Statistics and Probability Letters 202 (2023) 109901

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Conjugate priors and bias reduction for logistic regression models

Tommaso Rigon [a,*], Emanuele Aliverti [b]

[a] Department of Economics, Management and Statistics, University of Milano–Bicocca, 20126 Milano, Italy
[b] Department of Statistical Sciences, University of Padova, 35121, Padova, Italy

ARTICLE INFO

Article history:
Received 16 May 2022
Received in revised form 1 June 2023
Accepted 1 July 2023
Available online 11 July 2023

Dataset link: https://github.com/tommasorigon/logistic-bias-reduction

Keywords:
Bias reduction
Boundary estimate
Conjugate prior
Exponential family
Pseudo-counts

ABSTRACT

We address the issue of divergent maximum likelihood estimates for logistic regression models by considering a conjugate prior penalty which always produces finite estimates. We show that the proposed method is closely related to the reduced-bias approach of Firth (1993), and that the induced penalized likelihood can be expressed as a genuine binomial likelihood, replacing the original data with pseudo-counts.

1. Introduction

Logistic regression is arguably one of the most widely used generalized linear models in statistical practice. In such a model, it is assumed that each entry of the vector $y = (y_1, \dots, y_n)^\top$ is a realization of independent binomial random variables with number of trials $m_1, \dots, m_n$ and success probabilities $\pi_1, \dots, \pi_n$. Moreover, suppose that each response y_i is associated with a p-dimensional covariate vector $x_i = (1, x_{i2}, \dots, x_{ip})^\top$, for $i = 1, \dots, n$. Then, a logistic regression model for $i = 1, \dots, n$ has

$$(y_i \mid m_i, \pi_i) \sim \mathrm{Bin}(m_i, \pi_i), \qquad \pi_i = \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)}, \tag{1}$$

where $\beta = (\beta_1, \dots, \beta_p)^\top$ is a p-dimensional vector of unknown regression coefficients; clearly, binary regression is a special case of the binomial likelihood of Eq. (1), when $m_i = 1$ for $i = 1, \dots, n$. We customarily assume that the $n \times p$ design matrix X, whose rows are $x_1, \dots, x_n$, is of full rank, with $p \leq n$. Thus, the log-likelihood function is

$$\ell(\beta; y) = \sum_{i=1}^n y_i(x_i^\top \beta) - \sum_{i=1}^n m_i \log\{1 + \exp(x_i^\top \beta)\}, \tag{2}$$

up to an additive constant not depending on β; the maximum likelihood (ML) estimate $\hat{\beta}$ of β is the maximizer of (2). Unfortunately, the ML estimate may not exist, meaning that at least one component of the vector $\hat{\beta}$ is infinite. It is well

[*] Corresponding author.
E-mail addresses: tommaso.rigon@unimib.it (T. Rigon), emanuele.aliverti@unipd.it (E. Aliverti).

https://doi.org/10.1016/j.spl.2023.109901
0167-7152/© 2023 Elsevier B.V. All rights reserved.

known that such an unpleasant phenomenon occurs in presence of data separation, as carefully described in Albert and Anderson (1984). If separation occurs, standard optimization procedures may fail to converge, resulting in degenerate predicted success probabilities and misleading inferential conclusions.

In a seminal paper, Firth (1993) showed that the maximizer of a suitable penalized likelihood has a smaller asymptotic bias compared to the ML estimator, while Kosmidis and Firth (2021) proved that Firth's bias-reduced estimates for logistic regression always exist and solve the separability issue. Penalized methods have become increasingly popular in health and medical sciences, and numerous developments of the original approach of Firth (1993) have been proposed in such specialized literature (e.g., Greenland and Mansournia, 2015; Puhr et al., 2017).

