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

Comment: You should include the other software packages used in this analysis.

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. 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]. 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, maximum likelihood estimate (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 = 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

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

where β is an unknown parameter vector.

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, 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

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

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

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

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).

1.4 Mean-value Parameters

Comment: I changed the first paragraph, see if it makes sense.

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 linear model, we can easily obtain this expected value 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, $\mathrm{E}(Y|X=x) = \Pr(Y=1|X=x) = p_i$ while $\log(\frac{p_i}{1-p_i}) = x_i^T\beta$. Therefore, β does not offer easy interpretation about changes in the expected response as the covariates change.

To get the expected value from the logistic model, we can consider the mean-value parameterization directly. Instead of indirectly using the estimated coefficients of the logistic

regression model, we can have the estimated conditional probability of success given data by plugging in $\hat{\beta}$ into (1). The advantage of this parameterization is now our parameters of interest shifts to the mean-value parameters from the coefficients of the model. Consequently, we can provide a more informative and intuitive inference. For example, we can tell the expected probability of success at particular x which is what we desire from the statistical model (without mean-value parameterization, our interpretation on model is that as one unit of explanatory variable increases the expected change in log odds ratio of conditional probability of success is $\hat{\beta}$).

Comment: The above paragraph needs work, you mention plugin after stating problems with β. I think that the mean-value parameterization should be separated from the canonical parameterization. It is fine to mention that estimating the mean-value parameters begins with estimating β. But it is not as simple is plugin, at least when inference is concerned.

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 $x^T = [1, x_i]$. Then this data is completely separated with $b = [-50, 1]^T$. Moreover, we have $\hat{p} = 0$ for $x < 50$ and $\hat{p} = 1$ for $x > 50$.

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

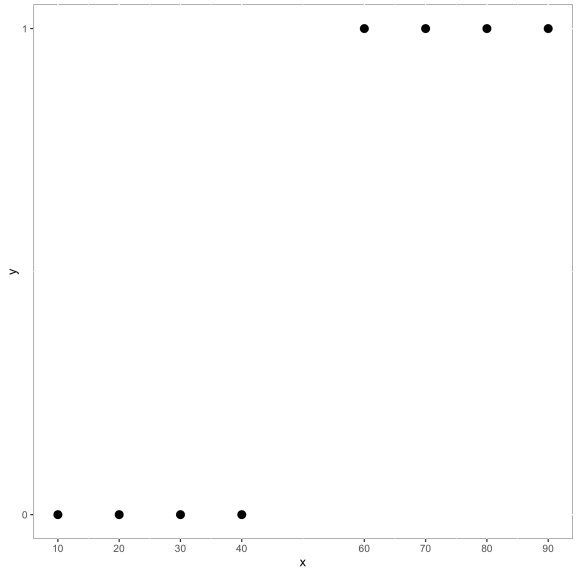

Figure 1: Example of complete separation from Section 6.5.1 of Agresti [2013]. The conventional MLE of a logistic model does not exist.

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 (**Provide specific references**). The `glmdr` software package (**I think a citation should go here**) 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

Comment: Choose mean value or mean-value throughout, do not mix and match.

