Statistics and Computing (2020) 30:43–59
https://doi.org/10.1007/s11222-019-09860-6

Mean and median bias reduction in generalized linear models

Ioannis Kosmidis[1,2] · Euloge Clovis Kenne Pagui[3] · Nicola Sartori[3]

Received: 5 September 2018 / Accepted: 2 February 2019 / Published online: 4 March 2019

Abstract
This paper presents an integrated framework for estimation and inference from generalized linear models using adjusted score equations that result in mean and median bias reduction. The framework unifies theoretical and methodological aspects of past research on mean bias reduction and accommodates, in a natural way, new advances on median bias reduction. General expressions for the adjusted score functions are derived in terms of quantities that are readily available in standard software for fitting generalized linear models. The resulting estimating equations are solved using a unifying quasi-Fisher scoring algorithm that is shown to be equivalent to iteratively reweighted least squares with appropriately adjusted working variates. Formal links between the iterations for mean and median bias reduction are established. Core model invariance properties are used to develop a novel mixed adjustment strategy when the estimation of a dispersion parameter is necessary. It is also shown how median bias reduction in multinomial logistic regression can be done using the equivalent Poisson log-linear model. The estimates coming out from mean and median bias reduction are found to overcome practical issues related to infinite estimates that can occur with positive probability in generalized linear models with multinomial or discrete responses, and can result in valid inferences even in the presence of a high-dimensional nuisance parameter.

Keywords Adjusted score equations · Data separation · Dispersion · Iterative reweighted least squares · Multinomial regression · Parameterization invariance

1 Introduction

The flexibility of generalized linear models (McCullagh and Nelder 1989) in handling count, categorical, positive and real-valued responses under a common modelling framework has not only made them a typical choice in applications but also the focus of much methodological research on their estimation and use in inference.

Suppose that $y_1, \ldots, y_n$ are observations on independent random variables $Y_1, \ldots, Y_n$, each with probability density or mass function of the exponential family form

$$f_{Y_i}(y; \theta_i, \phi) = \exp\left\{\frac{y\theta_i - b(\theta_i) - c_1(y)}{\phi/m_i} - \frac{1}{2}a\left(-\frac{m_i}{\phi}\right) + c_2(y)\right\}$$

for some sufficiently smooth functions $b(\cdot)$, $c_1(\cdot)$, $a(\cdot)$ and $c_2(\cdot)$, and fixed observation weights $m_1, \ldots, m_n$. The expected value and the variance of Y_i are then $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{var}(Y_i) = \phi b''(\theta_i)/m_i = \phi V(\mu_i)/m_i$, respectively, where $b'(\theta_i)$ and $b''(\theta_i)$ are the first two derivatives of $b(\theta_i)$. Compared to the normal distribution, exponential family models are generally heteroscedastic because the response variance depends on the mean through the variance function $V(\mu_i)$, and the dispersion parameter ϕ allows shrinking or inflating that contribution of the mean. A generalized linear model (GLM) links the mean μ_i to a linear predictor η_i through a monotone, sufficiently smooth link function

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11222-019-09860-6) contains supplementary material, which is available to authorized users.

✉ Ioannis Kosmidis
Ioannis.Kosmidis@warwick.ac.uk

Euloge Clovis Kenne Pagui
kenne@stat.unipd.it

Nicola Sartori
sartori@stat.unipd.it

[1] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[2] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

[3] Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 35121 Padova, Italy

Table 1 Clotting data

Parameter	B	RMSE	B^2/SD^2	PU	MAE	C
β_1	−0.33	16.15	0.04	50.42	12.87	89.26
						93.05*
β_2	0.36	23.09	0.02	49.61	18.46	88.87
						92.66*
β_3	0.06	4.69	0.01	49.73	3.74	89.62
						93.04*
β_4	−0.11	6.71	0.03	50.51	5.36	88.78
						92.47*
ϕ	−0.38	0.65	54.13	78.77	0.55	
	< 0.01*	0.67*	< 0.01*	55.61*	0.53*	

Estimated bias (B), root mean squared error (RMSE), percentage of underestimation (PU), mean absolute error (MAE) of maximum likelihood estimator and coverage of nominally 95% Wald-type confidence intervals (C), based on 10,000 samples under the ML fit. The summary B^2/SD^2 is the relative increase in mean squared error from its absolute minimum due to bias. The results include the same summaries of the moment-based estimator of ϕ (row marked with $*$). All reported figures are $\times 100$ of their actual value and < 0.01 is used for a value that is less than 0.01 in absolute value

$g(\mu_i) = \eta_i$ with $\eta_i = \sum_{t=1}^{p} \beta_t x_{it}$ where x_{it} is the (i,t)th component of a model matrix X, and $\beta = (\beta_1, \ldots, \beta_p)^\top$. An intercept parameter is typically included in the linear predictor, in which case $x_{i1} = 1$ for all $i \in \{1, \ldots, n\}$.

Estimation of the parameters of GLMs is commonly done using maximum likelihood (ML) because of the limiting guarantees that the ML estimator provides assuming that the model assumptions are adequate. Specifically, the ML estimator $(\hat{\beta}^\top, \hat{\phi})^\top$ is consistent, asymptotically unbiased and asymptotically efficient with a limiting normal distribution centred at the target parameter value and a variance–covariance matrix, given by the inverse of the Fisher information matrix, which is also the Cramér–Rao lower bound for the variance of unbiased estimators. These properties are used as reassurance that inferential procedures based on Wald, score or likelihood ratio statistics will perform well in large samples. Another reason that ML is the default estimation method for GLMs is that maximizing the likelihood can be conveniently performed by iteratively reweighted least squares (IWLS; Green 1984), requiring only standard algorithms for least squares and the evaluation of working weights and variates at each iteration.

Nevertheless, the properties of the ML estimator and of the associated inferential procedures that depend on its asymptotic normality may deteriorate for small or moderate sample sizes or, more generally, when the number of parameters is large relative to the number of observations.

