Statistics Surveys
Vol. 18 (2024) 139–162
ISSN: 1935-7516
https://doi.org/10.1214/24-SS148

A comprehensive review of bias reduction methods for logistic regression

Marieke Stolte[1], Swetlana Herbrandt[2] and Uwe Ligges[1]

[1] *Department of Statistics, TU Dortmund University, e-mail:* stolte@statistik.tu-dortmund.de; ligges@statistik.tu-dortmund.de

[2] *Statistical Consulting and Analysis, Center for Higher Education, TU Dortmund University, e-mail:* swetlana.herbrandt@tu-dortmund.de

Abstract: The maximum likelihood estimator (MLE) for the unknown parameter vector in logistic regression is well known to be biased. There are many different approaches to reduce this bias including bias correction, adjustment of the score function or of the data itself, jackknifing, penalizing the likelihood, exact logistic regression, and the discriminant function approach. These approaches, as well as many different simulation studies comparing them, are reviewed here. Since the studies use very different parameter settings and sometimes contradict each other, no general recommendations can be given. However, most studies find that the bias of the MLE is substantial for small to medium samples, that the bias-corrected estimators tend to overcorrect in very small samples, and that Firth's estimator, when considered, is the best choice.

MSC2020 subject classifications: 62J12, 62F12.
Keywords and phrases: Maximum likelihood, logistic regression, Firth's estimator.

Received December 2023.

1. Introduction

Although the maximum likelihood estimator (MLE) in logistic regression is asymptotically unbiased, it has been known for decades that its bias can be considerably large for small and moderate samples (Cox and Snell, 1968; Anderson and Richardson, 1979; Schaefer, 1983; Cordeiro and McCullagh, 1991). Accordingly, aiming at bias reduction, there are many proposals for alternative estimators and simulation studies comparing subsets of these estimators. Some studies shortly review results for selected approaches to correct for bias from certain points of view (Zorn, 2005; Heinze, 2006; Gao and Shen, 2007). We try to give a comprehensive review of the various approaches, estimators, and results of simulation studies, and a comparison of these.

2. Logistic regression model and maximum likelihood estimation

Consider the logistic regression model

$$Y_i|x_{i1},\dots,x_{ip} \sim Ber(\pi_i),$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

with the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n\times(p+1)}$ consisting of n observations of p covariates with $\boldsymbol{x}_i = (1\ x_{i1}, \ldots, x_{ip})^T$, Y_i the corresponding observations of a binary target variable $Y, i = 1, \ldots, n$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ the unknown parameter vector (Fahrmeir et al., 2013, pp. 270–273) which is usually estimated using maximum likelihood (ML). The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_{ML}$ is calculated by numerically solving the score equations

$$s(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \pi_i) = \mathbf{0}$$

where

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(\pi_i) - y_i \log(1-\pi_i) + \log(1-\pi_i)]$$

is the log-likelihood with likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

(Fahrmeir et al. 2013, pp. 279–283; Tutz 2011, pp. 63–66). The MLE exists if and only if there is no complete separation

$$\exists \boldsymbol{\beta} : \boldsymbol{x}_i^T \boldsymbol{\beta} > 0, \text{ if } y_i = 1 \text{ and } \boldsymbol{x}_i^T \boldsymbol{\beta} < 0, \text{ if } y_i = 0$$

or quasicomplete separation

$$\exists \boldsymbol{\beta} \neq \mathbf{0} : \boldsymbol{x}_i^T \boldsymbol{\beta} \geq 0, \text{ if } y_i = 1 \text{ and } \boldsymbol{x}_i^T \boldsymbol{\beta} \leq 0, \text{ if } y_i = 0, i = 1, \ldots, n,$$

(Albert and Anderson, 1984; Santner and Duffy, 1986).

3. Bias of the MLE

Under relatively weak regularity conditions (Fahrmeir and Kaufmann, 1986) the MLE $\hat{\boldsymbol{\beta}}_{ML}$ exists asymptotically and is a consistent estimator for $\boldsymbol{\beta}$ for $n \to \infty$. Moreover, under the same conditions, it holds that the MLE is asymptotically normally distributed and unbiased

$$\hat{\boldsymbol{\beta}}_{ML} \overset{a}{\sim} N\left(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\boldsymbol{\beta})\right)$$

where

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \pi_i(1-\pi_i) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

with the information matrix $\boldsymbol{W} = diag\{\pi_1(1-\pi_1), \ldots, \pi_n(1-\pi_n)\}$. However in finite samples the MLE is biased due to the combination of the unbiasedness of the score function $\mathbb{E}_{\boldsymbol{\beta}}(s(\boldsymbol{\beta})) = 0$ and its curvature $s''(\boldsymbol{\beta}) \neq 0$ (Firth, 1993). The bias

$$b(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}_{ML}\right) - \boldsymbol{\beta}$$

of the MLE is of order $\mathcal{O}(n^{-1})$ (McCullagh and Nelder, 1989, p. 119) and can't be expressed in closed form. There are many expressions for the first order term $b^{(1)}(\boldsymbol{\beta})/n$ in the representation

$$b(\boldsymbol{\beta}) = \frac{b^{(1)}(\boldsymbol{\beta})}{n} + \frac{b^{(2)}(\boldsymbol{\beta})}{n^2} + \frac{b^{(3)}(\boldsymbol{\beta})}{n^3} + \cdots$$

of the bias (Cox and Snell, 1968; Anderson and Richardson, 1979; McLachlan, 1980; Schaefer, 1983; Copas, 1988; McCullagh and Nelder, 1989; Cordeiro and McCullagh, 1991; O'neill, 1994; Park and Choi, 2008). Here we will use the expression given by Cordeiro and McCullagh (1991)

$$b(\boldsymbol{\beta}) = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\xi},$$

with $\boldsymbol{\xi}$ a vector with components

$$\xi_i = z_{ii}\left(\pi_i - \frac{1}{2}\right), i = 1, \ldots, n, \tag{1}$$

and z_{ii} the ith diagonal element of

$$\boldsymbol{Z} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T.$$

The expression is known to be equivalent to the expressions given by the other previously mentioned authors except for McLachlan (1980) and O'neill (1994). Bias approximations to a higher order are for example given by Bowman and Shenton (1965) and Cordeiro and Barroso (2007). These approximations will not be considered further since they are neither specifically determined for the logistic regression model nor compared in any simulation study for logistic regression.

Cordeiro and McCullagh (1991) also show that the bias vector $b(\boldsymbol{\beta})$ and the parameter vector $\boldsymbol{\beta}$ are approximately collinear which implies that the MLE is biased away from the origin. Under further assumptions, it holds that

$$b(\boldsymbol{\beta}) \approx \frac{p+1}{n}\boldsymbol{\beta},$$

so the bias is approximately proportional to $\boldsymbol{\beta}$ (Cordeiro and McCullagh, 1991).

A different approach is given by Sur and Candès (2019) who characterize the asymptotic bias of the MLE for normally distributed covariates under the assumption

$$\lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i^T\boldsymbol{\beta}) = \gamma^2$$

with fixed signal strength γ by an equation system with three unknowns $(\alpha, \sigma, \lambda)$ that is solvable if and only if the MLE exists. Then the bias of the MLE can be quantified by the solution $(\alpha_*, \sigma_*, \lambda_*)$ in a statistical sense:

$$\frac{1}{p+1}\sum_{j=0}^{p}\left(\hat{\beta}_{ML,j} - \alpha_*\beta_j\right) \overset{a.s.}{\to} 0.$$

This means that the $\hat{\beta}_{ML,j}$ are centered around $\alpha_*\beta_j$ (Sur and Candès, 2019).

4. Alternative estimators

There are different proposals for alternative estimators to reduce the bias of the MLE. Some of those have similar motivations and are therefore sorted together in the overview below. See Table 1 of online Appendix A (Stolte et al., 2024) for a complete overview of each estimator.