In high-dimensional settings, that is when p and n are both large, existing implementations of Firth (1993) may be computationally too demanding; refer for instance to Sur and Candès (2019). Thus, in this paper we propose a regularization approach that share several similarities with Firth's method but is much easier to implement, since it relies on a simple perturbation of the data. Specifically, we penalize the logistic regression likelihood by the conjugate prior of Diaconis and Ylvisaker (1979), resulting in the following penalized log-likelihood

$$\tilde{\ell}(\beta; y) = \ell(\beta; y) + \frac{p}{2m}\sum_{i=1}^{n} m_i(x_i^{\mathrm{T}}\beta) - \frac{p}{m}\sum_{i=1}^{n} m_i \log\{1+\exp(x_i^{\mathrm{T}}\beta)\}, \tag{3}$$

where $m = \sum_{i=1}^{n} m_i$. The penalized log-likelihood (3) can be rewritten, up to a scaling factor depending on p and m, in terms of a genuine log-likelihood, where the original data y are replaced with a vector of pseudo-counts $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)^{\mathrm{T}}$, defined as

$$\tilde{y}_i = \frac{p}{p+m}\frac{m_i}{2} + \frac{m}{p+m}y_i, \qquad i = 1, \dots, n. \tag{4}$$

Each pseudo-count $\tilde{y}_i \in (0, m_i)$ is a convex combination of the data y_i and $m_i/2$. This is equivalent to adding $(pm_i)/(2m)$ successes and $(pm_i)/m$ trials to each y_i and m_i, respectively, therefore shrinking the success proportions towards equiprobability. Importantly, it holds that $\tilde{\ell}(\beta; y) = (p/m+1)\ell(\beta; \tilde{y})$ and therefore well-established algorithms for logistic regression may be used to maximize (3) even in presence of large datasets. Adding small corrections to the response of a binomial model is a common regularization strategy. In the simplest case $p = n = 1$, correction (4) reduces to the familiar bias-reducing form of the empirical logit (Haldane, 1955; Anscombe, 1956). In more general settings, these penalties have been referred to as *data augmentation* priors (e.g., Greenland and Mansournia, 2015). The scheme of Clogg et al. (1991) is closely related to (4), albeit with several critical distinctions. Their correction amounts to adding $(p\sum_{i=1}^{n} y_i)/(nm)$ successes and p/n trials to each y_i and m_i, respectively. Thus, Clogg et al. (1991) approach shrinks the success proportions towards the mean $\sum_{i=1}^{n} y_i/m$ rather than $1/2$, which is a key aspect if one aims at reducing the bias (Cordeiro and McCullagh, 1991; Firth, 1993). Furthermore, the correction of Clogg et al. (1991) depends on the specific aggregation of the data and leads to a different amount of shrinkage compared to (4).

2. Conjugate Bayes for logistic regression

2.1. Diaconis and Ylvisaker conjugate priors

The penalized log-likelihood (3) has been obtained by leveraging conjugate priors for logistic regression. In this section, we introduce several key quantities and we then discuss the Bayesian interpretation of Eq. (3). We recall that Bayesian inference is based on the posterior law, obtained by updating the prior distribution through the likelihood $\exp\{\ell(\beta; y)\}$ via Bayes theorem. Let $\tau > 0$ and let $\beta_0 \in \mathbb{R}^p$ be a vector of hyperparameters. Moreover, let us define a vector of real numbers $\kappa = (\kappa_1, \dots, \kappa_n)^{\mathrm{T}}$ such that $\kappa_i = m_i \exp(x_i^{\mathrm{T}}\beta_0)/\{1+\exp(x_i^{\mathrm{T}}\beta_0)\} \in (0, m_i)$ for $i = 1, \dots, n$. Thus, the conjugate prior of Diaconis and Ylvisaker (1979) for a logistic regression model is

$$p(\beta) = C\exp\left[\tau\sum_{i=1}^{n}\kappa_i(x_i^{\mathrm{T}}\beta) - \tau\sum_{i=1}^{n} m_i\log\{1+\exp(x_i^{\mathrm{T}}\beta)\}\right], \tag{5}$$

