Leitgöb, H. (2019). Analysis of Rare Events. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), SAGE Research Methods Foundations. doi: 10.4135/9781526421036863804

SAGE researchmethods

Analysis of Rare Events

Foundation Entries

SAGE Research Methods Foundations

By: Heinz Leitgöb
Published:2019
Length: 10,000 Words
DOI: http://dx.doi.org/10.4135/9781526421036
Methods: Analysis of Rare Events
Online ISBN: 9781526421036
Disciplines: Anthropology, Business and Management, Criminology and Criminal Justice, Communication and Media Studies, Counseling and Psychotherapy, Economics, Education, Geography, Health, History, Marketing, Nursing, Political Science and International Relations, Psychology, Social Policy and Public Policy, Social Work, Sociology, Science, Technology, Computer Science, Engineering, Mathematics, Medicine
Access Date: January 23, 2020
Publishing Company: SAGE Publications Ltd
City: London

Abstract

Rare events represent a great analytical challenge. The maximum likelihood-based (ML) binary logit model as the workhorse model in the social sciences can generate heavily biased parameter estimates if events are rare. In detail, the finite sample bias in ML estimates may be substantially larger than that observed in cases with balanced data of the same sample size. Furthermore, the ML estimator is prone to overfitting rare event data even in low-dimensional models and not identified in cases of perfectly separated data. Starting with a brief introduction to the standard binary logit as a reference model, this entry discusses several design issues (e.g., selection on the dependent variable) and analytical approaches (e.g., first-order bias correction, exact conditional inference, penalized ML estimation, specification of cloglog models) to overcome these threats to valid inferences. Finally, the potential of Bayesian rare event modeling, which addresses some limitations of the frequentist probability perspective, is briefly introduced.

Introduction

Studying events is at the heart of research in various disciplines. The occurrence of an event is a situational concept, defined as change in an object's characteristic from one state or level to another (Blossfeld & Rohwer, 1997). Events cover a broad range of phenomena that differ substantially in their probability of occurrence. While some events are prevalent and occur on a regular basis, others occur very seldom. Numerous intensively investigated issues in contemporary societies belong to the latter type, including severe crimes, epidemiological infections, economic shocks, political upheavals, and natural disasters, denoted as *rare* events. As data on these events are highly imbalanced, "with dozens to thousands of times fewer ones (events *[…]*) than zeros ("nonevents")" (King & Zeng, 2001, p. 137), several design and analytical issues must be considered to obtain valid inferences. Otherwise, nonconvergence, substantial bias and uncertainty in parameter estimates, misinterpretation of results, and high prediction error are likely.

Starting with a brief introduction of the binary logit as a reference model for analyzing the occurrence of events in the social sciences, this entry subsequently addresses core analytical problems associated with data on rare events, provides solutions that avoid defective modeling results, and finally discusses some prospects for the analysis of rare events. This entry is limited to *binary explanatory models* of event occurrence from a *frequentist* probability perspective. Some remarks regarding *Bayesian inference* are outlined in the final section.

Binary Logit Model

(Non)occurrences of some event of interest can be captured by a binary random variable Y ($D_Y = \{0, 1\}$), with $Y = 1$ indicating that the event did occur and $Y = 0$ otherwise. Let Y follow a Bernoulli distribution, with

parameter π_i as individual event probability $\Pr(Y_i = 1)$ for $i = 1, \ldots, n$ observations under study. Assume further that π_i is subject to variation because observations differ with respect to a set of $k = 1, \ldots, K$ individual characteristics, captured by the $1 \times K + 1$ vector $\boldsymbol{x}_i^{\mathrm{T}} = (1, x_{i1}, \ldots, x_{iK})$.

A statistical model that adequately describes the data generating process (DGP) outlined in the previous paragraph can be formulated in terms of a generalized linear model (GLM; McCullagh & Nelder, 1989): The (i) systematic component in the form of the linear predictor

(1)

$$\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$$

introduces information about $\boldsymbol{x}$ to the model, with $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \ldots, \beta_K)$ as a vector of effect parameters, while the (ii) stochastic (or random) component reflects the probability distribution of Y by specifying some density function from the exponential family. In the present case, it is self-evident to choose the Bernoulli density $f(Y|\pi) = \pi^Y(1 - \pi)^{1 - Y}$. Furthermore, the model comprises a (iii) monotonic and invertible link function g, which serves as a connection between the systematic and stochastic components. Its inverse g^{-1} transforms the unbounded linear predictor $\eta(\boldsymbol{x})$ into the Bernoulli distribution's single parameter $\pi(\boldsymbol{x})$, now representing the event probability conditional on $\boldsymbol{x}$. To conform to the probability metric, function values of g^{-1} are expected to lie within the interval [0,1], forcing g to satisfy $[0, 1] \rightarrow]-\infty, \infty[$. Typically, g is defined by some cumulative distribution function (CDF). The well-known candidates include the logit $g_{\mathrm{l}} = \Lambda^{-1} = \ln\{\pi(\boldsymbol{x}) / [1 - \pi(\boldsymbol{x})]\}$ and probit $g_{\mathrm{p}} = \Phi^{-1}[\pi(\boldsymbol{x})]$ function, with Λ as the CDF of the standard logistic and Φ of the standard normal distribution. For several reasons (see, e.g., the arguments provided by Paul Allison on his blog https://statisticalhorizons.com/whats-so-special-about-logit), the canonical logit is preferred over the noncanonical probit link in the social sciences. Thus, the contribution refers to the logit as a reference model. However, most of the issues discussed in this entry apply equally to the probit model.

The logit model can be derived in conditional expected probability (CEP) form by formulating π_i as a logistic function of η_i:

(2)

$$E[\Pr(y_i = 1|\boldsymbol{x}_i)] = \pi_i = \Lambda(\eta_i) = \frac{1}{1 + \exp(-\eta_i)}$$

Assuming $y_1, \ldots, y_n$ are independent (y_i from individual i is not influenced by y_j from individual j) and identically distributed (possess the same underlying distribution), the model's likelihood function relates to the product of the n individual probabilities of observing the outcome y_i, represented by the Bernoulli density determined as a stochastic component:

(3)

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

Under some regularity conditions, the likelihood function is strictly concave, guaranteeing unique maximum likelihood (ML) estimates $\hat{\boldsymbol{\beta}}^{ML}$ that are consistent, asymptotically unbiased, efficient, and normal (Amemiya, 1985). Monotonicity of the logarithmic function allows identifying $\hat{\boldsymbol{\beta}}^{ML}$ through the more convenient log-likelihood function:

(4)

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

The first derivative of Equation 4 with respect to $\boldsymbol{\beta}$ is referred to as the score function $U(\boldsymbol{\beta})$, represented by the $(K + 1)$-dimensional vector $\boldsymbol{q}$:

(5)

$$U^{\mathrm{ML}}(\boldsymbol{\beta}) = \frac{\partial l^{\mathrm{ML}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{q}$$

Finally, $\hat{\boldsymbol{\beta}}^{ML}$—the $\boldsymbol{\beta}$ values that maximize the function value of Equation 4—can be obtained by setting $\boldsymbol{q} = \mathbf{0}$ and solving for $\boldsymbol{\beta}$. In most cases, the system of equations has no closed-form solution, necessitating numerical approximation methods such as the Newton–Raphson algorithm.

