

Solutions to problems of nonexistence of parameter estimates and sparse data bias in Poisson regression

Statistical Methods in Medical Research

2022, Vol. 31(2) 253–266

© The Author(s) 2021



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802211065405

journals.sagepub.com/home/smm

Ashwini Joshi¹, Angelika Geroldinger² , Lena Jiricka²,
Pralay Senchaudhuri³, Christopher Corcoran⁴,
and Georg Heinze² 

Abstract

Poisson regression can be challenging with sparse data, in particular with certain data constellations where maximum likelihood estimates of regression coefficients do not exist. This paper provides a comprehensive evaluation of methods that give finite regression coefficients when maximum likelihood estimates do not exist, including Firth's general approach to bias reduction, exact conditional Poisson regression, and a Bayesian estimator using weakly informative priors that can be obtained via data augmentation. Furthermore, we include in our evaluation a new proposal for a modification of Firth's approach, improving its performance for predictions without compromising its attractive bias-correcting properties for regression coefficients. We illustrate the issue of the nonexistence of maximum likelihood estimates with a dataset arising from the recent outbreak of COVID-19 and an example from implant dentistry. All methods are evaluated in a comprehensive simulation study under a variety of realistic scenarios, evaluating their performance for prediction and estimation. To conclude, while exact conditional Poisson regression may be confined to small data sets only, both the modification of Firth's approach and the Bayesian estimator are universally applicable solutions with attractive properties for prediction and estimation. While the Bayesian method needs specification of prior variances for the regression coefficients, the modified Firth approach does not require any user input.

Keywords

Count data, Firth's penalization, generalized linear models, penalized regression, Poisson regression, separation

1 Introduction

Poisson regression is widely used to model the distribution of count variables as functions of predictive covariates. This approach provides particular utility when accommodating differential follow-up times of study subjects,¹ as well as the modelling of 'non-events' or excess zeroes through so-called zero-inflated models.² As with other generalized linear models, such as logistic regression, Poisson regression can be especially challenging in the presence of rare events, making it more likely that particular covariate patterns in a given dataset result in the nonexistence of maximum likelihood (ML) estimates; for example, if no events are observed for one of two groups represented by a binary covariate. A necessary and sufficient condition for the nonexistence of ML estimates has been identified by Correia et al.³ This problem of 'separation' has been

¹Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

²Center for Medical Statistics, Informatics and Intelligent Systems, Section for Clinical Biometrics, Medical University of Vienna, Vienna, Austria

³Cytel Inc., Cambridge, MA, USA

⁴Jon M. Huntsman School of Business, Department for Data Analytics and Information Systems, Utah State University, Logan, UT, USA

Corresponding author:

Georg Heinze, Center for Medical Statistics, Informatics and Intelligent Systems, Section for Clinical Biometrics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.

Email: georg.heinze@meduniwien.ac.at

studied extensively for logistic regression,^{4–7} but comparatively little is known about how various approaches perform when separation arises in Poisson modelling, such as the bias reduction method of Firth^{8,9} and (in special cases) conditional exact Poisson (EP) regression.¹⁰ Bayesian alternatives, including the application of weakly informative priors for the distribution of the log relative risk, have likewise not been evaluated with regard to separation. Given the common use of Poisson regression, particularly for rare events and in other settings that are prone to separation, more definitive empirical studies are needed to assess the comparative performance of these modelling options. In particular, we need to better understand how various modelling choices impact the estimation of regression coefficients to ensure that analysts are confident in their predictions.

In this paper, we provide a more comprehensive evaluation of methods that provide finite estimates of regression coefficients when separation arises. We propose a modification of Firth's approach to improving its performance for predictions, without compromising its attractive bias-correcting properties for regression coefficients. Furthermore, we include the conditional median unbiased estimator¹¹ as implemented in the software package LOGXACT¹⁰ or in SAS's PROC GENMOD and a Bayesian estimator using weakly informative priors in our evaluation. To provide context, we first illustrate the issue of separation for rare event data with a dataset arising from the recent outbreak of COVID-19. We subsequently provide a brief review of Poisson regression and describe proposed solutions to the problem of separation. We then summarize the results of a comprehensive simulation study to compare these options under a variety of realistic scenarios. We further illustrate their relative performance with an additional example.

2 Motivating example

During the outbreak of the coronavirus (COVID-19) in winter and spring 2020, employees of supermarkets, nursing homes, and hospitals were considered key professionals as they are in contact with many clients each day even under the lockdowns of public life that were imposed by many countries in that time. Many of their clients such as older adults or persons with chronic diseases were considered to be at high risk of a severe course of the disease if they got infected. At some point, the question arose if more stringent control of virus spread should be imposed on one of these groups, in particular, considering the unknown probability of asymptomatic infections. Therefore, on 28 March 2020, and 30 March 2020, a representative series of 1161 tests for the presence of an infection with the novel coronavirus (COVID-19) was performed in Austria among randomly selected asymptomatic employees of supermarkets, nursing homes, and hospitals.¹²

The research question behind this series was how the risk of infection differs between employees of supermarkets, nursing homes, and hospitals. The question could be answered by Poisson regression to compute risk ratios, for example, comparing nursing homes and hospitals to supermarkets.

Among the 1161 persons tested, only six tested positive for COVID-19. Three of them were working in hospitals and three in nursing homes (Table 1).

Unfortunately, neither SAS/PROC GENMOD nor R/glm was able to provide risk ratios based on a Poisson regression analysis because ML analysis failed as there was a category with no events. However, the conditional median unbiased estimates for the two interesting risk ratios and associated exact¹³ 95% confidence intervals (CIs) were 3.05 (0.46, ∞) for nursing homes versus supermarkets and 3.71 (0.56, ∞) for hospitals versus supermarkets. Using Firth's bias reduction approach,⁸ which in the absence of covariates can be obtained by adding 0.5 events to each of the three observations, the analysis resulted in risk ratios (95% profile penalized likelihood (PPL) CI) of 5.55 (0.54, 746.13) and 6.75 (0.65, 907.60), respectively. Therefore, while it appears that employees of nursing homes or hospitals are at considerably higher risk of spreading the disease compared to supermarket employees, this claim is not fully supported by the study but still could be found in local media. Note that estimates of risk accompanied by 95% CI in each group, which could be obtained much easier than risk ratios, give a good summary of the data but do not answer the question of whether supermarket employees or health care workers are at higher risk of infection.

Table 1. Austrian COVID-19 test data.

Type of employment	Positive	Tested
Supermarket	0	352
Nursing home	3	444
Hospital	3	365

3 Methods

In this section, we will review the Poisson regression model with a special focus on nonexistence of ML estimates. We will further review Firth's correction in the context of the Poisson regression model and we will propose a modification that gives unbiased predictions. Finally, we will present exact conditional analysis and Bayesian estimation with weakly informative priors.

3.1 The Poisson model

The Poisson model assumes that the counts of events in a study, Y , follow a Poisson distribution with parameter μ : $Y \sim \text{Poisson}(\mu)$, where the logarithm of μ is modelled by a linear combination of covariates: $\log(\mu) = X\beta + Z$. Here, X describes a $n \times (k + 1)$ matrix of covariates, with n and k denoting the number of observations and covariates, respectively. By convention, X_0 , the first column of X , consists of 1s only to enable the estimation of an intercept. Z denotes an offset variable, for example, the logarithm of follow-up time or the number of people tested in total. The parameter μ is interpreted as incidence per unit of follow-up time, β_0 is the intercept and β_j , $j = 1, \dots, k$, is the log incidence rate ratio (IRR) between two individuals differing in X_j by one unit.