Example 1.1 To illustrate the differences between finite sample and limiting behaviour of the ML estimator and associate inferential procedures, consider the data in McCullagh and Nelder (1989, Sect. 8.4.2) of mean blood clotting times in seconds for nine percentage concentrations of normal plasma and two lots of clotting agent. The plasma concentrations are 5, 10, 15, 20, 30, 40, 60, 80, 100, with corresponding clotting times 118, 58, 42, 35, 27, 25, 21, 19, 18 for the first lot, and 69, 35, 26, 21, 18, 16, 13, 12, 12 for the second lot, respectively. We fit a gamma GLM with $\log \mu_i = \sum_{t=1}^{4} \beta_t x_{it}$, where μ_i is the expectation of the ith clotting time, $x_{i1} = 1$, x_{i2} is 1 for the second lot and 0 otherwise, x_{i3} is the corresponding (log) plasma concentration and $x_{i4} = x_{i2}x_{i3}$ is an interaction term. The ML estimates are $\hat{\beta} = (5.503, -0.584, -0.602, 0.034)$ and $\hat{\phi} = 0.017$. Table 1 shows the estimated bias, root mean squared error, percentage of underestimation and mean absolute error of the ML estimator from 10,000 simulated samples at the ML estimates, with covariates values fixed as in the original sample. The table also includes the same summaries of the moment-based estimator of ϕ (see, for example, McCullagh and Nelder 1989, Sect. 8.3, and the `summary.glm` function in R). The ML estimator of the regression parameters illustrates good bias properties, with distributions that have a mode around the parameter value used for simulation. On the other hand, the ML estimator of the dispersion parameter is subject to severe bias, which inflates the mean squared error by 54.13% from its absolute minimum, and has a severely right skewed distribution. Note here that the latter observation holds for any monotone transformation of the dispersion parameter. The moment-based estimator on the other hand has a much smaller bias, probability of underestimation closer to 0.5, and its use delivers a marked improvement to the coverage of standard confidence intervals for all model parameters.

Improvements in first-order inference based on ML can be achieved in several ways. For instance, bootstrap methods guarantee both correction of bias and higher-order accurate inference. Alternatively, analytical methods derived from higher-order asymptotic expansions based on the likeli-

hood (see, for instance, Brazzale et al. 2007) have been found to result in accurate inference on model parameters. Nevertheless, bootstrap methods typically require intensive computation, and analytical methods, typically, require tedious, model-specific algebraic effort for their implementation. Furthermore, both bootstrap and analytical methods rely on the existence of the ML estimate, which is not always guaranteed. Such an example is GLMs with multinomial or discrete responses (Heinze and Schemper 2002; Kosmidis 2014b).

This paper presents a unified approach for mean and median bias reduction (BR) in GLMs using adjusted score functions (Firth 1993; Kosmidis and Firth 2009; and Kenne Pagui et al. 2017, respectively). Specifically, Firth (1993) and Kosmidis and Firth (2009) achieve higher-order BR of the ML estimator through the additive adjustment of the score equation. Kenne Pagui et al. (2017) use a similar approach in order to obtain component-wise higher-order median BR of the ML estimator, i.e. each component of the estimator has, to third order, the same probability of underestimating and overestimating the corresponding parameter component. We illustrate how those methods can be implemented without sacrificing the computational simplicity and the first-order inferential properties of the ML framework, and illustrate that they provide simple and practical solutions to the issue of boundary estimates in models with categorical responses.

Explicit, general formulae are derived for the adjusted score equations that produce higher-order mean and median unbiased estimators for GLMs. It is shown that, like ML, both mean and median BR can be conveniently performed by IWLS after the appropriate adjustment of the working variates for ML. Extensive empirical evidence illustrates that such an adjustment of IWLS leads to a stable estimation procedure even in case in which standard IWLS for ML estimation diverges.

Each method possesses invariance properties that can be more useful or less desirable depending on the GLM under consideration; the estimators resulting from mean BR (mean BR estimators, in short) are exactly invariant under linear transformations of the parameters in terms of the mean bias of the transformed estimators, which is useful, for example, when estimation and inference on arbitrary contrasts of the regression parameters is of interest. These invariance properties do not extend, though, to more general nonlinear transformations. On the other hand, median BR delivers estimators that are exactly invariant in terms of their improved median bias properties under general component-wise transformations of the parameters, which is useful, for example, when a dispersion parameter needs to be estimated from data. However, estimators from median BR are not invariant in terms of the median bias properties under more general transformations, for example, parameter contrasts. In order to combine the desirable invariance properties of each method when modelling with GLMs, we exploit the Fisher orthogonality (Cox and Reid 1987) of the mean and dispersion parameters to formally derive a novel mixed adjustment approach that delivers estimators of the regression parameters with improved mean bias and estimators for any unknown dispersion parameter with improved median bias.

Examples and simulation studies for various response distributions are used to demonstrate that both methods for BR are effective in achieving their respective goals and improve upon maximum likelihood, even in extreme settings characterized by high-dimensional nuisance parameters. Particular focus is given on special cases, like estimation of odds ratios from logistic regression models and estimation of log odds ratios from multinomial baseline category models.

All methods and algorithms discussed in this paper are implemented in the `brglm2` R package (Kosmidis 2018), which has been used for all numerical computations and simulation experiments (see Supplementary Material).

The remaining of the paper is structured as follows. Section 2 gives a brief introduction to estimation using IWLS and shows how IWLS can be readily adjusted to perform mean or median BR. In particular, Sects. 2.1 and 2.2 review known results of ML estimation and explicit mean bias correction in generalized linear models. These subsections are useful to set up the notation and allow the introduction of mean and median bias-reducing adjusted score functions in Sects. 2.3 and 2.4, respectively. Inferential procedures based on the bias-reduced estimators are discussed in Sect. 3. Section 4 motivates the need for and introduces the mixed adjustment strategy for GLMs with a dispersion parameter. All methods are then assessed and compared through case studies and simulation experiments in Sects. 5 and 6. Section 6 also discusses how multinomial logistic regression models can be easily estimated with all methods using the equivalent Poisson log-linear model. Section 7 concludes the paper with a short discussion and possible extensions.

2 Bias reduction and iteratively reweighted least squares

2.1 Iteratively reweighted least squares

The log-likelihood function for a GLM is $\sum_{i=1}^{n} \log f_{Y_i}(y_i; g^{-1}(\eta_i), \phi)$, where $g^{-1}(\cdot)$ is the inverse of the link function. Suppressing the dependence of the various quantities on the model parameters and the data, the derivatives of the log-likelihood function with respect to the components of β and ϕ are

Table 2 Working variates for ML and additional quantities needed in mean and median BR, for the most popular combinations of distributions and link functions

