Robust model based prediction of gene expression in maize

Suyoung Park, Alex E. Lipka, Daniel J. Eck
University of Illinois at Urbana-Champaign

Month 2021

Abstract

Help us with the title Alex, you're our only hope!

Key Words: list of keywords

1 Materials and Method

1.1 Materials

We implemented our methodology in R package `glmdr`. We used R version 3.6.1 and the required R packages for `glmdr` is `nloptr` version 1.2.2.2. To compare its performance, we considered `arm` version 1.11-1, `brglm2` 0.7.0, `logistf` version 1.23 and `stats` version 3.6.1. To determine the optimal cut-off for the logistic regression, we used `PresenceAbsence` version 1.1.9. For visualization, data wrangling and experiments, we used `ggplot2` version 3.3.3, `gridExtra` version 2.3, `latex2exp` version 0.4.0, `foreach` version 1.4.7, `doParallel` version 1.0.15, and `tidyverse` version 1.2.1. Further details are included in the technical reports.

1.2 Data

We provide inference and prediction results for the maize data as well as an extensive set of examples. These include:

Complete separation: We first analyze the Agresti [2013] example discussed in Section 1.5.

Quasi-complete separation: We analyze the Agresti [2013] example with two points added, a success and a failure at $x = 50$.

Quadratic logistic regression model: This example comes from Section 2.2 of Geyer [2009]. In this example $y_i = 1$ for $12 < x_i < 24$ and $y_i = 0$, otherwise. In this case, maximum likelihood estimate (MLE) does not exist when we fit a quadratic 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 = 1) and 49 patients for grading III-IV (HG = 0). 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. [Place for description of the kernel color dataset] 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.

1.3 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." In logistic regression the conditional success probability at a particular x is modeled as

$$\Pr(Y = 1|X = x) = \frac{\exp(x^T\beta)}{1 + \exp(x^T\beta)} = p_x, \tag{1}$$

where β is an unknown canonical parameter vector (coefficient vector), X and Y are the predictor and response random variables, and x is an observed value.

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

$$g(p_x) = \log\left(\frac{p_x}{1-p_x}\right) = x^T\beta \tag{2}$$

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

Therefore, as in classical ordinary least squares (OLS) regression, we can estimate model parameters using maximum likelihood estimation. Statistical inferences about model parameters can be obtained from estimates of the Fisher information. Unlike in OLS regression, estimates for $\hat{\beta}$ are not given in closed form. The log-likelihood function for the logistic regression model is

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

one then obtains $\hat{\beta}$ by solving the score function equation

$$\frac{\partial \log L(\beta|Y)}{\partial \beta} = \sum_{i=1}^{N}(y_i - \log\left(p_{x_i}\right))x_i^T = \sum_{i=1}^{N}[y_i + \log\left(1+\exp\left(-x_i^T\beta\right)\right)] = 0. \tag{4}$$

Conventional softwares finds $\hat{\beta}$ through Fisher-scoring or iteratively reweighted least squares algorithms [Agresti, 2013, Chapter 4]. We then obtain inferences using an estimate of the Fisher information matrix evaluated at the MLE solution $\hat{\beta}$

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

Conventional software provides (5).

1.4 Mean-value Parameters

The parameter of primary interest is often the mean-value parameter on the scale of the response variable. This is the expected response expressed as a function of covariates. In the logistic regression model the mean-value parameter is the conditional success probability p_x at some particular x, and, unlike in linear regression, this parameter is not easily interpreted from β. Furthermore, the natural constraints on a conditional probability corresponding to a binary response variable require an alteration to the linear model.

In linear regression, we can easily obtain $\mathrm{E}(Y|X=x)$ from β since $\mathrm{E}(Y|X=x) = x^T\beta$. Plugging in $\hat{\beta}$ produces the MLE for this expectation $\widehat{\mathrm{E}}(Y|X=x) = x^T\hat{\beta}$ with x fixed. On the other hand, in the logistic model, $\mathrm{E}(Y|X=x) = \Pr(Y=1|X=x)$ where $\log(\frac{p_{x_i}}{1-p_{x_i}}) = x_i^T\beta$. Thus, β does not offer an easy interpretation about changes in the expected response as the covariates change, and it is therefore less useful as a parameter for understanding how p_x changes with x. The mean-value parametrization is the primary parameter of interest

in both regression contexts, but in linear regression the mean-value parameter and β are interchangeable.