The log-likelihood of the Poisson model is given by $\ell(\beta) = \sum_{i=1}^n [-\exp(x_i\beta + z_i) + (x_i\beta + z_i)y_i - \log(y_i!)]$. Estimates of β_j , $j = 0, \dots, k$, can be obtained by ML estimation, solving the score equations $U_j(\beta) = \sum_{i=1}^n x_{ij}[y_i - \exp(x_i\beta + z_i)] = 0$ for $j = 0, \dots, k$, where x_{ij} is the observed value of covariate j for subject i , x_i is the row vector of covariate values for subject i , and z_i is the value of the offset for that subject.

3.2 Conditions for nonexistence of ML estimates in Poisson regression and consequences

In the coronavirus testing study, no infections were detected among the 352 supermarket workers, while among the 365 hospital employees, three were infected. The risk ratio would be computed as $(3 \text{ of } 365)/(0 \text{ of } 352)$, which is not defined because of the division by zero. Similarly, there is no finite maximizer β of the corresponding Poisson likelihood. Correia et al.³ showed that in Poisson regression the ML estimate does not exist if and only if there is a non-zero $(k + 1)$ -dimensional vector γ^* such that $x_i\gamma^* = 0$ for i with $y_i > 0$ and $x_i\gamma^* \leq 0$ for i with $y_i = 0$, see Appendix 1 for a replication of their proof in the special case of Poisson regression. If such a linear combination exists, we say that the data are 'separated'. From a geometrical point of view, the data are separated if and only if there exists a hyperplane such that all observations with $y_i > 0$ lie on the plane and all observations with $y_i = 0$ lie on one side of the plane or also on the plane. For the corona virus testing study, multiplying the dummy variable for supermarket employees by -1 represents a linear combination satisfying the condition given by Correia et al. More generally, we observe that the ML estimate would exist if and only if for each category (supermarket, nursing home, and hospital) at least one person had been tested positive. The existence of the ML estimate only depends on the number of events (people tested positive), but not on the number of people tested in total (Table 1, last column). This observation highlights the difference to the concept of separation in logistic regression (Albert and Anderson, 1984): for the corona virus testing study, ML estimates in logistic regression would exist, if and only if at least one person had been tested positive and at least one person had been tested negative for each category. In the following, we will always use the term separation in the context of Poisson regression as defined above.

As in logistic regression, adding covariates will not remove the nonexistence in Poisson regression. While numerical ML algorithms may declare convergence when the log-likelihood cannot be improved by a further iteration,¹⁴ 'a spurious solution is characterized by a "perfect" fit for the observations with $y_i = 0$ ', that is, $\exp(x_i\hat{\beta}) \rightarrow 0$ for all $y_i = 0$.

3.3 Firth's likelihood penalization applied to the Poisson model

Generally, Firth⁸ suggested adding a penalty term to the log-likelihood of exponential family models that resembles the Jeffreys invariant prior such that the penalized log-likelihood becomes:

$$\ell^*(\beta) = \ell(\beta) + (1/2) \log |I(\beta)| \quad (1)$$

where $|I(\beta)|$ denotes the determinant of the Fisher information matrix. The modification is motivated by elimination of bias of order $O(n^{-1})$ in the ML estimates of β , and various empirical studies have proven the bias-preventive properties of Firth's correction. It also prevents the nonexistence of ML estimates of β and has become the default solution to solve this problem for logistic and Cox regression.^{5,15}

Already in Firth's seminal paper,⁸ an example with Poisson regression was included. However, Firth's likelihood penalization (FL) for the Poisson model has not been studied any further. FL estimates maximizing (1) can be obtained through

iteratively solving the modified score equations:

$$U_j^*(\beta) = \sum_{i=1}^n x_{ij}(y_i + h_i/2 - \exp(x_i\beta + z_i)), \quad j = 0, \dots, k \quad (2)$$

where h_i are the diagonals of the ‘hat’ matrix $H = XW^{1/2}(X'WX)^{-1}XW^{1/2}$, with $W = \text{diag}(\exp(x_i\beta + z_i))$. Generally, $\text{tr}(H) = k + 1$ and $h_i > 0$ for all i . Equation (2) can be written in a form revealing that FL estimates can be obtained by ML estimation on an augmented data set that consists of the original data in which observed outcomes y_i were augmented by $h_i/2$. While we expect that Firth’s correction will correct some of the small-sample bias of the ML estimates also in Poisson regression, it will supply predictions for the counts $\hat{\mu}_i = \exp(x_i\hat{\beta} + z_i)$, which are slightly too high because the modified score equation $U_0^*(\beta)$ implies $\sum_{i=1}^n (y_i + h_i/2) = \sum_{i=1}^n \hat{\mu}_i$.

3.4 A modified Firth correction to achieve unbiased prediction

Similarly, in logistic regression with rare events, Firth’s penalization provides predicted probabilities that are on average higher than the observed event rate. To solve this problem, Puhr et al.⁷ suggested two methods, ‘FL with intercept correction’ (FLIC) and ‘FL with added covariate’ (FLAC). Here we explore the performance of these methods in the setting of Poisson regression.

FLIC consists of first obtaining the FL solution and then to correct the intercept parameter by adding a constant δ such that $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i$. This is achieved by using the linear predictors from the FL solution, $\hat{\eta}_i = \sum_{j=0}^k x_{ij}\hat{\beta}_j + z_i$ as offsets in a second logistic regression that only estimates an intercept δ^{FLIC} by ML. The regression coefficients $\hat{\beta}_j$, $j = 1, \dots, k$, are left unchanged by FLIC, and the new intercept is given by $\hat{\beta}_0^{\text{FLIC}} = \hat{\beta}_0 + \delta^{\text{FLIC}}$.

FLAC starts by estimating the FL solution as well, but this is only done to compute the values of h_i . Subsequently, an augmented data set is constructed, adding a pseudo observation to each original observation with event count $h_i/2$, and defining an ‘added covariate’ G such that it distinguishes original from pseudo-observations by assuming values of 0 and 1 for them, respectively. The augmented data set with the added covariate is then subjected to ML Poisson regression where an additional coefficient γ^{FLAC} corresponding to G is estimated. For predictions, G is assigned a value of 0. As for FLIC, we have $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i$ (see Appendix 1).

While the two approaches generally give different results for logistic regression, it is remarkable that not only the FLIC estimates, but also FLAC estimates of β_1, \dots, β_k coincide with those obtained by FL for Poisson regression, and that $\hat{\beta}_0^{\text{FLIC}} = \hat{\beta}_0^{\text{FLAC}}$ (see Appendix 1).

Because FLIC and FLAC do not modify the FL regression coefficients, CI for the regression coefficients can be obtained out of the FL model, and in the case of data sparsity should preferably be computed by the PPL method.^{5,6} Here, a $(1 - \alpha) \times 100$ per cent CI for a parameter β_j is defined as the set of values β_j^* for which

$$2 \left[\ell^*(\hat{\beta}^{\text{FL}}) - \max_{\beta_0, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k} \ell^*((\beta_0, \dots, \beta_{j-1}, \beta_j^*, \beta_{j+1}, \dots, \beta_k)) \right] \leq \chi_1^2(1 - \alpha)$$

where $\max_{\beta_0, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k} \ell^*((\beta_0, \dots, \beta_{j-1}, \beta_j^*, \beta_{j+1}, \dots, \beta_k))$ is the PPL, that is, the penalized log-likelihood fixed at $\beta_j = \beta_j^*$ and maximized over $\beta_{j'}$; $j' \in \{0, \dots, k\} \setminus j$; and $\chi_1^2(1 - \alpha)$ is the $(1 - \alpha)$ quantile of the chi-squared distribution with one degree of freedom. PPL CI can become asymmetric and in such a case indicate the inadequacy of the Wald method for CI estimation.

