Robust model based prediction of gene expression in maize

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

Month 2021

Abstract

Check this out - use this as a template: `http://www.cbs.umn.edu/sites/default/files/public/downloads/Annotated_Nature_abstract.pdf`

Key Words: list of keywords

1 Introduction

Paragraph 1: Begin with a big-picture problem, and narrow things down to how binary response variables are critical for solving the big-picture problem

The logistic model is one of the most common statistical model for the classification problem. However, the coefficient of the logistic model is not finite (or unstable) when the separation exists. This separation issue occurs when one or more covariates perfectly predicts the response variable for the whole training data. In such case, the maximum likelihood estimate (MLE) does not exist because the likelihood value of the logistic model diverges. Therefore, any interpretations or conducting significance tests on coefficients is meaningless. Moreover, common statistical software do not provide sufficient information on what the problem is and how to handle this appropriately. Thus, a person who is not familiar with the separation issue may use the incorrect model unconsciously.

The easiest way to deal with the separation issue is to remove the problematic covariate until we have finite coefficients. However, this naïve approach often leads us to get rid of the highly correlated covariates with the response variable [Zorn, 2005]. Alternatively, Heinze and Schemper [2002] use the Firth's penalized maximum likelihood estimation which is initially proposed to reduce the bias of maximum likelihood estimator to obtain the finite parameter estimates. Kosmidis and Firth [2009] then generalize this method for the nonlinear exponential family. These bias reduction methods enable us to estimate coefficients when the coefficients of problematic covariates are at infinity. On the other hand, researchers try to assign various prior distributions on the coefficient to solve this problem in the Bayesian framework. For example, Heinze and Schemper's method can be seen as the application of the Jeffrey's invariant prior. Dunson et al. [2006] use the mixture prior distributions for the logistic model with large number of covariates and Genkin et al. [2007] consider the Laplace prior distribution. Gelman et al. [2008] suggest the Cauchy distribution with center 0 and scale 2.5 as the default choice, and this method shows faster and better performance in the prediction in comparison to the other methods. However, both bias reduction and Bayesian approaches handle the separation issue by switching the problem setting rather than solving the issue within the original model. Meanwhile, Geyer [2009] devises the method to find the MLE in the Barndorff-Nielsen completion of the exponential family from the original model even if the MLE does not exist in the conventional sense. Thus, his method and proposed one-sided confidence interval provide valid and the most accurate statistical inference in the separation problem. However, this method requires a massive computation cost which makes it hard to apply to the practical problems. Hence, Eck and Geyer [2021] propose new, faster and scalable method to find the MLE in the completion when MLE does not exist. In this study, we propose the prediction framework in the Eck and Geyer [2021]'s method. Considering the important role of statistical model is often to preform the inference and prediction, our work can make Eck and Geyer [2021]'s method more practical and useful to use when the separation issue presents.

Paragraph 4: Begin by explicitly stating what the purpose of the research presented in this paper is. After that, describe the objectives of the study. It would be a great idea to provide an overarching hypothesis, and what you would expect/predict ot see in your results if that hypothesis were true.

2 Materials and Methods

2.1 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.

2.2 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 analyzed the Agresti [2013] example discussed in Section 2.5. In this example, we had one binary response variable, $y \in \{0, 1\}$ and one covariate variable, z, with 8 data points. Specifically, $y_i = 1$ at $z = 10, 20, 30, 40$, and $y_i = 0$ at

$z = 60, 70, 80, 90$. Since y could be completely separable by z, we observed the complete separation in this example.

Quasi-complete separation: We analyzed the Agresti [2013] example with two points added, $y_i = 1$ and $y_i = 0$ at $z = 50$ to the previous example. Thus, we had 10 data points where $y_i = 1$ at $z = 10, 20, 30, 40, 50$ and $y_i = 0$ at $z = 50, 60, 70, 80, 90$. We observed the quasi-complete separation in this example.

Quadratic logistic regression model: This example comes from Section 2.2 of Geyer [2009]. We had one binary response variable, $y \in \{0, 1\}$ and two covariate variables, z which took a value from 1 to 30 by 1 and its quadratic term, z^2. The response variable was $y_i = 1$ for $12 < z_i < 24$ and $y_i = 0$, otherwise. In this case, we fitted a quadratic logistic model, and we observed the complete separation in z.

Endometrial Cancer Study: Heinze and Schemper [2002] 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). The response variable had 30 patients classified grading 0-II for histology (HG = 1) and 49 patients for grading III-IV (HG = 0). There were 13 patients who had 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 observed the quasi-complete separation in NV.

