Article

'Bring more data!' – a good advice? Removing separation in logistic regression by increasing sample size

Hana Šinkovec [1], Angelika Geroldinger [1] and Georg Heinze [1]*

[1] Institute of Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems (CEMSIIS), Spitalgasse 23, 1090 Vienna, Austria

* Correspondence: georg.heinze@meduniwien.ac.at

Abstract: The parameters of logistic regression models are usually obtained by the method of maximum likelihood (ML). However, in analyses of small data sets or data sets with unbalanced outcomes or exposures, ML parameter estimates may not exist. This situation has been termed 'separation' as the two outcome groups are separated by the values of a covariate or a linear combination of covariates. To overcome the problem of non-existing ML parameter estimates, applying Firth's correction (FC) was proposed. In practice, however, a principal investigator might be advised to 'bring more data' in order to solve a separation issue. We illustrate the problem by means of examples from colorectal cancer screening and ornithology. It is unclear if such an increasing sample size (ISS) strategy that keeps sampling new observations until separation is removed improves estimation compared to applying FC to the original data set. We performed an extensive simulation study where the main focus was to estimate the cost-adjusted relative efficiency of ML combined with ISS compared to FC. FC yielded reasonably small root mean squared errors and proved to be the more efficient estimator. Given our findings, we propose not to adapt the sample size when separation is encountered but to use FC as the default method of analysis whenever the number of observations or outcome events is critically low.

Keywords: maximum likelihood estimation, logistic regression, Firth's correction, separation, penalized likelihood, bias

1. Introduction

In medical research logistic regression is a popular method used to study the relationship between a binary outcome and a set of covariates. Regression coefficients can be interpreted as log odds ratios and are usually estimated by the method of maximum likelihood (ML). Moreover, individualized prognosis can be obtained by estimating the probability of an outcome given the covariates, making logistic regression indispensable in the era of personalized medicine.

Despite analytically attractive properties under some regulatory conditions as the sample size increases [1,2], in analyses of small or sparse data sets the properties of ML estimator become questionable: ML coefficient estimates are biased away from zero and very unstable [3] or may even not exist [4]. The situation in which the ML estimate of at least one regression coefficient does not exist, i.e., diverges to plus or minus infinity, has been termed 'separation' as the two outcome groups are separated by the values of a covariate or a linear combination of covariates. Therefore, perfect predictions for some (quasi-complete separation) or for all observations (complete separation) of the data set the model is fitted on are obtained. The simplest case of separation arises when an odds ratio should be estimated from a 2×2 contingency table with one zero cell count. Beside small sample size and sparsity various other factors, such as a small relative frequency of one of two levels of the outcome variable (unbalanced outcome), rare exposures and strong associations with the outcome can give rise to separation [5,6]. A large number of strongly correlated covariates is another such factor that has induced a possibly unjustified promotion of the 10 events per variable rule in biomedical literature [7,8].

In order to solve the problem of non-existing ML coefficient estimates, applying Firth's correction (FC) [9], originally intended for bias reduction in generalized linear models, was proposed [5]. FC was shown to be robust to the problem of separation, making it a default choice in situations where sampling artefacts that lead to separation are likely to occur. These artefacts, however, may always be removed by increasing sample size, and in practice a principal investigator is often advised to 'bring more data' in order to solve a separation issue. It is unclear if such a simple increasing sample size (ISS) strategy that keeps sampling new observations until separation is removed improves estimation compared to FC applied to the original data set.

The objective of this paper is therefore to investigate the performance of ML after separation has been removed by ISS compared to FC applied to the initial data. The paper is organized as follows. In the next section we briefly summarize the logistic regression model together with the ML and the FC estimators. In addition, we provide a simple example illustrating the separation problem and describe the setup for a simulation study. Subsequently we describe our findings evaluating the empirical performance of ML combined with ISS and FC. As illustrative examples we consider preliminary and final analyses of two studies from colorectal cancer screening and from ornithology. Finally, we summarize our most important findings.

2. Materials and Methods

2.1. General

Assume that we have N independent observations $(x_i, y_i), i = 1, \dots, N$, where $y_i \in \{0,1\}$ denotes a binary outcome with level 1 occurring with probability π_i and $x_i = (x_{i1} \dots x_{iK})$, $K < N$, denotes a vector of covariate values for the i-th subject. With logistic regression, designed to provide individualized prognosis $\hat{\pi}_i$ given x_i by ensuring that $\hat{\pi}_i \in [0,1]$ [10], the logit (i.e., the log-odds) of π_i is modeled as a linear combination of the covariate values:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_K x_{iK}, \quad (1)$$

where β_0 is an intercept and β_k, $k = 1, \dots K$, are regression coefficients for the covariates X_k. The regression coefficients for covariates X_k can be interpreted as log odds ratios, corresponding to one unit differences in X_k. They are usually obtained by ML estimation, maximizing the log-likelihood function

$$l(\beta) = \sum_{i=1}^{N} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i). \quad (2)$$

In the simplest case with one binary covariate X observations can be classified according to x and y in a 2×2 contingency table with observed frequencies $f_{00}, f_{01}, f_{10}, f_{11}$:

		x	
		0	1
y	0	f_{00}	f_{01}
	1	f_{10}	f_{11}

The ML estimate of β is then given by

$$\hat{\beta} = \log\left(\frac{f_{00} f_{11}}{f_{01} f_{10}}\right), \quad (3)$$

and predicted probabilities are equal to the proportions of events in the two groups.

Now suppose that one of the observed frequencies of the 2×2 table, e.g., $f_{11} = 0$, is zero. Accordingly, the ML estimate of the predicted probability for $x = 1$ is zero, while the ML regression coefficient $\hat{\beta}$ is not defined. According to (3), whenever one of the cell counts of a 2×2 table is zero $\hat{\beta}$ is not defined. To obtain a finite regression coefficient, we have to assume $\hat{\pi}_i \in (0,1)$. This assumption is in accordance with our general belief that a finite number of covariates can never be sufficient to perfectly predict an outcome in the underlying population. If so, then separation should be considered a sampling artefact that may always be removed by increasing sample size.

In FC the likelihood function is penalized by the Jeffreys invariant prior [9]. For 2 × 2 tables FC is equivalent to ML after adding a 0.5 to each of the observed frequencies [11]. Applying FC does not only remove the first-order bias from ML regression coefficient estimates in generalized linear models [9], which may be severe in small or sparse data sets. Moreover, FC is robust to the problem of separation [5] as predicted probabilities are pulled to 0.5 such that $\hat{\pi}_i \in (0,1)$. Note that other methods imposing shrinkage on the regression coefficients can as well overcome the separation issue [11–13].