We have written a SAS macro FLACPOISSON that performs the Firth correction and the FLAC modification based on iterated data augmentation using repeated calls to PROC GENMOD (<https://github.com/georgheinze/flicflac>). For simplicity, we approximate the PPL CI in FLACPOISSON by evaluating the profile likelihood (PL) of the augmented data fixing the event counts of the pseudo data at the values $h_i/2$ obtained at the FL solution. In LogXact,¹⁰ PPL for FL is available where the hat diagonals are iteratively updated when computing the confidence limits. We will illustrate the difference between the two methods in an example of the occurrence of complications in implant dentistry.

3.5 Alternative methods do deal with separation

3.5.1 Exact conditional Poisson regression with median unbiased estimation

In exact conditional Poisson regression, inference is based on the exact conditional likelihood of a parameter β_j conditional on the observed sufficient statistics $t_{j'}$ of all other parameters $\beta_{j'}$; $j' \in \{1, \dots, k\} \setminus j$, where a sufficient statistic $t_{j'}$ is given by $t_{j'} = \sum_{i=1}^n x_{ij'} y_i$.¹³ The maximum conditional likelihood estimate (MCLE) is the value of β_j that maximizes its exact

conditional likelihood. In case a finite MCLE does not exist, it can be replaced by a median unbiased estimate (MUE).¹⁶ If the exact distribution is degenerate, neither MCLE nor MUE can be computed.

Exact conditional Poisson regression does not provide an estimate of the intercept and thus cannot be used for prediction. The implementations in SAS/PROC GENMOD, in Cytel studio, and in Cytel’s PROC LOGXACT add-on for SAS¹⁰ allow for computation of exact and mid-p CIs and more details can be found in the software documentation.¹⁷ In the remainder, we will refer to this method as EP regression.

3.5.2 Bayesian data augmentation

Bayesian estimation with properly specified priors also solves the separation issue. To overcome problems of computing time and diagnostics in Bayesian analysis with Markov chain Monte Carlo algorithms, Sullivan and Greenland¹⁸ illustrate the use of data augmentation to specify normal priors, including an example for Poisson regression. The spread of the normal prior for a regression coefficient is determined by the width of a prior interval for the associated IRR. For example, if a 95% prior interval for the IRR of (1/1000, 1000) is specified, then based on a normal distribution the prior standard error is $\log(1000) / 1.96 = 3.52$, suggesting a prior variance ν of $3.52^2 = 12.39$. The prior distribution can be specified by adding pseudo-observations, one for each regression coefficient, with a value of $1/S$ for the associated covariate and 0 for all other covariates, where S denotes an approximation constant (higher values giving a better approximation). No pseudo-observations are specified for the intercept. The event count of each pseudo observation is set to $y = S^2 / \nu$, and the corresponding offset to $z = \log(y)$. While Sullivan and Greenland¹⁸ chose $S = 25$, we used $S = 10,000$ to obtain from the pseudo-observations 95% PL CI for the regression coefficients that were symmetric up to the third decimal place. After data augmentation, maximum posterior estimates can be computed by applying ML methods to the augmented data, and intervals for them from PL. Bayesian data augmentation (BDA) results in unbiased predicted counts in the sense that $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i$.

4 A simulation study

4.1 Methods

The methodology of our simulation study is described as recommended by Morris et al.¹⁹

Aims: We intended to explore and compare the performance of different estimation methods for Poisson regression with sparse data.

Data-generating mechanisms: To capture a plausible context, we considered a data-generating scheme as described by Binder et al.²⁰ and Zöller et al.²¹ First, we generated data sets of n observations on 10 covariates X_1, \dots, X_{10} of different, prespecified distributions that were obtained by applying certain transformations to normal random variables Z_1, \dots, Z_{10} sampled from a standard multivariate normal distribution with correlation matrix Σ (Table 2). In this way, we generated four binary covariates X_1, \dots, X_4 , two ordinal covariates X_5 and X_6 with three levels, and four continuous covariates X_7, \dots, X_{10} . The correlation structure of the variables Z_1, \dots, Z_{10} is transferred to the variables X_1, \dots, X_{10} in a somewhat attenuated way.

To avoid extreme values for the two log-normally distributed covariates X_8 and X_9 , we used truncated normal distributions (with truncation at the 99th percentile) to generate them. We generated the data sets using the R package simdata.²²

Table 2. Covariates generated in the simulation study. $I(x)$ is the indicator function that equals 1 if the argument x is true, and 0 otherwise. $[x]$ indicates that the non-integer part of the argument x is eliminated.

z_j	Correlation of z_j	Type	x_j	$E(x_j)$
z_1	$z_2(0.5), z_7(0.5)$	Binary	$x_1 = I(z_1 > 1.28)$	0.1
z_2	$z_1(0.5)$	Binary	$x_2 = I(z_2 > 0.35)$	0.36
z_3	$z_4(-0.5), z_5(-0.3)$	Binary	$x_3 = I(z_3 > 0)$	0.5
z_4	$z_3(-0.5), z_5(0.5), z_7(0.3), z_8(0.5), z_9(0.3)$	Binary	$x_4 = I(z_4 > 0)$	0.5
z_5	$z_3(-0.3), z_4(0.5), z_8(0.3), z_9(0.3)$	Ordinal	$x_5 = I(z_5 \geq -1.2) + I(z_5 \geq 0.75)$	1.11
z_6	$z_7(-0.3), z_8(0.3)$	Ordinal	$x_6 = I(z_6 \geq 0.5) + I(z_6 \geq 1.5)$	0.37
z_7	$z_1(0.5), z_4(0.3), z_6(-0.3)$	Continuous	$x_7 = [10 \cdot z_7 + 55]$	54.5
z_8	$z_4(0.5), z_5(0.3), z_6(0.3), z_9(0.5)$	Continuous	$x_8 = [\max(0, 100 \cdot \exp(z_8) - 20)]$	131.1
z_9	$z_4(0.3), z_5(0.3), z_8(0.5)$	Continuous	$x_9 = [\max(0, 80 \cdot \exp(z_9) - 20)]$	1.77
z_{10}		Continuous	$x_{10} = [10 \cdot z_{10} + 120]$	119.5

We considered a full factorial design, varying the number of covariates, $k \in \{2, 5, 10\}$, the events per variable (EPV) ratio, $EPV \in \{3, 5, 10\}$, and the true regression coefficient (\log IRR) of X_1 , $\beta_1 \in \{-\log(16), -\log(8), -\log(4), -\log(2), 0, \log(2), \log(4), \log(8), \log(16)\}$. We kept all other β_j fixed with $\beta_2, \beta_4 = 0.69$; $\beta_3 = -0.69$; $\beta_5 = 0.35$; $\beta_6 = -0.35$; $\beta_7, \beta_9 = 0.69 / \text{ISR}$; $\beta_8, \beta_{10} = -0.69 / \text{ISR}$, where ISR was the intersextile range (difference between fifth and first sextile) of the corresponding continuous covariate. The intercept β_0 was chosen such that the marginal event incidence was ~ 0.1 . We simulated a rate multiplier ψ following our example on the occurrence of complications in implant dentistry (see below) by sampling from a zero-truncated Poisson distribution (restricted to numbers greater than 0) with mean 1.6. The outcome (number of events) y_i was then drawn from a Poisson distribution with parameter $\mu_i = \exp(\eta_i) = \exp(\beta_0 + x_{i1}\beta_1 + \dots + x_{ik}\beta_k)\psi_i$. The sample size was determined by fixing the expected EPV ratio at desired values typical for sparse epidemiological data sets (3, 5, or 10). This resulted in 81 possible combinations of simulation parameters. We simulated 10,000 data sets with each of those combinations.