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

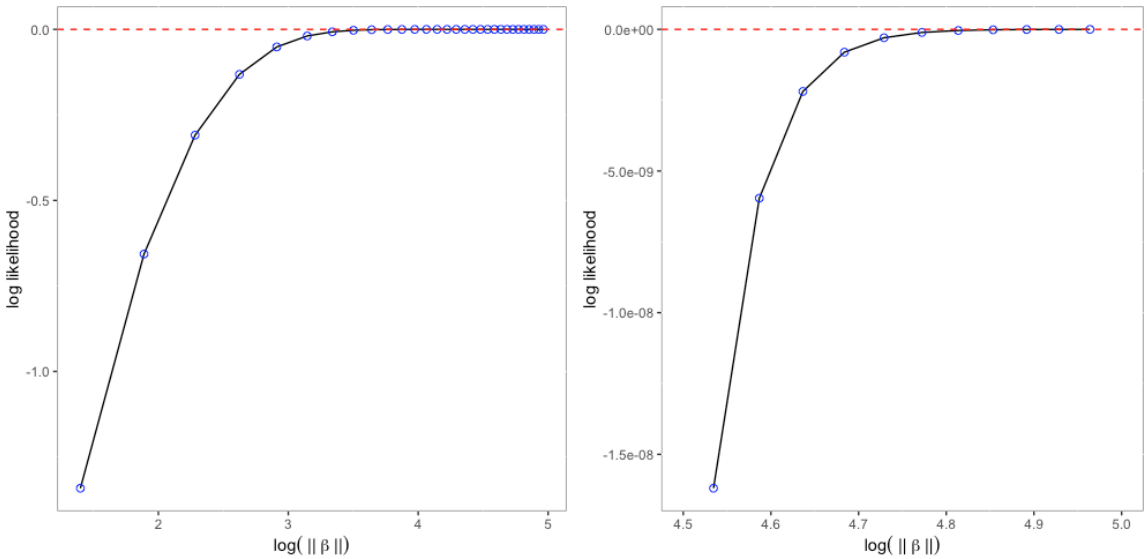

Figure 2: **Left panel**: Log likelihood values of logistic model at different working estimates. Blue dot represents the log likelihood value at each iteration. **Right panel:** Zoom in view of a log likelihood values of logistic model where log of norm of working estimates lie between 4.5 and 5.

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:

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

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 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 x increases until $x = 40$ then it starts to decrease as x increases from $x = 60$. Also, we have a widest interval where $x = 40$ and $x = 60$ with the length of intervals, $1 - \alpha$. It means our uncertainty on estimation keep increases from $x = 10$ to $x = 40$ 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).

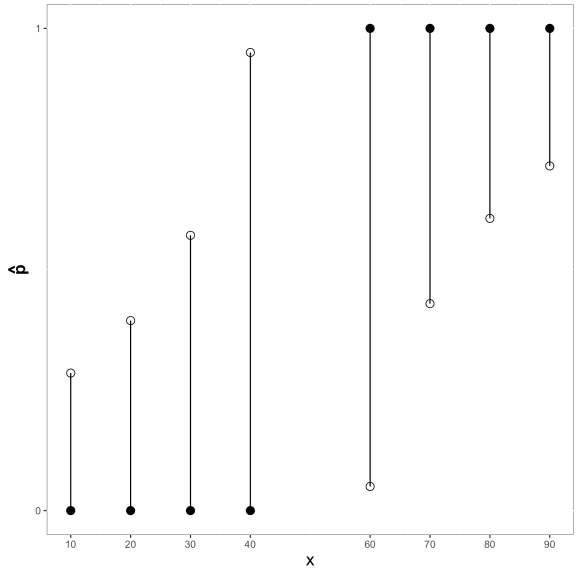

Figure 3: One-sided 95% confidence interval for the example of complete separation from Section 1.5. Solid dot represents the observed value and bar shows the interval. $\hat{p}$ is the estimated probability of a success given x.

1.7 Prediction

Comment: You need to use a consistent notation for subscripts. For example, x_{new} should be x_{new}. Maybe write $\hat{p}_1$ instead of $\widehat{p_1}$, consistency is what is important.

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 $\hat{p}_{pred} = (1 + \exp(-x_{new}^T \hat{\beta}))^{-1}$. 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, $\widehat{p^*} = \sum_{j=1}^{2} w_j \hat{p_j}$. This averaged estimate is especially useful for prediction in our framework because we can use all predicted probabilities from models we have. For IC, we recommend the Akaike information criteria corrected (AICc) because Akaike information criteria (AIC) is asymptotically equivalent to choice of model by leave-one-out cross validation [Stone, 1977].

Comment: More is needed on the choice of AICc. I think the primary reason for using it is that it is appropriate for small finite samples. This is likely the setting that one is in when there is complete separation. I see that you actually mention this point a little later, it should be mentioned first. Asymptotic considerations should go next.