4.1. Bias correction for the MLE

The most obvious way to reduce the bias of the MLE is a bias correction. The idea behind these estimators is to simply subtract an expression for the bias from the MLE:

$$\hat{\boldsymbol{\beta}}_{corr} = \hat{\boldsymbol{\beta}}_{ML} - \frac{b^{(1)}\left(\hat{\boldsymbol{\beta}}_{ML}\right)}{n}.$$

This removes the first-order bias. Estimators of this type are defined for each of the previously mentioned expressions for the first-order bias. Since the expressions are equivalent the same holds for the corresponding bias-corrected estimators and therefore we will once again only present the one defined by Cordeiro and McCullagh (1991)

$$\hat{\boldsymbol{\beta}}_{CM} = \hat{\boldsymbol{\beta}}_{ML} - \frac{1}{2}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\xi}.$$

The advantage of this form is that the bias vector can be calculated by taking it as the vector of coefficients of a weighted linear regression of $\boldsymbol{\xi}$ on $\boldsymbol{X}$ with weights given by $\boldsymbol{W}$.

Another bias-corrected estimator $\hat{\boldsymbol{\beta}}^*_{\gamma,Copas}$ is given by Copas (1988). The estimator is based on the same principle as the estimators presented so far, but instead of the MLE, a robust estimator $\hat{\boldsymbol{\beta}}_\gamma$ and its bias are considered. For this purpose, in addition to the logistic regression model, it is assumed that there is a swap of the values 0 and 1 in the target variable with a probability of γ and the MLE $\hat{\boldsymbol{\beta}}_\gamma$ is determined based on this assumption.

Sur and Candès (2019) also define a corrected estimator

$$\hat{\boldsymbol{\beta}}_{SC} = \frac{1}{\hat{\alpha}}\hat{\boldsymbol{\beta}}_{ML}$$

based on their ML theory where $\hat{\alpha}$ is the solution of their equation system with an estimator for the signal strength γ inserted. For details see Sur and Candès (2019).

4.2. Adjusting the score function

Firth A different approach for removing the first-order bias is given by Firth (1993). The idea here is to rather prevent the emergence of the bias than to correct the estimator afterwards. In Kosmidis (2014) this idea of preventing instead of correcting the bias is summarized as implicit methods in contrast to explicit methods.

To understand the adjustment proposed by Firth (1993) it is helpful to consider a one-dimensional setting $\beta \in \mathbb{R}$. Then a positive curvature of the score combined with its unbiasedness results in a positive bias. The idea of Firth (1993) is to reduce this bias by adding a small bias to the score. To achieve this, at each point β the score function has to be shifted down by the value $F(\beta)b(\beta)$ where $-F(\beta) = s'(\beta)$ is the derivative of the score function. This results in the modified score function

$$s^*(\beta) = s(\beta) - F(\beta)b(\beta), \tag{2}$$

whose zero $s^*(\beta) = 0$ provides the estimator $\hat{\beta}_{Firth}$. For a parameter vector $\boldsymbol{\beta}$ the modified score function is defined analogously to (2) where $F(\boldsymbol{\beta})$ is the information matrix. Substituting the expression given by Cordeiro and McCullagh (1991) for the bias in (2) yields

$$s^*(\boldsymbol{\beta}) = s(\boldsymbol{\beta}) - \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\xi}.$$

As shown in Firth (1993), this modification removes the first-order bias. Furthermore the estimator $\hat{\boldsymbol{\beta}}_{Firth}$ is the stationary point of the penalized likelihood

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|F(\boldsymbol{\beta})|^{\frac{1}{2}},$$

so the calculation of $\hat{\boldsymbol{\beta}}_{Firth}$ is equivalent to calculating the posterior mode using Jeffreys' prior (Jeffreys, 1946). For grouped data, Firth's estimator is also equivalent to adding a constant to the number of events and non-events for each group. In the saturated model this constant equals 1/2, in the more general case it equals $h_i/2$ where h_i is the ith hat value, i.e. the ith diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{\frac{1}{2}}$ (Firth, 1993; Galindo-Garre, Vermunt and Bergsma, 2004).

A property that makes Firth's estimator particularly popular is that it also exists in the case of separation (Heinze and Schemper, 2002; Heinze, 2006; Zorn, 2005). This has already been observed by Heinze and Schemper (2002) in practice. However, this property was only formally proven in Kosmidis and Firth (2021, Corollary 1). In addition, the shrinkage property of the estimator, which was also observed in some simulation studies Heinze and Schemper (2002), is

demonstrated there: compared with the MLE, Firth's estimator is shrunk towards $\boldsymbol{\beta} = \mathbf{0}$ relative to a metric based on the expected information matrix, so $\hat{\boldsymbol{\beta}}_{Firth}$ typically takes on smaller absolute values than $\hat{\boldsymbol{\beta}}_{ML}$ (Kosmidis and Firth, 2021, Theorem 2). From the same theorem follows another frequently observed property: confidence ellipsoids based on the asymptotic normal distribution of the estimators for the bias-reduced estimator have a smaller volume than for the MLE.

Firth's estimator is probably the best-known of the presented alternatives. It is recommended in various papers (Bull, Mak and Greenwood, 2002; Zorn, 2005; Heinze and Schemper, 2002; Heinze, 2006), in particular, because of its applicability on separated data. It is mentioned in several textbooks (Tutz, 2011; Agresti, 2012; Hosmer, Lemeshow and Sturdivant, 2013; Steyerberg, 2019) with increasing popularity during the last few years (Kosmidis and Firth, 2021).

Puhr et al. (2017) propose two modifications of Firth's estimator to reduce the bias in the estimated probabilities. For this purpose, the property of the MLE is obtained that the average predicted probability is equal to the proportion of ones in $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. The first modification $\hat{\boldsymbol{\beta}}_{FLIC}$ changes the estimate of the intercept so that this property applies again. For the second modification, an additional indicator variable, which distinguishes between the original and the artificially added observations, is introduced in the calculation of Firth's estimator by adjusting the data with $h_i/2$. The MLE is then determined using this additional covariable.

Kenne Pagui et al. Another estimator based on an adjustment for the score function is proposed in Kenne Pagui, Salvan and Sartori (2017). The motivation for this estimator is to preserve the equivariance of the MLE, which the estimators presented so far do not do. As a result, the entire estimation process does not have to be carried out again when the data is reparameterized. This is achieved by correcting for the median.

For the derivation of the estimator, the case $\beta \in \mathbb{R}$ is considered first. By subtracting a representation for the median $M(s(\beta))$ the score function is modified in such a way that the resulting modified score function $\tilde{s}(\beta)$ is median-unbiased to third order. Now, if $\hat{\beta}_{KPSS}$ is the unique solution of $\tilde{s}(\beta) = 0$, the events $\tilde{s}(\beta) \leq 0$ and $\hat{\beta}_{KPSS} \leq \beta$ are equivalent, since the score function is strictly decreasing in β as follows from the relationship

$$-s'(\beta) = F(\beta) = 1/\operatorname{Var}(\hat{\beta}_{ML}) > 0.$$

So it follows directly that $\hat{\beta}_{KPSS}$ is also median-unbiased to third order, i.e. it holds

$$\mathbb{P}_\beta\left(\hat{\beta}_{KPSS} \leq \beta\right) = \frac{1}{2} + \mathcal{O}\left(n^{-3/2}\right).$$

Since there is no definition of the multidimensional median, by means of which this approach could be generalized for multidimensional parameter vectors β, Kenne Pagui, Salvan and Sartori (2017) choose a different approach instead. They set up a system of estimation equations whose solution for each component

$\beta_j,\ j = 0, \ldots, p$, gives the same estimator as the estimate in the one-dimensional case when using the profile score function up to and including the terms of order $\mathcal{O}(n^{-1})$. This means that the estimate for β_j in the multidimensional case is approximately the same as the estimate in the one-dimensional case if the MLEs are inserted for the remaining parameters $\beta_r,\ r \neq j$, in the adjusted score equation. The resulting modified score vector then consists of the components

$$
\begin{aligned}
\tilde{s}_j &= s_j - \sum_{\substack{a=0 \\ a \neq j}}^{p} \gamma_{ja} s_a + M_j, j = 0, \ldots, p, \text{ where} \\
M_j &= -\kappa_{1j} + \kappa_{3j}/(6\kappa_{2j}), \\
\kappa_{1j} &= -\frac{1}{2} \sum_{\substack{a,b,c=0 \\ a,b,c \neq j}}^{p} \nu_{ab}^{-1} (\nu_{j,a,b} - \gamma_{jc} \nu_{a,b,c}), \\
\kappa_{2j} &= \nu_{j,j} - \sum_{\substack{a=0 \\ a \neq j}}^{p} \gamma_{ja} \nu_{j,a}, \\
\kappa_{3j} &= \nu_{j,j,j} \sum_{\substack{a,b,c=0 \\ a,b,c \neq j}}^{p} [-3\gamma_{ja}\nu_{j,j,a} + 3\gamma_{ja}\gamma_{jb}\nu_{j,a,b} - \gamma_{ja}\gamma_{jb}\gamma_{jc}\nu_{a,b,c}], \\
\gamma_{ja} &= \sum_{\substack{b=0 \\ b \neq j}}^{p} \nu_{ab}^{-1} \nu_{j,b}, \\
\nu_{j,s} &= \nu_{js} = -F_{js}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij} x_{is} \pi_i (1 - \pi_i) \text{ and} \\
\nu_{j,s,t} &= \sum_{i=1}^{n} x_{ij} x_{is} x_{it} \pi_i (1 - \pi_i)(1 - 2\pi_i), \\
j, s, t &\in \{0, \ldots, p\},\ a \in \{0, \ldots, p\} \setminus \{j\}.
\end{aligned}
$$