where the normalizing constant $0 < C < \infty$, albeit not available in closed form, is necessarily finite (Theorem 1, Diaconis and Ylvisaker, 1979); additional considerations about (5) can be found in Chen and Ibrahim (2003) and Greenland (2003). The location parameter β_0 is the mode of the prior distribution and hence each ratio $\kappa_i/m_i \in (0, 1)$ can be interpreted as the prior guess for the success probability π_i. Instead, the parameter τ controls the variability and quantifies the strength of our prior beliefs about β_0. More precisely, when $\tau \to 0$ then (5) reduces to a uniform improper prior for β, whereas as $\tau \to \infty$ the prior converges to a point mass at β_0. The choice $\tau = 1$ places equal weight to the prior and the likelihood, therefore one typically consider $\tau \in (0, 1)$. In the special case of a binomial model with $p = n = 1$, the law (5) induces the usual conjugate prior on the probability $\pi = \exp(\beta)/\{1+\exp(\beta)\}$, namely $\pi \sim \mathrm{Beta}\{\tau\sum_{i=1}^{n}\kappa_i, \tau(m - \sum_{i=1}^{n}\kappa_i)\}$, with $m = \sum_{i=1}^{n} m_i$.

Application of Bayes theorem to the binomial log-likelihood (2) under the prior (5) leads to the following posterior distribution

$$p(\beta \mid y) = C(y)\exp\left[(\tau+1)\sum_{i=1}^{n} y_i^*(x_i^{\mathrm{T}}\beta) - (\tau+1)\sum_{i=1}^{n} m_i \log\{1+\exp(x_i^{\mathrm{T}}\beta)\}\right], \tag{6}$$

where $C(y) > 0$ is the normalizing constant and $y^* = (y_1^*, \ldots, y_n^*)^{\mathrm{T}}$ is a vector of pseudo-counts such that $y_i^* = \kappa_i\tau/(\tau+1) + y_i/(\tau+1) \in (0, m_i)$, for $i = 1, \ldots, n$. The posterior is still in the class of Diaconis and Ylvisaker distributions, with updated location y^* and precision $\tau + 1$. As we shall discuss in Section 3, in light of a connection with Firth (1993), a natural default is $\beta_0 = (0, \ldots, 0)^{\mathrm{T}}$ and $\tau = p/m$, implying that $\kappa_i = m_i/2$, for $i = 1, \ldots, n$. This choice leads to the pseudo-counts in Eq. (4).

2.2. Posterior inference

The posterior distribution of Eq. (6) has several appealing properties, which are briefly discussed here. In the first place, note that the posterior law can be expressed as $p(\beta \mid y) = C(y)\exp\{(\tau+1)\ell(\beta; y^*)\}$, that is, the posterior distribution is a function of the log-likelihood evaluated on the set of pseudo-counts y^*. Hence, the maximum a posteriori coincides with the maximizer of the log-likelihood $\ell(\beta; y^*)$. Heuristically, the pseudo-counts regularize the estimation problem, as each y_i^* belongs to the open set $(0, m_i)$. Consequently, the mode of the posterior distribution always exists, effectively solving the separability issue. This is formalized in the following Theorem 1, which refers to a general set of parameters β_0 and τ, but it applies also to the penalized log-likelihood (3), occurring when $\beta_0 = (0, \ldots, 0)^{\mathrm{T}}$ and $\tau = p/m$; refer to the supplementary materials for a proof.

Theorem 1. *Let X be of full rank. Then the mode of the posterior distribution (6), corresponding to the maximizer of the penalized likelihood*

$$\ell^*(\beta; y) = (\tau+1)\ell(\beta; y^*) = \ell(\beta; y) + \tau\sum_{i=1}^{n}\kappa_i(x_i^{\mathrm{T}}\beta) - \tau\sum_{i=1}^{n} m_i\log\{1+\exp(x_i^{\mathrm{T}}\beta)\}$$

with respect to β, exists and is unique.

The posterior mode does not have, in general, a strong theoretical foundation. However, the maximizer of (6) has a much more solid justification, due to the properties of exponential families with conjugate priors. In fact, in this special case the posterior mode is a licit Bayesian estimator, being the minimizer of the posterior expectation of the entropy loss; refer to Robert (1996) for a detailed discussion. By construction, Bayesian estimators obtained under such a loss are invariant under reparametrizations. Thus, if $\hat\beta_{\mathrm{DY}}$ denotes the maximizer of (6), then $\exp(\hat\beta_{\mathrm{DY}})$ is the Bayesian estimator of the odds-ratios under the entropy loss.