2.2. Simulation study setup

The simulation study setup is described as recommended by Morris [14].

Aims: We intended to systematically investigate the performance of ML after removing separation by ISS (ML+ISS) in comparison to FC applied to the original data, focusing on scenarios where separation is likely to occur. We also investigated the performance of combining FC with ISS (FC+ISS).

Data-generating mechanisms: To capture a context plausible in medical research, we considered a data-generating scheme as described in Binder et al [15]. First we generated data sets of size N with a binary outcome variable Y and K covariates X_k of mixed types that were obtained by applying certain transformations to variables $Z_1, \ldots, Z_{10}$ sampled from a standard multivariate normal distribution with correlation matrix Σ, see Table 1. In this way, we obtained 4 binary covariates $X_1, \ldots, X_4$, two ordinal covariates, X_5 and X_6, with three levels, and four continuous covariates $X_7, \ldots, X_{10}$ (Table 1). In order to avoid extreme values, we truncated continuous covariates at the third quartile plus five times the interquartile distance of the corresponding distribution.

Table 1. Covariate structure applied in the simulation study. $I(\cdot)$ is the indicator function that equals 1 if the argument is true, and 0 otherwise. $[\cdot]$ indicates that a non-integer part of the argument is eliminated.

z_{ik}	Correlation of z_{ik}	Type	x_{ik}	$\mathrm{E}(x_{ik})$
z_{i1}	$z_{i2}(0.6), z_{i3}(0.5), z_{i7}(0.5)$	binary	$x_{i1} = I(z_{i1} < 0.84)$	0.8
z_{i2}	$z_{i1}(0.6)$	binary	$x_{i2} = I(z_{i2} < -0.35)$	0.36
z_{i3}	$z_{i1}(0.5), z_{i4}(-0.5), z_{i5}(-0.3)$	binary	$x_{i3} = I(z_{i3} < 0)$	0.5
z_{i4}	$z_{i3}(-0.5), z_{i5}(0.5), z_{i7}(0.3), z_{i8}(0.5), z_{i9}(0.3)$	binary	$x_{i4} = I(z_{i4} < 0)$	0.5
z_{i5}	$z_{i3}(-0.3), z_{i4}(0.5), z_{i8}(0.3), z_{i9}(0.3)$	ordinal	$x_{i5} = I(z_{i5} \geq -1.2) + I(z_{i5} \geq 0.75)$	1.11
z_{i6}	$z_{i7}(-0.3), z_{i8}(0.3)$	ordinal	$x_{i6} = I(z_{i6} \geq 0.5) + I(z_{i6} \geq 1.5)$	0.37
z_{i7}	$z_{i1}(0.5), z_{i4}(0.3), z_{i6}(-0.3)$	continuous	$x_{i7} = [10z_{i7} + 55]$	54.5
z_{i8}	$z_{i4}(0.5), z_{i5}(0.3), z_{i6}(0.3), z_{i9}(0.5)$	continuous	$x_{i8} = [\max 0, 100 \exp(z_{i8}) - 20]$	138.58
z_{i9}	$z_{i4}(0.3), z_{i5}(0.3), z_{i8}(0.5)$	continuous	$x_{i9} = [\max 0, 80 \exp(z_{i9}) - 20]$	106.97
z_{i10}	-	continuous	$x_{i10} = [10z_{i10} + 55]$	54.5

We considered a full factorial design, varying the number of covariates $K \in \{2, 5, 10\}$, the sample size $N \in \{80, 200, 500\}$, the expected value of Y, $\mathrm{E}(Y) \in \{0.1, 0.25\}$, and the value of $\beta_1 \in \{0, 0.35, 1.39, 2.77\}$. This resulted in 72 possible combinations of simulation parameters. We simulated 1000 data sets with each of those combinations. We held the true regression coefficients of covariates $X_2, \ldots, X_{10}$ constant across simulation scenarios, setting them to $\log 2 = 0.69$ for binary covariates, and to $\log\sqrt{2} = 0.35$ for ordinal covariates. For continuous covariates the effects were chosen such that the log odds ratio between the first and the fifth sextile of the corresponding distribution was 0.69. An intercept β_0 was determined for each simulation scenario such that the desired proportion of events was approximately obtained. Finally, we sampled the binary outcome y_i for subject i, $i = 1, \ldots N$, from a Bernoulli distribution as $y_i \sim \mathrm{Bern}(\pi_i)$ after calculating π_i by $\pi_i = 1/(1 + \exp(-\beta_0 - \beta_1 x_{i1} - \ldots - \beta_K x_{iK}))$. Whenever separation was encountered in the original data set

of size N, we added to these data new observations sampled from the same distribution to generate an increased data set of sample size N_{new} in which separation was removed.

Methods: We analyzed each simulated data set by fitting a logistic regression model (1) and estimating the regression coefficients by

1. ML after removing separation by ISS (ML+ISS),
2. FC applied to the original data and
3. FC after removing separation by ISS (FC+ISS).

For ML we estimated 95% confidence intervals (CI) by the Wald method, and for FC by penalized likelihood profiles [5]. ML and FC estimation was performed using the **logistf** [16] package in R, version 3.5.0. We checked for the presence of separation by the algorithm [17] implemented in **brglm2** [18] package.

Estimands: The true regression coefficient β_1 was the estimand in our study as X_1 was considered the target covariate (e.g., exposure to a risk factor). X_1 was simulated as an unbalanced binary covariate with expected value 0.8.

Performance measures: As for prediction separation might not necessarily be considered as a problem, we focused on estimation and evaluated bias ($\mathrm{E}(\hat{\beta}_1 - \beta_1)$), and mean squared error (MSE; $\mathrm{E}\left(\hat{\beta}_1 - \beta_1\right)^2$) of regression coefficient estimates as well as the probability that a 95% CI excludes $\beta_1 = 0$ (type I error rate or power), the probability that it includes the true value of β_1 (coverage) and the width of the 95% CIs. To compare the performance of ML+ISS and FC we defined the cost-adjusted relative efficiency (CARE) of ML+ISS relative to FC as

$$\mathrm{CARE} = \frac{\mathrm{MSE}(\hat{\beta}_{ML+ISS}) \times \bar{N}_{new}}{\mathrm{MSE}(\hat{\beta}_{FC}) \times N}, \quad (4)$$

where $\bar{N}_{new}$ is the increased sample size needed to remove separation averaged over 1000 simulation runs. This measure weighs the increased efficiency against the costs of increasing the sample size, compared to the efficiency of FC at the original sample size. By its definition, CARE > 1 suggests that FC is a more efficient estimator.