$\hat{\boldsymbol{\beta}}_{KPSS}$ is defined as the solution of $\tilde{s}(\boldsymbol{\beta}) = 0$. The resulting estimator is equivariant under reparametrizations that transform each component β_j separately. It holds that

$$
\mathbb{P}_{\boldsymbol{\beta}} \left(\hat{\beta}_{KPSS,j} \leq \beta_j \right) = \frac{1}{2} + \mathcal{O}\left(n^{-3/2} \right).
$$

Equivalently, $\hat{\boldsymbol{\beta}}_{KPSS}$ can be calculated by solving the equation system

$$
s(\boldsymbol{\beta}) + F(\boldsymbol{\beta}) M_1(\boldsymbol{\beta}) = 0 \tag{3}
$$

with vector $M_1(\boldsymbol{\beta})$ consisting of elements $M_{1j}(\boldsymbol{\beta}) = M_j/\kappa_{2j}, j = 0, \ldots, p$. The equation system (3) is asymptotically solvable if the MLE exists. Further general statements on the solvability are not known.

Properties, tests and estimation Both estimators $\hat{\boldsymbol{\beta}}_{Firth}, \hat{\boldsymbol{\beta}}_{KPSS}$ presented in this chapter preserve the asymptotic properties of the MLE, i.e. they are asymptotically unbiased, efficient and $N(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}))$-distributed. Therefore, the same tests and confidence intervals as for the MLE that utilize these properties can be used. The two estimators can each be calculated using Iteratively Weighted Least Squares (IWLS) or, equivalently, using a quasi-Fisher scoring algorithm (Kosmidis, Kenne Pagui and Sartori, 2020).

4.3. Penalizing the likelihood

In addition to bias, highly correlated covariates also lead to unstable or infinitely large estimates in ML estimation. For highly correlated covariates, a ridge estimator (Schaefer, Roi and Wolfe, 1984) is often used instead of the MLE. However, in general, the ridge estimator is no longer asymptotically unbiased. Gao and Shen (2007) therefore propose a combination of Firth's estimator and the ridge estimator. This estimator $\hat{\boldsymbol{\beta}}_{SG}$, which they call the *double penalized MLE*, is determined as the maximum of the double penalized likelihood, which results from the further addition of the ridge penalty to the penalized likelihood $L^*(\beta)$ according to Firth (1993). The corresponding double-penalized log-likelihood is given as

$$\ell^{**}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \frac{1}{2}\log|F(\boldsymbol{\beta})| + k||\boldsymbol{\beta}||_2^2,$$

where k is the ridge parameter. If, in addition to the assumptions of the logistic regression model, it is assumed that the entries of the design matrix are bounded by a constant, the estimator $\hat{\boldsymbol{\beta}}_{SG}$ is asymptotically consistent and has the same asymptotic distribution as the MLE. The estimator can be determined using a Newton-Raphson algorithm. The ridge parameter k is chosen such that the prediction error is minimized (Gao and Shen, 2007).

4.4. Adjusting the data

One approach to reduce the bias in estimation for grouped data is to adjust the data itself. The origin of this approach is described in Elgmati et al. (2015): Anscombe proposed the so-called *empirical logistic transformation*

$$Z_i = \frac{n_i\bar{Y}_i + \frac{1}{2}}{n_i - n_i\bar{Y}_i + \frac{1}{2}}$$

in 1956 with n_i and $\bar{Y}_i$ the group size and arithmetic mean of the target variable in the ith group, $i = 1, \ldots, G$, and noticed that this leads to almost unbiased estimates for β_i. Later Cox proved that $1/2$ is in fact precisely the constant whose addition removes the first-order bias in the empirical logistic transformation. The MLE after the addition of $1/2$ to the number of ones and zeros in the grouped case is denoted by $\hat{\boldsymbol{\beta}}_{Haldane}$ in the following. For large group sizes n_i, the bias of $\hat{\boldsymbol{\beta}}_{Haldane,i}$ is only of order $1/n_i^2$ (Agresti, 2012, p. 195).

An alternative transformation, which can also be found associated with bias reduction sometimes (Heinze and Schemper, 2002; Galindo-Garre, Vermunt and Bergsma, 2004), is given by Clogg et al. (1991). They suggest to use the MLE after adding $\bar{Y}(p+1)/G$ events and $(1-\bar{Y}(p+1))/G$ non-events in each of the G groups. However, this does not shrink the MLE in the direction of the equiprobability model $\beta = 0$, but in the direction of the independence model, i.e. the intercept is not shrunk.

4.5. Jackknifing

The bias of the MLE can not only be estimated by an analytic approximation but also by jackknifing. The multi-step jackknife estimator for $\boldsymbol{\beta}$ is given as

$$\hat{\boldsymbol{\beta}}_M = \hat{\boldsymbol{\beta}}_{ML} + \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\boldsymbol{\beta}}_{ML} - \hat{\boldsymbol{\beta}}_{-i} \right), \tag{4}$$

where $\hat{\boldsymbol{\beta}}_{-i}$ denotes the MLE if the ith observation is omitted (Bull, Hauck and Greenwood, 1994). Here, the MLE is corrected with the mean deviation of the MLEs when omitting one observation from the MLE in the total sample. The estimators $\hat{\boldsymbol{\beta}}_{-i}$ can be determined iteratively. If $\hat{\boldsymbol{\beta}}_{-i}^{(0)} = \hat{\boldsymbol{\beta}}_{ML}$ is used as the starting value, the estimate in the $(k+1)$-st iteration can be written as

$$\hat{\boldsymbol{\beta}}_{-i}^{(k+1)} = \hat{\boldsymbol{\beta}}_{ML} + \left(\boldsymbol{F}_{-i}^{(0)} \right)^{-1} \boldsymbol{X}_{-i}^T \boldsymbol{r}_{-i}^{(0)} + \sum_{t=1}^{k} \left(\boldsymbol{F}_{-i}^{(t)} \right)^{-1} \boldsymbol{X}_{-i}^T \boldsymbol{r}_{-i}^{(t)}$$

with $\boldsymbol{X}_{-i}$ the design matrix without the ith observation and $\boldsymbol{r}_{-i}^{(t)} = \boldsymbol{y}_{-i} - \hat{\boldsymbol{\pi}}_{-i}^{(t)}$ the residual vector in the t-th step, where the entry for the i-th observation is omitted and $\hat{\boldsymbol{\pi}}_{-i}^{(t)} \in \mathbb{R}^n$ the vector of predicted probabilities based on $\hat{\boldsymbol{\beta}}_{-i}^{(t)}$. The Fisher information matrix is $\boldsymbol{F}_{-i}^{(t)} = \boldsymbol{X}^T \boldsymbol{W}_{-i}^{(t)} \boldsymbol{X}$ with the weight matrix $\boldsymbol{W}_{-i}^{(t)} \in \mathbb{R}^{n \times n}$ based on $\hat{\boldsymbol{\pi}}_{-i}^{(t)}, t = 0, \ldots, k$. For each iteration and each of the n vectors $\hat{\boldsymbol{\beta}}_{-i}^{(t)}$, the data must be run through once for calculating $\hat{\boldsymbol{\pi}}_{-i}^{(t)}$ and $\boldsymbol{F}_{-i}^{(t)}$. The computational effort is therefore high. Hence, it makes sense to consider estimators after a few iterations. The correction that results after the first step, i.e. for the one-step estimator $\hat{\boldsymbol{\beta}}_{-i}^{(1)}$, is asymptotically negligible. Therefore, at least two iterations must be performed to reduce the bias. The quantities required to calculate $\hat{\boldsymbol{\beta}}_{-i}^{(1)}$ are the same for all i, since they are all based on the same starting value $\hat{\boldsymbol{\beta}}_{ML}$. However, the quantities $\left(\boldsymbol{F}_{-i}^{(1)} \right)^{-1}$ and $\boldsymbol{r}_{-i}^{(1)}$ needed for the calculation of $\hat{\boldsymbol{\beta}}_{-i}^{(2)}$ are based on the different estimators $\hat{\boldsymbol{\beta}}_{-i}^{(1)}$ for each $i \in \{1, \ldots, n\}$. This means that to calculate the two-step jackknife estimator, n matrices of size $(p+1) \times (p+1)$ must be calculated and inverted. In the representation