Another benefit of the mean-value parameterization over β in the logistic regression model is when complete separation exists. When complete separation exists β is estimated to be at infinity while p_x is estimated to be 0 or 1. We discuss complete separation and methods which address it in the next Section.

1.5 Complete Separation

Traditional maximum likelihood estimation for logistic regression does not work well when there is complete or quasi-complete separation in the data, a problem that is widespread in applications [Geyer, 2009].Agresti [2013] defines complete separation when there exists a vector b such that

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

That is, complete separation occurs when the one or more explanatory variables can perfectly predict the response variable [Albert and Anderson, 1984]. For example, as shown in the Figure 1, consider the following case that when x is less than 50, all corresponding y are 0 and when x is greater than 50, all corresponding y are 1. Suppose we are interested in a simple logistic regression model $x_i^T = [1, z_i]$. Then this data is completely separated with $b = [-50, 1]^T$. Moreover, we have $\hat{p} = 0$ for $z < 50$ and $\hat{p} = 1$ for $z > 50$.

Comment: I changed the notation around. In particular, I wrote $x_i^T = [1, z_i]$. This means that the axis labels for figures needs to be changed.

When there is complete separation, the parameter estimates $\hat{\beta}$ are "at infinity," the iteration based estimation algorithms provide a sequence of estimates that goes to infinity, and the log likelihood becomes flat when evaluated along this sequence. The left panel of Figure 2 shows the log likelihood of logistic model for this example with different working estimate from `glm` function in R. We can see that each iteration, norm of β becomes larger and asymptote of the log likelihood value goes to infinity. The right panel of Figure 2 is the zoomed part of the left panel of Figure 2 where the log of norm of working estimates is between 4.5 and 5. It displays the log likelihood value still approaches near zero although the left panel of Figure 2 looks flat in the same region. In complete separation, the usual statistical inference is not valid. The standard errors of predicted probabilities of success are very small, which leads to extremely narrow confidence intervals for each observation. 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 [R Core Team, 2020; SAS Institute Inc., 2003; Van Rossum and Drake Jr,

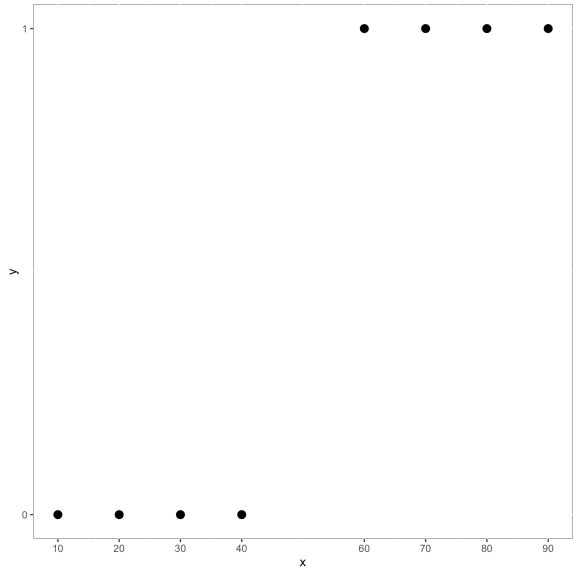

Figure 1: Example of complete separation from Section 6.5.1 of Agresti [2013]. The conventional MLE of a logistic model does not exist. **Maybe change the x-axis label to z.**

1995]. The `glmdr` software package [Geyer et al., 2021] is designed to provide users with a description of the complete separation problem when it occurs, and provide statistical inferences when it occurs.

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 $y_i = 1$ for $x \geq 50$. In this case, the maximized log likelihood is always negative and we experience same phenomenon as the complete separation case.

1.6 One-Sided Confidence Interval

We use one-sided confidence intervals for the logistic model's 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 [2021]. 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:

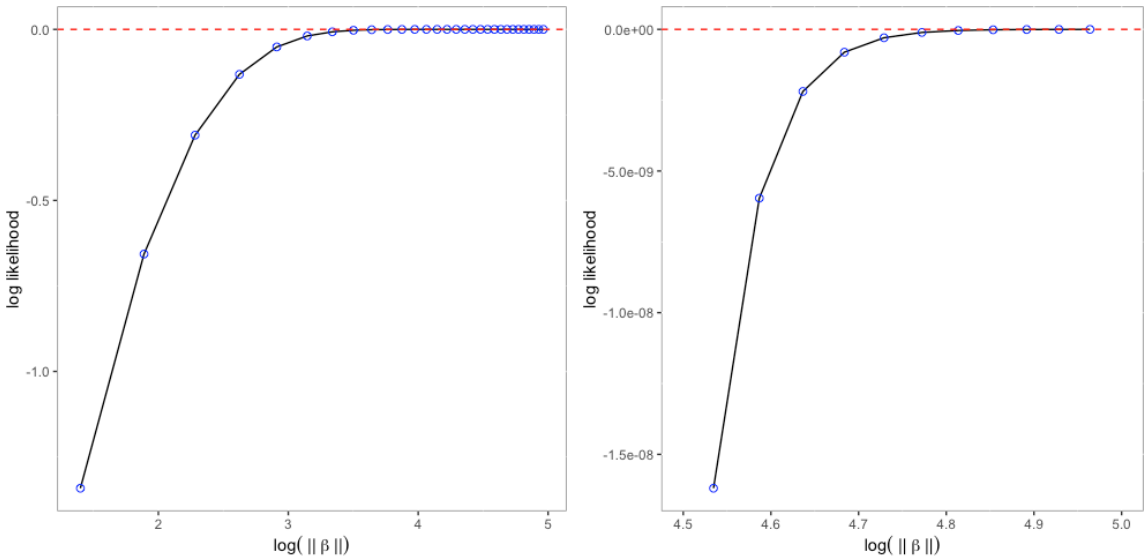

Figure 2: **Left panel**: Log likelihood values of logistic model at different working estimates. Blue dot represents the log likelihood value at each iteration. **Right panel:** Zoom in view of a log likelihood values of logistic model where log of norm of working estimates lie between 4.5 and 5.

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

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 parameter, and α is a significance level. For example, Figure 3 shows the one-sided confidence interval for the complete separation example we discussed in Section 1.5. We can see the confidence interval increases as z increases until $z = 40$ then it starts to decrease as z increases from $z = 60$. Also, we have a widest interval where $z = 40$ and $z = 60$ with the length of intervals, $1 - \alpha$. It means our uncertainty on estimation keep increases from $z = 10$ to $z = 40$ and we have the highest uncertainty near the separation occurs. Then it diminishes as it furthers away from the boundary of the separation. In `glmdr`, `inference` function provides this confidence intervals using the sequential quadratic programming (SQP) to solve the constrained nonlinear problem (7).

1.7 Prediction

Prediction in `glmdr` framework is different from that of the conventional statistical model because we do not have a finite estimate. Specifically, in the traditional sense, we

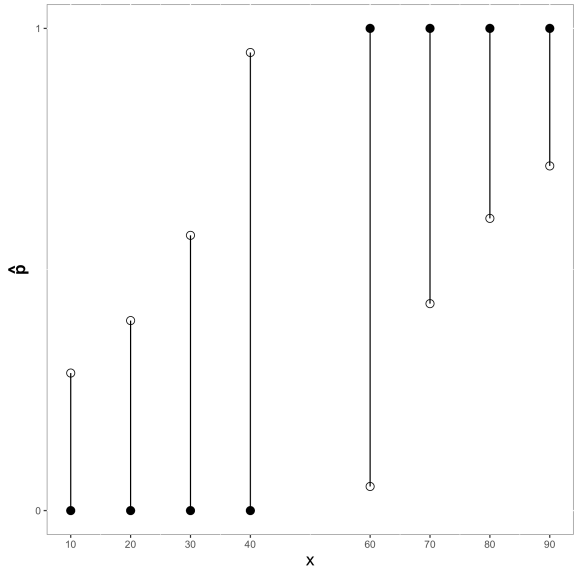

Figure 3: One-sided 95% confidence interval for the example of complete separation from Section 1.5. Solid dot represents the observed value and bar shows the interval. $\hat{p}$ is the estimated probability of a success given x.

can compute the predicted value for new data point from the logistic model using $\hat{p}_{\text{pred}} = (1 + \exp(-x_{\text{new}}^T \hat{\beta}))^{-1}$. However, when the complete separation presents, this approach does not work. Therefore, we propose a new method for the prediction that we fit two possible models for new data point with different value of a response variable then compute the weighted conditional probability of a success.