3. Results

This section is divided into two parts: first, we report the results from the simulation study described in *2.2.*; second, we illustrate the problem by two-real life data examples showing results from preliminary and final analyses. The examples are taken from colorectal cancer screening and from ornithology.

3.1. Results of simulation study

Before evaluating the performance of the methods we describe the mean new sample size $\bar{N}_{new}$ that was required to remove separation and its dependence on the prevalence of separation in the original data sets of size N. These results are shown by means of a nested loop plot [19] in Figure 1. Clearly, the mean sample size $\bar{N}_{new}$ was positively correlated with the prevalence of separation. Both, $\bar{N}_{new}$ and the prevalence of separation, were generally higher in scenarios with smaller sample sizes, in scenarios where $\mathrm{E}(Y) = 0.1$, and with larger effects of β_1. Moreover, they were lower for scenarios with two covariates compared to scenarios with five covariates. Interestingly, many scenarios with ten covariates had the fewest separated data sets, indicating that correlations between the covariates are often a more important factor than the number of covariates itself.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2019 doi:10.20944/preprints201910.0321.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 4658; doi:10.3390/ijerph16234658

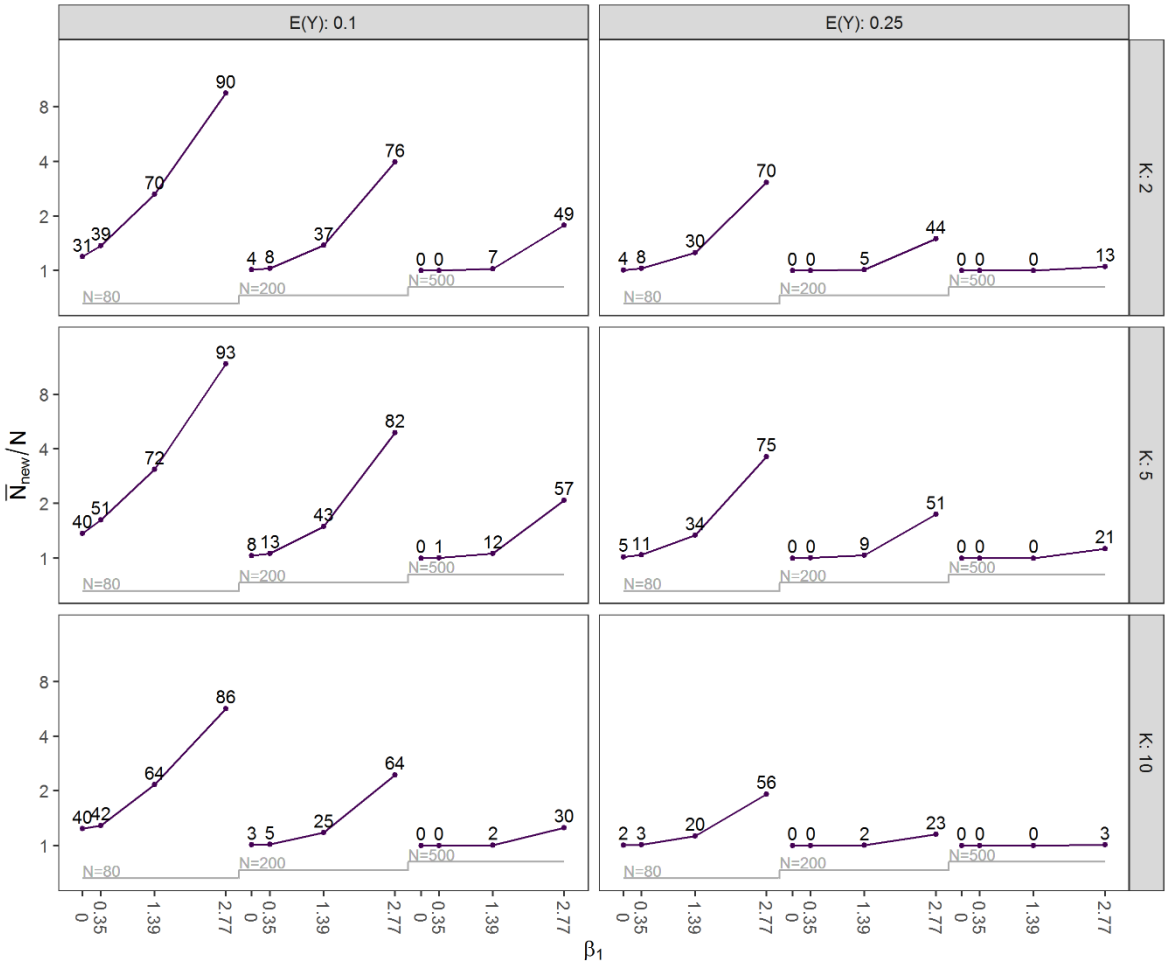

Figure 1. Nested loop plot of $\bar{N}_{\text{new}}/N$ by the expected value of Y, $\text{E}(Y) \in \{0.1, 0.25\}$, the number of covariates $K \in \{2, 5, 10\}$, the value of $\beta_1 \in \{0, 0.35, 1.39, 2.77\}$ and the sample size $N \in \{80, 200, 500\}$ for all simulated scenarios. The numbers indicate the prevalence of separation (%).

Bias of $\hat{\beta}_1$ is shown in Figure 2. For ML+ISS positive bias was generally observed in scenarios where separation occurred in less than a third of simulated data sets. By requiring samples to be non-separated, the ML+ISS strategy appeared to correct the small-sample bias of ML (away from zero), and in scenarios with high prevalence of separation its corrective nature even resulted in some negative bias. Applying FC, a small-sample bias-reduction method, indeed yielded almost unbiased regression coefficients in many scenarios; however, in extreme cases with small sample sizes and higher values of β_1 negative bias was non-negligible and could be severe especially if, in addition, the expected outcome prevalence was only 10%. Generally, bias towards zero increased with higher values of β_1. With the exception of some scenarios with strong effects applying FC to the increased samples (FC+ISS) additionally increased the bias towards zero compared to FC estimation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2019 doi:10.20944/preprints201910.0321.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 4658; doi:10.3390/ijerph16234658

