Title

Suyoung Park, Daniel Eck, and Alex Lipka
University of Illinois at Urbana-Champaign

Month 2021

Abstract

Space for the Abstract.

Key Words: list of keywords

1 Material

We implemented our methodology in R package **glmdr**. We used R version 3.6.1 and the required R packages for **glmdr** are **binom** version 1.1 and **nloptr** version 1.2.2.2. Further details are included in the technical reports.

2 Method

2.1 Logistic Regression

The logistic regression is the special case of the generalized linear model which the response variable follows Bernoulli distribution (i.e., $y \in \{0, 1\}$) [Nelder and Wedderburn, 1972]. By convention, we encode 1 as a "success" and 0 as a "failure." We model the conditional probability of observing one (success) given x:

$$\Pr(Y_i = 1|X_i = x_i) = p_i = \frac{\exp\left(x_i^T \beta\right)}{1 + \exp\left(x_i^T \beta\right)}, \tag{1}$$

where β is unknown parameters.

From the linear regression's point of view, this logistic regression is equivalent to:

$$g(p_i) = \log(\frac{p_i}{1 - p_i}) = x_i^T \beta \tag{2}$$

where $g(x) = \log(\frac{x}{1-x})$ is a logit link (log-odds ratio).

Therefore, like an ordinary least squares (OLS), we can estimate β using the score function and make a statistical inference using diagonal elements of the inverse of the Fisher information, which represent the estimated variance of parameters.

Specifically, given log-likelihood function of the logistic regression model,

$$\log L(\beta|Y) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i), \tag{3}$$

the score function is:

$$\frac{\partial \log L(\beta|Y)}{\partial \beta} = \sum_{i=1}^{N} (y_i - \log(p_i))X_i = \sum_{i=1}^{N} [y_i + \log(1 + \exp(-x_i\beta))] = 0, \tag{4}$$

and the variance-covariance matrix is the inverse of the Fisher information:

$$\widehat{\mathrm{Var}(\beta)} = [I(\beta)]^{-1} = \left[-E\left[\frac{\partial^2 \log L(\beta|Y)}{\partial \beta_i \partial \beta_j}\right] \right]^{-1}. \tag{5}$$

2.2 Mean-value Parameters

From the statistical model, what we actually want to know is the expected value of response variable given data. In the logistic regression, $\mathrm{E}(Y|X = x) = \mathrm{Pr}(Y = 1|X = x)$, meanwhile, $\mathrm{E}(Y|X = x) = x^T\beta$ for the linear regression. Hence, unlike the linear model, an interpretation on β from the logistic model is difficult because our response variable is log-odds ratio rather than the probability of a success. For example, we can say that as one unit of explanatory variable increases the expected change in log odds ratio of conditional probability of success is β. However, it is not very intuitive and informative. To overcome this, we can consider a mean-value parameterization. Instead of directly using the estimated coefficients of the logistic regression model, we can obtain the estimated probability of success given data by plugging in $\hat{\beta}$ into (1).

2.3 Complete Separation

From authoritative textbook [Agresti, 2013, Section 6.5.1], complete separation is defined as there exists a vector β such that

$$\begin{aligned} x_i^T\beta > 0 \text{ whenever } y_i = 1, \\ x_i^T\beta < 0 \text{ whenever } y_i = 0. \end{aligned} \tag{6}$$

That is, it occurs when the one or more explanatory variables can perfectly predict the response variable [Albert and Anderson, 1984]. For example, as shown on Figure 1, consider the following case that when x is less than or equal to 40, all corresponding y are 0 and when x is greater than or equal to 60, all corresponding y are 1. Suppose $x^T = [1, x_i]$ and $\beta = [-50, 1]^T$. Then, we have $\hat{p} = 0$ for $x < 50$ and $\hat{p} = 1$ for $x > 50$.