$$\hat{\boldsymbol{\beta}}_{-i}^{(2)} = \hat{\boldsymbol{\beta}}_{-i}^{(1)} + \left(\boldsymbol{F}_{-i}^{(0)} \right)^{-1} \boldsymbol{X}_{-i}^T \boldsymbol{r}_{-i}^{(1)} + \left[\left(\boldsymbol{F}_{-i}^{(1)} \right)^{-1} - \left(\boldsymbol{F}_{-i}^{(0)} \right)^{-1} \right] \boldsymbol{X}_{-i}^T \boldsymbol{r}_{-i}^{(1)}$$

it can be seen that if the change $\left(\boldsymbol{F}_{-i}^{(1)}\right)^{-1} - \left(\boldsymbol{F}_{-i}^{(0)}\right)^{-1}$ is small, this large effort, which arises from the inversion of the updated Fisher matrices, can be saved in the calculation. This is the motivation behind the *approximate two-step jackknife estimator* $\hat{\boldsymbol{\beta}}_{JA}$ proposed in Bull, Hauck and Greenwood (1994). To eliminate the additional calculations and inversions of the updated Fisher matrices, the matrix $\left(\boldsymbol{F}_{-i}^{(1)}\right)^{-1}$ is replaced by $\left(\boldsymbol{F}_{-i}^{(0)}\right)^{-1}$ in the calculation of $\hat{\boldsymbol{\beta}}_{-i}^{(2)}$. The resulting estimator can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{JA} = \hat{\boldsymbol{\beta}}_{ML} &+ \left[\frac{n-1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i(Y_i-\pi_i)(1-a_i)^{-1}\right] \\ &+ \frac{n-1}{n}\left[\sum_{i=1}^{n}\boldsymbol{x}_i(Y_i-\pi_{ii})(1-a_i)^{-1}\right. \\ &\left.+ \sum_{i=1}^{n}\left(\boldsymbol{I}+\pi_i(1-\pi_i)(1-a_i)^{-1}\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{F}^{-1})\right)\left(\sum_{k=1}^{n}\boldsymbol{x}_k(\pi_{ik}-\pi_k)\right)\right]\end{aligned}$$

with

$$\begin{aligned}\pi_{ik} &= \frac{\exp(m_{ik})}{1+\exp(m_{ik})}, \\ m_{ik} &= \boldsymbol{x}_k^T\left[\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{F}^{-1}\boldsymbol{x}_i(Y_i-\pi_i)(1-a_i)^{-1}\right] \text{ and} \\ a_i &= \pi_i(1-\pi_i)\boldsymbol{x}_i^T\boldsymbol{F}^{-1}\boldsymbol{x}_i,\ i,k=1,\ldots,n,\end{aligned}$$

where all quantities only depend on the MLE and $\boldsymbol{I}$ denotes the identity matrix. A prerequisite for the existence of the jackknife estimator is the existence of the MLE. In addition, the existence of the MLE after the removal of each observation is required, which means that deleting an observation must not result in separation of the data (Bull, Hauck and Greenwood, 1994).

4.6. Bias correction for the ridge estimator

Based on the same motivation as in Gao and Shen (2007), that with highly correlated covariables the ridge estimator provides more stable estimates with less variance than the MLE but is no longer asymptotically unbiased, there are also approaches for direct bias corrections of the ridge estimator.

Wu and Asar (2015) determine the bias of the ridge estimator

$$\hat{\boldsymbol{\beta}}_{Ridge} = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML}$$

according to Schaefer, Roi and Wolfe (1984) with ridge parameter k and the $(p+1)$-dimensional identity matrix $\boldsymbol{I}$, and define the bias-corrected ridge estimator

$$\hat{\boldsymbol{\beta}}_{WA}(k) = \left[\boldsymbol{I}-k^2\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+k\boldsymbol{I}\right)^{-2}\right]\hat{\boldsymbol{\beta}}_{ML}.$$

Its bias $b_{WA}(\boldsymbol{\beta})$ is smaller than the bias $b_{Ridge}(\boldsymbol{\beta})$ of the ridge estimator for every $k > 0$ (Wu and Asar, 2015, Theorem 3.1). In addition, conditions can be specified for k depending on the eigenvalues of the Fisher matrix, under which this estimator is superior to the MLE or the ridge estimator with regard to the mean squared error $\text{MSE} = \mathbb{E}\left[||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2\right]$ (Wu and Asar, 2015, Theorem 3.2 and 3.3). Wu and Asar suggest

$$k_{WA} = \frac{p+1}{\sum_{j=0}^{p}[\alpha_j^2/(1+(1+\lambda_j\alpha_j^2)^{1/2})]}$$

as a choice for k with

$$\begin{aligned}&\boldsymbol{\alpha} = \boldsymbol{Q}^T\boldsymbol{\beta} \in \mathbb{R}^{p+1}, \text{ where}\\ &\boldsymbol{\Lambda} = diag\{\lambda_0, \ldots, \lambda_p\} = \boldsymbol{Q}^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})\boldsymbol{Q}\end{aligned}$$

denotes the spectral decomposition of the Fisher matrix $\boldsymbol{F} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ of the MLE, with eigenvalues $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_p > 0$ and $\boldsymbol{Q}$ the matrix of eigenvectors.

In Özkale and Arıcan (2019) a different approach is chosen to reduce the bias of the ridge estimator, namely jackknifing. However, the usual ridge estimator is not considered here, but rather its one-step approximation

$$\hat{\boldsymbol{\beta}}_{Ridge}^{(1)} = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}_q\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\beta}}_{ML}^{(1)}$$

with $k > 0$ and $\hat{\boldsymbol{\beta}}_{ML}^{(1)}$ the MLE after the first iteration of Iteratively Weighted Least Squares (IWLS). For this, the weighted jackknife estimator is formed analogously to (4), with the difference that the estimates for the bias are not all given the same weight $1/n$, but each is weighted with $1 - h_{ii}(k)$,

$$h_{ii}(k) = W_{ii}\boldsymbol{x}_i^T\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}_q\right)^{-1}\boldsymbol{x}_i, i = 1, \ldots, n.$$

The *first-order approximated jackknifed ridge logistic estimator* is thus defined as

$$\hat{\boldsymbol{\beta}}_{OA} = \left[\boldsymbol{I}_q - k^2\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}_q\right)^{-2}\right]\hat{\boldsymbol{\beta}}_{ML}^{(1)}.$$

As choices for the ridge parameter, Özkale and Arıcan make three suggestions, of which they favor $\hat{k}_H$, which is equal to Wu and Asar's suggestion when using $\hat{\boldsymbol{\beta}}_{ML}^{(1)}$ instead of the iterated MLE. This choice of ridge parameter corresponds to the harmonic mean of the individual ridge parameters that minimize the component-wise MSEs

$$\text{MSE}_j = \mathbb{E}\left[\left(\hat{\beta}_j - \beta_j\right)^2\right], j = 0, \ldots, p.$$

With this choice, $\hat{\boldsymbol{\beta}}_{OA}$ corresponds to the estimator from Wu and Asar (2015) when replacing the usual MLE by its one-step approximation. Özkale and Arıcan

(2019) also work out analytically in which situations the proposed estimator is superior regarding the bias or the MSE compared to the estimators $\hat{\boldsymbol{\beta}}^{(1)}_{ML}$, $\hat{\boldsymbol{\beta}}^{(1)}_{Ridge}$ as well as the *first-order approximated principal components logistic regression estimator* and the *first-order approximated r-k class estimator* (Özkale and Arıcan, 2019, Theorems 1 to 12). The latter two estimators do not generally reduce the bias and are therefore not considered here.