Methods: We analysed each simulated data set by fitting a Poisson regression model including $\log \psi_i$ as an offset variable and estimating the regression coefficients by maximizing the likelihood (ML), using BDA based on prior intervals for the IRR of (1/1000, 1000) for binary and ordinal covariates and of (1/100, 100) for continuous covariates, using FL, and using FLAC. We also included EP regression in simulated scenarios where it was computationally feasible, that is, with $k \leq 5$ and $n \leq 250$.

We estimated 95% CI for regression coefficients by the Wald method for ML using PROC GENMOD, and by likelihood profiles applied to the augmented data for BDA, FL, and FLAC (FLACPOISSON macro). In the case of separation, the values for ML are those reported by PROC GENMOD at the last iteration.

Exact point and interval estimates and mid-p corrected CI were obtained by PROC LOGXACT.¹⁰

Estimands: The estimands in this study were the expected event counts μ_i and the regression coefficient β_1 . We also evaluated the frequency of nonexistence of ML estimates (separation).

Performance measures: For point estimates of β and predictions, we evaluated bias and root mean squared error (RMSE) $\times \sqrt{n}$. For predictions, the mean squared prediction error was obtained as $n^{-1} \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$ in each data set and then averaged over all simulated data sets in a scenario. The root of the averaged mean squared prediction error (root mean squared prediction error (RMSPE)) times \sqrt{n} is reported. For CI of β , we evaluated left/right-tailed one-sided coverage rates (nominal levels 0.975) and power (probability to exclude 0). We summarized the simulation results graphically using nested loop plots.^{23,24}

4.2 Results

4.2.1 Incidence of separation

The incidence of separation was generally higher in scenarios with smaller sample sizes and with larger negative values of β_1 , see Figure S1. The latter phenomenon is a consequence from the imbalance of X_1 , where with negative β_1 events were less likely in the less frequent group ($X_1 = 1$), and data sets with no events when $X_1 = 1$ occurred more frequently. Among the scenarios with a given EPV ratio and a given value of β_1 , scenarios with 10 covariates often had the fewest separated data sets. This might seem counterintuitive since, for a fixed non-separated data set, omitting covariates can never induce separation. However, in our simulation study, scenarios with 2, 5, and 10 covariates do not only differ in the number of covariates but also in the type of covariates, the magnitude of the intercept, and the sample size.

We only included EP in the comparison of methods for simulation scenarios with $n \leq 250$ and $k \leq 5$. For larger data sets, its application was computationally not feasible or not possible because with continuous covariates (when $k = 10$) degenerate distributions of sufficient statistics were encountered for which no inference is possible. The MCLE in EP did not exist and had to be replaced by the MUE for almost the same data sets where ML estimation failed, see Figure S1. Only with large positive values of β_1 and small-sample sizes, there were considerably more data sets where the MCLE had to be replaced by the MUE than separated data sets. Finally, there were a few datasets where neither the MCLE nor the MUE existed (at most 0.7% of data sets, as observed for the scenarios with $n = 60$, $k = 2$ and $\beta_1 = -\log(16)$ or $\beta_1 = -\log(4)$).

4.2.2 Predictions

The description of the prediction performance is restricted to the methods ML, FL, FLAC, and BDA, since XL does not allow for predictions. Across all methods, predictions were more accurate in terms of RMSPE(μ) for simulation scenarios with fewer variables or with higher EPV ratio. Throughout all evaluated scenarios, FLAC and BDA yielded the most accurate predictions, followed by ML and FL, which, because of the overprediction, performed worst, see Figure 1. With FL, the bias in predictions is fully characterized by the number of covariates, in fact the sum of predictions $\sum_{i=1}^n \hat{\mu}_i$ overestimates the sum of observed counts $\sum_{i=1}^n y_i$ by $(k + 1) / 2$. As described in Section 3, ML, FLAC, and BDA yield unbiased

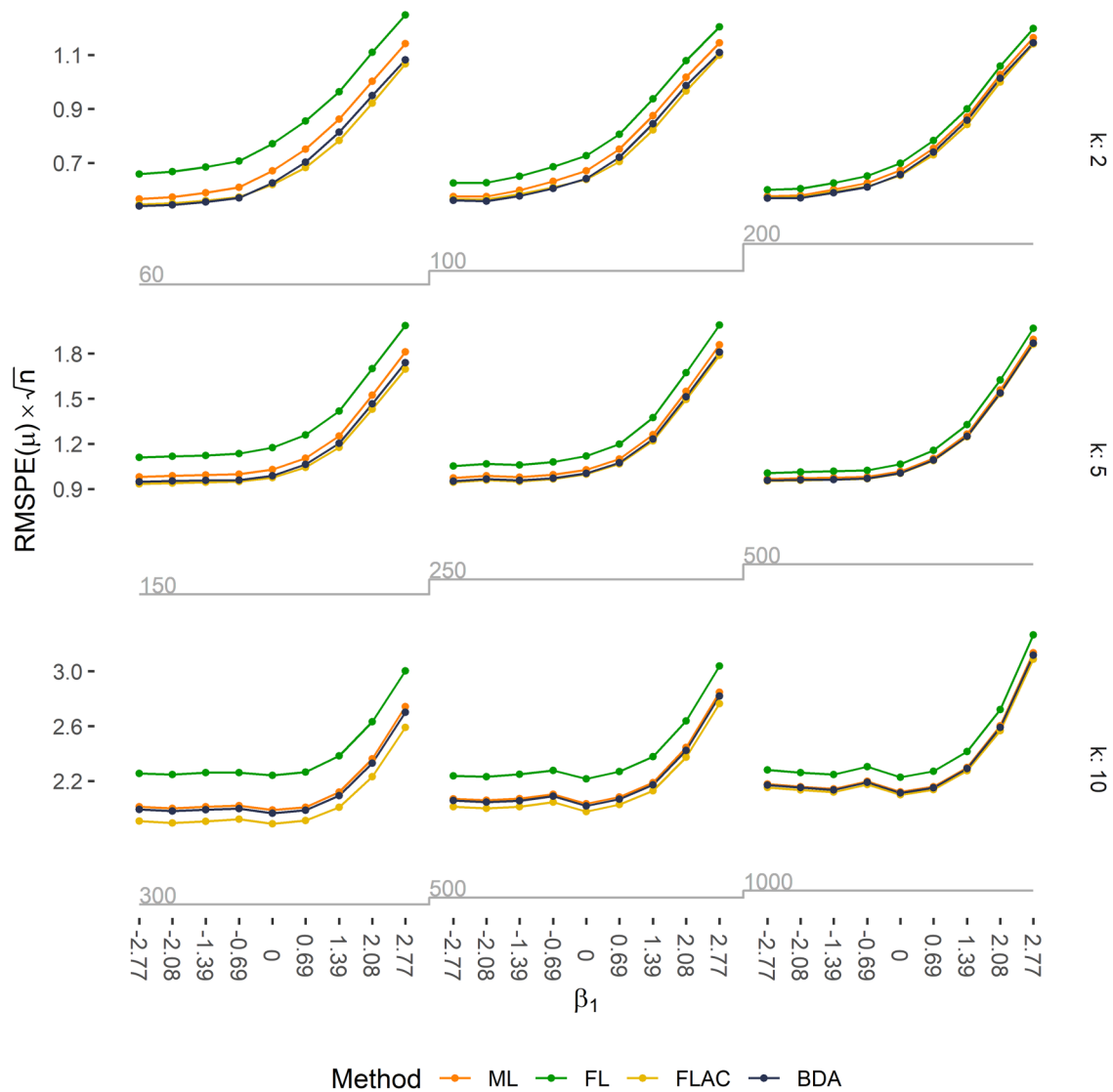


Figure 1. Predictive accuracy is expressed as $RMSPE(\mu)$ multiplied by the square root of the sample size n for all 81 simulation scenarios. Expected counts were obtained by ML, FL, FLAC, and BDA. Rows correspond to the number of covariates in the respective simulation scenario, columns to the EPV ratio, and ticks on the x-axis to the true value of β_1 . Grey step functions below the plots indicate the sample size.