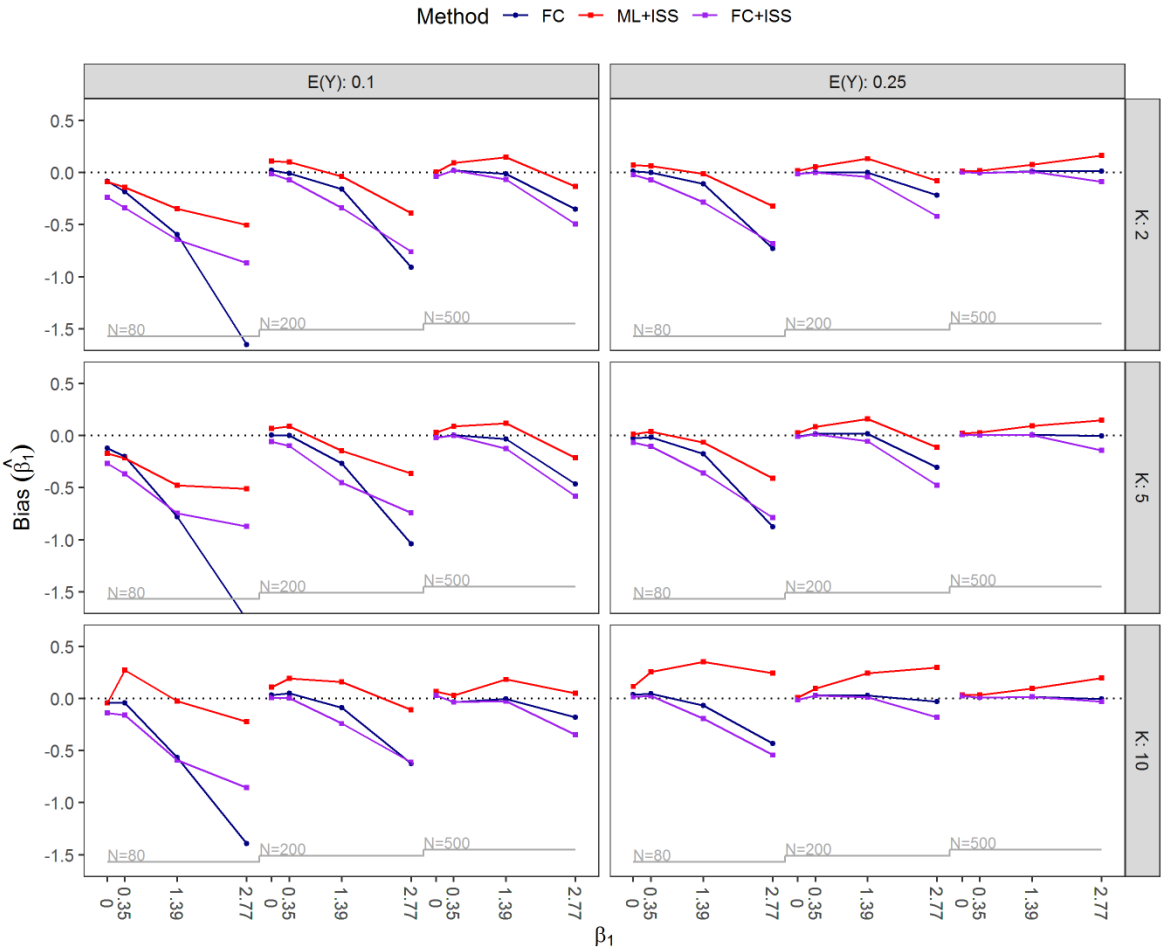

Figure 2. Nested loop plot of bias of $\hat{\beta}_1$ by the expected value of Y, $\mathrm{E}(Y) \in \{0.1, 0.25\}$, the number of covariates $K \in \{2, 5, 10\}$, the value of $\beta_1 \in \{0, 0.35, 1.39, 2.77\}$ and the sample size $N \in \{80, 200, 500\}$ for all simulated scenarios. FC, Firth's correction; ML+ISS, maximum likelihood combined with the increasing sample size approach; FC+ISS, Firth's correction combined with the increasing sample size approach.

Figure 3 presents the MSE of $\hat{\beta}_1$ and CARE of ML+ISS compared to FC. For most scenarios there were no considerable differences in terms of MSE of $\hat{\beta}_1$ between the methods. However, ML+ISS yielded in some cases much larger sample sizes N_{new} with the average sample size $\bar{N}_{new}$ up to 11.8-times higher than the original N and a maximum difference between N_{new} and N of 6424 observations. Therefore, FC was a more efficient estimator in all but four scenarios, and in these four scenarios CARE was very close to one. Interestingly, even with sample sizes of $N = 500$ FC usually performed slightly better than ML+ISS. In extreme situations with highest values of β_1 and small sample sizes of $N = 80$ the MSE by FC was poor in absolute terms and could only slightly be improved by ML+ISS. The improvement came at the cost of much higher sample sizes, such that FC was substantially more efficient (CARE >> 1). FC often achieved reasonably small MSE even with large proportions of separation but it failed considerably in small sample situations ($N = 80$) with strongest effects of β_1 and $\mathrm{E}(Y) = 0.1$, and in all small sample situations with $\mathrm{E}(Y) = 0.1$ and $K = 10$ covariates included in the model. In the latter situations ML+ISS as well yielded very unstable regression coefficient estimates with extremely large MSE. As expected, by increasing the sample size FC+ISS most often reduced the MSE of FC. However, it also occurred that the MSE was slightly increased despite the larger sample size: in some scenarios where bias towards zero was smaller for FC+ISS than FC, the latter yielded lower MSE, and, notably, in a few scenarios FC outperformed FC+ISS with respect to both bias and the MSE.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2019 doi:10.20944/preprints201910.0321.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 4658; doi:10.3390/ijerph16234658