4.7. Exact logistic regression

Another method for estimating the parameter vector $\boldsymbol{\beta}$ is exact logistic regression which goes back to Cox (Cox and Snell, 1989, pp. 27–30) and is often recommended as an alternative to the MLE for small samples (Hirji, Tsiatis and Mehta 1989; Agresti 2012, p. 267; Hosmer, Lemeshow and Sturdivant 2013, pp. 387, 395).

Exact inference with respect to $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)^T$ uses the permutation distribution of the sufficient statistics

$$T_0 = \sum_{i=1}^{n} Y_i \text{ and } \boldsymbol{T} = (T_1, \ldots, T_p)^T = \sum_{i=1}^{n} Y_i \tilde{\boldsymbol{x}}_i$$

with $\tilde{\boldsymbol{x}}_i = (x_{i1}, \ldots, x_{ip})^T$. To estimate a single parameter, the conditional probability for the associated sufficient statistic is determined given the values of the sufficient statistics associated with all other parameters. This probability only depends on the parameter of interest. Consider β_p without loss of generality. Then this probability is given as

$$\begin{aligned} f(t_p \mid \beta_p) &= \mathbb{P}(T_p = t_p \mid T_1 = t_1, \ldots, T_{p-1} = t_{p-1}) \\ &= \frac{c(t_1, \ldots, t_p) \exp(\beta_p t_p)}{\sum_{u: c(t_1, \ldots, t_{p-1}, u) \geq 1} \exp(\beta_p u)} \end{aligned}$$

with

$$\begin{aligned} &c(\boldsymbol{t}) = |S(\boldsymbol{t})|, \\ &S(\boldsymbol{t}) = \left\{ (y_1, \ldots, y_n)^T : \sum_{i=1}^{n} y_i = t_0, \sum_{i=1}^{n} y_i x_{ij} = t_j, j = 1, \ldots, p \right\}. \end{aligned}$$

That is, $c(\boldsymbol{t})$ gives the number of 0-1 vectors $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ such that the sufficient statistics equal the values $t_0, t_1, \ldots, t_p$. With this, a conditional maximum likelihood estimator (CMLE) can be determined as the value for β_p that maximizes $f(t_p \mid \beta_p)$. However, this only works as long as t_p is neither the smallest possible value t_{min} nor the largest possible value t_{max}. Alternatively, a median unbiased estimator is given as

$$\hat{\beta}_{exact,p} = \begin{cases} (\beta_+ + \beta_-)/2, & \text{if } t_{min} < t_p < t_{max} \\ \beta_+, & \text{if } t_p = t_{min} \\ \beta_-, & \text{if } t_p = t_{max} \end{cases}$$

with β_- such that

$$\mathbb{P}_{\beta_-}(T_p \geq t_p \mid T_1 = t_1, \ldots, T_{p-1} = t_{p-1}) = 0.5$$

and β_+ such that

$$\mathbb{P}_{\beta_+}(T_p \leq t_p \mid T_1 = t_1, \ldots, T_{p-1} = t_{p-1}) = 0.5.$$

This estimator exists for all values of t_p. For tests and confidence intervals see e.g. Mehta and Patel (1995). The calculation of the conditional distribution of the sufficient statistics is very computationally intensive and was almost infeasible until a recursive algorithm was developed by Hirji, Mehta and Patel (1987).

4.8. Discriminant function approach

The discriminant function approach to estimating odds ratios was known before logistic regression. The approach exploits the fact that the logistic regression model is equivalent to the usual linear discriminant analysis model if the covariables in the two groups with $Y_i = 0$ and $Y_i = 1$ are normally distributed with different expectations but the same variances. Using this relationship, estimators for $\boldsymbol{\beta}$ that only depend on estimators for the expectation and covariance matrix of the normal distribution in each group can be derived. The approach was no longer used over time due to the disadvantage of requiring a distribution assumption for the covariables in $\boldsymbol{X}$. In case of non-normally distributed covariables such as binary covariables, the estimates are again biased away from the origin (Hosmer, Lemeshow and Sturdivant, 2013, pp. 20–21, 45-46).

Lyles, Guo and Hill (2009) consider the approach again and show that the distributional assumptions can be relaxed for the case that only one of the regression parameters β_j is of interest. Let $\boldsymbol{X}_j$ denote the j-th covariable for this section, $\boldsymbol{X}_{-j}$ the remaining covariables and $\boldsymbol{x}_j$ resp. $\boldsymbol{x}_{-j}$ the associated realizations. Under the assumption that a linear regression model

$$\mathbb{E}[X_{ij} \mid Y_i = y_i, \boldsymbol{X}_{i,-j} = \boldsymbol{x}_{i,-j}] = \alpha^* + \beta^* y_i + (\boldsymbol{\gamma}^*)^T \boldsymbol{x}_{i,-j}, \tag{5}$$

holds for the j-th covariate with parameters $\alpha^*, \beta^* \in \mathbb{R}$ and $\boldsymbol{\gamma}^* \in \mathbb{R}^{p-1}$, where i.i.d. $N(0, \sigma^2)$ distributed errors are assumed, the estimator

$$\hat{\beta}_{LGH,samp,j} = \hat{\beta}_j^* / \hat{\sigma}^2$$

and the *Uniformly Minimum Variance Unbiased Estimator* (UMVUE)

$$\hat{\beta}_{LGH,UMVU,j} = \frac{n-p-5}{n-p-3} \hat{\beta}_j^* / \hat{\sigma}^2$$

are defined where $\hat{\beta}^*$ denotes the ordinary least squares (OLS) estimator for β^* and $\hat{\sigma}^2$ is the variance estimator in the linear model (5).

5. Comparison of estimators in the literature

The results of various simulation studies regarding the previously introduced estimators are presented below. The focus in this summary is on the results regarding bias, although this is not the main focus for some of the studies. Statements that something is observed for "small" n (sample size) or p (number of covariates) are always to be understood in the context of the respective simulation study. In addition, all findings apply conditionally to the fact that no separation is detected in the data set, as long as the opposite is not explicitly stated. An overview of the parameter settings used in the individual simulation studies can be found in Table 2 of the online Appendix B (Stolte et al., 2024). For more details, please refer to the corresponding publication.

Maximum likelihood estimator In addition to the theoretical considerations, numerous simulation studies deal with the bias of the MLE and, in particular, investigate possible influencing factors on the size of the bias. In addition to the bias, many of the simulation studies reveal other weaknesses of the MLE in finite samples like underestimation (Park and Choi, 2008; Sur and Candès, 2019) or overestimation (Peduzzi et al., 1996) of its true variance, bias in the resulting estimates of the probabilities π_i (King and Zeng, 2001; Lyles, Guo and Greenland, 2012), strong deviation of the distribution of the MLE from the normal (van Smeden et al., 2016) and conservativeness and very low power for tests like the Wald test (Peduzzi et al., 1996; Courvoisier et al., 2011). However, these problems will not be discussed further, as this would go beyond the scope of this review.

As early as 1979, Anderson and Richardson showed in a small simulation study that if the sample size n is small, a serious bias of the MLE can be observed. This dependency of the bias on the sample size is confirmed in all mentioned simulation studies. However, there is disagreement about which values of n are to be considered as "small", which is already reflected in the parameter settings of the individual studies (cf. Table 2, online Appendix B, Stolte et al., 2024). It seems to depend in particular on the specific data structure (Courvoisier et al., 2011). Schaefer (1983) observed shortly after Anderson and Richardson that a large number of covariables p also leads to a higher bias. Furthermore, he notes that the bias is higher at an angle between the parameter vector $\boldsymbol{\beta}$ and the eigenvector to the smallest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$ of 90° than at an angle of 0° and that collinearity increases the variability and bias of the MLE. The connection between bias and collinearity is also confirmed in other simulation studies (Bull, Greenwood and Hauck, 1997; Courvoisier et al., 2011), but this effect cannot be observed in others (van Smeden et al., 2016). Various studies confirm the theoretically derived fact from Cordeiro and McCullagh (1991) that the bias points away from the origin and that the magnitude of the bias is approximately proportional to the size of the true coefficients (Bull, Mak and Greenwood, 2002; Nemes et al., 2009; van Smeden et al., 2016).