Distribution	η	ML	Mean BR	Median BR
		$\eta + (y-\mu)/d$	$\phi\xi$	$dv'/(6v) - d'/(2d)$
Normal	μ	y	0	0
Binomial	$\log\dfrac{\mu}{1-\mu}$	$\eta + \dfrac{y-\mu}{\mu(1-\mu)}$	$\dfrac{h\{e^{\eta}-e^{-\eta}\}}{2m}$	$\dfrac{2(1-e^{\eta})}{3(1+e^{\eta})}$
	$\Phi^{-1}(\mu)$	$\eta + \dfrac{y-\mu}{\phi(\eta)}$	$-\dfrac{h\eta\{\Phi(\eta)(1-\Phi(\eta))\}}{2m\phi(\eta)^2}$	$\dfrac{\phi(\eta)(1-2\Phi(\eta))}{6\Phi(\eta)(1-\Phi(\eta))} + \dfrac{\eta}{2}$
	$\log\{-\log(1-\mu)\}$	$\eta + \dfrac{y-\mu}{e^{\eta-e^{\eta}}}$	$\dfrac{h\mu\{1-e^{\eta}\}}{2me^{2\eta-e^{\eta}}}$	$\dfrac{-e^{\eta-e^{\eta}}+2e^{\eta}+3e^{-e^{\eta}}-3}{6(1-e^{-e^{\eta}})}$
Gamma	$\dfrac{1}{\mu}$	$\eta - \dfrac{y-\mu}{\mu^2}$	$-\dfrac{h\eta\phi}{m}$	$\dfrac{2}{3\eta}$
	$\log\mu$	$\eta + \dfrac{y-\mu}{\mu}$	$\dfrac{h\phi}{2m\eta e^{2\eta}}$	$-\dfrac{1}{6}$
Poisson	$\sqrt{\mu}$	$\eta + \dfrac{y-\mu}{2\eta}$	$\dfrac{h\eta}{2m}$	$\dfrac{3}{2\eta}$
	$\log\mu$	$\eta + \dfrac{y-\mu}{\mu}$	$\dfrac{h}{2me^{\eta}}$	$-\dfrac{1}{3}$

$$s_\beta = \frac{1}{\phi} X^T W D^{-1}(y-\mu) \quad \text{and} \quad s_\phi = \frac{1}{2\phi^2}\sum_{i=1}^{n}(q_i - \rho_i), \tag{1}$$

respectively, with $y = (y_1, \ldots, y_n)^\top$, $\mu = (\mu_1, \ldots, \mu_n)^\top$, $W = \text{diag}\{w_1, \ldots, w_n\}$ and $D = \text{diag}\{d_1, \ldots, d_n\}$, where $w_i = m_i d_i^2 / v_i$ is the ith working weight, with $d_i = d\mu_i/d\eta_i$ and $v_i = V(\mu_i)$. Furthermore, $q_i = -2m_i\{y_i\theta_i - b(\theta_i) - c_1(y_i)\}$ and $\rho_i = m_i a_i'$ are the ith deviance residual and its expectation, respectively, with $a_i' = a'(-m_i/\phi)$, where $a'(u) = \mathrm{d}a(u)/\mathrm{d}u$.

The ML estimators $\hat\beta$ of β and $\hat\phi$ of ϕ can be found by solution of the score equations $s_\beta = 0_p$ and $s_\phi = 0$, where 0_p is a p-dimensional vector of zeros. Wedderburn (1976) derives necessary and sufficient conditions for the existence and uniqueness of the ML estimator of $\hat\beta$. Given that the dispersion parameter ϕ appears in the expression for s_β in (1) only multiplicatively, the ML estimate of β can be computed without knowledge of the value of ϕ. This fact is exploited in popular software like the `glm.fit` function in R (R Core Team 2018). The jth iteration of IWLS updates the current iterate $\beta^{(j)}$ for β by solving the weighted least squares problem

$$\left(X^\top W^{(j)} X\right)^{-1} X^\top W^{(j)} z^{(j)}, \tag{2}$$

where the superscript (j) indicates evaluation at $\beta^{(j)}$, and $z = (z_1, \ldots, z_n)^\top$ is the vector of "working" variates with $z_i = \eta_i + (y_i - \mu_i)/d_i$ (Green 1984). Table 2 reports the working variates for well-used combinations of exponential family models and link functions. The updated β from the weighted least squares problem in (2) is equal to the updated β from the Fisher scoring step

$$\beta^{(j)} + \left\{i_{\beta\beta}^{(j)}\right\}^{-1} s_\beta^{(j)},$$

where $i_{\beta\beta}$ is the (β,β) block of the expected information matrix about β and ϕ

$$i = \begin{bmatrix} i_{\beta\beta} & 0_p \\ 0_p^\top & i_{\phi\phi} \end{bmatrix} = \begin{bmatrix} \frac{1}{\phi} X^\top W X & 0_p \\ 0_p^\top & \frac{1}{2\phi^4}\sum_{i=1}^n m_i^2 a_i'' \end{bmatrix}, \tag{3}$$

with $a_i'' = a''(-m_i/\phi)$, where $a''(u) = \mathrm{d}^2 a(u)/\mathrm{d}u^2$.

2.2 Explicit mean bias reduction

Efron (1975) has shown that under the usual regularity conditions, the asymptotic mean bias of the ML estimator $\hat\gamma$ for a general parametric model $\mathcal{M}_\gamma$ can be reduced by the explicit correction of $\hat\gamma$ as $\tilde\gamma = \hat\gamma - b_\gamma(\hat\gamma)$, where $b_\gamma \equiv b_\gamma(\gamma)$ is the first term in the expansion of the mean bias of $\hat\gamma$. Kosmidis (2014a) provides a review of explicit and implicit methods for mean BR. The general form of b_γ is given in Cox and Snell (1968) in index notation and in Kosmidis and Firth (2010, Section 2) in matrix notation. For GLMs, $b_\beta = -i_{\beta\beta}^{-1} A_\beta^*$ and $b_\phi = -i_{\phi\phi}^{-1} A_\phi^*$ with

$$A_\beta^* = X^\top W \xi \quad \text{and} \quad A_\phi^* = \frac{(p-2)}{2\phi} + \frac{\sum_{i=1}^n m_i^3 a_i'''}{2\phi^2 \sum_{i=1}^n m_i^2 a_i''}, \tag{4}$$

where $\xi = (\xi_1, \ldots, \xi_n)^T$ with $\xi_i = h_i d_i'/(2d_i w_i)$ and $d_i' = \mathrm{d}^2\mu_i/\mathrm{d}\eta_i^2$, h_i is the "hat" value for the ith observation, obtained as the ith diagonal element of the matrix $H = X(X^\top WX)^{-1}X^\top W$, and $a_i''' = a'''(-m_i/\phi)$, with $a'''(u) = \mathrm{d}^3a(u)/\mathrm{d}u^3$. The derivation of b_ϕ above is done using Kosmidis and Firth (2010, expressions (4.8) in Remark 3) to write b_ϕ in terms of the first term in the expansion of the bias of $1/\hat{\phi}$, which is given in Cordeiro and McCullagh (1991).