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 $y_{\text{new}} = 0$ and the other has $y_{\text{new}} = 1$ for new data point. Based on these two datasets, we fit two logistic models to compute the estimated probability of a success for new data points, $\hat{p}_1$ and $\hat{p}_2$. Since we do not know which model is fitted from the true value of response variable, we compare the weight of evidence for each model based on the Akaike weights for the model selection [Burnham and Anderson, 2002]. Let w_j be the weight for model j defined by:

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

where IC_j is the information criteria of model j. Then we can calculate the model averaged estimate, $\hat{p}^* = \sum_{j=1}^{2} w_j \hat{p}_j$. This averaged estimate is especially useful for prediction in our framework because we can use all predicted probabilities from models we have. For IC, since our sample size is more likely to be small when the complete separation presents, we recommend the Akaike information criteria corrected (AICc). The primary reason is that

AICc does not have an overfit problem despite of small sample size [Sugiura, 1978]. Also, it converges to Akaike information criteria (AIC) when we have large sample size, and AIC is asymptotically equivalent to choice of model by leave-one-out cross validation [Stone, 1977]. Meanwhile, Bayesian information criteria (BIC) attempts to find the true model among the sets of candidate models which is not appropriate our prediction framework [Schwarz, 1978]. We then label 1 if $\hat{p}^* \geq C^*$ and 0 if $\hat{p}^* < C^*$ where C^* is the optimal cut-off that maximizes the overall accuracy. The main motivation of using optimal cut-off is that threshold of 0.5 produces unreliable and poor model accuracy when the response variable is highly unbalanced [Freeman and Moisen, 2008]. For prediction intervals, we construct the Wilson intervals [1927] for predicted probabilities. Wilson intervals show better coverage probability although $\hat{p}$ is near 0 and 1 boundaries in comparison to the standard binomial confidence interval because Wilson intervals are asymmetric [Brown et al., 2001]. Detailed implementation and examples are given in the supplementary materials.

2 Results

2.1 Inference

We report the in-sample accuracy for all observations and confidence intervals for observations that occur the (quasi) complete separation to compare each method. For `brglm`, it is theoretically equivalent to the `logistf` when `brglm` uses the maximum penalized likelihood with powers of the Jeffreys prior as penalty. However, `brglm` fails to converge for the maize example, meanwhile, `logistf` converges. Therefore, we use `logistf`'s result for `brglm` in maize example. For confidence intervals, we compute the average length of one-sided confidence interval for `glmdr` and average length of Wilson intervals for `bayesglm`, `brglm` (`logistf`) and linear models (since the predicted value of linear model does not have to fall into $[0, 1]$ range, we assign 1 for any predicted values greater than 1 and 0 for negative values). In Table 1, we can see all methods show the equivalent in-sample accuracy for the complete separation and quasi separation examples. Meanwhile, the logistic models, `glmdr`, `bayesglm`, and `brglm` (`logistf`), display the higher in-sample accuracy for quadratic, endometrial, and maize examples in comparison to the linear model. Within these examples, `glmdr` has the highest in-sample accuracy in maize example than other two logistic models. For confidence intervals, `glmdr` demonstrates the smallest length in all examples. Especially, in quadratic and endometrial examples, its lengths of confidence intervals are significantly smaller than other methods. Two logistic models, `bayesglm` and `brglm` (`logistf`) generally shows smaller lengths of confidence intervals but they are not highly different from that of linear model in all examples. This result suggests that linear model perform worse than logistic models, and `glmdr` which solves the complete separation within the MLE framework produces the most accurate inference for (quasi) complete separation problem.

Comment: I think the tables can formatted better. Perhaps something like the following would be better, especially when presenting the prediction results:

		Complete Separation	...
glmdr	accuracy	100%	
	length	0.55	
bayesglm	accuracy	100%	
	length	0.83	

Or maybe the following would be better:

		Complete Separation	...
accuracy	glmdr	100%	
	bayesglm	100%	
	⋮	⋮	⋮
length	glmdr	0.55	
	bayesglm	0.83	

I personally think the second format is better.