RMSPE: root mean squared prediction error; ML: maximum likelihood; FL: Firth’s likelihood penalization; FLAC: FL with added covariate; BDA: Bayesian data augmentation; EPV: events per variable.

predictions in the sense that $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i$. Figure S2 shows the scaled bias and $RMSPE(\mu)$ in relation to the true incidence exemplarily for one simulation scenario.

4.2.3 Regression coefficients: point estimates

When β_1 was evaluated as estimand, FLAC was not considered separately as it yields the same regression coefficients as FL. Concerning accuracy of regression coefficients, FL, EP and BDA performed similarly well, with some advantage of BDA for extreme, negative values of β_1 and some advantage of FL and EP for β_1 close to 0, see Figure 2. A major drawback of EP is that it is not applicable with larger sample sizes. The worse performance of ML is partly due to the occurrence of separation, but also to a generally higher inaccuracy for data with a low EPV ratio.

ML lead to a large negative bias because of divergent estimation caused by separation when β_1 was negative, see Figure S3. By contrast, FL, EP and, to a lesser extent, BDA lead to a positive bias in scenarios with large negative

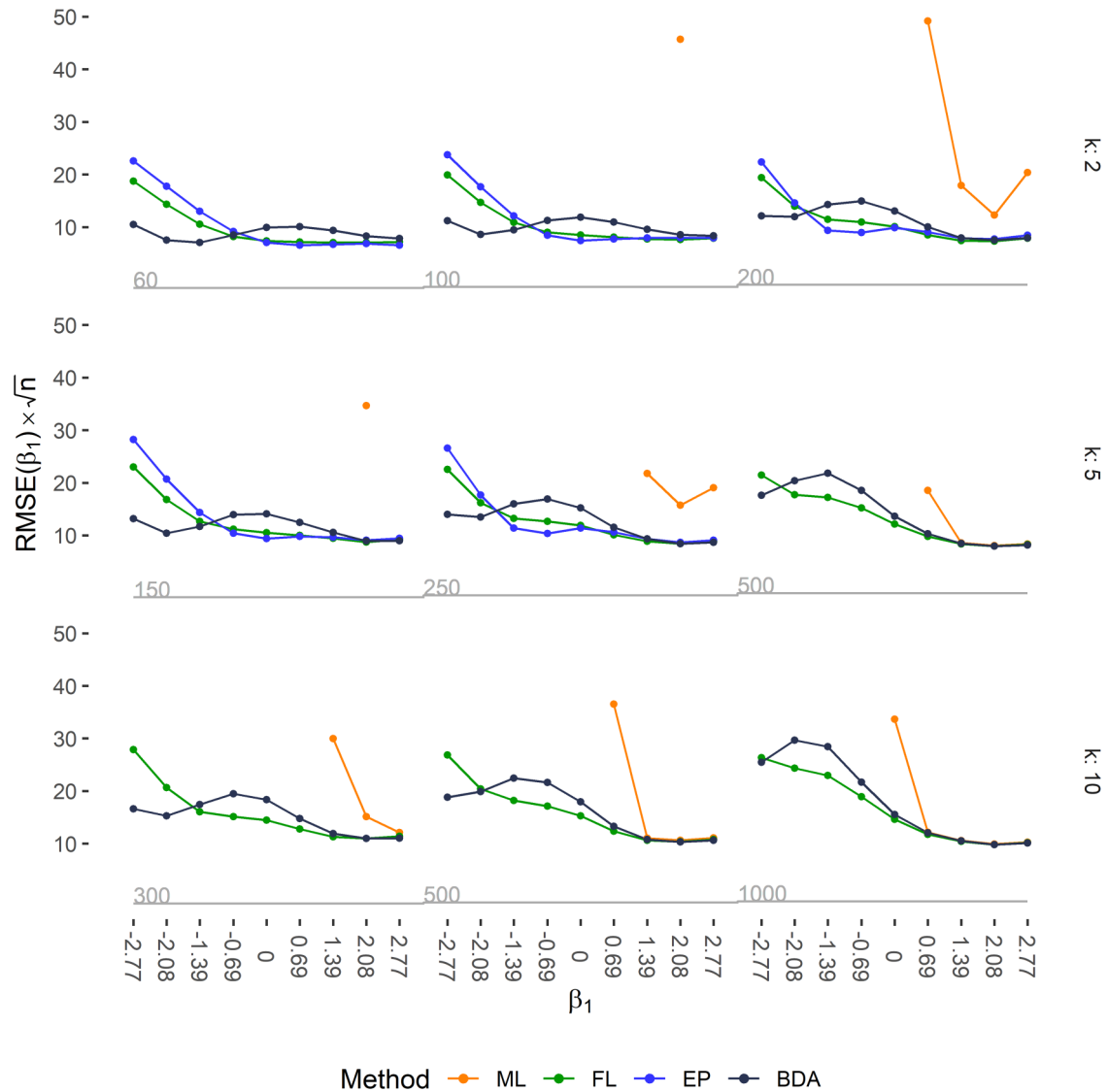


Figure 2. Accuracy of the estimated regression coefficient β_1 , evaluated as $RMSE(\beta_1)$ multiplied by the square root of the sample size n for all 81 simulation scenarios coefficients. The regression coefficient β_1 was estimated by ML, FL, EP, and by BDA. Rows correspond to the number of covariates in the respective simulation scenario (k), columns to sample size (see numbers in grey), and ticks on the x-axis to the true value of β_1 . Grey step functions below the plots indicate sample size. $RMSE(\beta_1)$ for ML occasionally exceeded the upper limit of the plotting range and was then omitted. RMSE: root mean squared error; ML: maximum likelihood; FL: Firth’s likelihood penalization; EP: exact Poisson; BDA: Bayesian data augmentation.

values of β_1 , meaning that some bias toward 0 was introduced. The behaviour of the methods was similar in estimating β_2 , see Figures S4 and S5.

4.2.4 Regression coefficients: CIs

Left-tailed and right-tailed coverage rates of 95% two-sided CIs are depicted in Figure 3, and the power to exclude $\beta_1 = 0$ is shown in Figure S6. For large negative β_1 , all methods yielded higher than nominal right-tailed coverage often in combination with low power. In these scenarios ML–Wald intervals showed undercoverage of their left tails despite the considerable amount of separated data sets (cf. Figure S1). This left-tailed undercoverage for large negative β_1 was even more severe with FL–PPL intervals. While exact CIs were over conservative, mid-p intervals could correct the conservatism,