Note here that neither $i_{\phi\phi}$ nor A_ϕ^* depend on β, and hence, the bias-reduced estimator for ϕ can be computed by knowledge of $\hat{\phi}$ only as

$$\hat{\phi}\left\{1 + \hat{\phi}\frac{\sum m_i^3\hat{a}_i'''}{\left(\sum m_i^2\hat{a}_i''\right)^2} + \hat{\phi}^2\frac{p-2}{\sum m_i^2\hat{a}_i''}\right\},$$

where $\hat{a}_i''' = a'''(-m_i/\hat{\phi})$. Some algebra gives that the bias-reduced estimator for β is

$$\left(X^\top\hat{W}X\right)^{-1}X^\top\hat{W}\left(\hat{z} + \hat{\phi}\hat{\xi}\right), \tag{5}$$

where $\hat{B}$ denotes evaluation of B at the ML estimator. Equivalently, and as also noted in Cordeiro and McCullagh (1991), the explicit correction $\hat{\beta} - b_\beta(\hat{\beta}, \hat{\phi})$ can be performed by IWLS as in (2) up to convergence, and then making one extra step, where the working variate z is replaced by its adjusted version $z + \phi\xi$. Table 2 gives the quantity $\phi\xi$ for some well-used GLMs.

2.3 Mean bias-reducing adjusted score functions

Firth (1993) shows that the solution of the adjusted score equations

$$s_\beta + A_\beta^* = 0_p \quad \text{and} \quad s_\phi + A_\phi^* = 0 \tag{6}$$

with A_β^* and A_ϕ^* as in (4) result in estimators β^* and ϕ^* with mean bias of smaller asymptotic order than the ML estimator.

A natural way to solve the adjusted score equations is through quasi-Fisher scoring (see Kosmidis and Firth 2010, for the corresponding quasi-Newton–Raphson iteration), where at the jth step the values for β and ϕ are updated as

$$\begin{aligned}\beta^{(j+1)} &\leftarrow \beta^{(j)} + \left\{i_{\beta\beta}^{(j)}\right\}^{-1}s_\beta^{(j)} - b_\beta^{(j)},\\ \phi^{(j+1)} &\leftarrow \phi^{(j)} + \left\{i_{\phi\phi}^{(j)}\right\}^{-1}s_\phi^{(j)} - b_\phi^{(j)}.\end{aligned} \tag{7}$$

The term "quasi-" here reflects the fact that the expectation of the negative second derivatives of the scores, instead of the adjusted scores, is used for the calculation of the step size. Setting $\phi^{(j+1)} - \phi^{(j)} = 0$ in the above iteration shows that it has the required stationary point. Furthermore, if the starting values $\beta^{(0)}$ and $\phi^{(0)}$ for iteration (7) are the ML estimates, then $\beta^{(1)}$ and $\phi^{(1)}$ are the estimates from explicit BR, because $s_\beta^{(0)} = 0_p$ and $s_\phi^{(0)} = 0$. Figure 1 illustrates the quasi-Fisher scoring iterations for an one-parameter problem, starting from the ML estimate.

A similar calculation to that in Sect. 2.2 can be used to show that (7) can be written in terms of an IWLS step for β and an appropriate update for ϕ. In particular,

$$\begin{aligned}\beta^{(j+1)} &\leftarrow \left(X^\top W^{(j)}X\right)^{-1}X^\top W^{(j)}\left(z^{(j)} + \phi^{(j)}\xi^{(j)}\right),\\ \phi^{(j+1)} &\leftarrow \phi^{(j)}\left\{1 + \phi^{(j)}\frac{\sum\left(q_i^{(j)} - \rho_i^{(j)}\right)}{\sum m_i^2 a_i''^{(j)}}\right.\\ &\qquad \left. + \phi^{(j)}\frac{\sum m_i^3 a_i'''^{(j)}}{\left(\sum m_i^2 a_i''^{(j)}\right)^2} + \left(\phi^{(j)}\right)^2\frac{p-2}{\sum m_i^2 a_i''^{(j)}}\right\}.\end{aligned} \tag{8}$$

Expression 8 makes apparent that, in contrast to ML, solving the mean bias-reducing adjusted score functions in GLMs with unknown dispersion parameter involves updating β and ϕ simultaneously. This is because b_β generally depends on ϕ.

Despite that the stationary point of the iterative scheme (8) is the mean BR estimates, there is no theoretical guarantee for its convergence for general GLMs. However, substantial empirical studies have shown no evidence of divergence, even in cases in which standard IWLS (2) fails to converge. Some of those empirical studies are presented in Sects. 4, 5 and 6 of the present paper.

2.4 Median bias-reducing adjusted score functions

Kenne Pagui et al. (2017) introduce a family of adjusted score functions whose solution has smaller median bias than the ML estimator. Specifically, the solution $\gamma^\dagger$ of $s_\gamma + A_\gamma^\dagger = 0$ is such that each of its components has probability $1/2$ of underestimating the corresponding component of the parameter γ with an error of order $O(n^{-3/2})$, as opposed to the error of order $O(n^{-1/2})$ for $\hat{\gamma}$. A useful property of the method is that it is invariant under component-wise monotone reparameterizations in terms of the improved median bias properties of the resulting estimators.

Some tedious but straightforward algebra starting from Kenne Pagui et al. (2017, expression (10)) gives that the median bias-reducing adjustments $A_\beta^\dagger$ and $A_\phi^\dagger$ for GLMs have the form