The optimal value $\hat\beta_{\mathrm{DY}}$ maximizing $\ell^*(\beta; y) = (\tau+1)\ell(\beta; y^*)$ can be found through standard Fisher scoring or expectation–maximization (Durante and Rigon, 2019) algorithms for logistic regression, considering the pseudo-counts y^* in place of the original binomial data y. Alternative algorithms could be exploited, including quasi-Newton or conjugate gradient ascent methods (e.g., Nocedal and Wright, 2006). In all settings, the existence and uniqueness of $\hat\beta_{\mathrm{DY}}$ are guaranteed by Theorem 1. Furthermore, $\ell^*(\beta; y)$ is a concave function (Diaconis and Ylvisaker, 1979), which means that its maximization is a fairly regular problem. For instance, concavity implies that there exists a single stationary point, i.e. the global maximum.

3. Penalized maximum likelihood

3.1. Shrinkage, bias reduction and data aggregation

The conjugate prior for logistic regression in (5) requires the elicitation of κ and τ. We now show that a careful choice of these hyperparameters leads to a connection with Firth (1993) method for bias reduction.

We recall that the ML estimates of β in a logistic regression model are biased away from the point $\beta = 0$; refer for example to McCullagh and Nelder (1989) and Cordeiro and McCullagh (1991). Thus, bias correction requires a certain amount of shrinkage towards that point. In Firth (1993) this is achieved through a Jeffrey's prior penalty, whose mode is indeed equal to 0 (Kosmidis and Firth, 2021). In a similar fashion, we set the mode $\beta_0 = (0, \ldots, 0)^{\mathrm{T}}$ in the prior specification (5), which implies that $\kappa_i = m_i/2$, for $i = 1, \ldots, n$. The amount of shrinkage is regulated by the precision parameter τ, whose choice is more delicate. Heuristically, one may set τ proportional to m^{-1}, so that for m large enough the contribution of the prior becomes negligible compared to the weight of the likelihood in (6). In particular, if $\tau = p/m$ then the amount of shrinkage is proportional to the model complexity, which seems desirable. These choices lead to the penalized log-likelihood (3) and the pseudo-counts (4). Such an intuitive choice resonates with the principles behind reduced-bias estimators of Firth (1993), and ensures that the resulting prior distribution is invariant under different aggregation of the data, in contrast with (Clogg et al., 1991). A more precise argument to this claim is offered in the next section.

3.2. Penalized score equations and connection with Firth (1993)

The relationship between Firth (1993) method and our proposal can be formally investigated by comparing the corresponding score functions. Firth's estimate is obtained as the solution of the system of penalized score equations $U_{r,\text{FI}}(\beta) = 0$ for $r = 1, \dots, p$, where

$$U_{r,\text{FI}}(\beta) = \sum_{i=1}^{n}(y_i - m_i\pi_i)x_{ir} - p\sum_{i=1}^{n}\left(\frac{h_i}{p}\right)(\pi_i - 1/2)x_{ir},$$

and where $h_1, \dots, h_n$ represent the diagonal elements of the $n \times n$ projection matrix

$$H(\beta) = W(\beta)^{1/2}X\{X^{\mathrm{T}}W(\beta)X\}^{-1}X^{\mathrm{T}}W(\beta)^{1/2},$$

with $W(\beta) = \text{diag}\{m_1\pi_1(1-\pi_1), \dots, m_n\pi_n(1-\pi_n)\}$. On the other hand, the maximizer of the penalized log-likelihood (3) corresponds to the solution of the system of penalized score equations $U_{r,\text{DY}}(\beta) = 0$ for $r = 1, \dots, p$, where

$$U_{r,\text{DY}}(\beta) = \frac{\partial}{\partial\beta_r}\tilde{\ell}(\beta; y) = \sum_{i=1}^{n}(y_i - m_i\pi_i)x_{ir} - p\sum_{i=1}^{n}\left(\frac{m_i}{m}\right)(\pi_i - 1/2)x_{ir}.$$

Thus, the penalized scores $U_{r,\text{FI}}(\beta)$ and $U_{r,\text{DY}}(\beta)$ differ only in their penalty term. Broadly speaking, the two approaches lead to similar estimates whenever the following approximation holds

$$\sum_{i=1}^{n}\left(\frac{h_i}{p}\right)(\pi_i - 1/2)x_{ir} \approx \sum_{i=1}^{n}\left(\frac{m_i}{m}\right)(\pi_i - 1/2)x_{ir}. \tag{7}$$

Note that the matrix $H(\beta)$ has rank p and is idempotent and symmetric, implying that the sum of its diagonal terms is $\sum_{i=1}^{n} h_i = p$. Hence, our method considers the mean of the values $(\pi_i - 1/2)x_{ir}$, with weights m_i/m, instead of the mean of the same quantities but with weights h_i/p, as in Firth (1993). The following Theorem sheds some light on the driving factors of this approximation; refer to the supplementary materials for a proof.

