemma 1. *For each* $x_i \in \mathbf{R}^n$, $\theta \in \mathbf{R}^n$, *and* $c \in [0,1]$, *the LL loss function* $\ln L(\theta,c)$ *based on 1-form modification* (9) *for floor and ceiling effects are the same and equal to*

$$\ln L(\theta,c) = -\sum_i y_i \theta^T x_i - \ln\left(1+\exp\left(\theta^T x_i\right)\right) - \ln(1-c). \quad (9)$$

Corollary 2. For $c=0$, (9) results in (3).

3.2. 2-form modification

This modification, in contrast, employs the same definition for presenting floor and ceiling effects

$$\ln L(\theta,c) = -\sum_i y_i \ln\left(g^{-1}\left(p(x_i,\theta),c\right)\right) + (1-y_i)\ln\left(1-g^{-1}\left(p(x_i,\theta),c\right)\right). \quad (10)$$

The lemma that is posed below seems to clearly underline the difference between the proposed modifications of LL.

Lemma 3. *For each* $x_i \in \mathbf{R}^n$, $\theta \in \mathbf{R}^n$, *and* $c \in [0,1]$, *the LL loss function* $\ln L(\theta,c)$ *based on 2-form modification* (10) *for floor and ceiling effects are equal to*

$$\ln L(\theta,c)=\begin{cases} -\sum_i y_i \ln\left(c+\exp\left(\theta^T x_i\right)\right)+\left(1-y_i\right)\ln\left(1-c\right)-\ln\left(1+\exp\left(\theta^T x_i\right)\right) & \text{if floor;} \\ -\sum_i y_i \ln\left(\left(1-c\right)\exp\left(\theta^T x_i\right)\right)+\left(1-y_i\right)\ln\left(1+c\exp\left(\theta^T x_i\right)\right)-\ln\left(1+\exp\left(\theta^T x_i\right)\right) & \text{if ceiling.} \end{cases} \quad (11)$$

Corollary 4. For $c=0$, (11) gives (3).

The similarity between (3), (9), and (11) invites the following comparison: while *2-form* modification (11) implies both the inclusion of the estimates θ and the parameter c to penalize LL, *1-form* modification (9) includes only one extra term $-\ln\left(1-c\right)$ compared to the known definition (3). It should be noted, though, that the problem (7) produces the estimates θ based on (6), i.e. the parameter c modifies the coefficients θ and, thus, is implicitly present in both modifications regardless of the form or the type of effect to be introduced (floor or ceiling). Comparing the proposed modifications with the regularized LL (5), it can be suggested that the latter approach to penalizing LL rather similar to *1-form* modification, but *1-form* modified LL produces less biased estimates. The detailed analysis of the estimates' bias is beyond the scope of this paper, but still seems promising direction for further research. In the present study, to confirm the theoretical outcomes, we conducted a series of computational experiments the results of which are given in the following section.

4. Results

To highlight the benefits of the proposed solution to the complete separation problem, the present study is intended to compare the results proposed in this paper with previously reported in the literature. For this purpose, we considered ridge regression with the penalty $P(\theta)=\sum_j \theta_j^2$ in (5) to introduce the regularized LL. Before we go any further, let us first describe a dataset chosen to support these comparisons.

4.1. Dataset description

The dataset *heart* was taken from UCI Machine Learning Repository (Statlog (Heart)) [14] X^m ($m=270, n=13$) and divided into the training subset X^{l_1} and the validation subset X^{l_2} using 4-fold cross validation. To increase a chance of identifying the separation problem, the experiments suggested varying the limited number of observations, i.e. $m=\{25, 50\}$. Taking into account the fact that we had to deal with small subsets, we carried out a series of additional experiments to provide statistically significant estimates. Thus, each presented result is the average of $N=50$ computed values.

4.2. Computational experiments

The computational experiments were designed to: 1) model validation curves; 2) find the optimal values of regulation parameters c/λ; 3) estimate the accuracy of classification with regard to a form of LL and $m=\{25, 50\}$. Fig. 1, 2, and 3 depict the corresponding validation curves subject to $m=25$. As can be seen, the validation curves based on regularized LL (see Fig. 1) are similar to those presented in Fig. 2 a) and Fig. 3 a) based on *1-form* modified LL. This fully complies with the theoretical results.