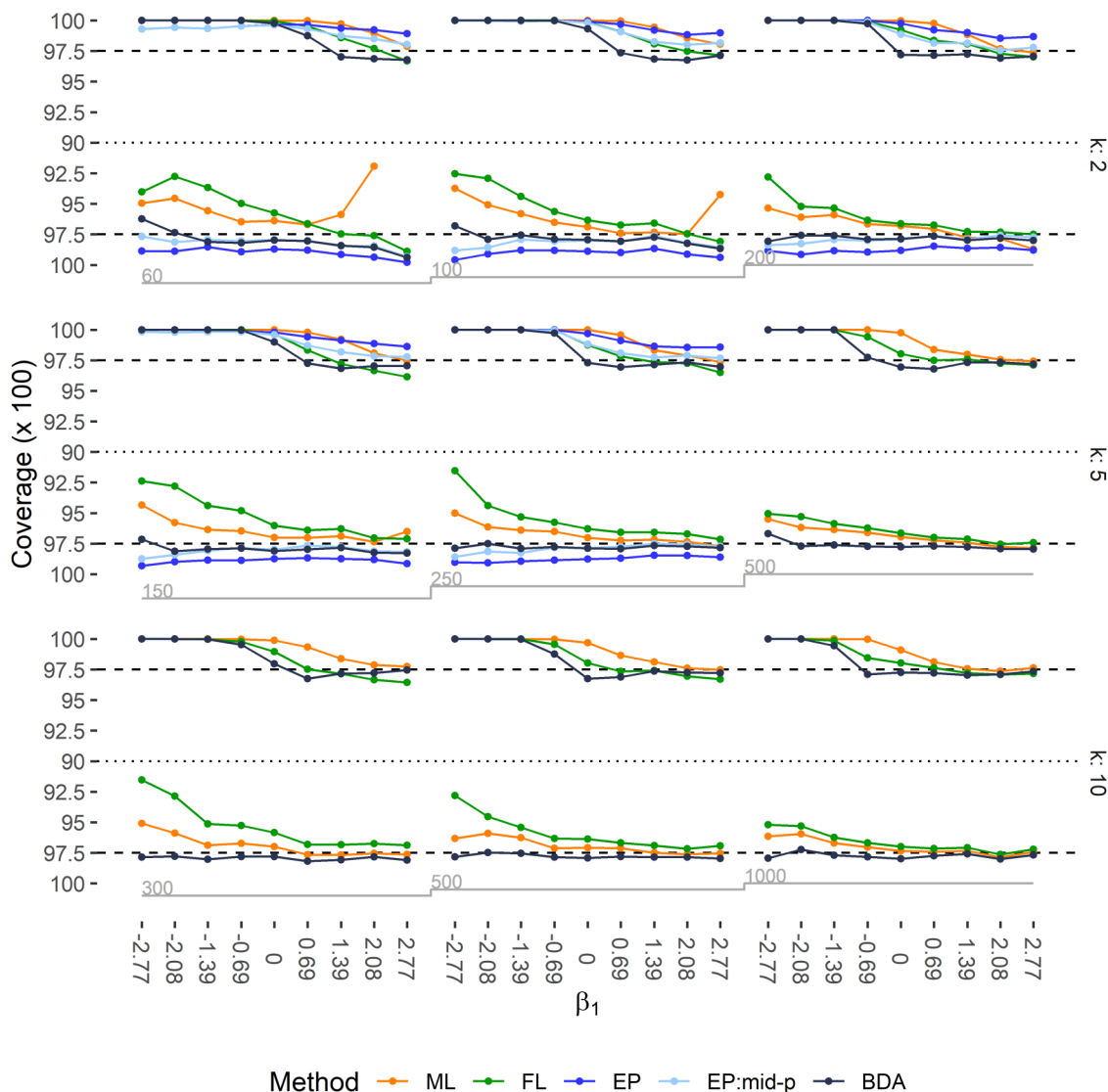


Figure 3. Left-tailed and right-tailed coverage of 95% two-sided CIs for regression coefficient β_1 for all 81 simulation scenarios. CIs were estimated using the Wald method with ML estimation, likelihood profiles with FL, exact interval estimates (EP), and mid-p corrected CIs (EP:Mid-p) with EP regression and likelihood profiles with BDA. The upper halves of the plots describe the right-tailed coverage, lower halves the left-tailed coverage. Nominal levels of 97.5% are marked by dashed lines. Rows correspond to the number of covariates in the respective simulation scenario, columns to the EPV ratio, and ticks on the x-axis to the true value of β_1 . Grey step functions below the plots indicate sample size. For one scenario ($n = 60$, $k = 2$, $\beta_1 = \log(16)$) the left-tailed coverage of the CIs for ML exceeded the lower limit of the plotting range. CI: confidence interval; ML: maximum likelihood; BDA: Bayesian data augmentation; EP: exact Poisson; FL: Firth’s likelihood penalization; EPV: events per variable.

resulting in increased power and coverage close to the nominal one-sided level of 97.5%. BDA–PL intervals generally were preferable, both in terms of coverage and power. Median width of CIs is described in Figure S7.

5 Implant dentistry study

Feher et al.²⁵ studied risk factors for complications in implant dentistry. Risk factors were assessed in 1133 patients undergoing 2405 implantations. We used Poisson regression to model the number of haematological complications by the risk factors age (in decades), smoking (no smoking, light smoking, and heavy smoking), and diabetes mellitus and considered the number of implantations performed per patient as the rate multiplier. Smoking was coded using ‘ordinal coding’, that is,

two dummy variables were defined contrasting light smokers from non-smokers, and heavy smokers from light smokers. Table S1 contains the basic descriptives for these variables.

ML analysis failed to converge because no haematological complications were observed for light smokers (Table 3). This caused the regression coefficients of the two dummy variables associated with smoking to diverge. PROC GENMOD reported arbitrarily large estimates and standard errors for the corresponding regression coefficients but rather than reflecting the large standard errors, the reported Wald CI was collapsed at the point estimates.

By contrast, FL, BDA, and EP using the MUE gave plausible estimates for all variables (Table 3). These methods also supplied 95% CI, which provided some evidence that light smokers experienced fewer haematological complications than non-smokers or heavy smokers. When comparing the two methods of estimating CI for FL, as expected, fixing the pseudo data with weights computed at the maximum penalized likelihood estimate ('PL CI' from the augmented data, used in the simulation study) led to slightly narrower intervals than iterating the weights ('PPL CI'). We also compared the impact of specifying different priors with BDA. Compared to weakly informative priors ('BDA 100/1000': 95% prior intervals for the IRR extending to 100 for age and to 1000 for the binary covariates, used in the simulation study), with narrower priors ('BDA 5/50': 95% prior intervals extending to 5 for age and to 50 for the binary covariates), all point estimates and nearly all CI were pulled towards unity. The effect of changing the prior was particularly strong for the two IRRs corresponding to smoking which caused the separation problem. Employing weakly informative priors resulted in CI supporting the hypothesis that light smokers experienced fewer complications than non-smokers and heavy smokers, while with narrower priors the CI included unity. EP generally resulted in IRR and CI closer to the estimates by BDA 5/50 than to the estimates by other methods.

While 37 hematologic complications were observed, 39.5 complications were predicted by FL, which estimated the intercept at -4.3728 . Applying FLAC changed the intercept to -4.4382 , and recalibrated the total number of predicted complications to the observed number of 37. Because age was centred at 50 years, $\exp(\beta_0)$ expresses the risk of a complication with one implantation for a 50-year-old non-smoking non-diabetic person. This risk was estimated at 1.26% by FL and at 1.18% by FLAC, and for a 70-year-old diabetic at 21.0% or 19.7%, respectively.

6 Discussion

We proposed and investigated Poisson regression methods to deal with the problem of separation, which leads to nonexistence of ML estimates. We adapted two modifications of FL, namely FLIC and FLAC, which were originally developed to debias predictions in logistic regression, for Poisson regression. It turned out that in Poisson regression FLIC and FLAC lead to the same estimation method. This method, which we refer to as FLAC, competed well with alternative approaches such as EP regression, which needs special software and is only applicable for small-sized problems, or BDA, which is easy to implement but crucially depends on the choice of the width of the prior distribution. A possible advantage of BDA could arise if a model with many covariates should be fitted. It is essentially equal to ridge regression, and unlike FL can handle situations where the number of covariates exceeds the number of events by regularizing parameter estimates. In our

Table 3. Results for implant dentistry study: point estimates of IRRs and 95% CI.