Theorem 2. *Let X be of full rank and let $x_{ir} \in [-1, 1]$ for $i = 1, \dots, n$ and $r = 1, \dots, p$. Then it holds:*

$$|U_{r,\text{FI}}(\beta) - U_{r,\text{DY}}(\beta)| \le p\, d_{\text{TV}}\left(\boldsymbol{w}_{\text{FI}}, \boldsymbol{w}_{\text{DY}}\right) = \frac{p}{2}\sum_{i=1}^{n}\left|\frac{h_i}{p} - \frac{m_i}{m}\right|,$$

for $r = 1, \dots, p$, where d_{TV} denotes the total-variation distance and $\boldsymbol{w}_{\text{FI}} = (h_1/p, \dots, h_n/p)^T$ and $\boldsymbol{w}_{\text{DY}} = (m_1/m, \dots, m_n/m)^T$. Moreover, the following inequality holds:

$$d_{\text{TV}}(\boldsymbol{w}_{\text{FI}}, \boldsymbol{w}_{\text{DY}}) \le \left(1 - p\frac{m_{(1)}}{m}\right),$$

where $m_{(1)} = \min\{m_1, \dots, m_n\}$.

Theorem 2 clarifies that the quality of the approximation is directed by the closeness between the two set of weights $\boldsymbol{w}_{\text{FI}}$ and $\boldsymbol{w}_{\text{DY}}$. In other terms, in the binary regression case, i.e. $m_i = 1$, the discrepancy is low whenever the leverages $h_i \approx p/m$ are approximately constant. The latter assumption is known as *approximate quadratic balance* and has been exploited by Cordeiro and McCullagh (1991) to obtain the reduced bias estimate $(1 - p/m)\hat{\beta}$, which indeed deflates the maximum likelihood estimate towards zero by the factor $1 - p/m$. This assumption appears in seemingly unrelated settings, for example for the definition of the generalized cross-validation index (Craven and Wahba, 1978). The quadratic balance condition holds exactly true in some special cases, as clarified by Theorem 2. For instance, when $p = n = 1$ then both Firth (1993) and our approach correspond to a Beta(1/2, 1/2) prior penalty for π and coincide with the empirical logit correction. More generally, in any saturated model with $p = n$ and with balanced number of trials $m_1 = \cdots = m_n$ the two approaches formally agree, i.e. $|U_{r,\text{FI}}(\beta) - U_{r,\text{DY}}(\beta)| = 0$. Finally, Theorem 2 shows that the absolute difference between $U_{r,\text{FI}}(\beta)$ and $U_{r,\text{DY}}(\beta)$ is upper bounded by p.

From a computational perspective, Firth's modified score equations can be solved numerically via quasi-Fisher scoring (e.g. Kosmidis and Firth, 2010), in which at each iteration one must compute an adjusted response based on the current parameter value, and update the coefficients accordingly (Kosmidis and Firth, 2021, Section 4). This approach can be computationally challenging especially in settings with large n and p, since each iteration requires to re-obtain the weights $h_1, \dots, h_n$. In contrast, the proposed penalized score requires to compute the pseudo-responses only once and it can be directly solved with any algorithm for standard logistic likelihood optimization, without further adjustments. Hence, it provides a computationally convenient option to approximate Firth's scores in large dimensions.

Table 1
Estimated regression coefficients on the endometrial cancer study (with standard errors in parentheses). ML: maximum likelihood.