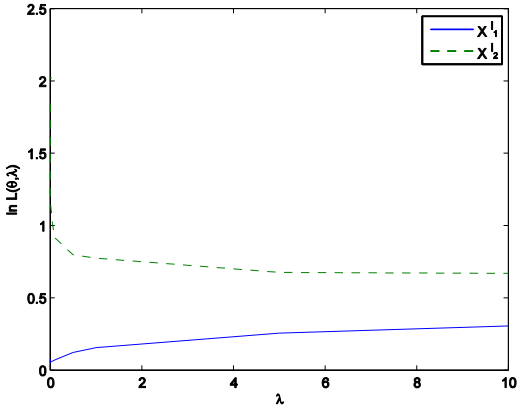
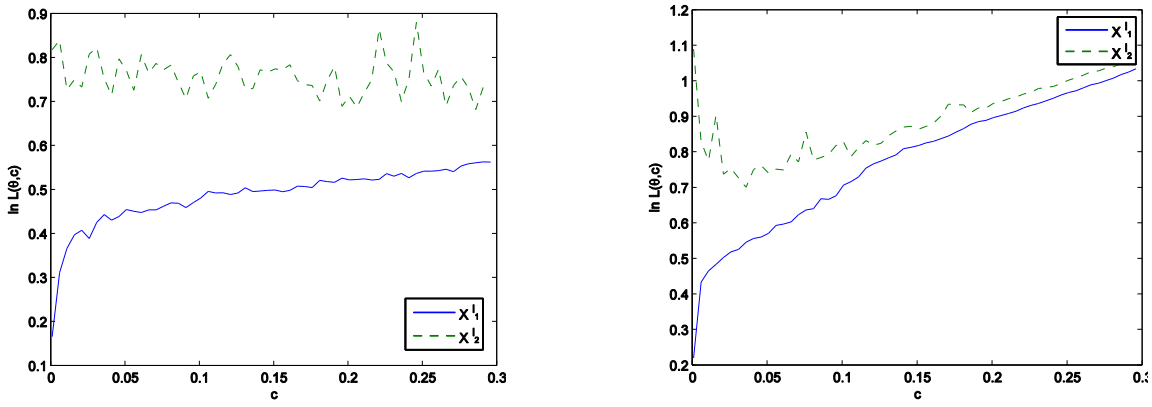


Fig.1. The validation curves based on regularized LL subject to $m=25$.

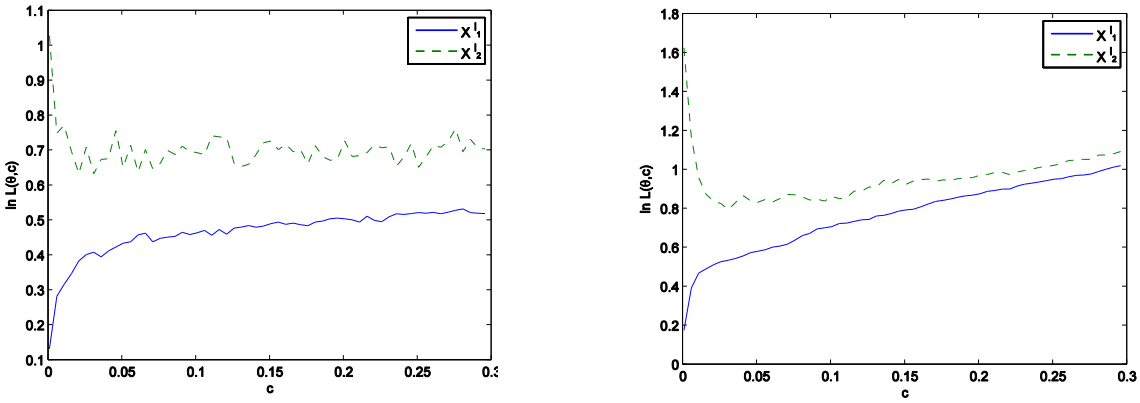
The validation curves that demonstrate the estimates of *2-form* modified LL (see Fig. 2 b) and Fig. 3 b)) seem different to others, but are more attractive: it is easier to point out the optimal value of c.

The values of classification accuracy based on different modifications of LL subject to $m=\{25, 50\}$ are presented in Table 1 and 2, respectively. If we look at these values, we can see that all the penalized LL (5), (9), (11) produced better results – an increase in accuracy is up to 5% – than the ordinary LL (3). In addition, the proposed modified LL permitted to reach higher accuracy on the validation subset in comparison with the regularized LL. As the values of accuracy are not high enough for the chosen dataset, the modifications that describe a floor effect allowed us to yield more marked improvement on the results than the modifications of LL for a ceiling effect. Moreover, *2-form* modified LL performed better than *1-form* modified LL that confirmed the findings of this research. Analyzing the presented results subject to $m=50$, we observe that the proposed modifications brought us little or no advantage over the regularized LL, but we can still state that the desired outcome is achieved: the modified LLs are designed to measure floor and ceiling effects intrinsic to smaller datasets.



a) b)

Fig.2. The validation curves based on modified LL subject to $m = 25$ (floor effect): a) 1-form; b) 2-form.



a) b)

Fig.3. The validation curves based on modified LL subject to $m = 25$ (ceiling effect): a) 1-form; b) 2-form.