	Age (per decade)	Light versus no smoking	Heavy versus light smoking	Diabetes versus no diabetes
ML, IRR	1.657	0*	$20.3 \times 10^{9*}$	6.012
Wald 95% CI	1.298, 2.116			2.895, 12.486
FL/FLAC, IRR	1.646	0.147	9.045	6.154
PL 95% CI	1.309, 2.101	0.001, 1.047	0.950, 1110.5	2.944, 12.231
PPL 95% CI	1.3, 2.084	0.001, 1.065	0.928, 1209.8	2.901, 12.384
BDA 100/1000, IRR	1.665	0.113	10.339	5.977
PL 95% CI	1.315, 2.143	0.003, 0.867	1.020, 373.4	2.792, 12.105
BDA 5/50, IRR	1.656	0.24	4.686	5.860
PL 95% CI	1.317, 2.130	0.029, 1.055	0.793, 41.133	2.754, 11.788
EP, IRR	1.657	0.213	5.160	6.027
Exact 95% CI	1.296, 2.147	0, 1.217	0.627, ∞	2.630, 13.058
Mid-p 95% CI	1.305, 2.130	0, 0.972	0.856, ∞	2.806, 12.395

ML: maximum likelihood; FL: Firth's likelihood penalization; FLAC: FL with added covariate; BDA: Bayesian data augmentation; 100/1000 and 5/50: upper prior limit for IRR for age/binary covariates; EP: exact Poisson; PL: profile likelihood; PPL: profile penalized likelihood; CI: confidence interval; IRR: incidence rate ratio.

*Not converged; Wald CIs reported by PROC GENMOD collapsed at point estimate.

application and simulations we fixed the variance of the prior distribution, which is inversely proportional to the penalty parameter in ridge regression. Optimizing that penalty parameter by, for example, cross-validation invalidates inference about regression coefficients, is not robust to separation¹⁴ and can lead to instable results.²⁶ To sum up, for data sets fitting into the framework of this study, that is, with an EPV ratio of 3 or higher and moderate correlation between covariates, we advocate using FL as it does not need any user input or optimization of a penalty parameter and is computationally feasible, while showing good performance.

In some situations, count outcomes can be naturally reinterpreted as dichotomous outcomes, for example, in the coronavirus testing study where we can either count the number of infections per working place or determine whether a person is infected or not. These data allow for analysis via Poisson regression as well as via logistic regression. It was not the aim of this study to provide guidance on which analysis method to prefer with sparse data, but primarily the decision should depend on the estimand of interest, that is, whether risk ratios or odds ratios should be estimated. Our study employed a realistic design for simulations that resembled data one could typically see in epidemiological studies. Such a design is suitable to draw conclusions on the relative performance of methods in practically relevant situations. While in the simulation study we used R for data generation and summarizing results, we focused on the SAS software for fitting the Poisson models. SAS is widely used among epidemiologists, and with the Cytel SAS procedures, robust and efficient software was available to include EP regression in our comparisons. Hence we employed Cytel's PROC LOGXACT for EP regression even if an implementation of EP regression is readily available in PROC GENMOD. We provide a SAS macro to apply the Firth correction and its modification FLAC in Poisson regression. The median unbiased estimator implemented in PROC LOGXACT has been described to suffer from extreme shrinkage in case the exact conditional distribution of the sufficient statistic is nearly degenerate. Therefore, an alternative estimator based on maximizing the conditional penalized likelihood was proposed.²⁷ We did not include it in our comparison as Heinze and Puhr²⁷ found it to be very similar to the FL estimator. We also did not consider the median bias-corrected estimator of Kosmidis et al.²⁸

R code for fitting a Poisson model with FL is also available,²⁹ however, without the possibility to invoke the FLAC extension or the estimation of PPL CI. We are also not aware of software packages in R which allow fitting EP regression models. Our SAS macro FLACPOISSON, a SAS macro to implement BDA and further SAS macros implementing FL and FLAC for logistic, conditional logistic and Cox regression are available on the GitHub repository <https://github.com/georgheinze/flicflac>. A further public repository, <https://github.com/georgheinze/PoissonF>, contains the aggregated data set of the implant dentistry study and code to reproduce its analysis, all codes used to conduct the simulation study, and an R markdown file which summarizes its results.

The FLAC method lends itself to several extensions. Most naturally, it can easily accommodate overdispersion by including the estimation of a dispersion parameter.³⁰ This can already be achieved with our SAS macro which is based on PROC GENMOD. Further work could be done to investigate the advantages of considering FLAC for this and other extensions, such as zero-inflated Poisson models and Poisson hurdle models.

Acknowledgements

AJ acknowledges support by the research mobility grant (No. 331377) awarded by the Academy of Finland. AJ carried out parts of this work at Cytel Statistical Software and Services, Pune, India, and at Cytel Inc., Cambridge, MA, USA. PS and CC acknowledge the support by NIH, grant SBIR-9R44 GM104597-02A1. We thank Drs Kuchler, Gruber, and Feher from the University Clinic of Dentistry, Medical University of Vienna, for providing the data of the implant dentistry study which partly motivated this research.

Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: GH's, AG's, and LJ's work were partly supported by the Austrian Science Fund (FWF), award I-2276-N33. AJ was supported by the research mobility grant (No. 331377) awarded by the Academy of Finland. PS and CC were supported by NIH, grant SBIR-9R44 GM104597-02A1.

ORCID iDs

Angelika Geroldinger  <https://orcid.org/0000-0002-4659-4911>

Georg Heinze  <https://orcid.org/0000-0003-1147-8491>

Supplemental Material

Supplemental material for this article is available online.

References

- Dunn P and Smyth G. *Generalized linear models with examples in R*. New York: Springer, 2018.
- Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.
- Correia S, Guimarães P and Zylkin T. Verifying the existence of maximum likelihood estimates for generalized linear models. *ArXiv*. 2019. <https://arxiv.org/abs/1903.01633> (accessed 18 October 2021).
- Albert A and Anderson JA. On the existence of maximum-likelihood estimates in logistic-regression models. *Biometrika* 1984; **71**: 1–10.
- Heinze G and Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409–2419.
- Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med* 2006; **25**: 4216–4226.
- Puhr R, Heinze G, Nold M, et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med* 2017; **36**: 2302–2317.
- Firth D. Bias reduction of maximum-likelihood-estimates. *Biometrika* 1993; **80**: 27–38.
- Kosmidis I and Firth D. Bias reduction in exponential family nonlinear models. *Biometrika* 2009; **96**: 793–804.
- Cytel Inc. LogXact 12 PROCs for SAS users. Cambridge, MA2019. Software Manual.
- Read CB. Median unbiased estimators. In: Kotz S, Balakrishnan N, Read CB, Vidakovic B and Johnson NL (eds) *Encyclopedia of statistical sciences*. New York: John Wiley & Sons, 2006, pp.4713–4715.
- Al-Youssef M, Taschwer K and Thaler S. Stichprobentests sollen Klarheit bringen. *Der Standard*. 3 April 2020.
- Cummings P. Exact Poisson regression. In: *Analysis of incidence rates. Biostatistics series*. New York: Chapman and Hall/CRC, 2019, pp.395–402.
- Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic regression: causes, consequences, and control. *Am J Epidemiol* 2018; **187**: 864–870.
- Heinze G and Schemper L. A solution to the problem of monotone likelihood in Cox regression. *Biometrics* 2001; **57**: 114–119.
- Hirji KF, Tsiatis AA and Mehta CR. Median unbiased estimation for binary data. *Am Stat* 1989; **43**: 7–11.
- SAS Institute Inc. *SAS/STAT 15.1 user's guide*. Cary, NC: SAS Institute Inc., 2018.
- Sullivan SG and Greenland S. Bayesian Regression in SAS software. *Int J Epidemiol* 2013; **42**: 308–317.
- Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; **38**: 2074–2102.
- Binder H, Sauerbrei W and Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Stat Med* 2013; **32**: 2262–2277.
- Zöllner D, Wockner LF and Binder H. Automatic variable selection for exposure-driven propensity score matching with unmeasured confounders. *Biom J* 2020; **62**: 868–884.
- Kammer M. simdata: An R package to create simulated datasets. R package, 2021. <https://github.com/matherealize/simdata> (accessed 18 October 2021).
- Kammer M. looplot: An R package for creating nested loop plots, version 0.5.0.9001. R package, 2021. <https://github.com/matherealize/looplot> (accessed 18 October 2021).
- Rucker G and Schwarzer G. Presenting simulation results in a nested loop plot. *BMC Med Res Methodol* 2014; **14**: 129.
- Feher B, Lettner S, Heinze G, et al. An advanced prediction model for postoperative complications and early implant failure. *Clin Oral Implants Res* 2020; **31**: 928–935.
- Van Calster B, van Smeden M, De Cock B, et al. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Stat Methods Med Res* 2020; **29**: 3166–3178.
- Heinze G and Puhr R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Stat Med* 2010; **29**: 770–777.
- Kosmidis I, Pagui ECK and Sartori N. Mean and median bias reduction in generalized linear models. *Stat Comput* 2020; **30**: 43–59.
- Kosmidis I. brglm2: Bias reduction in generalized linear models, version 0.7.0.2020. R package, 2021. <https://cran.r-project.org/package=brglm2> (accessed 18 October 2021).
- Lawless JF. Negative binomial and mixed Poisson regression. *Can J Stat* 1987; **15**: 209–225.