Modeling Rare Event Binary Data: Problems and Promising Solutions

Analyzing binary rare event data is afflicted with various (partially related) challenges, covering the key issues of unbiased parameter estimation, susceptibility to overfitting, handling of separated data, and appropriate specification of the link function.

Finite Sample and Rare Event Bias

The linear model's well-known ordinary least squares estimator has—given the Gauss–Markov assumptions are satisfied—a set of desirable finite sample properties including unbiasedness, consistency, and efficiency. In contrast, the ML estimator is only asymptotically unbiased (e.g., Schaefer, 1983), supposing the expectation of some parameter estimate $\hat{\theta}$ to converge to the underlying population parameter θ as n converges to infinity: $\lim_{n \to \infty} E\left(\hat{\theta}_n\right) = \theta$. Thus, ML estimates are biased in finite samples, and the amount of bias may be substantial in small- to moderate-sized samples.

Generally, bias in ML estimates is induced by small (total) Fisher information (FI; Cordeiro & McCullagh, 1991). In case of the binary logit model, FI measures the amount of information the random variable Y of

size n contains about the set of unknown parameters $\boldsymbol{\beta}$ and is captured in $\mathrm{I}(\boldsymbol{\beta})$, the $K + 1 \times K + 1$ expected FI matrix:

(6)

$$\mathrm{I}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta \partial \beta'}\right] = \sum_{i=1}^{n} \pi_i(1-\pi_i)\boldsymbol{x}_i\boldsymbol{x}_i'$$

Equation 6 indicates that FI increases with sample size (summation over n) and individuals close to $\pi_i = .5$ tend to contribute more to FI than those with rather low/high event probabilities because the term $\pi_i(1-\pi_i)$ represents a strictly concave function, having its maximum at $\pi_i = .5$. By definition, rare events are characterized by an event probability π_i close to zero for a vast majority of individuals in the population. Thus, for some given n, FI is expected to be considerably lower for rare event data than that for data balanced in Y, implying that finite sample bias is *amplified* by rare events (King & Zeng, 2001). It should be noted, however, that the same holds true for (very) frequent events because $\pi_i(1-\pi_i)$ is symmetric in shape and converges to zero as π_i tends to its upper boundary one. Furthermore, as the inverse of the *observed* (not the *expected*) FI matrix $\hat{\mathrm{I}}\left(\hat{\boldsymbol{\beta}}\right)^{-1}$ represents the covariance matrix for $\hat{\boldsymbol{\beta}}^{ML}$, with the square roots of the elements from the main diagonal as respective standard errors, the obviously negative relationship between FI and the uncertainty in parameter estimates becomes visible.

Generally, asymptotic bias in some arbitrary ML estimate $\hat{\theta}$, given sample size n, can be expressed as a Taylor series expansion:

(7)

$$b\left(\hat{\theta}_n\right) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \ldots$$

By definition, $b\left(\hat{\theta}\right)$ (for notational simplicity, subscript n will be skipped) constitutes the divergence between the expectation of $\hat{\theta}$ and the corresponding population parameter θ: $b\left(\hat{\theta}\right) \equiv E\left(\hat{\theta}\right) - \theta$ (e.g., McCullagh & Nelder, 1989). As the sample size increases, the bias converges to zero because $\lim_{n \to \infty} b_1(\theta)n^{-1} = 0$. For the logit model's set of ML estimates $\hat{\boldsymbol{\beta}}^{ML}$ (Section *Binary Logit Model*), the bias equals

(8)

$$b\left(\hat{\boldsymbol{\beta}}^{ML}\right) \equiv E\left(\hat{\boldsymbol{\beta}}^{ML}\right) - \boldsymbol{\beta}$$

To account for $b\left(\hat{\beta}^{ML}\right)$, two strategies are commonly applied: *correction* and *prevention*.

Correction

Correction approaches have in common the generation of ML estimates in the first step and removal of first-

order bias $O\left(n^{-1}\right) = b_1(\boldsymbol{\beta})n^{-1}$ (the first term on the right-hand side of Equation 7) in the second step. Two standard procedures are discussed in the literature: The (i) *analytical* or *Taylor series* approach and the (ii) *simulation* approach.

(i) The *analytical* approach is based on the idea of substituting $\hat{\boldsymbol{\beta}}^{ML}$ for $\boldsymbol{\beta}$ in $b_1(\boldsymbol{\beta})$ and calculating the first-order corrected ML estimates $\hat{\boldsymbol{\beta}}^{cML}$ by

(9)

$$\hat{\boldsymbol{\beta}}^{\text{cML}} = \hat{\boldsymbol{\beta}}^{ML} - \frac{b_1\left(\hat{\boldsymbol{\beta}}^{ML}\right)}{n}$$

This presupposes the existence of finite $\hat{\boldsymbol{\beta}}^{ML}$ and a derivation of $b_1\left(\hat{\boldsymbol{\beta}}^{ML}\right)$, which is provided by Peter McCullagh and John A. Nelder (1989) as

(10)

$$b_1\left(\hat{\boldsymbol{\beta}}^{ML}\right) = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{\xi}$$

where $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ is the *observed* FI matrix (the right-hand side of Equation 6 in matrix notation when π_i is replaced by its estimate $\hat{\pi}_i$), $\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}^{ML}$, $\boldsymbol{W} = \text{diag}\left\{m_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)\right\}$ is a diagonal matrix of weights with m_i as the number of observations that share a common covariate vector $\boldsymbol{x}_i^{\mathsf{T}}$, and $\boldsymbol{\xi}_i = Q_{ii}\left(\hat{\pi}_i - 1 \,/\, 2\right)$ indicates the elements of $\boldsymbol{\xi}$ with $\boldsymbol{Q} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ as an asymptotic covariance matrix of $\hat{\eta}$ (for further details, see Cordeiro & McCullagh, 1991). Again, Equation 10 reveals the inverse relation between FI and bias. Regarding corrections for higher order terms ($b_2(\boldsymbol{\beta})n^{-2}$, …), although there may be some additional adjustment effect if n is small, they remain unconsidered due to computational impracticability (Schaefer, 1983).

Owing to the seminal contributions of Gary King and Langche Zeng (2001) and their provision of the program ReLogit for statistical software packages Gauss, R (as part of the Zelig-package), and Stata, the analytical approach gained remarkable popularity in the scientific community, indicated by almost 3,000 citations for King and Zeng (2001; source: Google Scholar, December 19, 2018).

(ii) The computer-intensive *simulation* approach is based on *resampling methods*, particularly variants of the jackknife and bootstrapping. It does not require the estimation of $b_1(\boldsymbol{\beta})$ for its implementation (for an overview, see Kosmidis, 2014).

Prevention

Prevention approaches propose (i) *alternative estimation strategies* to avoid parameter estimates relying on asymptotic properties or a (ii) *modified* or *penalized ML estimator* by manipulating the score function instead

of correcting ML estimates a posteriori.