Table 1: Model performances for all examples.

glmdr denotes Generalized Linear Model Done Right [Geyer et al., 2021], bayesglm denotes Generalized Linear Model with Student-t prior distribution [Gelman et al., 2008], brglm denotes Bias Reduction in Generalized Linear Models [Kosmidis and Firth, 2009], logistf denotes Logistic model with Firth's modified score function [Heinze and Schemper, 2002], and linear denotes the multiple linear model using ordinary least squares.

	in-sample accuracy				average length of confidence intervals			
	glmdr	bayesglm	brglm2 / logistf	linear	**glmdr**	bayesglm	brglm2 / logistf	linear
Complete Separation	100 %	100 %	100 %	100 %	0.55	0.83	0.84	0.83
Quasi Separation	90 %	90 %	90 %	90 %	0.31	0.83	0.83	0.83
Quadratic	100 %	100 %	100 %	90.00 %	0.20	0.82	0.81	0.86
Endometrial	88.61 %	88.61 %	88.61 %	86.08 %	0.19	0.80	0.81	0.81
Maize	87.14 %	87.07 %	87.01 %	86.81 %	0.56	0.81	0.83	0.84

2.2 Prediction

To compare the performance of prediction, we compare out-of-sample accuracy, prediction intervals and computational cost. We use the leave-one-out cross validation (LOOCV) for out-of sample accuracy, Wilson intervals for the prediction intervals, and `proc.time` function in R to measure the execution time. In Table 2, we can see all methods except for `glmdr` correctly classify all 8 data points in complete separation example. Meanwhile, `glmdr` misclassifies one data point near the boundary where the complete separation occurs. On the other hands, `glmdr` shows the highest out-of-sample accuracy in quasi complete separation

Table 2: Prediction results for all examples.

glmdr denotes Generalized Linear Model Done Right [Geyer et al., 2021], bayesglm denotes Generalized Linear Model with Student-t prior distribution [Gelman et al., 2008], brglm denotes Bias Reduction in Generalized Linear Models [Kosmidis and Firth, 2009], logistf denotes Logistic model with Firth's modified score function [Heinze and Schemper, 2002], and linear denotes the multiple linear model using ordinary least squares.

	out-of-sample accuracy				average length of prediction intervals			
	glmdr	bayesglm	brglm2 / logistf	linear	**glmdr**	bayesglm	brglm2 / logistf	linear
Complete Separation	87.5 %	100 %	100 %	100 %	0.821	0.839	0.843	0.833
Quasi Separation	90 %	80 %	80 %	80 %	0.865	0.845	0.847	0.844
Quadratic	93.33 %	93.33 %	100 %	90.00 %	0.807	0.828	0.813	0.861
Endometrial	87.34 %	86.08 %	86.08 %	86.08 %	0.839	0.843	0.844	0.851
Maize	86.04 %	86.36 %	86.30 %	86.55 %	0.836	0.837	0.837	0.839

and endometrial examples where other three methods perform the same. In quadratic example, `brglm` performs the best followed by other two logistic models and linear model, but linear model is better than the logistic models in maize data although their differences are not large. This result is surprising because the linear model is generally not recommended for binary classification, yet it shows a better performance than the logistic models. For prediction intervals, overall there is no significant difference between each method. We notice that `glmdr` has the smallest lengths of prediction intervals in all examples but for the quasi complete separation example where the linear model displays the smallest length of prediction intervals.

We present the computational cost of each method in Table 3. In all examples, linear model is much faster than logistic models. Although there is no significant difference in complete separation, quasi complete separation, quadratic, and endometrial examples, computational cost of `glmdr` increases much in maize example because execution time for `glmdr` increases as it requires more computations to solve the optimization problem if the data point to be predicted occur the separation. Similarly, `brglm` is notably slow because it needs to handle optimization problem to find the penalized MLE for each iteration. However, `bayesglm` does not suffer this issue because it does not carry the computation for the optimization problem in their method.

Considering all aspects, linear model performs well despite of the binary response. It shows comparable or better out-of-sample accuracy, small prediction intervals and lowest computational cost. Among logistic models, `bayesglm` presents the most rounded performance with fairly low computational cost. `glmdr` sometimes shows the highest out-of-sample accuracy and often smallest length of prediction intervals. But, it may not be scalable to the large datasets due to high computational cost. `brglm` rarely achieve the highest out-of-sample accuracy but its computational cost in large datasets is major issue.

Comment: The last paragraph should be changed. It looks like all of the methods are pretty comparable, but there are some noticeable differences. For example, glmdr is the only method that performs well on the quasi separation dataset (remember that 95% is the

target). The other difference is in computational costs and the brglm2/logistf seemingly fails when applied on the Maize dataset. It also looks like glmdr offers the smallest prediction regions across the board except in the quasi separation example. This is interesting since the glmdr method is the only method close to the desired accuracy in this example.