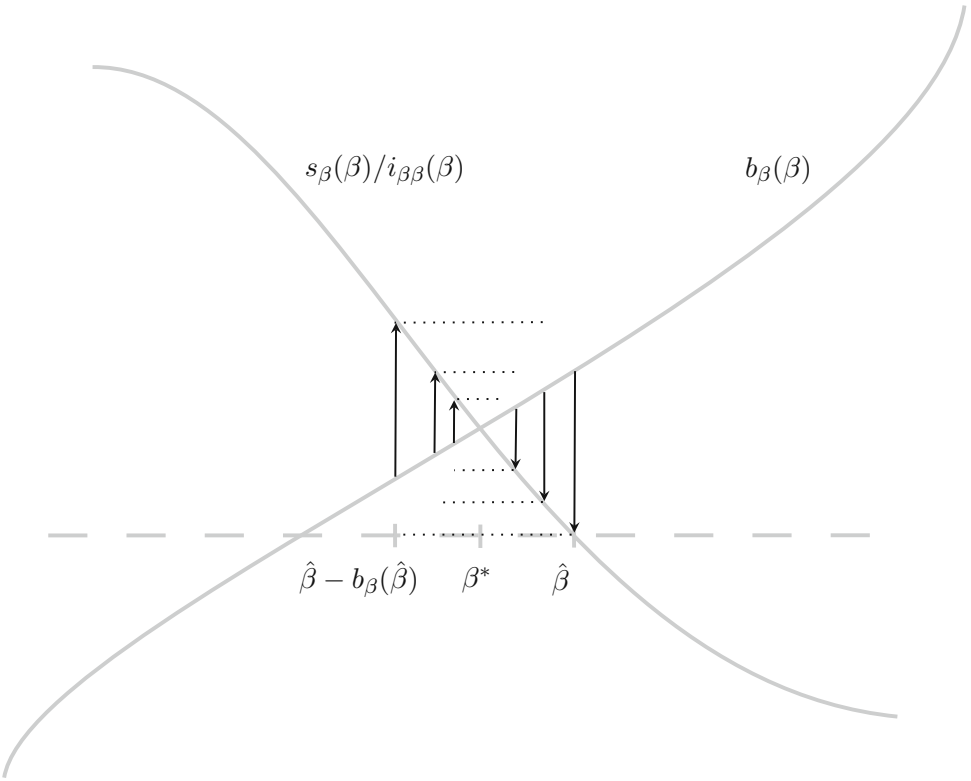

Fig. 1 Illustration of the quasi-Fisher scoring iterations for a model with a scalar parameter β, starting at the maximum likelihood estimate $\hat\beta$. One step gives the explicit mean reduced-bias estimator $\hat\beta - b_\beta(\hat\beta)$ of Sect. 2.2, and iterating until convergence results in the solution β^* of the mean bias-reducing adjusted score equation

$$A_\beta^\dagger = X^\top W(\xi + Xu) \quad \text{and}$$
$$A_\phi^\dagger = \frac{p}{2\phi} + \frac{\sum_{i=1}^n m_i^3 a_i'''}{6\phi^2 \sum_{i=1}^n m_i^2 a_i''}, \tag{9}$$

where $u = (u_1, \ldots, u_p)^\top$ with

$$u_j = [(X^\top W X)^{-1}]_j^\top X^\top \begin{bmatrix} \tilde h_{j,1}\left\{d_1 v_1'/(6v_1) - d_1'/(2d_1)\right\} \\ \vdots \\ \tilde h_{j,n}\left\{d_n v_n'/(6v_n) - d_n'/(2d_n)\right\} \end{bmatrix}. \tag{10}$$

In the above expressions, $[B]_j$ denotes the jth row of matrix B as a column vector, $v_i' = \mathrm{dV}(\mu_i)/\mathrm{d}\,\mu_i$, and $\tilde h_{j,i}$ is the ith diagonal element of XK_jX^TW, with

$$K_j = [(X^\top W X)^{-1}]_j [(X^\top W X)^{-1}]_j^\top / [(X^\top W X)^{-1}]_{jj},$$

and where $[B]_{jj}$ denotes the $(j, j)th$ element of a generic matrix B.

Similarly to the case of mean BR, the median bias-reducing adjusted score equations can be solved using quasi-Fisher scoring or equivalently IWLS, where at the jth iteration

$$\beta^{(j+1)} \leftarrow \left(X^\top W^{(j)} X\right)^{-1} X^\top W^{(j)} \left(z^{(j)} + \phi^{(j)}\xi^{(j)}\right) + \phi^{(j)} u^{(j)},$$
$$\phi^{(j+1)} \leftarrow \phi^{(j)} \left\{ 1 + \phi^{(j)} \frac{\sum \left(q_i^{(j)} - \rho_i^{(j)}\right)}{\sum m_i^2 a_i''^{(j)}} + \phi^{(j)} \frac{\sum m_i^3 a_i'''^{(j)}}{3\left(\sum m_i^2 a_i''^{(j)}\right)^2} + \left(\phi^{(j)}\right)^2 \frac{p}{\sum m_i^2 a_i''^{(j)}} \right\}. \tag{11}$$

Note here that the working variate for median BR is the one for mean BR plus the extra term ϕXu. Equivalently, and since the extra term is in the column space of X, the median BR IWLS update for β consists of a mean BR update for β as in (8), and a translation of the result by ϕu. Figure 2 illustrates that procedure. The core quantities in the definition of u are $d_i v_i'/(6v_i) - d_i'/(2d_i)$ in expression (10), and Table 2 includes their expressions for some well-used GLMs.

Similarly to (8), there is no theoretical guarantee for the convergence of the iterative scheme (11) for general GLMs. However, even in this case, our extensive empirical studies have produced no evidence of divergence.

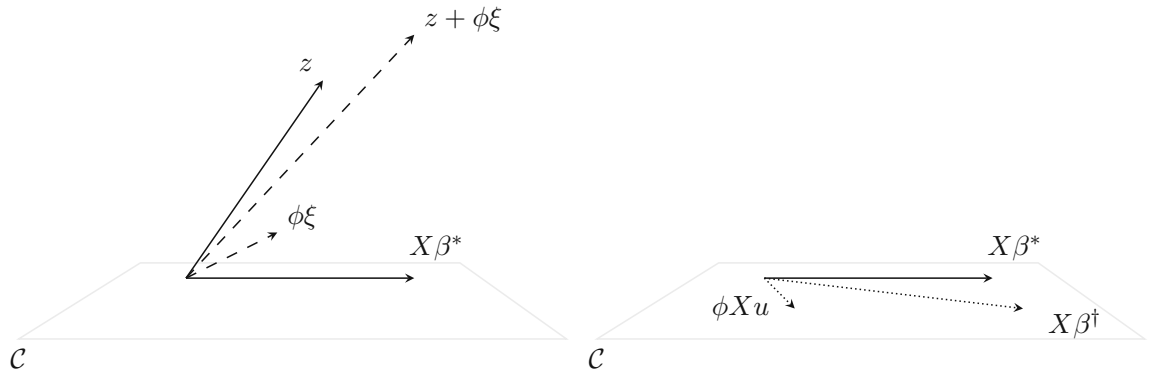

Fig. 2 Illustration of the IWLS update for computing the iterates of β for a given ϕ when performing mean BR and median BR . All quantities in the figure should be understood as being pre-multiplied by $W^{1/2}$. The left figure shows the addition of $\phi\xi$ to the maximum likelihood working variates z and the subsequent projection onto $\mathcal{C}$ (the column space of $W^{1/2}X$) that gives the updated value for the mean BR estimates β^*. The right figure illustrates the addition of ϕu on β^* to give the updated value for the median BR estimates $\beta^\dagger$

3 Inference with mean and median bias reduction

3.1 Wald-type inference by plug-in

According to the results in Firth (1993) and Kenne Pagui et al. (2017), both θ^* and $\theta^\dagger$ have the same asymptotic distribution as the ML estimator and hence are all asymptotically unbiased and efficient. Hence, the distribution of those estimators for finite samples can be approximated by a normal with mean θ and variance–covariance matrix $\{i(\theta)\}^{-1}$, where $i(\theta)$ is given in (3). The derivation of this result relies on the fact that both A^*_θ and $A^\dagger_\theta$ are of order $O(1)$ and hence dominated by the score function as information increases.