Also, 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]. Furthermore, AICc works well despite of small sample size and converges to AIC when we have large sample size [Sugiura, 1978]. We construct the prediction intervals based on Wilson intervals given model averaged predicted value. Wilson intervals [1927] are asymmetric unlike the standard binomial confidence interval. Thus, Wilson intervals show better coverage probability although $\hat{p}$ is near 0 and 1 boundaries [Brown et al., 2001].

Comment: I think that you should mention the coverage properties of Wilson intervals before mentioning the asymmetric nature.

Lastly, we label the mean of this Wilson intervals. In our method, we label 1 if $\widehat{p^*} \geq C^*$ and 0 if $\widehat{p^*} < C^*$ where C^* is the optimal cut-off that maximizes the overall accuracy. Detailed implementation and examples are given in the supplementary materials.

1.8 Optimal Cut-off

Comment: I am not sure if this needs to be its own subsection.

In the logistic regression, the most common threshold for mapping predicted probabilities of a success to a 0 or 1 label is 0.5. However, this threshold is not often an optimal cut-off. Freeman and Moisen [2008] find that the model accuracy is heavily affected by the

choice of this threshold. Especially, they point out that threshold of 0.5 produces unreliable and poor model accuracy when the response variable is highly unbalanced. In the literature, many criteria are suggested to optimize the threshold with regards to different metrics such as sensitivity, specificity and so on [Congalton, 1991; Fielding and Bell, 1997; Cantor et al., 1999; Manel et al., 2001]. Therefore, one can choose the most appropriate criteria for their goal. In our method, we select the threshold criteria such that maximizes the overall accuracy, called maximize percent correctly classified (MaxPCC). For the comparison of criteria and their implementation, see Freeman and Moisen [2008] and the references therein.

2 Results

Comment: Should the title of the following subsection be Inference instead of Estimation?

2.1 Estimation

Comment: The paragraph below is jarring, and seems to come from nowhere. The estimation section should be about coverage probabilities, length of confidence intervals, etc. The story should be about all of the datasets and not just the maize dataset. You can then discuss estimation challenges with respect to the maize dataset.

In maize data, 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 was 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. **The following sentence would be a better start to this section:** We report the average length of one-sided confidence interval for `glmdr` and average length of Wilson intervals for each predicted probability from `bayesglm`, `logistf` and linear models for fair comparison (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 model performs comparably but `glmdr` provides narrowest length of confidence intervals which indicate the inference result of `glmdr` is the most certain.

Comment: Put the explanation of the acronyms in the caption, not as a footnote. I am

not sure that we need to indicate that glmdr is our method.

Table 1: Model performances for all examples.

	in-sample accuracy				average length of confidence intervals			
	glmdr[1] (Ours)	bayesglm[2]	logistf[3]/ brglm2[4]	linear[5]	**glmdr (Ours)**	bayesglm	logistf / brglm2	linear
Complete Separation	100 %	100 %	100 %	100 %	0.55	0.83	0.84	0.83
Quasi Separation	90 %	90 %	90 %	90 %	0.42	0.84	0.84	0.84
Quadratic	100 %	100 %	100 %	90 %	0.20	0.82	0.81	0.86
Endometrial	88.61 %	88.61 %	88.61 %	86.08 %	0.74	0.84	0.84	0.86
Maize	87.14 %	87.07 %	87.01 %	86.81 %	0.82	0.84	0.84	0.84

2.2 Prediction

Comment: Like the previous section, this section should be about all of the datasets.

We use the leave-one-out cross validation (LOOCV) for prediction. This setting is the most suitable as we want to predict the kernel color of new maize inbreed given our data. Table 2 displays the out-of-sample accuracy, execution time and optimal cut-off from a MaxPCC for each data and method. Out-of-sample accuracy is calculated by sum of number of true positive and true negative divided by total number of samples in testing set and the execution time is measured by difference between the starting and ending time of the computation using `proc.time` function in R. For accuracy, we can see all methods perform similarly in complete separation, quasi-complete separation and maize data. Also, logistic models perform better than linear model in quadratic and endometrial data. Meanwhile, for execution time, `linear` model performs fastest followed by `bayesglm`, `glmdr`, and `logistf`/`brglm2` is significantly slow.