- (i) *Exact conditional inference*, in the given context also known as *exact logistic regression*, represents an alternative to ML estimation not relying on distributional assumptions. Rather, exact estimation procedures are based on the idea of constructing completely determined "permutational distributions of the sufficient statistics that correspond to the regression parameters of interest, conditional on fixing the sufficient statistics of the remaining parameters at their observed values" (Mehta & Patel, 1995, p. 2143). The approach is also known as *conditional ML* (CML). However, despite the development of efficient iterative algorithms that derive these distributions, exact inferences require—even for moderate-sized samples—enormous memory resources often not feasible with standard RAM. Thus, exact logistic regression can be considered as an alternative only if the sample size and the number of covariates are rather small. Furthermore, continuous covariates may lead to a loss of effective information due to *overconditioning*, arising when "the number of points of support in the conditional distribution becomes very small" (Barndorff-Nielsen & Cox, 1994, p. 44).
- (ii) The *modified score function approach* accounts for the fact that $\hat{\boldsymbol{\beta}}^{ML}$ is biased if the unbiased score function $U(\boldsymbol{\beta})$ (Equation 5) has some curvature (Firth, 1993), and the direction of bias depends on its direction: In case of the logit model, $\partial^2 U^{ML}(\boldsymbol{\beta}) \,/\, \partial \boldsymbol{\beta}^2 > 0$, and thus, $\hat{\boldsymbol{\beta}}^{ML}$ is systematically upward-biased *away from zero* (for a graphical demonstration, see Firth, 1993). To reduce bias, David Firth (1993) proposed the introduction of a small bias via some *penalty term* into the score function that shrinks parameter estimates toward zero. The modified score function $U^{PML}(\boldsymbol{\beta})$ is then defined as

(11)

$$U^{PML}(\boldsymbol{\beta}) = U^{ML}(\boldsymbol{\beta}) - \mathrm{I}(\boldsymbol{\beta})b_1(\boldsymbol{\beta}) = \boldsymbol{q} - \mathrm{I}(\boldsymbol{\beta})b_1(\boldsymbol{\beta})$$

with the penalty term as the product of the FI matrix $\mathrm{I}(\boldsymbol{\beta})$ from Equation 6 and first-order bias $b_1(\boldsymbol{\beta})$ from Equation 10. Because $b_1(\boldsymbol{\beta}) = \mathrm{I}(\boldsymbol{\beta})^{-1}\boldsymbol{X}'\boldsymbol{W}\xi$, Equation 11 can be reformulated as $U^{PML}(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) - \boldsymbol{X}'\boldsymbol{W}\xi$ (Firth, 1993).

The penalized parameter estimate $\hat{\boldsymbol{\beta}}_k^{PML}$ ($k = 0, \ldots, K$)—free from first-order bias—is obtained by solving

(12)

$$U^{*}(\boldsymbol{\beta}_k) \equiv U(\boldsymbol{\beta}_k) + \frac{1}{2}\mathrm{trace}\left[\mathrm{I}(\boldsymbol{\beta})^{-1}\frac{\partial \mathrm{I}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k}\right] = 0$$

The corresponding penalized ML (PML) functions $l^{PML}(\boldsymbol{\beta})$ and $L^{PML}(\boldsymbol{\beta})$ are

(13)

$$l^{PML}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + \frac{1}{2}\ln|\mathrm{I}(\boldsymbol{\beta})|$$

(14)

$$L^{PML}(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|\mathrm{I}(\boldsymbol{\beta})|^{1/2}$$

where the penalty function $|\mathrm{I}(\boldsymbol{\beta})|^{1/2}$ represents Jeffrey's invariant prior for the problem, also well known as noninformative data-driven prior in Bayesian modeling. As Georg Heinze and Michael Schemper (2002) indicate, its influence is asymptotically negligible. Thus, $\hat{\boldsymbol{\beta}}^{ML}$ and $\hat{\boldsymbol{\beta}}^{PML}$ are asymptotically equivalent.

PML estimation is usually based on solving Equation 12 for all $k = 0, \ldots, K$ by applying an iteratively weighted least-squares algorithm.

Comparing the Performance of Approaches

Although most systematic comparisons between the introduced approaches do not focus on rareness of event occurrence but rather on obtaining unbiased estimates under varying sample size conditions, some general conclusions can be drawn that also apply to rare event situations. In an evaluation of *correction approaches*, S. B. Bull and colleagues (1994) contrasted the Taylor series approach with different variants of the resampling-based jackknife method (one-step, two-step, fully iterated) through Monte Carlo (MC) simulations. While the common one-step jackknife does not effectively reduce bias, the two other approaches perform quite well in moderate-sized samples but tend to overcorrect in small samples. However, jackknife methods are not recommended when the number of events per covariate (independent variable) is fewer than 20. Furthermore, more complex resampling-based approaches are not yet readily available for applied studies in statistical standard software packages, such as R or Stata. The Taylor series approach is similar in behavior, generating accurate parameter estimates in samples of large to moderate size but also tends to overcorrect bias in small samples, particularly when the probability of event occurrence is low (for similar evidence, see Leitgöb, 2013).

Regarding prevention approaches, Elizabeth N. King and Thomas P. Ryan (2002) investigated the small sample performance of CML estimation and asserted that, although clearly outperforming ML, the estimation error is larger than expected. In combination with severe estimation problems in large data sets with many (continuous) covariates—which is the standard case in, for example, sociological, epidemiological, and political research—CML does not represent a viable alternative to avoid biased parameter estimates. In contrast, Firth-based PML estimation appears to work well in small n and/or rare event conditions (e.g., Leitgöb, 2013; Rainey & McCaskey, 2015). Furthermore, it is much more efficient than the ML estimator, "thus, researchers do not face a bias-variance tradeoff when choosing between the ML and PML estimators—the PML estimator has a smaller bias *and* a smaller variance" (Rainey & McCaskey, 2015, p. 1). However, Rainer Puhr and colleagues (2017) indicated that PML introduces bias toward 1/2 in predicted probabilities and proposed simple modifications to obtain unbiased estimates (see Heinze and Schemper, 2002, regarding adequate statistical inference).

Finally, Heinz Leitgöb (2013) focused explicitly on the interplay of sample size and rare event bias when conducting MC simulations comparing the behavior of the Taylor series correction approach, ML estimation, and PML estimation. While even ML produces negligible bias in intercept and effect parameter estimates in case of $n = 5,000$ and $\pi \geq .01$, the approaches' performance differs greatly in more unfavorable situations.

This indicates that the low variance in Y is not the core problem of rare events bias but the low amount of absolute FI. Again, Taylor series correction results in marginally to moderately overcorrected estimates for $n \leq 1,000$; $\pi \leq .01$ and $n \leq 100$; $\pi \leq .1$. In contrast, the Firth-based PML approach can produce virtually unbiased estimates under these extreme data conditions. Thus, its application appears recommendable for rare event binary logit modeling. The procedure is available in R (packages *logistf* and *brglm*) and Stata (command *firthlogit*).