The implication of the above results is that standard errors for the components of θ^* and $\theta^\dagger$ can be computed as for the ML estimator, using the square roots of the diagonal elements of $\{i(\beta^*, \phi^*)\}^{-1}$ and $\{i(\beta^\dagger, \phi^\dagger)\}^{-1}$, respectively. As a result, first-order inference like standard Wald tests and Wald-type confidence intervals and regions are constructed in a plug-in fashion, by replacing the ML estimates with the mean BR or median BR estimates in the usual procedures in standard software.

Of course, for finite samples, Wald-type procedures based on the use of ML, mean and median bias reduction will yield different results. Such differences will disappear as the samples size increases. Sect. 3.2 explores those differences in normal linear regression models.

3.2 Normal linear regression models

Consider a normal regression model with $y_1, \ldots, y_n$ realizations of independent random variables $Y_1, \ldots, Y_n$ where Y_i has a $N(\mu_i, \phi/m_i)$ $(i = 1, \ldots, n)$ with $\mu_i = \eta_i = \sum_{t=1}^p \beta_t x_{it}$. The adjustment terms A^*_β and $A^\dagger_\beta$ are zero for this model. As a result, the ML, mean BR and median BR estimators of β coincide with the least squares estimator $(X^\top MX)^{-1}X^\top My$, where $M = \text{diag}\{m_1, \ldots, m_n\}$. On the other hand, the ML, mean BR and median BR estimators for ϕ are $\hat\phi = \sum_{i=1}^n (y_i - \hat\mu_i)^2/n$, $\phi^* = \sum_{i=1}^n (y_i - \hat\mu_i)^2/(n-p)$ and $\phi^\dagger = \sum_{i=1}^n (y_i - \hat\mu_i)^2/(n - p - 2/3)$.

The estimator ϕ^* is mean unbiased for ϕ, and for this reason, it is the default choice for estimating the precision parameter in normal linear regression models. On the other hand, and as shown by Theorem 3.1, the use of $\phi^\dagger$ for Wald-type inference about β_j based on asymptotic normality leads to inferences that are closer to the exact ones, based on the Student t_{n-p} distribution, than when ϕ^* is used, for all practically relevant values of $n - p$ and α.

Let $\hat I_{1-\alpha} = \{\hat\beta_j \pm z_{1-\alpha/2}\,(\kappa_j\,\hat\phi)^{1/2}\}$, $I^*_{1-\alpha} = \{\hat\beta_j \pm z_{1-\alpha/2}\,(\kappa_j\,\phi^*)^{1/2}\}$ and $I^\dagger_{1-\alpha} = \{\hat\beta_j \pm z_{1-\alpha/2}\,(\kappa_j\,\phi^\dagger)^{1/2}\}$ be the Wald-type confidence intervals for β_j of nominal level $1 - \alpha$, based on the asymptotic normal distribution of $\hat\beta$, β^* and $\beta^\dagger$, respectively, where z_α is the quantile of level α of the standard normal and $\kappa_j = [(X^\top MX)^{-1}]_{jj}$. Let also $I^E_{1-\alpha} = \{\hat\beta_j \pm t_{n-p;1-\alpha/2}\,(\kappa_j\,\phi^*)^{1/2}\}$ be the confidence interval of exact level $1 - \alpha$ for β_j, where $t_{n-p;\alpha}$ is the quantile of level α of the Student t distribution with $n - p$ degrees of freedom, and define $\text{Len}(I)$ to be the length of interval I.

Theorem 3.1 *For $n - p \geq 1$ and $\alpha \in (0, 1)$, $\hat I_{1-\alpha} \subset I^*_{1-\alpha} \subset I^E_{1-\alpha}$ and $I^*_{1-\alpha} \subset I^\dagger_{1-\alpha}$. Moreover, for $n - p \geq 1$ and $0 < \alpha < 0.35562$, $I^\dagger_{1-\alpha} \subset I^E_{1-\alpha}$.*

Finally, for $n - p > 1$ and $\alpha \in (0, 1)$

$$\left|\text{Len}(I^\dagger_{1-\alpha}) - \text{Len}(I^E_{1-\alpha})\right| < \left|\text{Len}(I^*_{1-\alpha}) - \text{Len}(I^E_{1-\alpha})\right|.$$

Table 3 Alternative, equivalent parameterizations of a gamma regression model with independent responses $Y_1, \ldots, Y_{12}$ where, conditionally on covariates, each Y_i has a gamma distribution with mean $\mu_i = \exp(\eta_i)$ and variance $\phi\mu_i^2$

Parameterization	Predictor η_i	Dispersion ϕ	Parameter vector
I	$\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 t_i$	ϕ	$(\beta_1, \beta_2, \beta_3, \beta_4, \phi)^\top$
II	$\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 t_i$	e^ζ	$(\beta_1, \beta_2, \beta_3, \beta_4, \zeta)^\top$
III	$\gamma_1 + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \beta_4 t_i$	ϕ	$(\gamma_1, \gamma_2, \gamma_3, \beta_4, \phi)^\top$

The covariates x_{i1}, x_{i2} and x_{i3} encode the levels of a three-level categorical covariate s_i as follows: x_{i1} is 1 for $i = 1, 2, 3, 4$ and 0, otherwise, x_{i2} is 1 for $i = 5, 6, 7, 8$ and 0, otherwise, and x_{i3} is 1 for $i = 9, 10, 11, 12$ and 0, otherwise. The covariate values $t_1, \ldots, t_{12}$ are generated from an exponential distribution with rate 1

If $n - p = 1$, the latter inequality holds for any $0 < \alpha < 0.62647$.

The proof of Theorem 3.1 is given in Appendix.

Exact inferential solutions are not generally available for other GLMs with unknown dispersion parameter. It is therefore of interest to investigate whether the desirable behaviour of inference based on the median BR estimator, as demonstrated in Theorem 3.1 for the normal linear regression model, is preserved, at least approximately, in other models. Section 5.2 considers an example with gamma regression.

4 Mixed adjustments for dispersion models

In contrast to ML, mean BR is inherently not invariant to general transformations of the model parameters, in terms of its smaller asymptotic mean bias properties. This imposes a level of arbitrariness when carrying out inference on β in GLMs with unknown dispersion parameters, mainly because ϕ appears as a factor on the variance–covariance matrix $\{i(\beta, \phi)\}^{-1}$ of the estimators. For example, standard errors for β^* will be different if the bias is reduced for ϕ or $1/\phi$. The mean BR estimates are exactly invariant under general affine transformations, which is useful in regressions that involve categorical covariates where invariance under parameter contrasts is, typically, required. On the other hand, median BR is invariant, in terms of smaller asymptotic median bias, under component-wise monotone transformations of the parameters, but it is not invariant under more general parameter transformations, like parameter contrasts.