	β_1	β_2	β_3	β_4
Maximum likelihood $\hat{\beta}$	4.305 (1.637)	$+\infty$ ($+\infty$)	−0.042 (0.044)	−2.903 (0.846)
Penalized ML $\hat{\beta}_{\text{DY}}$	3.579 (1.459)	3.431 (1.893)	−0.034 (0.040)	−2.458 (0.748)
Clogg et al. (1991)	3.622 (1.471)	3.223 (1.722)	−0.034 (0.040)	−2.511 (0.761)
Firth (1993)	3.775 (1.489)	2.929 (1.551)	−0.035 (0.040)	−2.604 (0.776)
Kenne Pagui et al. (2017)	3.969 (1.552)	3.869 (2.298)	−0.039 (0.042)	−2.708 (0.803)

Table 2
Simulation study on the low birthweight study. RMSE: root mean squared error; ML: maximum likelihood.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7
Bias	Maximum likelihood $\hat{\beta}$	−1.42	−0.01	0.09	−0.04	−0.18	−0.12	0.34
	Penalized ML $\hat{\beta}_{\text{DY}}$	−0.08	0.00	−0.01	0.03	−0.01	0.03	0.00
	Clogg et al. (1991)	−0.22	0.00	0.00	0.02	−0.01	0.03	0.05
	Firth (1993)	−0.08	0.00	0.01	0.00	0.00	0.00	0.02
	Kenne Pagui et al. (2017)	−0.38	0.00	0.03	−0.01	−0.06	−0.03	0.10
RMSE	Maximum likelihood $\hat{\beta}$	6.88	0.06	0.64	0.65	0.81	1.12	1.49
	Penalized ML $\hat{\beta}_{\text{DY}}$	5.71	0.05	0.54	0.56	0.71	0.96	1.22
	Clogg et al. (1991)	5.83	0.05	0.55	0.57	0.70	0.97	1.25
	Firth (1993)	5.94	0.05	0.57	0.58	0.71	0.95	1.28
	Kenne Pagui et al. (2017)	6.12	0.06	0.58	0.60	0.78	1.01	1.31

4. Illustrations

4.1. Endometrial cancer study

To empirically assess the performance of our proposal, we consider the endometrial cancer grade dataset (Heinze and Schemper, 2002), a study on $n = 79$ patients which aims at evaluating the relationship between the histology of the endometrium (low against high), and three risk factors. A logistic regression model was fitted with parameter vector $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^{\intercal}$, with the first coefficient corresponding to an intercept term and the remaining to neovasculation, pulsatility index of arteria uterina and endometrium height, respectively. As shown in Heinze and Schemper (2002), the ML estimate does not exist since the estimated value for the coefficient β_2 associated with neovasculation is divergent. The omission of the neovasculation information from the set of covariates is inappropriate, as the other factors would not be properly adjusted for this highly informative risk factor. We therefore compare the proposed penalized estimate $\hat{\beta}_{\text{DY}}$ with the approaches of Clogg et al. (1991) and Firth (1993). We also consider the median unbiased estimators of Kenne Pagui et al. (2017).

Estimates for the regression coefficients $(\beta_1, \beta_2, \beta_3, \beta_4)^{\intercal}$ and the corresponding Wald standard errors are reported in Table 1. As expected, all the reducing-bias methods produce finite estimates for β_2 and indicate a strong effect of neovasculation on the response probability. Point estimates and standard errors for the remaining coefficients are also quite similar. The approach of Clogg et al. (1991) and our penalized method lead to similar estimates since the former shrinks the estimated probabilities towards the sample proportion, which in this case is $\sum_{i=1}^{n} y_i/m = 0.38$. Despite the stated aim of Clogg et al. (1991) was not performing bias reduction, it performs reasonably well in this example, and the reasons for its good empirical performance are likely due to its similarity with our approach, which in turn approximates the one of Firth (1993).

4.2. Infant birthweight study

In this second empirical study, we replicate an example presented in Kosmidis et al. (2020), which considers a study of low birthweight. Data comprises $n = 100$ births and the binary outcome of interest is a dichotomization of infant birthweight (below or above 2.5 kilograms). The probability of low birthweight is modeled as a function of an intercept and six covariates about the mother. The maximum likelihood estimate $\hat{\beta}$ of the regression coefficients $\beta = (\beta_1, \ldots, \beta_7)^{\intercal}$ exists and is finite. We simulate 10 000 datasets from a logistic regression model with parameter $\hat{\beta}$ and evaluate the inferential properties of the proposed estimator, comparing them with popular bias-correction methods in a regular scenario.