In terms of MLE, its value goes to the infinity or does not exist and the shape of log likelihood is flat over different values of estimate. Figure 2 shows the log likelihood of logistic
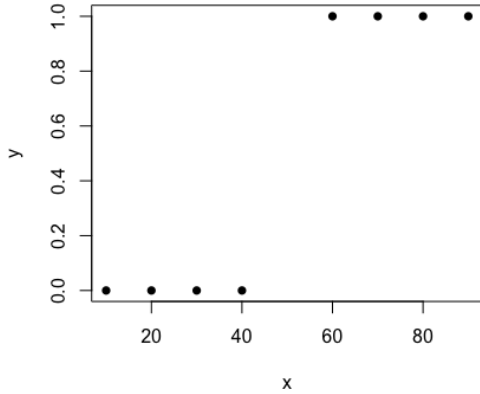
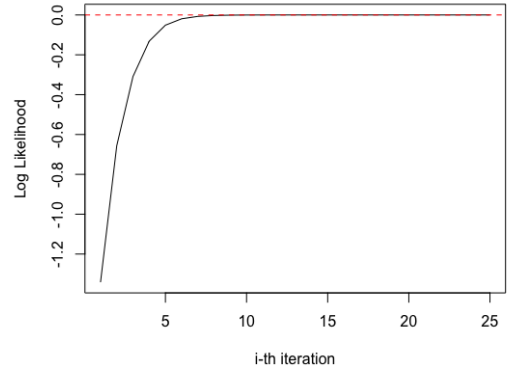Figure 1: Example of Complete Separation in Logistic Model

Figure 2: Log Likelihood Values of Logistic Model at each iteration

model for this example with different working estimate from **glm** function in R. We can see that likelihood value goes to infinity after a few iteration. As a result, the variance of $\hat{\beta}_j$ becomes very large because it comes from the inverse of second derivatives of the likelihood function (5) and it leads us to unreasonable statistical inference. Unfortunately, none of common statistical software such as R, SAS and Python can handle the separation issue properly and uninformed users sometimes uses the wrong model without knowing it.

2.4 One-Sided Confidence Interval

We use one-sided confidence intervals for the logistic model mean value parameters to explain the uncertainty of estimation. Original concept can be found in Section 3.16 of Geyer's paper [2009] and implementation details can be found in Section 4.3 of Eck and Geyer's work [2020]. Briefly, we construct confidence interval for mean value parameters such that one endpoint is observed response variable (i.e., lower bound if $y_i = 0$ and upper bound if $y_i = 1$) and the other endpoint is obtained by solving the optimization problem:

$$
\begin{aligned}
&\text{minimize} \quad -\theta_k \\
&\text{subject to} \quad \sum_{i \in I} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] - \log(\alpha) \geq 0.
\end{aligned}
\tag{7}
$$

where $\theta_k = x_k^T \beta$ for any $k \in I$, I is a index of problematic points that cause the separation, p is a mean value parameters, and α is a significance level. In **glmdr**, **inference** function provides this confidence intervals using the sequential quadratic programming (SQP) to solve the constrained nonlinear problem (7).

3

2.5 Prediction

Prediction in **glmdr** framework is different from that of the traditional statistical model because we estimate the mean value parameters rather than coefficient. Therefore, we firstly find the prediction intervals given new data point then take an average of them for predicted value.

Given new data x_{new} and training set x_{train}, we generate testing set by combing training set and each observation from new data. That is, $x_{i,\text{test}} = x_{\text{train}} \cup x_{i,\text{new}}$ where i is a index of whole new data. Then, we construct two testing labels that one has $y = 0$ and the other has $y = 1$ for new data point. Based on these two datasets, we fit two logistic models to compute the estimated probability of success for new data points, $\hat{p}_1$ and $\hat{p}_2$. Since we do not know which model is true model, we calculate the weighted probability, $\widehat{p^*} = w_1\hat{p}_1 + w_2\hat{p}_2$ where w_j is Akaike weights:

$$w_j = \frac{\exp(-\frac{IC_j}{2})}{\exp(-\frac{IC_1}{2}) + \exp(-\frac{IC_2}{2})},$$