Number of Events per Variable

Additional bias in parameter estimates and respective standard errors arises when specifying ML-based logit models with an unfavorably small number of events per independent variable (EPV; e.g., Harrell et al., 1985). The phenomenon—very likely to occur when analyzing rare event outcomes—is called *overfitting* (or *overparameterization*). Generally, an overfitted model is more complex (in terms of specified parameters) than can be justified by the information available in the applied data and tends to fit idiosyncratic random noise besides reflecting the underlying DGP at population level. Furthermore, the ML estimator contributes to overfitting because it maximizes the likelihood of observing the data at hand. Although it is technically more accurate to consider the number of events per model parameter as relevant for bias rather than EPV (Wynants et al., 2015), here EPV is considered for reasons of consistency with pertinent terminology.

To counteract bias, Frank Harrell and colleagues (1985) were the first to propose a general guideline for the minimum number of EPVs. Based on theoretical considerations, they recommended 10–20 EPVs as the cutoff range. In contrast, Peter Peduzzi and colleagues (1996) relied on MC simulations of epidemiological data to derive a threshold value of 10 EPVs for dichotomous covariates. For fewer than 10 EPVs, they identified heavily biased parameter estimates in both directions as well as under- and overestimation of sampling variances, accompanied with poor confidence interval coverage. In a subsequent simulation study, Eric Vittinghoff and Charles E. McCulloch (2006) utilized a quasi-experimental design as underlying DGP to study the causal effect bias arising from the implementation of a large set of covariates necessary to adjust for confounding bias. Compared to Peduzzi and colleagues (1996), the simulation design allowed for higher generalizability through artificially generated data and the application of dichotomous as well as continuous covariates. For most simulation conditions, results indicate that substantial bias is uncommon with 5–9 EPVs and widely comparable in size with 10–16 EPVs. Thus, Vittinghoff and McCulloch (2006) concluded that "systematic discounting of results, in particular statistically significant associations, from any model with five to nine EPV does not appear to be justified" (p. 717) and advocated relaxing the rule of 10 EPVs to five EPVs. However, they also recognized that a low number of EPVs is not an isolated cause of bias but rather dependent on or interrelated with other factors such as the model's dimensionality, scale and variance of covariates (particularly low prevalences in dichotomous covariates), and total sample size. More recent MC studies have followed this line of argument and focused on the interplay between EPV and further factors: Delphine S. Courvoisier and colleagues (2011) demonstrated that even if the number of EPVs exceeds 10, bias is very likely to be substantial in case of high absolute values of effect parameters and high correlations between covariates. Beyond that, even for 20 or more EPVs, statistical power proved to be quite to extremely

low. L. Wynants and colleagues (2015) investigated the impact of a clustered data structure (sample elements are nested within contextual units, e.g., pupils are nested within classes) and total sample size on bias in parameter estimates and predictive performance for given EPV levels. While the amount of clustering appears to have no systematic influence on bias and the logit model's predictive performance (based on discrimination and calibration), simulations suggest that for a given EPV, larger sample sizes provide more accurate estimation and prediction results. This last finding is also in line with the simulation-based evidence reported by M. Van Smeden and colleagues (2016).

From the outlined state of research, it can be deduced that no general guideline for the minimum number of EPVs can guarantee accurate ML estimates of parameters in logit models. Rather, as Courvoisier and colleagues (2011) advocated, thorough a priori consideration of the expected data structure regarding the proportion of events under study, the number of covariates necessary for adequate model specification, distributions of and correlations between covariates as well as absolute values of effect parameters is necessary during the design phase to avoid severe bias. Although some approaches for sample size calculations that consider the number of and correlations between covariates exist (e.g., Væth & Skovlund, 2004), adequate sample size determination from a parameter bias perspective appears accomplishable only via MC simulations. The simulation approach allows for theory- and evidence-based generation of artificial data that correspond in structure to the expectable data. The objective is to identify the minimum sample size n' that maintains bias in parameter estimates stable below some defined acceptance threshold (e.g., 1% or 5% in terms of relative bias). From a frequentist perspective, it can further be investigated whether n' provides sufficient power to conduct adequate hypothesis testing. If this is not the case, sample size can be increased until both criteria are met.

From an analytical perspective, *changing the estimation method* appears effective to attenuate EPV bias without introducing confounding bias due to nonconsideration of relevant covariates. Van Smeden and colleagues (2016) compare the performance of the ML and Firth-based PML estimators under varying EPV conditions in a quasi-experimental simulation setting similar to that done by Vittinghoff and McCulloch (2006). Even for fewer than 10 EPVs and high absolute values in the effect parameter of a normally distributed covariate, PML yields estimates with relative bias considerably below 5%. With binary lasso (*l*east *a*bsolute *s*hrinkage and *s*election *o*perator) and ridge regression, Qingxia Chen and colleagues (2016) tested two further penalization methods that shrink effect parameter estimates toward zero. While the lasso was originally developed for covariate selection in high-dimensional settings, the primary scope of ridge regression is enhancing the stability, and therefore, precision of parameter estimates under multicollinearity. Both methods regularize estimates by introducing some constraint on the overall size of effect parameter estimates. Thus, maximizing the log-likelihood function from Equation 3 under restriction $\sum_{k=1}^{K} |\boldsymbol{\beta}_k| \leq t$ (L1 norm) results in lasso estimates and under $\sum_{k=1}^{K} \beta_K^2 \leq t$ (L2 norm) in ridge estimates, with $k = 1, \ldots, K$ labeling the K effect parameters specified in the linear predictor and $t \leq 0$ determines the threshold that controls the amount of shrinkage via its correspondence with the penalty or tuning parameter λ. From a

functional perspective, the two methods differ in the way they shrink parameter estimates. While the lasso tends to shrink some of the parameter estimates exactly to zero, the ridge estimator allows only for shrinkage to nonzero values. According to the results of simulating a quasi-experimental design with a dichotomous exposure variable under case–control sampling reported by Chen and colleagues (2016), both estimators perform well regarding relative bias in situations with EPV < 10, particularly when all but the effect parameter of the exposure variable are penalized. However, the shrinkage estimator's unnecessary model overfitting is not without consequences but leads to increased uncertainty in parameter estimates, reducing statistical power to detect nonzero exposure effects.

Problem of Separation

Researchers analyzing rare event data with small to moderate sample size may also be confronted with a phenomenon called the problem of *complete separation* (Albert & Anderson, 1984) or *monotone likelihood* (Bryson & Johnson, 1981). It arises in case of highly predictive covariates that perfectly separate occurrences and nonoccurrences of the event under study or some nontrivial linear combination of covariates and affects parameter estimation (Heinze & Schemper, 2002). In these situations, the likelihood function given in Equation 2 is no longer concave but rather a monotonically increasing function of (at least) one parameter that approaches its maximum as the respective parameter converges to infinity (Allison, 2008). Depending on whether the convergence threshold will be reached within the determined number of iterations, the maximization algorithm (e.g., Newton-Raphson) may even terminate but provide some arbitrary parameter estimate of finite size. Thus, the problem of separation can be considered as the inability of identifying finite population parameters under specific data conditions because of the nonexistence of the ML estimate (Heinze & Schemper, 2002).

Let the problem of separation be demonstrated by a simple numerical example. Assume the following 2 × 2 contingency table between some rare event outcome variable Y and an unbalanced covariate X, which may be interpreted as *risk* factor, indicating that an increase in X is associated with an increase in $P(Y = 1|X)$ (when changing columns, one can think of X as a *protective* factor):

Table 1. 2×2 Contingency table with perfect separation.