A well-known rule for the sample size in relation to the number of covariates, which can also be found in the context of the bias of the MLE, is the so-called

rule of ten. This goes back to Peduzzi et al. (1996) and states that from a number of ten events per covariable (*events per variable*, EPV) problems in ML estimation are negligible. According to Vittinghoff and McCulloch (2007), five EPV are sufficient. The authors in van Smeden et al. (2016), on the other hand, observe that a noticeably larger number is required depending on the data structure. For $\boldsymbol{\beta} \neq \mathbf{0}$, the bias of the MLE decreases in terms of EPV, but it does not quite reach zero even for 150 EPV. According to the authors, the way in which separation is handled also has a major impact on simulation results. Courvoisier et al. (2011) also observe that 10 EPV is generally not sufficient, but that the sample size, number, and correlation of the covariates and the true size of $\boldsymbol{\beta}$ are also decisive. According to King and Zeng (2001), in close connection with the rule of ten, a strong bias can generally be observed if the proportion of ones in the target variable Y is small.

Another factor that seems to affect the bias is the distribution of the covariates. In general, larger problems occur with binary covariates than with continuous (Vittinghoff and McCulloch, 2007). In addition, imbalance in binary covariates seems to increase the bias (Hirji, Tsiatis and Mehta, 1989; Heinze and Schemper, 2002; Vittinghoff and McCulloch, 2007).

Bias-corrected estimators As mentioned before there are different equivalent expressions for the first-order bias and thus different equivalent bias-corrected estimators. Various studies have used different of these estimators often consistent with a first-order bias expression derived in the respective article. The first simulation results for bias-corrected MLEs are provided by Anderson and Richardson (1979), where the authors compare the estimator they proposed based on their bias expression with the MLE. However, due to the limited computing power at the time, the simulation study was small. The authors find that the correction is effective except in extreme cases (like n is small and few zeros are expected for Y) and achieves good results from $n = 60$ on. If n is small, they observe overcorrection. Schaefer (1983) also compares the estimator resulting from his bias expression with the MLE. He observed that the bias-corrected estimator no longer differs from the MLE for $n = 200$ or more. On the other hand, for small n, large p, or a true parameter vector that is orthogonal to the eigenvector for the smallest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$, a strong improvement can be observed using the correction. The corrected estimator is also less influenced by collinearity. The latter is also observed in Bull, Greenwood and Hauck (1997). Bull, Hauck and Greenwood (1994) and Bull, Greenwood and Hauck (1997) also observe that the bias-corrected estimator overcorrects when n is small so that the mean bias for the corrected estimator tends to have the opposite sign to that of the MLE. King and Zeng (2001) focus on the bias in the estimated probabilities instead of the bias of the parameter estimates, but the latter is also considered. For $\hat{\boldsymbol{\beta}}_{CM}$ an improvement over the MLE is observed. In Bull, Mak and Greenwood (2002) the correction according to Cox and Snell (1968) is examined. Once again, an overcorrection in small samples is observed. In contrast to Schaefer (1983), however, even with an n of 200 there is still a noticeable improvement in terms of bias compared to the MLE. In a

very small-scale simulation, Matin (2006) investigates under which conditions the difference between the MLE and $\hat{\boldsymbol{\beta}}_{Schaefer}$ is large for a dataset where it is known what the parameters stand for from a medical point of view. Differences in the odds ratios (OR) that are considerable for small samples are found. In particular, the difference is large if $1-2\hat{\pi}_i$ is small on average or if the interquartile range of these terms is large. Park and Choi (2008) observe for their estimator that its bias is negligible and that this estimator also has a lower variance than the MLE, with the gain being higher the smaller n is. However, the variance estimator for this corrected estimator usually overestimates its true variance. Maiti and Pradhan (2008) recommend $\hat{\boldsymbol{\beta}}_{CM}$ among the compared estimators $\hat{\boldsymbol{\beta}}_{ML}$, $\hat{\boldsymbol{\beta}}_{CM}$ and $\hat{\boldsymbol{\beta}}_{Firth}$ they compared since it has the smallest bias and usually also the smallest MSE. No simulation results are known for comparisons of the estimators according to Copas (1988) and Sur and Candès (2019).

Firth's estimator The behavior of Firth's estimator was also frequently examined. Heinze and Schemper (2002) recommend Firth's estimator because it has the lowest bias among the ones they considered (MLE, Firth's estimator, adjustment according to Clogg et al. (1991), exact logistic regression). However, they observe that Firth's estimator also slightly overcorrects in some cases. Bull, Mak and Greenwood (2002) also observe the overcorrection in small samples but state that this is weaker than for the bias-corrected estimators. From a sample size n of 200, the bias-corrected and Firth's estimator can no longer be distinguished. Firth's estimator always shows a smaller MSE than the MLE, with the difference increasing for decreasing n or increasing p. The bias of $\hat{\boldsymbol{\beta}}_{Firth}$ is on average closer to zero on all datasets than only on the ones without separation. On the data sets with separation, Firth's estimator provides large but finite estimates. Overall, Firth's estimator is recommended for routine use, since it is less biased and more efficient than the MLE and the bias-corrected estimator for small n and can always be computed. With medium n it is biased to a similar extent as the bias-corrected estimator and is therefore still less biased than the MLE, and for large n Firth's estimator becomes equivalent to the MLE. Firth's estimator is also recommended in Heinze (2006) since it can be used for all data sets, has a low variance, and all nominal test levels are observed to be respected. Maiti and Pradhan (2008), on the other hand, observe a larger bias for Firth's estimator for coefficients that are far from zero ($|\beta_j| > 2$) than for the bias-corrected estimator and generally also a higher MSE. They confirm the result of Bull, Mak and Greenwood (2002) that Firth's estimator shows a large bias on data sets with separation. In contrast, Park and Choi (2008) summarize that the bias of Firth's estimator is negligible and that it has a smaller variance than the MLE. In addition, they notice that the bias of its variance estimator is also negligible. Shen and Gao (2008) also observe a lower bias and MSE for Firth's estimator than for the MLE, especially for small n. However, they conclude that the estimation of the variance via the approximate information matrix overestimates the true variance for small n. In van Smeden et al. (2016) for $\hat{\boldsymbol{\beta}}_{Firth}$ a bias close to zero and a lower MSE than for the MLE was detected. In addition, the mean width of the confidence intervals is observed to be systematically smaller

compared to those for the MLE. Puhr et al. (2017) also confirm that Firth's estimator almost always shows the smallest bias among those considered (MLE, ridge estimator, Firth's estimator without and with FLIC or FLAC adjustment, bayesian estimators according to Elgmati et al. (2015), Greenland and Mansournia (2015) and Gelman et al. (2008)). Even in the most unfavorable situations, the mean standardized bias is less than 1%. However, the root MSE (RMSE) is considerably higher than that of the ridge estimator. Kosmidis, Kenne Pagui and Sartori (2020) observe a very good behavior of Firth's estimator with regard to the mean bias and a noticeable reduction in MSE and bias when estimating the ORs.

Estimator according to Kenne Pagui, Salvan and Sartori (2017) There are not many results on the behavior of $\hat{\boldsymbol{\beta}}_{KPSS}$. Kenne Pagui, Salvan and Sartori (2017) themselves show in a small simulation that the estimator is nearly median-unbiased but has a bigger mean bias than Firth's estimator. Kosmidis, Kenne Pagui and Sartori (2020) also confirm that the estimator achieves median unbiasedness very well and also shows a noticeably lower mean bias than the MLE. However, the mean bias is again stronger than that of Firth's estimator.

Estimator according to Gao and Shen (2007) Shen and Gao (2007) show that the estimator $\hat{\beta}_{SG}$ generally has a lower bias and MSE than the MLE, especially in smaller samples. However, the bias is higher than that of Firth's estimator, especially in small or medium-sized samples. On the other hand, according to Gao and Shen (2007), the estimator has the lowest MSE and in larger samples, the two penalized estimators can no longer be distinguished. As with Firth's estimator, the approximate information matrix overestimates the variance for small n.