and IC_j is the information criteria of model j [Burnham and Anderson, 2002]. We recommend the AICc for IC because Akaike Inofrmation Criteria corrected (AICc) works well in small sample size and converges to AIC when we have large sample size. Lastly, we construct the Wilson intervals because it handles the extreme probability (i.e. $\hat{p} = 0$ or 1) better than Wald based intervals [Brown et al., 2001]. We can provide the predicted label based on the mean of this Wilson intervals. In our method, we label 1 if $\widehat{p^*} \geq 0.5$ and 0 if $\widehat{p^*} < 0.5$. Detailed implementation and examples are given in the supplementary materials.

3 Results

3.1 Examples

Complete Separation We revisit the problem we discussed in Section 2.3. In this example, we can find the hyperplane such that one side of this hyperplane contains all successes and the other side has all failures. Our logistic model says that its MLE lie on the infinity and its standard error goes to infinity as well.

Quasi Complete Separation

Quasi-complete separation is another case of separation that there are both a success and a failure on the hyperplane that separates the successes from the failures [Lesaffre and Albert, 1989]. For instance, we can consider additional two points that $x = 50$ with $y = 1$ and $y = 0$ to the previous complete separation example. That is, we have $y_i = 0$ for $x \leq 50$ and $y_i = 0$ for $x \geq 50$. In this case, the maximized log likelihood is always negative and we experience same phenomenon as the complete separation case.

Quadratic

This example comes from Section 2.2 of Geyer [2009]. Let $y_i = 1$ for $12 < x_i < 24$ and

$y_i = 0$, otherwise. Also, consider the following quadratic model:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2.$$

In this case, MLE does not exist when we fit the logistic model using **glm** and it complains that the algorithm did not converge. We demonstrate how to compute the one-sided confidence intervals for mean value parameters for this example in the supplementary material.

Endometrial Cancer Study

Heinze and Schemper (2002) firstly investigated the endometrial data set (n = 79), which was originally provided by Dr. Asseryanis from the Vienna University Medical School. The main purpose of this study was to describe histology of cases (HG) in terms of three risk factors: neovasculation (NV), endometrium height (EH) and pulsatility index of arteria uterina (PI). 30 patients was classified grading 0-II for histology (HG = 0) and 49 patients for grading III-IV (HG = 1). There are 13 patients who has neovasculization (NV = 1) and absent for 66 patients (NV = 0). Pulsatility index (PI) ranges from 0 to 49 with mean of 17.38 and median of 16.00, and endometirum height (EH) ranges from 0.27 to 3.61 with mean of 1.662 and median of 1.640. In this example, we observe the quasi-complete separation in NV.

Maize data

To predict the kernel color of maize, we merged two datasets on accession's name. One dataset comes from Romay et al. (2013)'s work that investigates the genetic constitution of 2,815 maize inbred accessions with 7 types of population structures. [Where does the kernel color dataset come from?] The other dataset contains the kernel color of accession where 1 indicates yellow kernel and 0 for white kernel. It has 24 marker genotypes for the DNA surrounding a biologically relevant gene for kernel color. Each marker has value from 0 to 1. In the final dataset, 309 observations have a white kernel and 1,238 for yellow kernel. We have 6 types of population structures: 115 non-stiff stalk, 54 popcorn, 120 stiff stalk, 116 sweet corn, 159 tropical, and 983 unclassified. In this example, there is no separation issues when we use single marker for explanatory variable. However, we have a separation issue for saturated model. In the later part, we mainly focus on this example.

3.2 Estimation

We fit the logistic model where the response variable is kernel color. We dropped 5 markers out of 24 markers due to the collinearity issue and thus we use the population structures and 19 markers for the explanatory variables. We compare the in-sample accuracy and total length of confidence intervals among **glmdr**, **bayesglm** [Gelman et al., 2008], **logistf** [Heinze and Schemper, 2002] (we tested **brglm2** [Kosmidis and Firth, 2009] but algorithm did not converge. Instead, we used **logistf**, which is equivalent to **brglm2** with type of score adjustment as a maximum penalized likelihood with powers of the Jeffreys prior as a penalty.) and multiple linear model. We use Wilson interval to compute the total length of confidence intervals for logistic models and Wald type confidence intervals for the linear model. In Table 1, we can see all model performs comparably.