Results are presented in Table 2. The values for the maximum likelihood are computed excluding samples with data separation, occurring in 100 replications out of 10 000. The ML estimator performs significantly worse than all the reduced-bias methodologies in terms of bias and root mean squared error. Moreover, we observe a striking empirical similarity in terms of bias between the proposed penalized estimator $\hat{\beta}_{\text{DY}}$ and the one of Firth (1993); this finding is in line with the considerations discussed in Section 3.2. Empirical results also confirm a limitation of Clogg et al. (1991), which fails

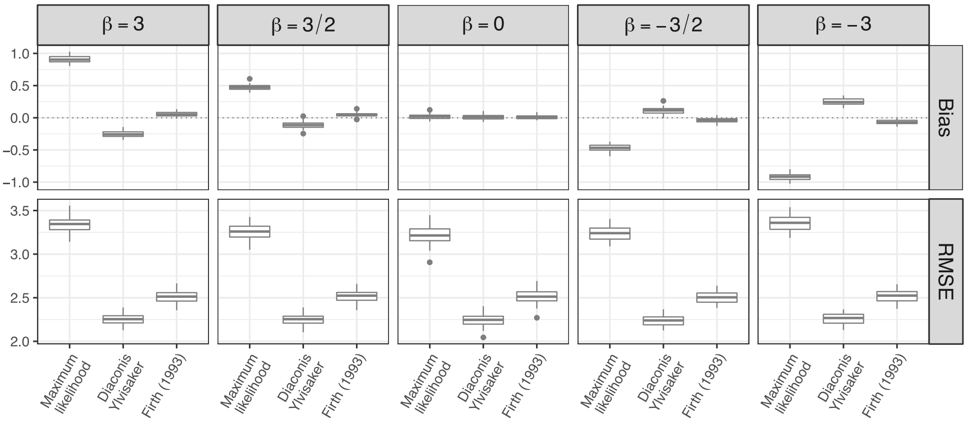

Fig. 1. Bias and root mean squared error (RMSE) of maximum likelihood and reduced-bias estimates, for the parameters of a logistic regression model. Boxplots indicate variability across groups of regression coefficients.

Table 3
Computational times (milliseconds) of different estimation procedures using different implementations, on a single high-dimensional synthetic dataset with $n = 1000$ and $p = 200$. The reported computational times corresponds to the average times over 100 executions.

R implementation	`RcppNumerical`		`glm`		`brglm2`
Estimate	ML	DY	ML	DY	Firth (1993)
Computational time (ms)	1.746	1.658	150.875	126.308	1168.211

at reducing the bias for the intercept parameter β_1, since it shrinks the response towards the empirical proportion of successes rather than 1/2. Finally, the approach of Kenne Pagui et al. (2017) is designed for correcting median bias and therefore is expected to perform worse, in terms of bias, than (Firth, 1993). The proposed penalized estimator $\hat{\beta}_{\text{DY}}$ slightly improves the root mean squared error compared to Firth (1993). This aspect seems to be empirically confirmed also in the next illustrative example.

4.3. High dimensional synthetic dataset

We finally consider a synthetic dataset that mimicks the high dimensional scenario described in Sur and Candès (2019). In particular, we simulate 5000 datasets from a binary logistic regression model with $n = 1000$ observations and $p = 200$. Covariates are sampled from a normal distribution with mean 0 and variance $1/n$. Moreover, regression coefficients are divided in 5 blocks of size 40, whose values are $\{-3, -3/2, 0, 3/2, 3\}$. The purpose of this study is to investigate the performance of the proposed estimator in computationally challenging scenarios.

In first place, we shall note that the execution times of the estimation procedure for $\hat{\beta}_{\text{DY}}$, obtained via scalable algorithms for logistic regression, such as the limited-memory BFGS implemented in the package `RcppNumerical` (Qiu et al., 2019), are orders of magnitude faster than the `brglm2` implementation of Firth (1993). These considerations are valid even when compared with the standard `glm` R function. For instance, a single replication with $n = 1000$ and $p = 200$ required an average elapsed time of 1.7 ms in the former case, against approximately a second in the latter, on a 2019 Macbook Pro. A summary of the computational times is provided in Table 3. Moreover, the differences are even more marked for larger values of n and p, to the extent that Firth (1993) estimates could not be obtained within hours of running time with $n = 10\,000$, $p = 2000$ and correlated design matrix X.