Adjusting the data Various methods for data adjustment (see Section 4.4 and Table 2 of online Appendix B, Stolte et al., 2024) are compared in Whaley (1991) for two binary covariates. No general best method can be identified. This depends on how frequently the event of interest occurs and whether bias or MSE is considered more important. However, the worst choice in terms of bias is always ACAC, i.e. the bias-corrected estimator according to Schaefer (1983) with previous $+1/2$ adjustment of the data. If n is small, all methods are observed to overcorrect. Especially with small n and small β_0 none of the methods performs well. Heinze and Schemper (2002) consider the adjustment of the data according to Clogg et al. (1991). They note a tendency to slightly underestimate the true effect and a larger bias than for Firth's estimator. However, the bias of the estimator on the adjusted data is smaller than that of the ordinary MLE.

Jackknifing Jackknife estimators are compared with other methods in Bull, Hauck and Greenwood (1994) and Bull, Greenwood and Hauck (1997). Both studies come to qualitatively the same results: the multi-step and the exact two-step jackknife estimator overcorrect for small n, while the approximate two-step jackknife estimator behaves similarly to the bias-corrected estimator and does

not overcorrect that much. From a sample size n of 75, the different correction methods can no longer be distinguished.

Bias-corrected ridge estimators Wu and Asar (2015) compare their estimator with the MLE and the ridge estimator. They sum up that $\hat{\boldsymbol{\beta}}_{WA}$ always has a lower squared bias and a smaller MSE than the MLE and ridge estimator. Özkale and Arıcan (2019) also compare $\hat{\boldsymbol{\beta}}_{OA}$ with the previously mentioned estimators and with their one-step approximations. They find that $\hat{\boldsymbol{\beta}}_{OA}$ is always superior to the MLE and ridge estimator and their one-step approximations with regard to bias and usually also to $\hat{\boldsymbol{\beta}}_{WA}$, except in rare cases such as high multicollinearity, small n and large ridge parameter k. In addition, $\hat{\boldsymbol{\beta}}_{WA}$ is more strongly biased than the approximate ridge estimator. Regarding the MSE, the MLE is the worst and the ridge estimator is the best. The two corrected ridge estimators lie in between, with their order depending on the parameter settings. $\hat{\boldsymbol{\beta}}_{OA}$ tends to be better for large n and $\hat{\boldsymbol{\beta}}_{WA}$ for small n, large k and moderate collinearity.

Exact logistic regression The median-unbiased estimator (MUE) from the exact logistic regression is recommended by Hirji, Tsiatis and Mehta (1989) for small and medium-sized samples as well as for sparse data based on their findings that the MUE is *absolutely more accurate* than the MLE in all situations considered, i.e. $\mathbb{P}_\beta\left(\left|\hat{\beta}-\beta\right|<\delta\right)$ is higher for all $\delta>0$. In addition, the MUE also has a smaller MSE on average if both estimators exist. King and Ryan (2002) state that the conditional MLE (CMLE) of exact logistic regression overestimates the effect less than the ordinary MLE. However, they sum up that in the case they are considering, both methods do not provide a particularly good estimate. In Heinze and Schemper (2002) the exact logistic regression is also considered, whereby the CMLE is calculated if no separation is detected and the MUE otherwise. They conclude that the exact logistic regression, like the adjustment according to Clogg et al. (1991), has a bias greater than that of Firth's estimator, but smaller than that of the MLE. Heinze and Schemper (2002) state that the exact logistic regression is often not applicable, due to the large proportion of the simulated datasets with degenerated conditional distribution of the sufficient statistics, especially for small n and large p.

Discriminant function approach According to Lyles, Guo and Hill (2009), the discriminant function approach leads to considerably less biased estimates than the MLE. The UMVUE is even said to be almost unbiased. However, under the same conditions, Lyles, Guo and Greenland (2012) find a large bias in the estimated ORs.

6. Conclusion

We have given a comprehensive overview of numerous proposals for estimators to reduce bias in logistic regression and also of many simulation studies that compare subsets of those estimators.

There are several reported properties where studies contradict each other, e.g., how collinearity influences the bias or what sample size and number of events per variable (EPV) can be seen as sufficiently large. Except for the frequent recommendation of Firth's estimator, there is no clear unique recommendation on which of the (other) estimators to use. There are many combinations of estimators presented in this article that are not directly compared in any simulation study. The results of various studies are not comparable to each other due to very different parameter settings (cf. Table 2 of online Appendix B, Stolte et al., 2024). For more specific recommendations more research comparing all estimators under unified conditions is needed.

Fortunately, some promising findings are confirmed by many of these studies, e.g., the bias-corrected estimators overcorrect in very small samples but are superior to the MLE for moderate sample sizes. Most of the reviewed studies recommend Firth's estimator if considered, due to small bias and applicability with separable data.

Funding

The first author was partly supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, A3.

Supplementary Material

Supplement
(doi: 10.1214/24-SS148SUPP; .pdf). Table containing additional information on all mentioned estimators.
Table containing additional information on all mentioned simulation studies and the respective parameter settings.

References

Agresti, A. (2012). *Categorical Data Analysis*. John Wiley & Sons, Somerset, USA. MR3087436

Albert, A. and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* **71** 1–10. https://doi.org/10.1093/biomet/71.1.1 MR0738319

Anderson, J. A. and Richardson, S. C. (1979). Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. *Technometrics* **21** 71–78. https://doi.org/10.1080/00401706.1979.10489724

Bowman, K. O. and Shenton, L. R. (1965). *Biases and Covariances of Maximum Likelihood Estimators*. Union Carbide Corporation, Nuclear Division.

Bull, S. B., Greenwood, C. M. T. and Hauck, W. W. (1997). Jackknife Bias Reduction for Polychotomous Logistic Regression. *Statistics in Medicine* **16** 545–560. https://doi.org/10.1002/(SICI)1097-0258(19970315)16:5<545::AID-SIM421>3.0.CO;2-3

Bull, S. B., Hauck, W. W. and Greenwood, C. M. T. (1994). Two-Step Jackknife Bias Reduction for Logistic Regression MLEs. *Communications in Statistics - Simulation and Computation* **23** 59–88. https://doi.org/10.1080/03610919408813156

Bull, S. B., Mak, C. and Greenwood, C. M. T. (2002). A Modified Score Function Estimator for Multinomial Logistic Regression in Small Samples. *Computational Statistics & Data Analysis* **39** 57–74. https://doi.org/10.1016/S0167-9473(01)00048-2 MR1895558

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression. *Journal of the American Statistical Association* **86** 68–78. https://doi.org/10.1080/01621459.1991.10475005

Copas, J. B. (1988). Binary Regression Models for Contaminated Data. *Journal of the Royal Statistical Society. Series B (Methodological)* **50** 225–265. MR0964178

Cordeiro, G. M. and Barroso, L. P. (2007). A Third-Order Bias Corrected Estimate in Generalized Linear Models. *TEST* **16** 76–89. https://doi.org/10.1007/s11749-006-0002-1 MR2368455

Cordeiro, G. M. and McCullagh, P. (1991). Bias Correction in Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)* **53** 629–643. MR1125720

Courvoisier, D. S., Combescure, C., Agoritsas, T., Gayet-Ageron, A. and Perneger, T. V. (2011). Performance of Logistic Regression Modeling: Beyond the Number of Events per Variable, the Role of Data Structure. *Journal of Clinical Epidemiology* **64** 993–1000. https://doi.org/10.1016/j.jclinepi.2010.11.012

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society. Series B (Methodological)* **30** 248–275. MR0237052

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, 2 ed. Chapman and Hall/CRC, London. https://doi.org/10.1201/9781315137391 MR1014891

Elgmati, E., Fiaccone, R. L., Henderson, R. and Matthews, J. N. S. (2015). Penalised Logistic Regression and Dynamic Prediction for Discrete-Time Recurrent Event Data. *Lifetime Data Analysis* **21** 542–560. https://doi.org/10.1007/s10985-015-9321-4 MR3397505

Fahrmeir, L. and Kaufmann, H. (1986). Asymptotic Inference in Discrete Response Models. *Statistische Hefte* **27** 179–205. https://doi.org/10.1007/BF02932567 MR0865487