5

Table 1: Model comparisons for Maize example.

model	in-sample accuracy	total length of confidence intervals
glmdr	87.14 %	1293.01
bayesglm	87.07 %	1294.97
logistf	87.01 %	1296.37
linear	86.81 %	152.32 (should we report this?)

3.3 Prediction

We use the leave-one-out cross validation (LOOV) for prediction. This setting is the most suitable as we want to predict the kernel color of new maize given our data. In Table 2, we can see all models show similar performance.

Table 2: Prediction for Maize example.

model	out-of-sample accuracy
glmdr	85.13 %
bayesglm	86.23 %
logistf	86.03 %
linear	86.29 %

3.4 Discussion

In the classification problem, the logistic model is one of the most common statistical model we can attempt. Although linear model is attractive option to use because of its easiness and handiness, the binary response variable makes the linear model violate some of Gauss-Markov assumptions. Therefore, even though results from Section 3.2 and 3.3 display that the performance of linear model is comparable to the logistic models, we can not fully utilize asymptotic properties of linear model and make an inference such as significance tests for coefficients.

On the other hands, we can see all logistic models in Section 3.2 and 3.3 perform similarly despite of different approaches and techniques. **glmdr** solves the separation problem within the original model settings unlike other methods. Also, it estimates the probability of success rather than coefficients by finding the MLE in the Barndorff-Nielsen completion based on approximate null eigenvectors of the Fisher information matrix. The **bayesglm** uses a Bayesian approach which scales the data first and then placing Cauchy distribution as a prior distribution on the coefficients and **logistf** (similar to **brglm2**) modifies the score function to produce finite coefficients when MLE does not exist or lies on infinity.

In conclusion, when separation issue present in the logistic model, one can consider using the **glmdr** which has an advantage in the inference because it provides the estimated probability of success as a default with smallest length of confidence intervals. **bayesglm**

will be most suitable for prediction thanks to its low computational cost yet high accuracy. **logistf** or **brglm2** may be least preferable methods because they are computationally unstable and expensive.

References

Alan Agresti. *Categorical data analysis.* Wiley series in probability and statistics. Wiley, 3rd ed edition, 2013. ISBN 9780470463635.

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 04 1984. ISSN 0006-3444. doi: 10.1093/biomet/71.1.1. URL https://doi.org/10.1093/biomet/71.1.1.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101 – 133, 2001. doi: 10.1214/ss/1009213286. URL https://doi.org/10.1214/ss/1009213286.

Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference - 2nd ed.: a practical information-theoretic approach.* Springer-verlag new york Inc., 2002.

Daniel J. Eck and Charles J. Geyer. Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist, 2020.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008. doi: 10.1214/08-aoas191.

Charles J. Geyer. Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3:259–289, 2009. doi: 10.1214/08-ejs349.

Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002. doi: 10.1002/sim.1047.

I. Kosmidis and D. Firth. Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804, 2009. doi: 10.1093/biomet/asp055.

E. Lesaffre and A. Albert. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):109–116, 1989. doi: 10.1111/j.2517-6161.1989.tb01752.x.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238. URL http://www.jstor.org/stable/2344614.

Maria C Romay, Mark J Millard, Jeffrey C Glaubitz, Jason A Peiffer, Kelly L Swarts, Terry M Casstevens, Robert J Elshire, Charlotte B Acharya, Sharon E Mitchell, Sherry A

Flint-Garcia, Michael D McMullen, James B Holland, Edward S Buckler, and Candice A Gardner. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, 14(6):R55, 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-6-r55. URL https://doi.org/10.1186/gb-2013-14-6-r55.