Comment: Should we add average length of the prediction regions?

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 eas-

[1]Our model, Generalized Linear Model Done Right from `glmdr` package.
[2]Generalized Linear Model with Student-t prior distribution from `arm` package [Gelman et al., 2008].
[3]Logistic model with Firth's modified score function from `logistf` package [Heinze and Schemper, 2002].
[4]Bias Reduction in Generalized Linear Models from `brglm2` package [Kosmidis and Firth, 2009].
[5]Multiple Linear Model using ordinary least squares.

Table 2: Prediction results for all examples.

	out-of-sample accuracy				execution time				cut-off
	glmdr[1] (Ours)	bayesglm[2]	logistf[3]/ brglm2[4]	linear[5]	**glmdr (Ours)**	bayesglm	logistf / brglm2	linear	
Complete Separation	87.5 %	100 %	100 %	100 %	0.13 secs	0.11 secs	0.19 secs	0.07 secs	0.465
Quasi Separation	90 %	80 %	80 %	80 %	0.27 secs	0.12 secs	0.19 secs	0.06 secs	0.510
Quadratic	93.33 %	93.33 %	93.33 %	83.33 %	0.31 secs	0.35 secs	0.44 secs	0.09 secs	0.450
Endometrial	87.34 %	86.08 %	86.08 %	81.01 %	1.06 secs	0.31 secs	0.59 secs	0.14 secs	0.500
Maize	86.04 %	86.30 %	86.04 %	86.36 %	4.74 mins	45.35 secs	2.26 hours	4.63 secs	0.540

See the footnote on the bottom of page 9 for the meaning of each acronym.

iness 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 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, we can see all logistic models in Section 2.1 and 2.2 perform similarly despite of different approaches and techniques.

Comment: I think glmdr performs best overall, no? It is by far the best in inference. It is often the top performer in prediction while being only slightly worse than the others in the maize dataset. We should think about how to word this... On one hand, MLE should be the best in inference and it is expected that glmdr kills the competition on this front. On the other hand, it exhibits comparable and often favorable predictive performance although it is quite slow.

The main difference between `glmdr` and other methods is that `glmdr` is only model that solves the separation problem within the maximum likelihood estimation framework under the subset of the original model, called limiting conditional model (LCM). It 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. Hence, the way `glmdr` handles the separation problem is the true remedy to the traditional `glm`'s issue causing from a separation problem. Meanwhile, all other methods solve the separation problem by switching the problem settings. For example, `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. As a result, it is hard to see their outputs as a true solution for separation problem of `glm`.

In conclusion, when separation issue present in the logistic model, one can consider using the `glmdr` which has the advantage in inference because it performs maximum likelihood estimation under the specified model. We see that this corresponds to the smallest confidence intervals in our examples, as expected. `bayesglm` is suitable for prediction thanks to its low computational cost yet high accuracy. `logistf` or `brglm2` may be least preferable

method because they are computationally unstable and expensive.

Comment: The bulk of the Discussion section is very good overall

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.

S. B. Cantor, C. C. Sun, G. Tortolero-Luna, R. Richards-Kortum, and M. Follen. A comparison of c/b ratios from studies using receiver operating characteristic curve analysis. *Journal of Clinical Epidemiology*, 52(9):885–892, 1999. ISSN 0895-4356. doi: https://doi.org/10.1016/S0895-4356(99)00075-X. URL https://www.sciencedirect.com/science/article/pii/S089543569900075X.

R. G. Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1):35–46, 1991. ISSN 0034-4257. doi: https://doi.org/10.1016/0034-4257(91)90048-B. URL https://www.sciencedirect.com/science/article/pii/003442579190048B.

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.

A. H. Fielding and J. F. Bell. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1):38–49, 1997. doi: 10.1017/S0376892997000088.

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.

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.

S. Manel, H. C. Williams, and S. Ormerod. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38 (5):921–931, 2001. doi: https://doi.org/10.1046/j.1365-2664.2001.00647.x. URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2664.2001.00647.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`.

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`.

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.

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.