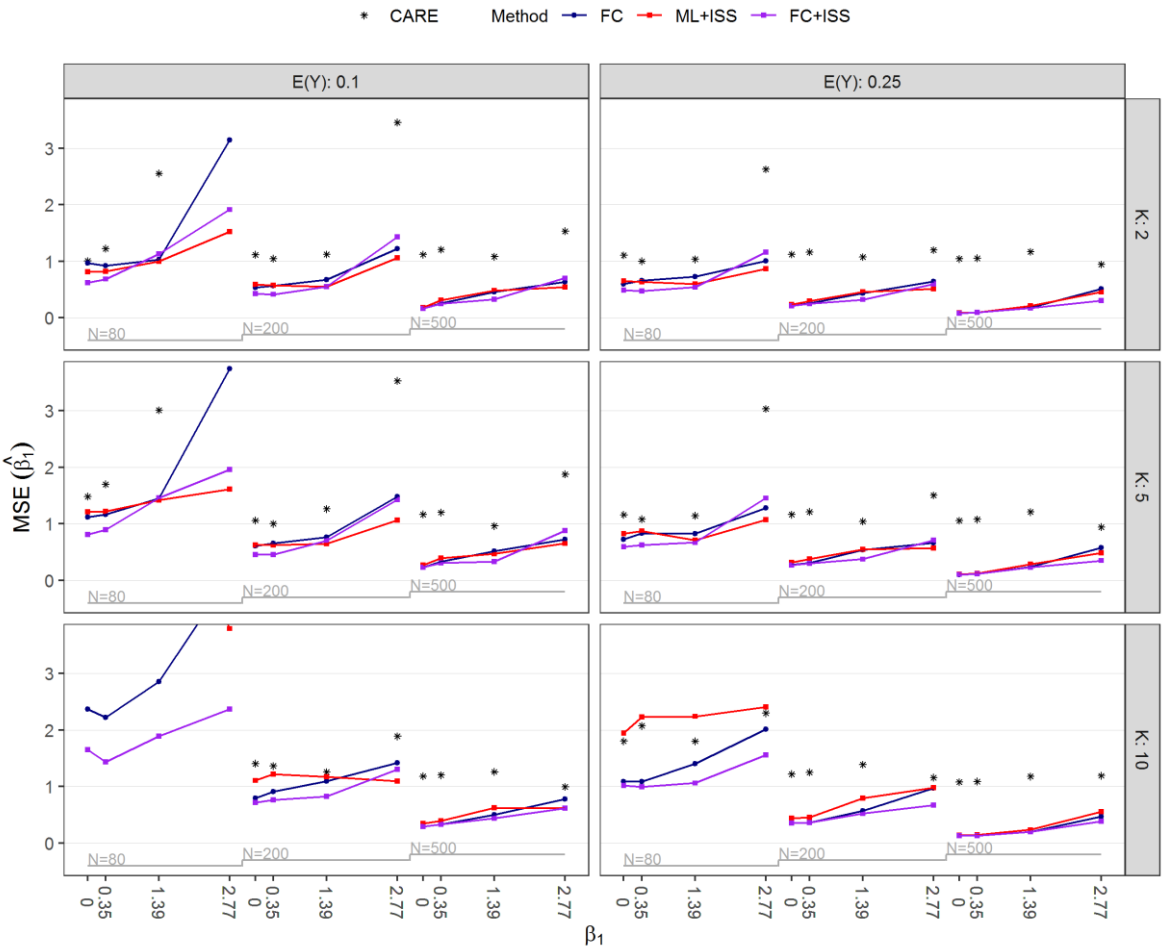

Figure 3. Nested loop plot of the mean squared error (MSE) of $\hat{\beta}_1$ by the expected value of Y, $\mathrm{E}(Y) \in \{0.1, 0.25\}$, the number of covariates $K \in \{2, 5, 10\}$, the value of $\beta_1 \in \{0, 0.35, 1.39, 2.77\}$ and the sample size $N \in \{80, 200, 500\}$ for all simulated scenarios. In addition, CARE as defined in equation (4) is shown by $*$, where $\mathrm{CARE} > 1$ suggests that ML+ISS is a less efficient estimator than FC. The results for ML+ISS for $\mathrm{E}(Y) = 0.1$, $K = 10$ and $N = 80$ are outside of the plot range. The CARE for some scenarios with $\mathrm{E}(Y) = 0.1$ and $N = 80$ also lies outside of the plot range. FC, Firth's correction; ML+ISS, maximum likelihood combined with the increasing sample size approach; FC+ISS, Firth's correction combined with the increasing sample size approach; CARE, cost-adjusted relative efficiency.

Throughout scenarios with $\beta_1 = 0$ and $\mathrm{E}(Y) = 0.25$, the type I error rate for FC was closest to the nominal level of 5% while by ML+ISS and FC+ISS the null hypothesis was on average rejected with slightly lower probability. The mean type I error rate over all these scenarios was 4.8% for FC, and 4.6% for ML+ISS and FL+ISS. In scenarios with $\mathrm{E}(Y) = 0.1$ all three approaches tended to be conservative; the type I error rate was on average 3.6% for FC and FC+ISS and 3.3% for ML+ISS. FC+ISS had the largest power throughout scenarios with non-zero effect of β_1. However, in scenarios with larger sample sizes of $N = 500$, all approaches resulted in similar power, ML+ISS performing slightly worse than FC and FC+ISS whenever $\mathrm{E}(Y) = 0.1$. The power of FC was distinctively lower than the power of FC+ISS or ML+ISS in small sample situations with strong effects; throughout all scenarios with a sample size of $N = 80$ and $\mathrm{E}(Y) = 0.1$ its power stayed close to the α level of 5% irrespective of β_1. The mean power to detect $\beta_1 \neq 0$ over all scenarios with $\mathrm{E}(Y) = 0.25$ was 49.7% for FC, 49.8% for ML+ISS and 52% for FC+ISS, and for $\mathrm{E}(Y) = 0.1$ the mean power was 25.9% for FC, 33.7% for ML+ISS and 38.2% for FC+ISS, respectively. See Figure S1 in Supplementary Materials for detailed results on power and type I error rate for β_1 over all simulated scenarios.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2019 doi:10.20944/preprints201910.0321.v1

Peer-reviewed version available at *Int. J. Environ. Res. Public Health* **2019**, *16*, 4658; doi:10.3390/ijerph16234658

The empirical coverage levels and the width of the 95% CIs for β_1 are reported in Figure 4 and Figure S2, respectively. In scenarios with $E(Y) = 0.25$ the average coverage of the profile penalized likelihood CIs for FC and FC+ISS was close to nominal. For both approaches there was only one out of 36 scenarios where the coverage was outside the 'plausible' interval $0.95 \pm 2.57\sqrt{\frac{0.95 \cdot 0.05}{1000}}$, into which the empirical coverage rate of a method with perfect coverage of 0.95 falls with a probability of 0.99 when the coverage is estimated from the analysis of 1000 simulated data sets. The ML+ISS Wald CIs were in contrast slightly conservative and outside the plausible range in 7 out of 36 scenarios. The mean coverage (and the average width) over all scenarios with $E(Y) = 0.25$ was 95.5% (3.2) for FC, 96% (3) for ML+ISS and 95.5% (2.8) for FC+ISS. The estimates obtained in scenarios with $E(Y) = 0.1$ were far less precise than for $E(Y) = 0.25$, with a mean coverage (and average width) of 96% (4.9) for FC, 96% (13.6 and 3.6 after excluding scenarios with $N = 80$ and $K = 10$) for ML+ISS and 95.4% (3.5) for FC+ISS. Notably, the coverage of the CIs of FC and FC+ISS was out of the plausible range in 9 and the one of ML+ISS in 18 out of 36 scenarios. These CIs for all three methods were generally conservative, however, in two scenarios with $N = 80$ and $\beta_1 = 2.77$ the coverage of the profile penalized likelihood CIs for FC+ISS was below 89%. Throughout all scenarios, the width increased with larger effect sizes of β_1. FC's CIs were on average the widest, followed by ML+ISS; FC+ISS resulted in the narrowest CIs.