In order to best exploit the invariance properties of each method, we propose the default use of a mixed adjustment that combines the mean bias-reducing adjusted score for β with the median bias-reducing adjusted score for ϕ by jointly solving

$$s_\beta + A_\beta^* = 0_p \quad \text{and} \quad s_\phi + A_\phi^\dagger = 0$$

with A_β^* and $A_\phi^\dagger$ as in expressions (4) and (9), respectively. For GLMs with known ϕ, like Poisson or Binomial models, the mixed adjustment results in mean BR. On the contrary, for the normal linear models of Sect. 3.2 the mixed adjustment results in median BR because $A_\beta^* = A_\beta^\dagger = 0_p$.

For general GLMs with unknown ϕ, the mixed adjustment provides the estimators $\beta^\ddagger$ and $\phi^\ddagger$, which are asymptotically equivalent to third order to β^* and $\phi^\dagger$, respectively. The proof of this result is a direct consequence of the orthogonality (Cox and Reid 1987) between β and ϕ and makes use of the expansions in Appendix of Kenne Pagui et al. (2017). Specifically, parameter orthogonality implies that terms up to order $O(n^{-1})$ in the expansion of $\beta^\ddagger - \beta$ are not affected by terms of order $O(1)$ in $s_\phi + A_\phi^\dagger$. As a result, and up to order $O(n^{-1})$, the expansion of $\beta^\ddagger - \beta$ is the same as that of $\beta^* - \beta$. The same reasoning applies if we switch the roles of β and ϕ, i.e. the expansion of $\phi^\ddagger - \phi$ is the same to the expansion of $\phi^\dagger - \phi$, up to order $O(n^{-1})$. Hence, $\beta^\ddagger$ has the same mean bias properties as β^* and $\phi^\ddagger$ has the same median bias properties as $\phi^\dagger$. For this reason, we use the term mixed BR to refer to the solution of adjusted score functions resulting from the mixed adjustment.

In order to illustrate the stated invariance properties of the estimators coming from the mixed adjustment, we consider a gamma regression model with independent response random variables $Y_1, \ldots, Y_{12}$, where, conditionally on covariates s_i and t_i, each Y_i has a gamma distribution with mean $\mu_i = \exp(\eta_i)$ and variance $\phi\mu_i^2$. The predictor η_i is a function of regression parameters and the covariates, s_i is a categorical covariate with values $L1$, $L2$ and $L3$, and $t_1, \ldots, t_{12}$ are generated from an exponential distribution with rate 1. Consider the three alternative parameterizations in Table 3. The identities $\beta_1 = \gamma_1$, $\beta_2 = \gamma_1 + \gamma_2$ and $\beta_3 = \gamma_1 + \gamma_3$ follow directly.

We simulate 1000 independent response vectors from the parameter value $(\beta_1, \beta_2, \beta_3, \beta_4, \phi)^\top = (-1, -0.5, 3, 0.2, 0.5)^\top$ and estimate the three parameter vectors in Table 3 for each sample using the ML estimator, and the estimators resulting from the mean, median and mixed bias-reducing adjusted scores. The estimates for parameterizations I and III are used to estimate the probability $P(|\tilde\beta_2 - \tilde\gamma_1 - \tilde\gamma_2| > \epsilon_1)$, and those for parameterizations I and II are used to estimate the probability $P(|\tilde\phi - \exp(\tilde\zeta)| > \epsilon_2)$ for various values of ϵ_1 and ϵ_2, using the various estimators in place of $\tilde\beta_2$, $\tilde\gamma_1$, $\tilde\gamma_2$, $\tilde\phi$ and $\tilde\zeta$. The results are displayed in Table 4. As expected, the probability $P(|\tilde\beta_2 - \tilde\gamma_1 - \tilde\gamma_2| > \epsilon_1)$ is zero for ML and mean BR, but not for median BR. Similarly, the probability $P(|\tilde\phi - \exp(\tilde\zeta)| > \epsilon_2)$ is zero for ML and median BR, but

Table 4 Probability $P(|\tilde{\beta}_2 - \tilde{\gamma}_1 - \tilde{\gamma}_2| > \epsilon_1)$ for parameterizations I and III, and $P(|\tilde{\phi} - \exp(\tilde{\zeta})| > \epsilon_1)$ for parameterizations I and II for various values of ϵ

ϵ_1	$P(\lvert\tilde{\beta}_2 - \tilde{\gamma}_1 - \tilde{\gamma}_2\rvert > \epsilon_1)$				ϵ_2	$P(\lvert\tilde{\phi} - \exp(\tilde{\zeta})\rvert > \epsilon_2)$			
	ML	Mean BR	Median BR	Mixed BR		ML	Mean BR	Median BR	Mixed BR
0.01	0	0	0.656	0	0.02	0	0.978	0	0
0.02	0	0	0.162	0	0.04	0	0.771	0	0
0.03	0	0	0.034	0	0.06	0	0.454	0	0
0.04	0	0	0.010	0	0.08	0	0.181	0	0
0.05	0	0	0.003	0	0.10	0	0.061	0	0

The ML estimator, the estimators from the mean, median and mixed bias-reducing adjusted scores are used in place of the tilded quantities. The figures are based on 1000 simulated response vectors from the gamma regression model of Table 3 with $(\beta_1, \beta_2, \beta_3, \beta_4, \phi)^\top = (-1, -0.5, 3, 0.2, 0.5)^\top$

not for mean BR. In contrast, the mixed adjustment strategy inherits the relevant properties of mean and median BR, and delivers estimators that are numerically invariant under linear contrasts of the mean regression parameters, and monotone transformations of the dispersion parameter.

Section 5.2 further evaluates the use of the mixed adjustment in the estimation of gamma regression models.

5 Illustrations and simulation studies

5.1 Case studies and simulation experiments

In this section, we present results from case studies and confirmatory simulation studies that provide empirical support to the ability of mean and median BR to achieve their corresponding goals, i.e. mean and median bias reduction, respectively. In particular, in Sect. 5.2 we consider gamma regression, in which we also evaluate the mixed adjustment strategy of Sect. 4, while in Sect. 5.3 we consider logistic regression, showing how both mean and median BR provide a practical solution to the occurrence of infinite ML estimates. Finally, Sect. 5.4 evaluates the performance of mean and median BR in a logistic regression setting characterized by the presence of many nuisance parameters. In this case, ML estimation and inference are known to be unreliable, while both mean and median BR practically reproduce the behaviour of estimation and inference based on the conditional likelihood, which, in this particular case, is the gold standard.