Maize data: To predict the kernel color of maize, we used the data set from Romay et al. [2013] that consists of 2,815 maize lines. The binary response variable we considered was the kernel color, where 1 indicated non-white kernel color and 0 indicated white kernel color. We fitted various models with kernel color as the response variable and 24 DNA markers surrounding the *psy1* gene. Each marker had a value from 0 to 1. In the final data set, 309 lines had a white kernel and 1,238 had non-white kernel color. These maize lines were subdivided into six subpopulaitons, namely 115 non-stiff stalk, 54 popcorn, 120 stiff stalk, 116 sweet corn, 159 tropical, and 983 unclassified. In this example, there was no separation issues when we used a single marker for covariate. However, we had a separation issue for saturated model with 24 DNA markers and subpopulations.

2.3 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

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

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

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

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).

2.4 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 $\log(\frac{p_x}{1-p_x}) = x^T\beta$. 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.

2.5 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 $\hat{p}_x = 0$ for $z < 50$ and $\hat{p}_x = 1$ for $z > 50$.

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

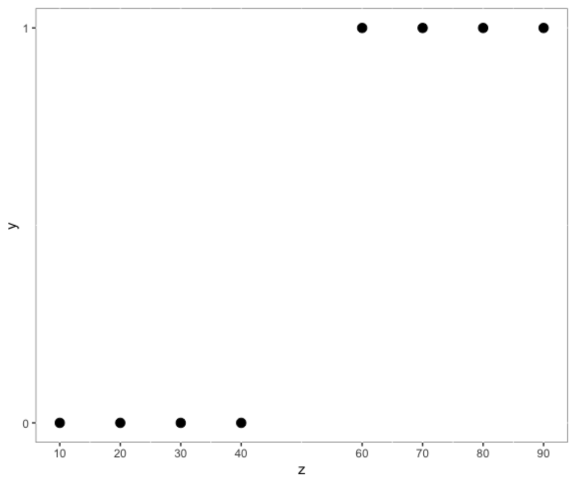

Figure 1: Example of complete separation from Section 6.5.1 of Agresti [2013]. The conventional MLE of a logistic model does not exist.

knowing it [R Core Team, 2020; SAS Institute Inc., 2003; Van Rossum and Drake Jr, 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 $z = 50$ with $y = 1$ and $y = 0$ to the previous complete separation example. That is, we have $y_i = 0$ for $z \leq 50$ and $y_i = 1$ for $z \geq 50$. In this case, the maximized log likelihood is always negative and we experience same phenomenon as the complete separation case.

2.6 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:

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

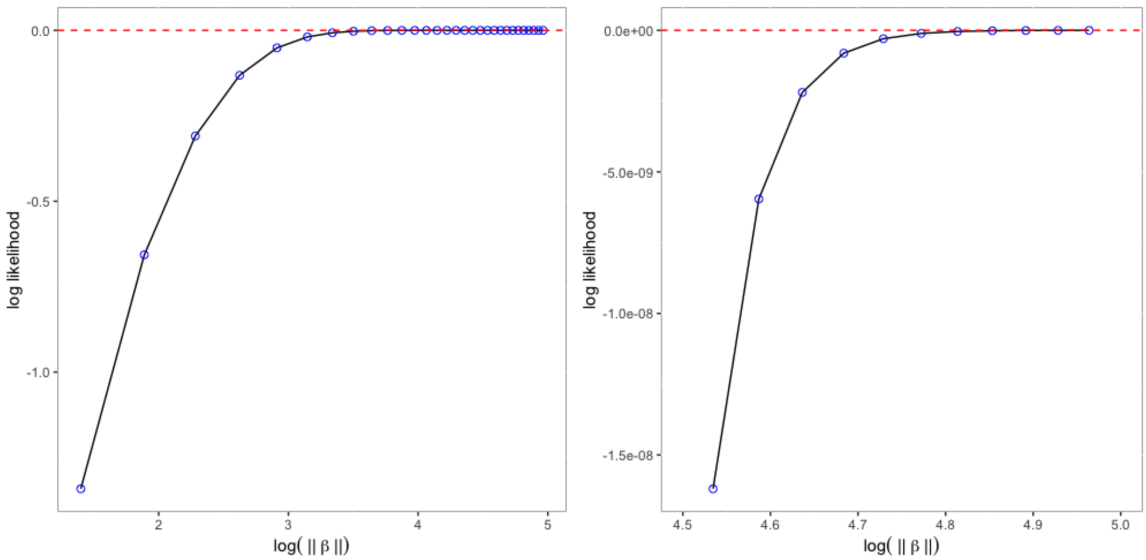

Figure 2: **Left panel**: Log likelihood values of logistic model at different working estimates. Blue dot represents the log likelihood value at each iteration. **Right panel:** Zoom in view of a log likelihood values of logistic model where log of norm of working estimates lie between 4.5 and 5.

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 2.5. 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).