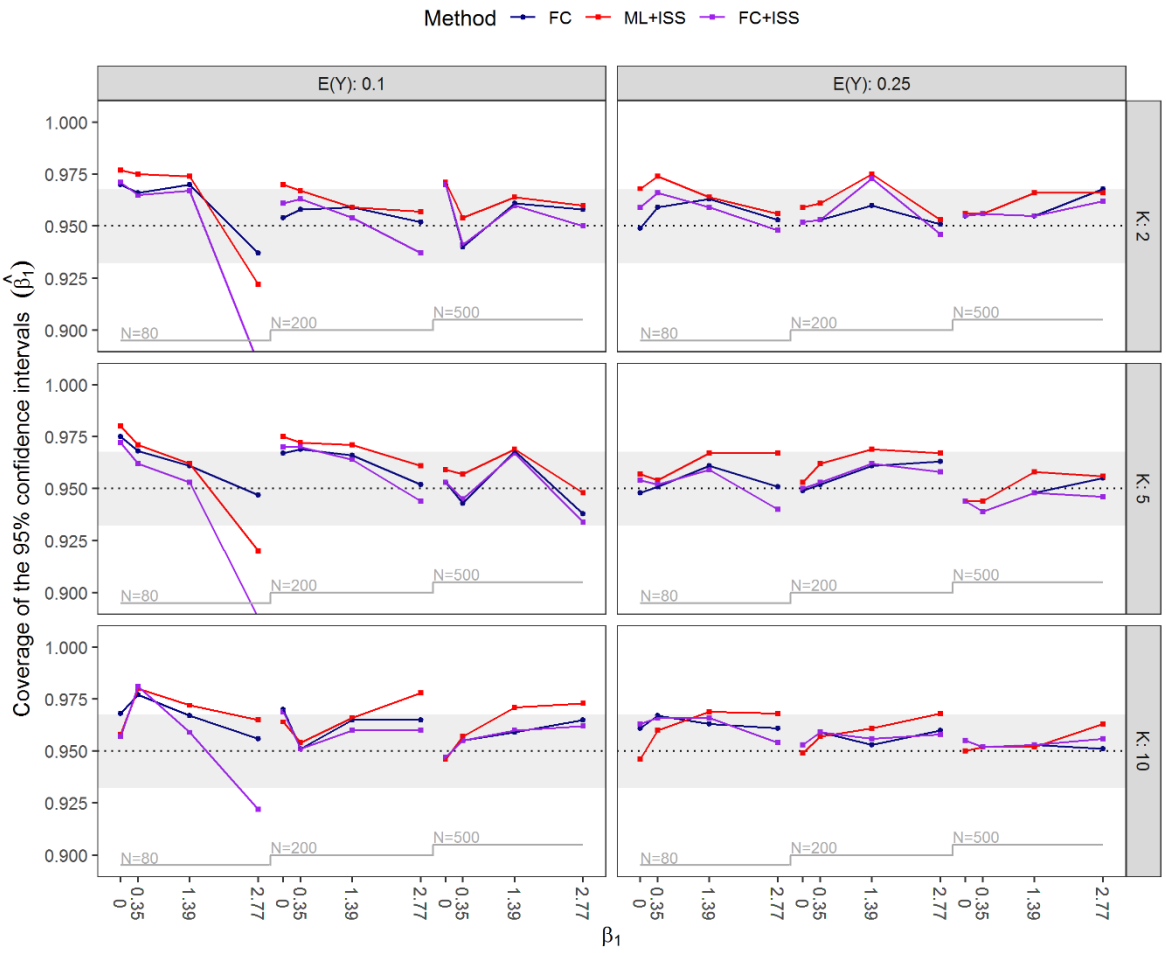

Figure 4. Nested loop plot of coverage of the 95% confidence intervals for $\hat{\beta}_1$ by the expected value of Y, $E(Y) \in \{0.1, 0.25\}$, the number of covariates $K \in \{2, 5, 10\}$, the value of $\beta_1 \in \{0, 0.35, 1.39, 2.77\}$ and the sample size $N \in \{80, 200, 500\}$ for all simulated scenarios. The grey shaded area marks the 'plausible' interval $0.95 \pm 2.57\sqrt{\frac{0.95 \cdot 0.05}{1000}}$, into which the result of a method with perfect coverage falls with a probability of 0.99. FC, Firth's correction; ML+ISS, maximum likelihood combined with the increasing sample size approach; FC+ISS, Firth's correction combined with the increasing sample size approach.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2019 doi:10.20944/preprints201910.0321.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 4658; doi:10.3390/ijerph16234658

3.2. Examples

3.2.1. Bowel preparation study

Our first example is a study comparing four different bowel purgatives (A, B, C, D) and their effect on the quality of colonoscopy. The procedure was called a 'success' if the cecum was intubated by the endoscopist, or 'failure' otherwise (see [20] for more details). The data set used in a preliminary analysis consisted of 4132 patients and suffered from a separation issue as the cecum of patients that used purgative B, which was the smallest of the four groups, was always successfully intubated. The final data set consisted of 5000 patients, and there was now a single patient in group B whose cecum could not be intubated. Table 2 shows the regression coefficients estimated by ML and FC in the preliminary and the final analysis. All the analyses presented here were adjusted for age and sex of patients. In line with our simulation results, the point estimates for purgative B vs. A by FC decreased after increasing the sample size, and the confidence interval of FC in the final data set was narrower than that of ML.

Table 2. Logistic regression coefficient estimates obtained by ML and FC estimation in the preliminary and final analysis of bowel preparation study. All analyses were adjusted for age and sex of the patients.