		X		
		1	0	∑
Y	1	5	0	5
	0	15	80	95
	∑	20	80	100

The absolute frequency in cell c_{12} ($X = 0 \wedge Y = 1$) equals zero. Thus, no individuals without exposure to risk factor X experience the event of interest. In contrast, all five individuals who experience the event are also exposed to X (located in cell c_{11} with $X = 1 \wedge Y = 1$). In case of a 2 × 2 contingency table, the logit model is saturated and the ML estimates $\hat{\beta}_0^{ML}$ and $\hat{\beta}_1^{ML}$ possess a *closed-form solution*:

(15)

$$\ln\left[\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right] = \underbrace{\ln\left[\frac{\Pr(Y=1|X=0)}{\Pr(Y=0|X=0)}\right]}_{\hat{\beta}_0^{ML}} + \underbrace{\ln\left[\frac{\Pr(Y=1|X=1)}{\Pr(Y=0|X=1)} \Big/ \frac{\Pr(Y=1|X=0)}{\Pr(Y=0|X=0)}\right]}_{\hat{\beta}_1^{ML}} X$$

Inserting the values from Table 1 into $\hat{\beta}_1$ from Equation 15 leads to $\hat{\beta}_1 = \ln\left(\frac{5/20}{15/20} \Big/ \frac{0/80}{80/80}\right) = \ln\left(\frac{80}{0}\right)$. Because a division by zero is not defined, $\hat{\beta}_1^{ML}$ does not exist.

However, whether perfectly separated data actually constitute an estimation problem depends on the nature of X. It is the case if β_1 is finite in size because individuals with $X = 0 \wedge Y = 1$, although not included in the sample, do exist in the population. In such a situation, the rationale behind separated data is that this group must inevitably be very small because rare event occurrence without being exposed to X —given that X indeed constitutes a relevant causal factor of Y —is extremely unlikely. Then, the sample inclusion probability of at least one of these individuals is close to zero and (random) sampling of separated data becomes very likely. Nonexistence of an ML-based solution for $\hat{\beta}_1$ is therefore *sampling induced* and gathering (much) more data from the population can resolve the problem.

Now assume that X satisfies the INUS condition proposed by J. L. Mackie (1965) and represents a cause defined as *insufficient* but *necessary* part of an *unnecessary* but *sufficient* condition. Let Y, the rare event under study, for example, be adolescent pregnancy. Then, any kind of sexual intercourse hypothetically possible to result in fertilization is in itself insufficient but necessary for pregnancy. Thus, adolescent pregnancy already in the population will be separated along sexual intercourse status because young females without such experiences belong deterministically to the nonpregnant group (given that artificial insemination is legally not permissible for adolescents). In other words, their probability of being pregnant is equal to zero. Thus, they do not contribute any variance to Y. Referring to count data modeling terminology, these young females have *structural zeros* in Y. Compared to the situation described previously, the population parameter β_1 is now actually infinite (or undefined) and so is its ML estimate.

Differentiating between situations with (i) sampling induced and (ii) structural causes of separated data allows a tailored analytical response to the phenomenon. Simply omitting the relevant covariate is in neither case a solution because leaving an obviously relevant part of the DGP unconsidered will introduce nothing but specification bias to the model (Zorn, 2005).

(i) In case of *sampling-induced separation*, gathering more data raises the probability of having only nonzero cells in the 2 × 2 contingency table. However, increasing the sample size under random sampling does

not guarantee nonseparated data. To overcome the problem, sampling designs based on selecting on the dependent variable (known as *choice-based* or *endogenous stratified* designs in econometrics and *case–control* designs in epidemiology) appear as a viable alternative (King & Zeng, 2001). The strategy implies stratifying the population with respect to Y and drawing random samples from the *cases* ($Y = 1$) as well as from the *controls* ($Y = 0$). As cases are more informative than controls in rare event data (for details, see King & Zeng, 2001), it may appear plausible to collect all cases available for gaining maximum efficiency. The oversampling of cases will result in a sampling distribution of Y that is systematically deviating from the underlying population distribution. Nonetheless, given that the sampling process for cases and controls is based on random or complete selection, which generates representative samples for the two subpopulations of the well-defined total population, selection on Y does not introduce selection bias in covariates $\boldsymbol{X}$ and the estimates of the logit model's effect parameters are still consistent (e.g., Breslow, 1996). Rather, this is not the case for $\hat{\beta}_0$, the estimate of the intercept parameter. Thus, if researchers are interested not only in $\hat{\beta}_k$ or the respective odds ratio $\exp\left(\hat{\beta}_k\right)$ but also in estimated event probabilities $\Pr\left(Y = 1 \mid \hat{\beta}\right)$ or in marginal effects $\partial \Pr\left(Y = 1|\hat{\boldsymbol{\beta}}\right) / \partial X_k$, bias correction for $\hat{\beta}_0$ becomes indispensable. King and Zeng (2001) proposed applying $\hat{\beta}_0^c$, a corrected estimate of $\hat{\beta}_0$, which is consistent for β_0, based on a priori information about τ, the fraction of events in the population

(16)

$$\hat{\beta}_0^c = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

with $\bar{y}$ as the observed fraction of ones in the sample. If information about τ is available, $\hat{\beta}_0^c$ can be easily computed according to Equation 16 after having obtained $\hat{\beta}_0$; otherwise, King and Zeng (2002) developed methods that allow valid inferences when the respective information is partially or completely absent. Note finally that ignorability of stratified sampling with respect to Y for $\hat{\beta}_k$ under the stated conditions is a unique property of the logit model and does not hold for other binary dependent variable models such as the probit or the cloglog.

As outlined, the strategy of selecting on the dependent variable requires strong a priori information. However, this information is often not accessible. For example, if the focus of research is on explaining sexual victimization as a dramatic life event, a considerable and highly selective number of offenses remain undetected because the victims refrain from reporting the incidents to the police for various reasons. Then, victimization status is not accessible for this specific subgroup of victims and selection based solely on accessible police records will introduce selection bias. In such a situation, generating large random samples from the total population of interest appears without alternatives. If sampling-induced separation still occurs, Heinze and Schemper (2002) proposed applying the Firth-based PML estimator as an analytical solution because it guarantees finite estimates in any case and, thus, eliminates the problem of separation. Although

exact methods based on CML estimation also provide finite parameter estimates in case of separated data, Heinze (2006) demonstrated the superiority of the PML estimator with respect to applicability, precision, and statistical inference.

(ii) For *structurally separated data*, sample homogenization by eliminating all *structural zeros* in Y ($Y = 0 \mid X = 0$ for INUS-based *risk* and $Y = 0 \mid X = 1$ for *protective* factors) appears as the method of choice. These cases are uninformative (Equation 6), as they do not contribute to variance in Y and exclusion rules out potential confounding bias in effect parameter estimates of other covariates in the model. The strategy appears particularly credible if the causal mechanism responsible for separation is well understood, such as the physiological processes in the adolescence pregnancy example provided earlier.