Appendix I

A.1 Conditions for the existence of ML estimates in Poisson regression

Correia et al.³ provide conditions governing the existence of ML estimates for a wide class of generalized linear models. Here, we apply their proof to the special case of Poisson models. We use the notation from the main text.

Theorem 1. Let the $n \times (k + 1)$ design matrix X be of full rank. ML estimation does not give a finite solution for the Poisson regression model if and only if there exists a non-zero vector $\gamma^* \in \mathbb{R}^{k+1}$ with

$$\begin{aligned} x_i \gamma^* &= 0 && \text{for } i \text{ with } y_i > 0 \text{ and} \\ x_i \gamma^* &\leq 0 && \text{for } i \text{ with } y_i = 0. \end{aligned} \tag{A1}$$

Proof. The log-likelihood is given by $\ell(\beta) = \sum_{i=1}^n -\exp(x_i \beta + z_i) + (x_i \beta + z_i) y_i - \log(y_i!)$. Let $\beta, \gamma \in \mathbb{R}^{k+1}$ be arbitrary vectors with $\gamma \neq 0$ and let $k > 0$ be a positive scalar. Differentiating $\ell(\beta + k\gamma)$ with respect to k yields

$$\frac{\partial \ell(\beta + k\gamma)}{\partial k} = \sum_{i=1}^n -\exp(x_i \beta + x_i k\gamma + z_i) x_i \gamma + x_i \gamma y_i$$

We will first show that if there exists γ^* fulfilling properties A1, then the log-likelihood function $\ell(\beta)$ is always increasing in the direction of γ^* , in other words there does not exist a finite β maximizing the log-likelihood. For γ^* satisfying conditions A1, the directional derivative reduces to

$$\frac{\partial \ell(\beta + k\gamma^*)}{\partial k} = \sum_{i: y_i=0} -\exp(x_i \beta + x_i k\gamma^* + z_i) x_i \gamma^*$$

This expression is greater than 0 since the exponential function is positive and $x_i \gamma^* \leq 0$ for all i with $y_i = 0$ and $x_i \gamma^* < 0$ for some i with $y_i = 0$. (The inequality $x_i \gamma^* \leq 0$ has to be strict for at least one observation because otherwise the full rank assumption would be violated.) This proves that for any β and for any $k > 0$ we have $\ell(\beta + k\gamma^*) > \ell(\beta)$, that is, we cannot find an ML solution maximizing ℓ .

Next, we will show that if there does not exist γ^* with the properties A1, then the log-likelihood has a maximum. Assume that there does not exist γ^* satisfying A1. Then for every $\gamma \in \mathbb{R}^{k+1}$ at least one of the following two conditions must be true:

1. there is j such that $x_j \gamma \neq 0$ and $y_j > 0$,
2. there is j such that $x_j \gamma > 0$ and $y_j = 0$.

Denote by $\ell_i(\beta) = -\exp(x_i \beta + z_i) + (x_i \beta + z_i) y_i - \log(y_i!)$ the summand in the log-likelihood function corresponding to the i th observation. One can show that if any of the two conditions above holds, then we have $\lim_{k \rightarrow \infty} \ell_j(\beta + k\gamma) = -\infty$. Since the components $\ell_i(\beta)$ have a common upper bound we can find a positive scalar \bar{k} such that $\ell(\beta + k\gamma) = \sum \ell_i(\beta + k\gamma) < \sum \ell_i(\beta) = \ell(\beta)$ for any $\beta, \gamma \in \mathbb{R}^{k+1}$ and $k > \bar{k}$. This means that following the log-likelihood from any starting point β in any direction γ will finally yield a decrease in the log-likelihood, guaranteeing the existence of a finite ML solution. \square

A.2 Properties of FLAC

A.2.1 The sum of the predicted counts with FLAC, $\sum_i \hat{\mu}_i$, is equal to $\sum_i y_i$

With FLAC, the score equation U_0^{FLAC} corresponding to the intercept reads

$$U_0^{\text{FLAC}}(\beta, \gamma) = \sum_i (y_i - \exp(x_i \beta + z_i)) + \sum_i (h_i / 2 - \exp(x_i \beta + z_i + \gamma)) = 0$$

and the score equation U_{k+1}^{FLAC} corresponding to the additional covariate γ reads

$$U_{k+1}^{\text{FLAC}}(\beta, \gamma) = \sum_i (h_i / 2 - \exp(x_i \beta + z_i + \gamma)) = 0.$$

Combining the two equations, we conclude that $\sum_i y_i = \sum_i \exp(x_i \beta + z_i) = \sum_i \hat{\mu}_i$.

A.2.2 FLAC estimates of β_1, \dots, β_k agree with the FL estimates in Poisson regression

Consider the score equations, $U_0^{\text{FL}}, \dots, U_k^{\text{FL}}$, of FL given as follows:

$$\begin{aligned} U_j^{\text{FL}}(\beta) &= \sum_i x_{ij}(y_i + h_i / 2 - \exp(x_i \beta + z_i)) \\ &= \sum_i x_{ij}y_i + \sum_i x_{ij}h_i / 2 - \exp(\beta_0) \sum_i x_{ij} \exp(x_{i1}\beta_1 + \dots + x_{ik}\beta_k + z_i) = 0 \end{aligned} \quad (A1)$$

The corresponding equations for FLAC are given by:

$$\begin{aligned} U_j^{\text{FLAC}}(\beta, \gamma) &= \sum_i x_{ij}(y_i - \exp(x_i \beta + z_i)) + \sum_i x_{ij}(h_i / 2 - \exp(x_i \beta + z_i + \gamma)) \\ &= \sum_i x_{ij}y_i + \sum_i x_{ij}h_i / 2 - (\exp(\beta_0) + \exp(\beta_0 + \gamma)) \sum_i x_{ij} \exp(x_{i1}\beta_1 + \dots + x_{ik}\beta_k + z_i) = 0. \end{aligned} \quad (A2)$$

This shows that estimates of β_1, \dots, β_k by FL and FLAC are identical.

A.2.3 FLIC and FLAC give identical estimates β_0, \dots, β_k in Poisson regression

We have shown above that FLAC estimates of β_1, \dots, β_k agree with the FL estimates. The same is true for FLIC by definition. Moreover, for both methods the sum of the predicted counts $\sum_i \hat{\mu}_i$ is equal to $\sum_i y_i$. This implies that FLIC and FLAC also give identical estimates for the intercept β_0 .