Secondly, we computed the bias and the root mean squared error of ML and reduced-bias estimators, which are depicted in Fig. 1. Current empirical findings confirm the poor behavior of ML estimates, consistently with the existing literature (Kosmidis and Firth, 2021). The proposed correction provides an important improvement in terms of bias reduction compared to the ML estimator, although the bias is slightly larger than that of Firth (1993). This is expected and consistent with the findings of Sections 3.2 and 4.2. Furthermore, our penalized procedure achieves a lower mean squared error compared to both the maximum likelihood and the approach of Firth (1993). This does not come as a contradiction, since Firth (1993) approach does not explicitly reduce the mean squared error. This empirical finding is likely due to the different tail behavior of Diaconis and Ylvisaker (1979) priors compared to Jeffrey's priors, the latter being more dispersed and displaying heavier tails.

Data availability

Data and code are available at the repository: https://github.com/tommasorigon/logistic-bias-reduction.

Acknowledgments

The authors wish to thank David Dunson, Sonia Migliorati and Aldo Solari for their insightful comments on an early draft of this manuscript. We thank the Editor, the Associate Editor and the anonymous reviewer for the insightful suggestions which improved this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2023.109901.

References

Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71 (1), 1–10.

Anscombe, F.G., 1956. On estimating binomial response relations. Biometrika 43, 461–464.

Chen, M.H., Ibrahim, J.G., 2003. Conjugate priors for generalized linear models. Stat. Sinica 13 (2), 461–476.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., Weidman, L., 1991. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. J. Amer. Statist. Assoc. 86 (413), 68–78.

Cordeiro, G.M., McCullagh, P., 1991. Bias correction in generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 53 (3), 629–643.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. 31 (4), 377–403.

Diaconis, P., Ylvisaker, D., 1979. Conjugate prior for exponential families. Ann. Statist. 7 (2), 269–292.

Durante, D., Rigon, T., 2019. Conditionally conjugate mean-field variational Bayes for logistic models. Statist. Sci. 34 (3), 472–485.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.

Greenland, S., 2003. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. Biometrics 59 (1), 92–99.

Greenland, S., Mansournia, M.A., 2015. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Stat. Med. 34 (23), 3133–3143.

Haldane, J.S.B., 1955. The estimation and significane of the logarithm of a ratio of frequencies. Ann. Hum. Genet. 20, 309–311.

Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21 (16), 2409–2419.

Kenne Pagui, E.C., Salvan, A., Sartori, N., 2017. Median bias reduction of maximum likelihood estimates. Biometrika 104 (4), 923–938.

Kosmidis, I., Firth, D., 2010. A generic algorithm for reducing bias in parametric estimation. Electron. J. Stat. 4, 1097–1112.

Kosmidis, I., Firth, D., 2021. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika 108 (1), 71–82.

Kosmidis, I., Kenne Pagui, E.C., Sartori, N., 2020. Mean and median bias reduction in generalized linear models. Stat. Comput. 30 (1), 43–59.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Springer.

Nocedal, J., Wright, S., 2006. Conjugate Gradient Methods. Springer New York, New York, NY.

Puhr, R., Heinze, G., Nold, M., Lusa, L., Geroldinger, A., 2017. Firth's logistic regression with rare events: accurate effect estimates and predictions? Stat. Med. 36 (14), 2302–2317.

Qiu, Y., Balan, S., Beall, M., Sauder, M., Okazaki, N., Hahn, T., 2019. RcppNumerical: "Rcpp" integration for numerical computing libraries. URL https://CRAN.R-project.org/package=RcppNumerical, R package version 0.4-0.

Robert, C.P., 1996. Intrinsic losses. Theory and Decision 40 (2), 191–214.

Sur, P., Candès, E.J., 2019. A modern maximum-likelihood theory for high-dimensional logistic regression. Proc. Natl. Acad. Sci. 116, 14516–14525.