Table 3: Computational cost for all examples.

glmdr denotes Generalized Linear Model Done Right [Geyer et al., 2021], bayesglm denotes Generalized Linear Model with Student-t prior distribution [Gelman et al., 2008], brglm denotes Bias Reduction in Generalized Linear Models [Kosmidis and Firth, 2009], logistf denotes Logistic model with Firth's modified score function [Heinze and Schemper, 2002], and linear denotes the multiple linear model using ordinary least squares.

	glmdr	bayesglm	brglm2 / logistf	linear
Complete Separation	0.13 secs	0.11 secs	0.19 secs	0.07 secs
Quasi Separation	0.27 secs	0.12 secs	0.19 secs	0.06 secs
Quadratic	0.31 secs	0.35 secs	0.44 secs	0.09 secs
Endometrial	1.06 secs	0.31 secs	0.59 secs	0.14 secs
Maize	4.74 mins	45.35 secs	2.26 hours	4.63 secs

3 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 necessary assumptions such as homoscedasticity and linearity (i.e. Gauss-Markov assumptions) as well as normality. Therefore, even though results from Section 2.1 and 2.2 display that the performance of linear model is comparable or better than the logistic models, we can not fully utilize asymptotic properties of linear model and make a proper inference such as significance tests for coefficients.

On the other hand, `glmdr` is considered to be the most preferable logistic model based on its overall performance in the inference and prediction. The main strength of `glmdr` is it provides the best inference because the way that `glmdr` handles the separation problem is the true remedy to the traditional `glm`'s separation issue. It solves the separation issue within the maximum likelihood estimation framework unlike other two logistic models and estimates the probability of success by finding the MLE in the Barndorff-Nielsen completion [1978] based on approximate null eigenvectors of the Fisher information matrix. Meanwhile, other two logistic models solve the separation problem by switching the problem settings. For example, `bayesglm` adopts a Bayesian approach which scales the data first and then places Cauchy distribution as a prior distribution on the coefficients and `brglm` modifies the score function to produce finite coefficients. As a result, not only are both models'

results in inference not the best, but it is also hard to see their outputs as a true solution for separation problem of `glm`. In prediction, `glmdr` often shows the highest out-of-sample accuracy with the narrowest length of prediction intervals with acceptable computational cost. It may take much time when we have a large number of observations, but the complete separation is likely to occur when we have a small sample size. Thus, high computational cost in large sample size should not be the major issue in `glmdr`.

In conclusion, when separation issue present in the logistic model, one can consider using the `glmdr` which has the advantage in inference and the comparable prediction power. `bayesglm` is suitable for prediction in large datasets thanks to its low computational cost yet high accuracy. `brglm` or `logistf` may be least preferable method because they are computationally unstable and expensive.

References

A. 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`.

O. Barndorff-Nielsen. *Information and exponential families: in statistical theory*. J. Wiley & Sons, 1978.

L. D. Brown, T. T. Cai, and A. DasGuptal. 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`.

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

D. J. Eck and C. J. Geyer. Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist. *Electronic Journal of Statistics*, 15(1), 2021. doi: 10.1214/21-ejs1815.

E. A. Freeman and G. G. Moisen. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, 217 (1):48–58, 2008. ISSN 0304-3800. doi: https://doi.org/10.1016/j.ecolmodel.2008.05.015. URL `https://www.sciencedirect.com/science/article/pii/S0304380008002275`.

A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distributionfor logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360 – 1383, 2008. doi: 10.1214/08-AOAS191. URL `https://doi.org/10.1214/08-AOAS191`.

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

C. J. Geyer, D. J. Eck, and S. Park. *glmdr: Exponential Family Generalized Linear Models Done Right*, 2021. URL `https://github.com/DEck13/complete_separation`. R package version 0.3.

G. Heinze and M. 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`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

M. C. Romay, M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts, T. M. Casstevens, R. J. Elshire, C. B. Acharya, S. E. Mitchell, S. A. Flint-Garcia, M. D. McMullen, J. B. Holland, E. S. Buckler, and C. 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`.

SAS Institute Inc. *SAS/STAT Software, Version 9.1*. Cary, NC, 2003. URL `http://www.sas.com/`.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 1978. doi: 10.1214/aos/1176344136.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47, 1977. doi: 10.1111/j.2517-6161.1977.tb01603.x.

N. Sugiura. Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978. doi: 10.1080/03610927808827599.

G. Van Rossum and F. L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.