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34333-9_1 MR3075546

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80** 27–38. https://doi.org/10.1093/biomet/80.1.27 MR1225212

Galindo-Garre, F., Vermunt, J. K. and Bergsma, W. P. (2004). Bayesian Posterior Estimation of Logit Parameters With Small Samples. *Sociological Methods & Research* **33** 88–117. https://doi.org/10.1177/0049124104265997 MR2086481

Gao, S. and Shen, J. (2007). Asymptotic Properties of a Double Penalized Maximum Likelihood Estimator in Logistic Regression. *Statistics & Probability Letters* **77** 925–930. https://doi.org/10.1016/j.spl.2007.01.004 MR2380656

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics* **2** 1360–1383. https://doi.org/10.1214/08-AOAS191 MR2655663

Greenland, S. and Mansournia, M. A. (2015). Penalization, Bias Reduction, and Default Priors in Logistic and Related Categorical and Survival Regressions. *Statistics in Medicine* **34** 3133–3143. https://doi.org/10.1002/sim.6537 MR3402580

Heinze, G. (2006). A Comparative Investigation of Methods for Logistic Regression With Separated or Nearly Separated Data. *Statistics in Medicine* **25** 4216–4226. https://doi.org/10.1002/sim.2687 MR2307586

Heinze, G. and Schemper, M. (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine* **21** 2409–2419. https://doi.org/10.1002/sim.1047

Hirji, K. F., Mehta, C. R. and Patel, N. R. (1987). Computing Distributions for Exact Logistic Regression. *Journal of the American Statistical Association* **82** 1110–1117. https://doi.org/10.2307/2289388 MR0922176

Hirji, K. F., Tsiatis, A. A. and Mehta, C. R. (1989). Median Unbiased Estimation for Binary Data. *The American Statistician* **43** 7–11. https://doi.org/10.2307/2685158 MR0997504

Hosmer, D. V., Lemeshow, S. and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Hoboken, New Jersey. https://doi.org/10.1002/9781118548387.fmatter

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186** 453–461. https://doi.org/10.1098/rspa.1946.0056 MR0017504

Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017). Median Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **104** 923–938. https://doi.org/10.1093/biomet/asx046 MR3737312

King, E. N. and Ryan, T. P. (2002). A Preliminary Investigation of Maximum Likelihood Logistic Regression Versus Exact Logistic Regression. *The American Statistician* **56** 163–170. https://doi.org/10.1198/00031300283 MR1963262

King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis* **9** 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kosmidis, I. (2014). Bias in Parametric Estimation: Reduction and Useful Side-

Effects. *WIREs Computational Statistics* **6** 185–196. https://doi.org/10.1002/wics.1296

Kosmidis, I. and Firth, D. (2021). Jeffreys-Prior Penalty, Finiteness and Shrinkage in Binomial-Response Generalized Linear Models. *Biometrika* **108** 71–82. https://doi.org/10.1093/biomet/asaa052 MR4226190

Kosmidis, I., Kenne Pagui, E. C. and Sartori, N. (2020). Mean and Median Bias Reduction in Generalized Linear Models. *Statistics and Computing* **30** 43–59. https://doi.org/10.1007/s11222-019-09860-6 MR4057470

Lyles, R. H., Guo, Y. and Hill, A. N. (2009). A Fresh Look at the Discriminant Function Approach for Estimating Crude or Adjusted Odds Ratios. *The American Statistician* **63** 320–327. MR2751748

Lyles, R. H., Guo, Y. and Greenland, S. (2012). Reducing Bias and Mean Squared Error Associated With Regression-Based Odds Ratio Estimators. *Journal of Statistical Planning and Inference* **142** 3235–3241. https://doi.org/10.1016/j.jspi.2012.05.005 MR2956808

Maiti, T. and Pradhan, V. (2008). A Comparative Study of Bias Corrected Estimates in Logistic Regression. *Statistical Methods in Medical Research* **17** 621–34. https://doi.org/10.1177/0962280207084156 MR2654668

Matin, M. (2006). Effect of Using Bias-Corrected Estimators in Logistic Regression Model in Small Samples: Prostate-Specific Antigen (PSA) Data. *Data Science Journal* **5** 100–107. https://doi.org/10.2481/dsj.5.100

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2 ed. Chapman & Hall, London. MR3223057

McLachlan, G. J. (1980). A Note on Bias Correction in Maximum Likelihood Estimation With Logistic Discrimination. *Technometrics* **22** 621–627. https://doi.org/10.1080/00401706.1980.10486214

Mehta, C. R. and Patel, N. R. (1995). Exact Logistic Regression: Theory and Examples. *Statistics in Medicine* **14** 2143–2160. https://doi.org/10.1002/sim.4780141908

Nemes, S., Jonasson, J. M., Genell, A. and Steineck, G. (2009). Bias in Odds Ratios by Logistic Regression Modelling and Sample Size. *BMC Medical Research Methodology* **9** 56. https://doi.org/10.1186/1471-2288-9-56

O'neill, T. J. (1994). The Bias of Estimating Equations With Application to the Error Rate of Logistic Discrimination. *Journal of the American Statistical Association* **89** 1492–1498. https://doi.org/10.1080/01621459.1994.10476888 MR1310238

Park, M. and Choi, B. (2008). Bias Corrected Maximum Likelihood Estimator Under the Generalized Linear Model for a Binary Variable. *Communications in Statistics - Simulation and Computation* **37** 1507–1514. https://doi.org/10.1080/03610910802063772 MR2542406

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. and Feinstein, A. R. (1996). A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology* **49** 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Puhr, R., Heinze, G., Nold, M., Lusa, L. and Geroldinger, A. (2017). Firth's Logistic Regression With Rare Events: Accurate Effect Estimates and

Predictions? *Statistics in Medicine* **36** 2302–2317. https://doi.org/10.1002/sim.7273 MR3660132

Santner, T. J. and Duffy, D. E. (1986). A Note on A. Albert and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* **73** 755–758. https://doi.org/10.1093/biomet/73.3.755 MR0897873

Schaefer, R. L. (1983). Bias Correction in Maximum Likelihood Logistic Regression. *Statistics in Medicine* **2** 71–78. https://doi.org/10.1002/sim.4780020108

Schaefer, R. L., Roi, L. D. and Wolfe, R. A. (1984). A Ridge Logistic Estimator. *Communications in Statistics - Theory and Methods* **13** 99–113. https://doi.org/10.1080/03610928408828664

Shen, J. and Gao, S. (2008). A Solution to Separation and Multicollinearity in Multiple Logistic Regression. *Journal of data science* **6** 515–531.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Statistics for Biology and Health.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-16399-0_2

Stolte, M. Herbrandt, S. Ligges, U. (2024). Supplement to "A comprehensive review of bias reduction methods for logistic regression". https://doi.org/10.1214/24-SS148SUPP

Sur, P. and Candès, E. J. (2019). A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression. *Proceedings of the National Academy of Sciences* **116** 14516–14525. https://doi.org/10.1073/pnas.1810420116 MR3984492

Tutz, G. (2011). *Regression for Categorical Data. Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511842061 MR2856629

van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C. and Reitsma, J. B. (2016). No Rationale for 1 Variable per 10 Events Criterion for Binary Logistic Regression Analysis. *BMC Medical Research Methodology* **16** 163. https://doi.org/10.1186/s12874-016-0267-3

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology* **165** 710–718. https://doi.org/10.1093/aje/kwk052

Whaley, F. S. (1991). Comparison of Different Maximum Likelihood Estimators in a Small Sample Logistic Regression With two Independent Binary Variables. *Statistics in Medicine* **10** 723–731. https://doi.org/10.1002/sim.4780100507

Wu, J. and Asar, Y. (2015). On Almost Unbiased Ridge Logistic Estimator for the Logistic Regression Model. *Hacettepe Journal of Mathematics and Statistics* **45** 989–998. https://doi.org/10.15672/HJMS.20156911030 MR3588237

Zorn, C. (2005). A Solution to Separation in Binary Response Models. *Political Analysis* **13** 157–170. https://doi.org/10.1093/pan/mpi009

Özkale, M. R. and Arican, E. (2019). A First-Order Approximated Jackknifed Ridge Estimator in Binary Logistic Regression. *Computational Statistics* **34** 683–712. https://doi.org/10.1007/s00180-018-0851-6 MR4142785