2.7 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 can compute the predicted value for new data point from the logistic model using $\hat{p}_{x_{\text{pred}}} = (1 + \exp(-x_{\text{new}}^T \hat{\beta}))^{-1}$. 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

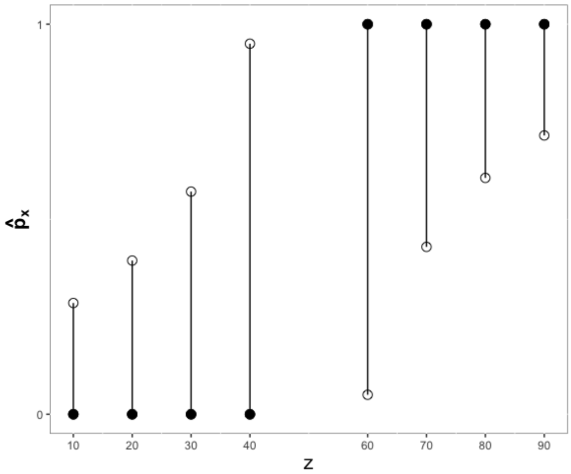

Figure 3: One-sided 95% confidence interval for the example of complete separation from Section 2.5. Solid dot represents the observed value and bar shows the interval. $\hat{p}_x$ is the estimated probability of a success given z.

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, $\hat{p}_{x_0}$ and $\hat{p}_{x_1}$. 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, $\hat{p}_x^* = \sum_{j=0}^{1} w_j \hat{p}_{x_j}$. 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 $\hat{p}_x^* \geq C^*$ and 0 if $\hat{p}_x^* < C^*$ 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 $\hat{p}_x$ 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.

3 Results

3.1 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.

3.2 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 show the same accuracy for the complete separation and quasi separation examples. `glmdr` shows the highest out-of-sample accuracy in endometrial example 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 example although their differences are not large. This result is surprising because the linear

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.

		Complete Separation	Quasi Separation	Quadratic	Endometrial	Maize
accuracy	glmdr	100 %	90 %	100 %	88.61 %	87.14 %
	bayesglm	100 %	90 %	100 %	88.61 %	87.07 %
	brglm / logistf	100 %	90 %	100 %	88.61 %	87.01 %
	linear	100 %	90 %	90 %	86.08 %	86.81 %
length	glmdr	0.550	0.308	0.199	0.194	0.563
	bayesglm	0.828	0.827	0.823	0.804	0.814
	brglm / logistf	0.835	0.831	0.811	0.808	0.826
	linear	0.829	0.829	0.859	0.806	0.838

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.

Table 2: Prediction results and 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.

		Complete Separation	Quasi Separation	Quadratic	Endometrial	Maize
accuracy	glmdr	100 %	80 %	93.33 %	87.34 %	86.04 %
	bayesglm	100 %	80 %	93.33 %	86.08 %	86.36 %
	brglm / logistf	100 %	80 %	100 %	86.08 %	86.30 %
	linear	100 %	80 %	90 %	86.08 %	86.55 %
length	glmdr	0.822	0.859	0.807	0.839	0.836
	bayesglm	0.839	0.845	0.828	0.843	0.837
	brglm / logistf	0.843	0.847	0.813	0.844	0.837
	linear	0.833	0.844	0.861	0.851	0.839
cost	glmdr	0.13 secs	0.27 secs	0.31 secs	1.06 secs	4.74 mins
	bayesglm	0.11 secs	0.12 secs	0.35 secs	0.31 secs	45.35 secs
	brglm / logistf	0.19 secs	0.19 secs	0.44 secs	0.49 secs	2.26 hours
	linear	0.07 secs	0.06 secs	0.09 secs	0.14 secs	4.63 secs

We present the computational cost of each method in Table 2. 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 the 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, all of four methods demonstrate comparable out-of-sample accuracy and length of prediction intervals. However, there are several notable differences. glmdr provides the smallest lengths of prediction intervals except in the quasi separation example. It also shows better performance in endometrial example. But, it may not be scalable to the large datasets due to relatively high computational cost. bayesglm performs well on all examples with the lowest computational cost, which indicates the bayesglm is suitable for prediction on large data. brglm achieves the highest out-of-sample accuracy in the quadratic example, but brglm fails to converge in maize example and alternative method, logistf, is very costly. Meanwhile, the linear model performs well despite of the binary response. It shows comparable or better out-of-sample accuracy with small prediction intervals and the lowest computational cost.

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 necessary assumptions such as homoscedasticity and linearity (i.e. Gauss-Markov assumptions) as well as normality. Therefore, even though results from Section 3.1 and 3.2 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 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 as 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 shows similar or better out-of-sample accuracy when the quasi-complete separation presents, and comparable performance when the complete separation exists 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. B. Dunson, A. H. Herring, and S. A. M. Engel. Bayesian selection and clustering of polymorphisms in functionally-related genes. *J. AM. STATIST. ASSOC*, 2006.

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`.

A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007. doi: 10.1198/004017007000000245. URL `https://doi.org/10.1198/004017007000000245`.

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.

C. Zorn. A solution to separation in binary response models. *Political Analysis*, 13(2):157–170, 2005. doi: 10.1093/pan/mpi009.