	Bowel purgative	N	$\widehat{\beta}_{ML}$ **[95% CI]**	$\widehat{\beta}_{FC}$ **[95% CI]**
Preliminary data set ($N = 4132$)	A	2149	reference	
	B	239	not available	1.95 [–0.01, 6.8]
	C	596	0.98 [–0.21, 2.18]	0.85 [–0.13, 2.16]
	D	1148	–0.83 [–1.33, –0.34]	–0.83 [–1.32, –0.34]
Final data set ($N = 5000$)	A	2648	reference	
	B	267	1.4 [–0.59, 3.39]	1.01 [–0.62, 2.64]
	C	799	0.83 [–0.11, 1.76]	0.74 [–0.15, 1.64]
	D	1286	–0.83 [–1.28, –0.39]	–0.83 [–1.27, –0.39]

3.2.2. European passerine birds study

The second example is a study investigating the influence of migration (migratory, non-migratory) and type of diet (granivorous, insectivorous and omnivorous) on the presence of intestinal parasites in the faeces of European passerine birds (see [21] for more details). The data set used in the initial analysis consisted of 366 birds, and no intestinal parasites were present in any of 17 granivorous birds. After the principal investigator was advised to 'bring more data' in order to solve the separation issue, the final data consisted of 385 birds, and intestinal parasites were present in 2 out of 30 granivorous birds. Table 3 shows the regression coefficient estimates, adjusted for migration, obtained by ML and FC estimation in the preliminary and the final analysis. Again, the FC estimates decreased dramatically after increasing the sample size.

Table 3. Logistic regression coefficient estimates obtained by ML and FC estimation in preliminary and final analyses of the European passerine birds study. All analyses were adjusted for migration.

	Diet	N	$\widehat{\beta}_{ML}$ **[95% CI]**	$\widehat{\beta}_{FC}$ **[95% CI]**
Preliminary data set ($N = 366$)	Granivorous	17	reference	
	Insectivorous	274	not available	1.53 [–0.7, 6.43]
	Omnivorous	75	not available	2.17 [–0.02, 7.06]
Final (ISS) data set ($N = 385$)	Granivorous	32	reference	
	Insectivorous	276	0.75 [–0.82, 2.33]	0.57 [–0.73, 2.26]
	Omnivorous	77	1.42 [–0.15, 2.98]	1.24 [–0.05, 2.91]

4. Discussion

In medical research, studies investigating a binary outcome often focus on estimating the effect of an independent variable adjusted for others by logistic regression, and ML is the most commonly used method to estimate such an effect. However, in the analysis of small or sparse data sets ML coefficient estimates are biased away from zero and very unstable, or ML estimation is even impossible. This occurs when covariates can perfectly separate the observations into the groups defined by the levels of the outcome variable as in our examples. (In fact, this may also happen when there is no separation in the data set. Rareness of events together with numerical instability or inaccuracy may make fitted probabilities already undistinguishable from zero or one by software packages.) Researchers confronted with log-odds ratio estimates diverging to $\pm\infty$ may question the plausibility of their analysis results, as it is not assumed that the effect of a variable can be truly 'infinite'. Rather, a nonexisting estimate is the result of an extreme small sample bias and a consequence of 'bad luck in sampling' [6].

Software packages do not always warn a user of the non-convergence of the iterative ML fitting procedure, and an alternative estimation method to plug-in as a remedy of the problem is usually not suggested. Although seldom reported, we are aware that especially in observational studies increasing the sample size is a common practice as it offers the most straightforward and intuitive solution albeit its properties have not been examined thoroughly yet and were perhaps rated overoptimistically. In this paper we investigated the empirical statistical properties of such a strategy as opposed to FC, that has been described as an 'ideal solution to the problem of separation' [5], making it the most widely available penalty in software packages [12]. We compared the performance of both approaches in terms of MSE and bias of regression coefficients, and also incorporated the cost of increased sample size when evaluating CARE of ML+ISS relative to FC applied on the original data set. To better understand the impact of ISS, we additionally included in our study a strategy where first sample size is increased until separation is removed and then FC is applied.

For finite sample sizes the sampling distribution of ML regression coefficients consists of a sub-distribution of finite values and a proportion of diverging estimates that can be explained as a consequence of an extreme small sample bias. The ML+ISS analysis compensates the bias away from zero of ML estimation by the bias towards zero that is induced by sampling until the extreme bias is removed. This can be exemplified by assuming that the original data set and the added observations are analyzed separately. The FC estimate obtained from the model fitted on the added observations has a strong negative bias: in the colonoscopy study, e.g., $\hat{\beta}_B$ for the added observations was equal to $-0.84\,[-2.34, 1.41]$, while the estimate in the original data was $1.95\,[-0.01, 6.8]$. In parallel, FC estimation can be represented by ML estimation on augmented data consisting of the original and pseudo-observations [5,11]. The model fitted on the pseudo observations has $\hat{\beta}_{pseudo} = 0$. Such the ML+ISS strategy indeed results in acceptable biases, but sometimes comes at the cost of highly inflated sample sizes. FC applied to data sets of (much) smaller sample sizes, in contrast, generally yields almost unbiased estimates. While the bias away from zero in ML estimation is a result of poor applicability of large-sample properties in small samples, the ISS strategy actually adds bias towards zero and so the two biases may cancel out. If combined with the unbiased FC strategy, however, ISS leads to a bias towards zero. When the true effect size is very large, FC can be biased towards zero [5], and ISS can amplify this effect as demonstrated in our simulation study.

In terms of cost-corrected efficiency, the ML+ISS strategy is clearly outperformed by FC applied to the original, pre-planned sample size. Therefore, whenever the event rate is low or exposures are rare, a good advice to a researcher planning a study would be to consider FC as the method of analysis. However, in studies where not only effect estimates but also predictions are of importance, e.g., if differences in outcomes attributable to different covariates should be described not only on a relative scale but also on the absolute scale of event probabilities, which is often more relevant, caution is advised – FC leads to predicted probabilities biased towards 0.5. This is a consequence of the unbiasedness of the linear predictor, which naturally cannot translate into unbiasedness of its nonlinear transformation on the probability scale. While pulling predicted probabilities towards 0.5 is not problematic when the outcome levels are approximately balanced, in situations of rare events

this bias becomes apparent. Recently, two methods to overcome this shortcoming – FLIC and FLAC – were proposed by Puhr et al [11], which both yield average predicted probabilities equal to the observed event rate. Therefore, the bias in predicted probabilities is no longer of concern if one of those methods is applied. Alternatively, one can resort to Bayesian methods and use weakly informative priors centered around 0 (so-called shrinkage priors) to cope with separation, e.g. Cauchy priors [22] or log-*F* priors [12]. Other more heuristic corrections based on a redefinition of the outcome variable have also been proposed [13]. Despite the usefulness of these other procedures in some instances, unlike FC, these methods are mostly justified by their empirical behavior, and less by theoretical considerations. Moreover, they are not invariant to linear transformations of the design matrix. For brevity we have not included their evaluation in the present paper as head-to-head comparisons with FC were already performed [11]. In further simulations not included in this report we also investigated the performance of FLAC, log-*F* and the combined strategies FLAC+ISS and log-*F*+ISS. FLAC and log-*F* showed similar patterns of behavior as FC, and can yield MSEs even smaller than FC as previously shown [11]. The addition of ISS to FLAC and log-*F* had the same impact as with FC.

