# 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 $x$:

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

where $\beta$ 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 $\beta$ 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:

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

## 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, $\mathrm{E}(Y|X=x) = \Pr(Y=1|X=x)$, meanwhile, $\mathrm{E}(Y|X=x) = x^T\beta$ for the linear regression. Hence, unlike the linear model, an interpretation on $\beta$ 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 $\beta$. 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 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 $\beta$ 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

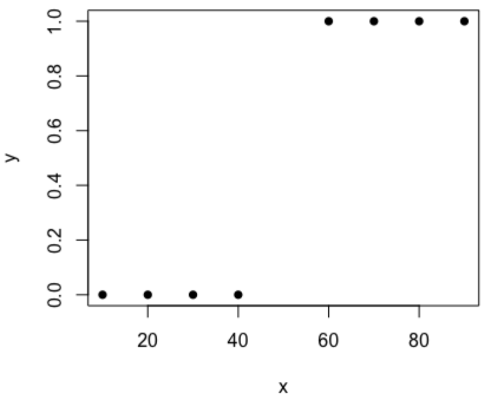

Figure 1: Example of Complete Separation in Logistic Model

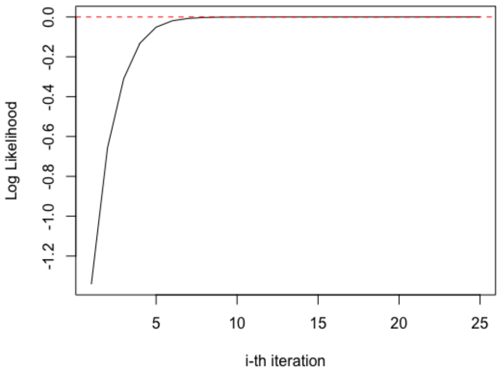

Figure 2: Log Likelihood Values of Logistic Model at each iteration

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 handles the separation issue 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 $y = 0$ and upper bound if $y = 1$) 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 $\alpha$ is a significance level.
In **inference** function from **glmdr**, we used the sequential quadratic programming (SQP) to solve the constrained nonlinear problem (7).

### 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 a average of them for predicted value.

Given new data $x_{\text{new}}$ and training set $x_{\text{train}}$, 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 $y = 0$ for new data point 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 point, $\hat{p_1}$ 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**

**Quasi Complete Separation**

**Quadratic**

**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.

**Maize data**

## 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.

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.

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`.