Model Specification

The logit model holds the symmetry property

(17)

$$g[\pi(\boldsymbol{x})] = -g[1 - \pi(\boldsymbol{x})]$$

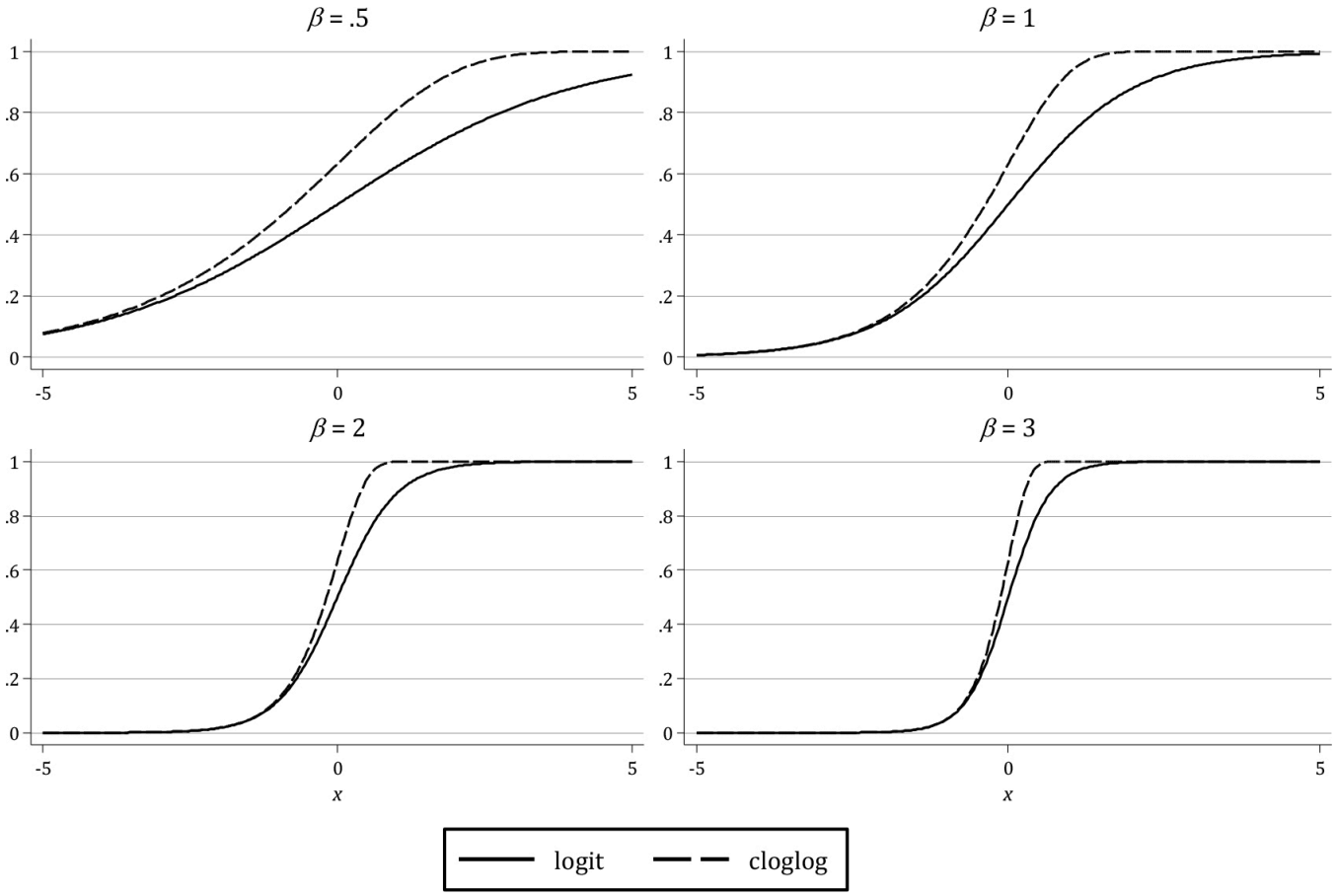
Thus, it has a symmetric sigmoid response curve for $\pi(\boldsymbol{x}_i)$, approaching the lower and upper limits at the same rate from the inflection point at $\pi = .5$ (Figure 1). However, several authors (e.g., Calabrese & Osmetti, 2013; Wang & Dey, 2010) have argued that the symmetry property, which implies that cases with zeros and ones contribute the same amount of information to the explanation of Y, may not be appropriate in case of highly imbalanced data. If the event under study is rare, the observed ones are more informative than the zeros. To optimize model fit and predictive accuracy, this fact should be considered explicitly in the estimation of the model (see Calabrese & Osmetti, 2013). This may be achieved by changing the link function from the logit to some noncanonical asymmetric function, such as the complementary log-log (cloglog) function with $g_{\text{cll}} = \ln\{-\ln[1 - \pi(\boldsymbol{x})]\}$. In CEP form, the cloglog model can be expressed in terms of the CDF of the standardized minimum extreme value (EV) Type I distribution (also referred to as Gumbel or log-Weibull distribution):

(18)

$$E[\Pr(y_i = 1|\boldsymbol{x}_i)] = \pi_i = 1 - \exp[-\exp(\eta_i)]$$

The response curve is specific in the sense that $\pi(\boldsymbol{x})$ asymptotically converges toward zero at a rather slow rate, while it approaches one quite sharply (Figure 1). With increasing absolute effect size β, it closely approximates the logit model's response curve in the lower tail. Hence, the cloglog model is more discriminatory particularly in the upper tail than the logit model and tends to produce higher probability estimates for given x (or η, respectively). For estimation purposes, Equation 18 can be inserted for π_i in Equation 3 instead of Equation 2.

Figure 1. Logit and cloglog model comparison ($\eta = \beta x$).



As the EV Type I distribution is a special case of the generalized EV (GEV) distribution when ξ = 0— $\xi \in R$ represents the GEV distribution's shape parameter that governs its tail behavior—Xia Wang and Dipak Dey (2010) proposed the application of the inverse of the standardized minimum GEV distribution's CDF with $g_{GEVmin} = -\ln[1-\pi(\boldsymbol{x})]^{-\xi} - 1 \,/\, \xi$ (for $\xi \neq 0$) as link function to introduce more flexibility to the response function's skewness to improve the model fit (see also Calabrese & Osmetti, 2013). The model has the following CEP specification:

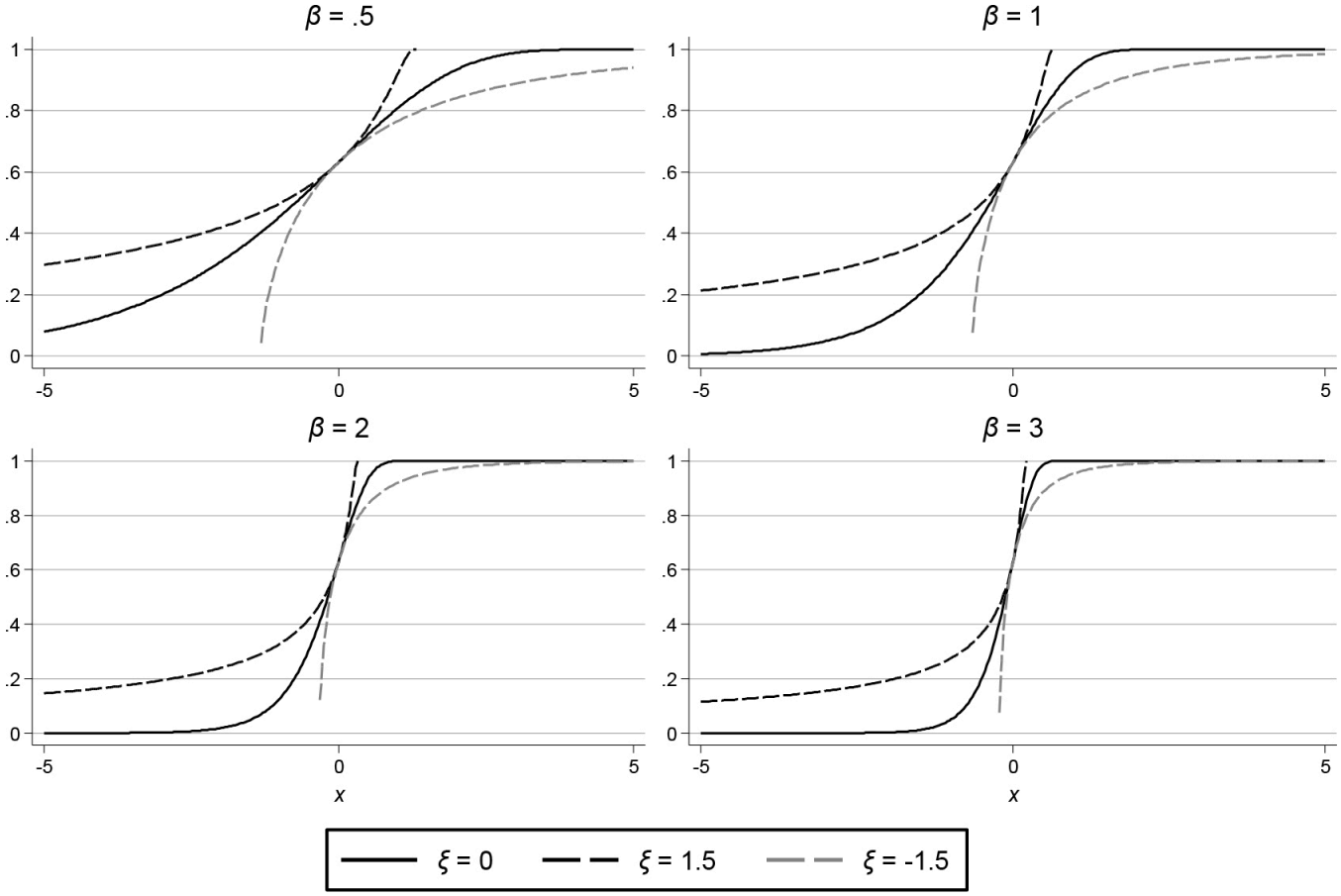
(19)

$$E\left[\Pr\left(y_i = 1|\boldsymbol{x}_i\right)\right] = \pi_i = 1 - \exp\left[-\left(1-\xi\eta_i\right)^{-\frac{1}{\xi}}\right]$$

The way variations in the shape parameter ξ affect the GEV model's response curve is demonstrated in Figure 2. Compared to the cloglog with ξ = 0 as reference model, its slope is steeper in the lower tail and flatter in the upper tail if ξ < 0, while the opposite is the case if ξ > 0.

Although it is—under strong regularity conditions—possible to obtain an ML estimator for ξ with asymptotic properties, Raffaella Calabrese and Silvia Angela Osmetti (2013) treated ξ as fixed. They proposed fitting various models with different prespecified values of ξ and selected the model with the highest predictive accuracy. In contrast, Wang and Day (2010) relied on the Bayesian approach to estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\xi}$.

Figure 2. GEV model for different combinations of β and ξ (η = β*x*).



Whether specifying some asymmetric response curve leads to a model that actually performs better than the logit model regarding model fit and predictive accuracy is finally an empirical question that has to be resolved through model comparison. Nonetheless, changing the link from logit to some asymmetric function has substantial interpretative implications. The term $\exp(\beta_k)$ then no longer represents an odds ratio, but in case of the cloglog model, a rate ratio (ratio between two incidence rates). However, because $\pi_i = 1 - \exp[-\exp(\eta_i)]$ and $1 - \pi_i = \exp[-\exp(\eta_i)]$, the cloglog model also allows for a computation of odds ratios.

Prospects

Substantial gains in the trustability of conclusions drawn from rare events analysis can be achieved by thorough research design considerations. Under given financial and time constraints, the primary design objective is to maximize the amount of FI in collected data to obtain parameter estimates as precise as possible. Respective measures include accurate sample size determination taking into account the number of EPVs to avoid overfitting and—if possible—the systematic reduction of imbalance in *Y* by oversampling events through the implementation of some case–control design and sample homogenization.

In addition, several analytical strategies support (or at least compensate design limitations when) drawing inferences from rare event data. As worked out, particularly the PML approach proposed by Firth (1993) appears promising as it (i) tends to adequately remove first-order finite sample bias even under rather

extreme sample size and rare event conditions, (ii) significantly reduces problems associated with an unfavorably small number of EPV, and (iii) fully eliminates the problem of (sampling-induced) separation. However, Ioannis Kosmidis and Firth (2009) indicated that PML functions do not exist for all GLMs with noncanonical links, for example, the probit and cloglog models. Furthermore, Andrew Gelman and colleagues (2008), Sander Greenland and Mohammad Ali Mansournia (2015), and Carlisle Rainey (2016) demonstrated that Jeffrey's data-driven prior may not represent the optimal choice because it contains too little or too much prior information, finally resulting in defective inferences from the binary logit model. From a frequentist perspective, Jeffrey's prior in PLM serves as a stabilization device to improve the repeated-sampling performance of an estimator (Cole et al., 2014). As formally derived by Rainey (2016), however, PML is equivalent to a Bayesian estimation approach with Jeffrey's prior as specified noninformative prior. Switching from frequentist to Bayesian probability theory allows for higher flexibility in prior specification to overcome the limitations associated with Jeffrey's prior in the PML approach. This matters particularly in case of separated data because for some large β, the posterior distribution—from which inferences are drawn—is dominated not by the data at hand but by the chosen prior distribution (Rainey, 2016). Despite the proposition of weakly informed priors based on the Cauchy (Gelman et al., 2008) or log-F distribution (Greenland & Mansournia, 2015) or empirical Bayes (objective) priors as superior alternatives to Jeffrey's prior (e.g., Rainey, 2016), it appears fruitful to learn more about the specification of appropriate priors in rare event situations—and about the potential of Bayesian rare event modeling in general. Notably, in the extreme case of having observed no (potentially occurring) events at all, a phenomenon discussed as the *zero-numerator problem* in the literature (e.g., Winkler et al., 2002), there is no alternative to the application of the Bayesian probability theory. Given $\hat{\pi} = 0$, the estimator of the frequentist standard error $\hat{\sigma}_{\hat{\pi}} = \sqrt{\hat{\pi}\left(1 - \hat{\pi}\right) \Big/ n} = 0$ and fails to reflect the uncertainty that the parameter estimate $\hat{\pi}$ is afflicted with. On the basis of the frequentist statistical inference, one would then assume *with certainty* that the *true* probability of event occurrence in the population equals zero.