In order to produce valid and informative research results it is desirable, already at the design stage of a study, to not fully rely on approximations based on large-sample properties of ML methods. Moreover, the possibility of separation, or more generally, of effects of sparsity in discrete outcome data should be taken into consideration and a suitable analysis strategy chosen. Although FC can be seen as an advancement of the traditional ML analysis, it cannot solve all problems of data sparsity [23]. Caution is needed when applying FC in extreme situations where the sample sizes are (too) small and the outcomes unbalanced at the same time: it should be taken into account that in such situations FC performs poorly and lacks power to detect even very strong effects. Finally, researchers should not believe that collecting more data could save a study after analysis has failed, but are advised to consult and involve experienced biostatisticians already at the design stage of their studies.

5. Conclusions

By means of a simulation study and the analysis of two real data examples, we compared two strategies for dealing with separation in logistic regression with respect to their empirical performance. We conclude that sampling observations until the problem of separation is removed has adequate performance in terms of precision and inference, but is relatively inefficient compared to Firth's correction applied to a data set of original size.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Type I error rate and power of $\hat{\beta}_1$, Figure S2: Width of the 95% confidence intervals for $\hat{\beta}_1$.

Author Contributions: Conceptualization, G.H. and A.G.; methodology, H.Š., A.G. and G.H.; software, H.Š.; validation, H.Š.; formal analysis, H.Š.; investigation, H.Š.; resources, G.H.; data curation, H.Š..; writing—original draft preparation, H.Š.; writing—review and editing, A.G. and G.H.; visualization, H.Š.; supervision, G.H. and A.G.; project administration, G.H.; funding acquisition, G.H.

Funding: This research was funded by the Austrian Science Fund (FWF), grant number I 2276.

Acknowledgments: We thank Michael Kammer for programming the nested loop plot function which will be publicly available soon on the Comprehensive R Archive Network (CRAN). The bowel preparation study data were provided by Elisabeth Waldmann and Monika Ferlitsch from the Medical University of Vienna, and the European passerine birds study data by Petra Bandelj and Rok Blagus from the University of Ljubljana.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cox, D.R.; Snell, E.J. *Analysis of binary data*, 2rd ed.; Chapman and Hall/CRC: London, UK, 1989.
2. Kosmidis, I. Bias in parametric estimation: reduction and useful side-effects. *WIREs Comput Stat* 2014, *6*(3), 185–196. doi:10.1002/wics.1296
3. King, G.; Zeng, L. Logistic regression in rare events data. *Political Anal.* 2001, *9*(2), 137–163.

4. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984, *71*(1), 1–10.
5. Heinze, G; Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* 2002, *21*(16), 2409–2419. doi:10.1002/sim.1047
6. Mansournia, M.A.; Geroldinger, A.; Greenland, S.; Heinze, G. Separation in logistic regression: Causes, consequences, and control. *Am. J. Epidemiol.* 2018, *187*(4), 864–870. doi:10.1093/aje/kwx299
7. Courvoisier, D.S.; Combescure, C.; Agoritsas, T.; Gayet-Ageron, A.; Perneger, T.V. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol.* 2011, *64*(9), 993–1000. doi:10.1016/j.jclinepi.2010.11.012
8. van Smeden, M; de Groot, J.A.H.; Moons, K.G.M.; Collins, G.S.; Altman, D.G.; Eijkemans, M.J.C.; Reitsma, J.B. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med. Res. Methodol.* 2016, *16*(1), 163. doi:10.1186/s12874-016-0267-3
9. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993, *80*(1), 27–38. doi:10.1093/biomet/80.1.27
10. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer New York Inc.: New York, USA, 2001; pp. 119–128.
11. Puhr, R.; Heinze, G.; Nold, M.; Lusa, L.; Geroldinger, A. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat. Med.* 2017, *36*(14), 2302–2317. doi:10.1002/sim.7273.sim.7273
12. Greenland, S.; Mansournia, M.A. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat. Med.* 2015, *34*(23), 3133–3143. doi:10.1002/sim.6537
13. Rousseeuw, P.J.; Christmann, A. Robustness against separation and outliers in logistic regression. *Comput. Stat. Data Anal.* 2003, *43*(3), 315–332. doi:10.1016/s0167-9473(02)00304-3
14. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Stat. Med.* 2019, *38*(11), 2074–2102. doi:10.1002/sim.8086
15. Binder, H.; Sauerbrei, W.; Royston, P. Multivariable model-building with continuous covariates: 1. Performance measures and simulation design. Technical Report FDM-Preprint 105, University of Freiburg Germany, 2011.
16. Heinze, G.; Ploner, M.; Dunkler, D.; Southworth, H. Logistf: Firth's Bias Reduced Logistic Regression. 2014. R package version 1.22.
17. Konis K. safeBinaryRegression: Safe Binary Regression. 2013. R package version 0.1-3.
18. Kosmidis, I.: Brglm2: Bias Reduction in Generalized Linear Models. 2018. R package version 0.1.8.
19. Rücker, G.; Schwarzer, G. Presenting simulation results in a nested loop plot. *BMC Med. Res. Methodol.* 2014, *14*, 129. doi:10.1186/1471-2288-14-129
20. Waldmann, E.; Penz, D.; Majcher, B.; Zagata, J.; Šinkovec, H.; Heinze, G.; Dokladanska, A.; Szymanska, A.; Trauner, M.; Ferlitsch, A.; Ferlitsch, M. Impact of high-volume, intermediate-volume and low-volume bowel preparation on colonoscopy quality and patient satisfaction: An observational study. *United European Gastroenterol J.* 2019, *7*(1), 114–124. doi:10.1177/2050640618809842
21. Bandelj, P.; Blagus, R.; Trilar, T.; Vengust, M.; Vergles Rataj, A. Influence of phylogeny, migration and type of diet on the presence of intestinal parasites in the faeces of European passerine birds (Passeriformes). *Wildlife Biol.* 2015, *21*(4), 227–233. doi:10.2981/wlb.00044
22. Gelman, A.; Jakulin, A.; Pittau, M.G.; Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2008; *2*(4), pp. 1360-1383. doi:10.1214/08-AOAS191
23. Greenland S.; Mansournia M.A.; Altman, D.G. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016; 353:i1981. doi:10.1136/bmj.i1981