Table 1. The accuracy of classification based on different modifications of LL subject to $m = 25$

A form of LL	c / λ	X^{l_1}	X^{l_2}
LL	0/0	100.000	70.5000
Regularized LL	0/10	91.3158	72.5000
1-form modified LL (floor)	0.036/0	99.9474	**74.5000**
2-form modified LL (floor)	0.2/0	100.000	**75.5000**
1-form modified LL (ceiling)	0.031/0	99.8421	70.8333
2-form modified LL (ceiling)	0.031/0	100.000	71.3333

Table 2. The accuracy of classification based on different modifications of LL subject to $m = 50$

A form of LL	c / λ	X^{l_1}	X^{l_2}
LL	0/0	97.1579	72.5000
Regularized LL	0/5	89.7368	76.0000
1-form modified LL (floor)	0.016/0	97.0526	73.5833
2-form modified LL (floor)	0.016/0	96.3421	**76.2500**
1-form modified LL (ceiling)	0.006/0	97.3684	73.0000
2-form modified LL (ceiling)	0.001/0	97.8947	74.0000

5. Conclusion

The present study was aimed at proposing a proper measure based on LL to directly address the issue of floor and ceiling effects in classification problems. For this reason, we offered two promising modifications: *1-form* modification and *2-form* modification. In support of these modifications, we provided a theoretical and experimental basis for comparison with the known ways of penalizing of LL reported in the literature, in particular, the regularization method. From these comparisons we may draw the following conclusions: the proposed modifications produced less biased estimates and reached higher accuracy on prediction compared to the regularized LL. Therefore, the purpose, stated in this paper, is accomplished.

Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation, grant 074-U01.

References

[1] Agresti, A. Foundations of linear and generalized linear models / A. Agresti – Wiley Series in Probability and Statistics, 2015. – 472 p.
[2] Hastie, T. The elements of statistical learning: Data mining, inference, and prediction: 2nd ed. / T. Hastie, R. Tibshirani, J. Friedman – Springer Series in Statistics, 2013. – 745 p.
[3] McCulloch, C.E. Generalized, linear, and mixed models: 2nd ed. / C.E. McCulloch, S.R. Searle, J.M. Neuhaus – New York: John Wiley, 2009. – 424 p.
[4] Zaidi, N.A. ALR^n: accelerated higher-order logistic regression / N.A. Zaidi, G.I. Webb, M.J. Carman, F. Petitjean, J. Cerquides // Machine Learning. – 2016. – Vol. 104. – P. 151-194.
[5] Fort, G. Classification using partial least squares with penalized logistic regression / G. Fort, S. Lambert-Lacroix // Bioinformatics. – 2005. – Vol. 21(7). – P. 1104-1111.
[6] Heinze, G. A solution to the problem of separation in logistic regression / G. Heinze, M. Schemper // Statistics in Medicine. – 2002. – Vol. 21. – P. 2409-2419.
[7] Donnelly, S. Empirical logit analysis is not logistic regression / S. Donnelly, J. Verkuilen // Journal of Memory and Language. – 2017. – Vol. 94. – P. 28-42.
[8] Gelman, A. A weakly informative default prior distribution for logistic and other regression models / A. Gelman, A. Jakulin, M.G. Pittau, Y.-S. Su // The Annals of Applied Statistics. – 2008. – Vol. 2(4). – P. 1360-1383.
[9] Ding, B. Classification using generalized partial least squares / B. Ding, C.M. Gentleman // Graphical Statistics. – 2005. – Vol. 14(2). – P. 280-298.
[10] Firth, D. Bias reduction, the Jeffreys prior and GLIM / D. Firth // Advances in GLIM and Statistical Modelling. – New York: Springer-Verlag, 1992. – P. 91-100.
[11] Firth, D. Generalized linear models and Jeffreys priors: An iterative weighted least-squares approach / D. Firth // Computational Statistics. – Vienna: Physica-Verlag, 1992. – P. 553-557.
[12] Fan, Y. Asymptotic equivalence of regularization methods in thresholded parameter space / Y. Fan, J. Lv // Journal of the American Statistical Association. – 2013. – Vol. 108(503). – P. 1044-1061.
[13] Park, M.Y. L_1-regularization path algorithm for generalized linear models / M.Y. Park, T. Hastie // Journal of the Royal Statistical Society. Series B. – 2007. – Vol. 69(4). – P. 659-677.
[14] UCI Machine Learning Repository. Statlog (Heart) [Electronic resource]. — Access mode: http://archive.ics.uci.edu/ml/datasets/Statlog+(Heart) (29.01.2017)