Finally, the potential of modern computer-intensive methods, such as machine learning techniques including random forests, boosting, and support-vector machines, for analyzing rare events must be evaluated thoroughly.

Further Readings

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: Wiley.

Badi, N. H. S. (2017). Properties of the maximum likelihood estimates and bias reduction for logistic regression model. *Open Access Library Journal*, 4, e3625.

Beaujean, A. A. (2014). Sample size determination for regression models using Monte Carlo methods. *Practical Assessment, Research & Evaluation*, 19. Retrieved from http://pareonline.net/getvn.asp?n=12&v=19

Coles, G. S. (2004). *An introduction to statistical modeling of extreme values*. London, England: Springer.

Efron, B., & **Hastie, T.** (2016). *Computer age statistical inference*. New York, NY: Cambridge University Press.

Firth, D. (1992). Generalized linear models and Jeffreys priors: An iterative weighted least-squares approach. In **Y. Doge** & **J. Whittaker** (Eds.), *Computational statistics* (Vol. 1, pp. 553–557). Heidelberg, Germany: Physica-Verlag.

Gelman, A. (2009). Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24, 176–178.

Gelman A., **Carlin, J. B.**, **Stern, H. S.**, **Dunson, D. B.**, **Vehtari, A.**, & **Rubin, D. B.** (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

Hosmer, D. W., **Lemeshow, S.**, & **Sturdivant, R. X.** (2013). *Applied logistic regression*. New York, NY: Wiley.

MacKinnon, J. G., & **Smith, A. A.** (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, 85, 205–230.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72, 67–90.

References

Albert, A., & **Anderson, J. A.** (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1–10. doi:10.1093/biomet/71.1.1

Allison, P. D. (2008). Convergence failures in logistic regression [SAS Global Forum, Paper 360-2008]. Retrieved from http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Barndorff-Nielsen, O. E., & **Cox, D. R.** (1994). *Inference and asymptotics*. London, England: Chapman & Hall.

Blossfeld, H. P., & **Rohwer, G.** (1997). Causal inference, time and observation plans in the social sciences. *Quality & Quantity*, 31, 361–384. doi:10.1023/A:1004289932598

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91, 14–28.

Bryson, M. C., & **Johnson, M. E.** (1981). The incidence of monotone likelihood in the Cox model. *Technometrics*, 23, 381–383.

Bull, S. B., **Hauck, W. W.**, & **Greenwood, C. M. T.** (1994). Two-step jackknife bias reduction for logistic regression MLEs. *Communication in Statistics—Simulation and Computation*, 23, 59–88. doi:10.1080/03610919408813156

Calabrese, R., & **Osmetti, S. A.** (2013). Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model. *Journal of Applied Statistics*, 40, 1172–1188. doi:10.1080/02664763.2013.784894

Chen, Q., **Nian, H.**, **Zhu, Y.**, **Talbot H. K.**, **Griffin M. R.**, & **Harrell, F. E.** (2016). Too many covariates and too few cases?—A comparative study. *Statistics in Medicine*, 35, 4546–4558. doi:10.1002/sim.7021

Cole, S. R., **Chu, H.**, & **Greenland, S.** (2014). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, 179, 252–260.

Cordeiro, G. M., & **McCullagh, P.** (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society (Series B)*, 53, 629–643.

Courvoisier, D. S., **Combescure, C.**, **Agoritsas, T.**, **Gayet-Ageron, A.**, & **Perneger, T. V.** (2011). Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, 64, 993–1000. doi:10.1016/j.jclinepi.2011.06.013

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27–38. doi:10.1093/biomet/80.1.27

Gelman, A., **Jakulin, A.**, **Pittau, G. M.**, & **Su, Y.-S.** (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2, 1360–1383.

Greenland, S., & **Mansournia, M. A.** (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34, 3133–3143. doi:10.1002/sim.6537|

Harrell, F., **Lee, K. L.**, **Matchar, D. B.**, & **Reichert, T. A.** (1985). Regression models for prognostic prediction: Advantages, problems and suggested solutions. *Cancer Treatment Reports*, 69, 1071–1077.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, 25, 4216–4226.

Heinze, G., & **Schemper, M.** (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409–2419.

King, E. N., & **Ryan, T. P.** (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, 56, 163–170. doi:10.1198/00031300283

King, G., & **Zeng, L.** (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.

King, G., & **Zeng, L.** (2002). Estimating risk and rate levels, ratios and differences in case-control studies.

Statistics in Medicine, 21, 1409–1427.

Kosmidis, I. (2014). Bias in parametric estimation: Reduction and useful side-effects. *WIREs Computational Statistics*, 6, 185–196. doi:10.1002/wics.1296|

Kosmidis, I., & **Firth, D.** (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, 96, 793–804. doi:10.1093/biomet/asp055

Leitgöb, H. (2013). *The problem of modeling rare events in ML-based logistic regression—Assessing potential remedies via MC simulations*. Presentation given at the Conference of the European Survey Research Association in Ljubljana. Retrieved from https://www.europeansurveyresearch.org/conf/uploads/494/678/167/PresentationLeitg_b.pdf

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, 2, 261–264.

McCullagh, P., & **Nelder, J. A.** (1989). *Generalized linear models*. Boca Raton, FL: Chapman & Hall.

Mehta, C. R., & **Patel, N. R.** (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine*, 14, 2143–2160.

Peduzzi, P., **Concato, J.**, **Kemper, E.**, **Holford, T. H.**, & **Feinstein, A. R.** (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49, 1373–1379.

Puhr, R., **Heinze, G.**, **Nold, M. Lusa, L.** & **Geroldinger, A.** (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics in Medicine*, 36, 2302–2317. doi:10.1002/sim.7273

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis*, 24, 339–355. doi:10.1093/pan/mpw014

Rainey, C., & **McCaskey, K.** (2015). Estimating logit models with small samples. Retrieved from http://www.carlislerainey.com/papers/small.pdf

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, 2, 71–78.

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.** (2016). No rational for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16, 163. doi:10.1186/s12874-016-0267-3

Væth, M., & **Skovlund, E.** (2004). A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine*, 23, 1781–1792. doi:10.1002/sim.1753

Vittinghoff, E., & **McCulloch, C. E.** (2006). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165, 710–718. doi:10.1093/aje/kwk052

Wang, X., & **Dey, D. K.** (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoptions. *The Annals of Applied Statistics*, 4, 2000–2023.

Winkler, R. L., **Smith, J. E.**, & **Fryback, D. G.** (2002). The role of informative priors in zero-numerator problems: Being conservative versus being candid. *The American Statistician*, 56, 1–4. doi:10.1198/000313002753631295

Wynants, L., **Bouwmeester, W.**, **Moons, K. G. M.**, **Moerbeek, M.**, **Timmerman, D.**, **Van Huffel, S.**, … **Vergouwe, Y.** (2015). A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *Journal of Clinical Epidemiology*, 68, 1406–1414. doi:10.1016/j.jclinepi.2015.02.002

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, 13, 157–170. doi:10.1093/pan/mpi009