All numerical computations are performed in R using the `brglm2` R package (Kosmidis 2018). The `brglm2` R package provides the `brglmFit` method for the `glm` R function that implements mean and median BR for any GLM using the quasi-Fisher scoring iteration introduced in Sect. 2.

5.2 Gamma regression model for blood clotting times

The regression model for the clotting data in Example 1.1 is fitted, here, using the mean, median and mixed bias-reducing adjusted score functions of Sects. 2.3, 2.4 and 4, respectively. The estimates and the corresponding estimated standard errors are reported in Table 5. The estimates of regression parameters are practically the same for all methods. More marked differences between ML and the three adjusted score methods are noted in the estimates of the dispersion parameter. In particular, the estimates from the adjusted score methods result in notable inflation of the estimated standard errors for the regression parameters, with the median and mixed bias-reducing adjustments resulting in the largest inflation.

Table 5 Clotting data

Method	β_1	β_2	β_3	β_4	ϕ
ML	5.503	−0.584	−0.602	0.034	0.017
	(0.161)	(0.228)	(0.047)	(0.066)	
Mean BR	5.507	−0.584	−0.602	0.034	0.022
	(0.183)	(0.258)	(0.053)	(0.075)	
Median BR	5.505	−0.584	−0.602	0.034	0.024
	(0.187)	(0.265)	(0.054)	(0.077)	
Mixed BR	5.507	−0.584	−0.602	0.034	0.024
	(0.187)	(0.265)	(0.054)	(0.077)	

Estimates and estimated standard errors (in parentheses) for the parameters of the model in Example 1.1

In order to assess the quality of the estimates in Table 5, the simulated data sets in Example 1.1 are used to estimate the bias, the root mean squared error, the percentage of underestimation, and the mean absolute error of the various estimators, and the coverage of nominally 95% Wald-type confidence intervals. Table 6 reports the results. A comparison with the results of ML in Table 1 shows that the ML, mean BR, median BR and mixed BR estimators of $\beta_1, \ldots, \beta_4$ have similar bias and variance properties. On the other hand, the mean BR estimator of the dispersion parameter almost fully compensates for the mean bias of the ML estimator, while median BR and mixed BR give almost exactly 50% probability of underestimation. Furthermore, all BR methods deliver marked improvements in terms of empirical coverage

Table 6 Clotting data

Method	Parameter	B	RMSE	B^2/SD^2	PU	MAE	C
Mean BR	β_1	−0.04	16.15	< 0.01	49.65	12.87	93.12
	β_2	0.36	23.09	0.02	49.59	18.46	92.69
	β_3	0.02	4.69	< 0.01	49.92	3.74	93.08
	β_4	−0.11	6.71	0.03	50.50	5.36	92.26
	ϕ	< 0.01	0.67	< 0.01	55.00	0.53	
Median BR	β_1	−0.15	16.15	0.01	49.93	12.87	93.67
	β_2	0.36	23.09	0.02	49.60	18.46	93.27
	β_3	0.03	4.69	0.01	49.88	3.74	93.73
	β_4	−0.11	6.71	0.03	50.50	5.36	93.05
	ϕ	0.09	0.71	1.67	49.99	0.55	
Mixed	β_1	−0.02	16.15	< 0.01	49.65	12.87	93.66
	β_2	0.36	23.09	0.02	49.59	18.46	93.28
	β_3	0.02	4.69	< 0.01	49.95	3.74	93.71
	β_4	−0.11	6.71	0.03	50.50	5.36	93.06
	ϕ	0.09	0.71	1.68	49.93	0.55	

Simulation results based on 10,000 samples under the ML fit. The quantities in the table are described in the caption of Table 1. The estimators considered are those from mean BR (Sect. 2.3), median BR (Sect. 2.4) and mixed BR (Sect. 4). All reported figures are ×100 of their actual value and < 0.01 is used for a value that is less than 0.01 in absolute value

over ML, and the confidence intervals based on the estimates from the median and mixed bias-reducing adjustments are behaving the best. Finally, all confidence intervals appear to be liberal in terms of coverage, most probably due to the small sample size and the need to estimate the dispersion parameter. Note here that the superior coverage when using estimates from median and mixed bias-reducing adjustments of the scores are similar to what is expected in the case of the normal linear model; see Sect. 3.2.

5.3 Logistic regression for infant birthweights

We consider a study of low birthweight using the data given in Hosmer and Lemeshow (2000, Table 2.1), which are also publicly available in the `MASS` R package. The focus here is on the 100 births for which the mother required no physician visits during the first trimester. The outcome of interest is a proxy of infant birthweight (1 if $\geq 2500g$ and 0 otherwise), whose expected value μ_i is modelled in terms of explanatory variables using a logistic regression model with $\log\{\mu_i/(1-\mu_i)\} = \sum_{t=1}^{7} \beta_t x_{it}$, where $x_{i1} = 1$, x_{i2} and x_{i3} are the age and race (1 if white, 0 otherwise) of the mother, respectively, x_{i4} is the mother's smoking status during pregnancy (1 if yes, 0 if no), x_{i5} is a proxy of the history of premature labour (1 if any, 0 if none), x_{i6} is history of hypertension (1 if yes, 0 if no) and x_{i7} is the logarithm of the mother's weight at her last menstrual period.

Table 7 gives the parameter estimates from ML, mean BR and median BR. Both mean BR and median BR deliver estimates that are shrunken versions of the corresponding ML estimates, with mean BR delivering the most shrinkage. This shrinkage translates to smaller estimated standard errors for the regression parameters. Kosmidis and Firth (2018) provide geometric insights for the shrinkage induced by mean BR in binary regression and prove that the mean BR estimates are always finite for full rank X.

The frequency properties of the resulting estimators are assessed by simulating 10,000 samples at the ML estimates in Table 7, with covariates fixed as in the observed sample, and re-estimating the model from each simulated sample. A total of 103 out of the 10,000 samples result in ML estimates with one or more infinite components due to data separation (Albert and Anderson 1984). The detection of infinite estimates was done prior to fitting the model using the linear programming algorithms in Konis (2007), as implemented in the `detect_separation` method of the `brglm2` R package (Kosmidis 2018). The separated data sets were excluded when estimating the bias and coverage of Wald-type confidence intervals for the ML estimator. In contrast, the estimates from mean and median BR estimates were finite in all cases. For this reason, the corresponding summaries are based on all 10,000 samples.

Table 8 shows the results. Both mean BR and median BR have excellent performance in terms of mean bias and probability of underestimation, respectively. Table 8 also includes summaries for the estimators $\hat{\psi}_t = e^{\hat{\beta}_t}$, $\psi_t^* = e^{\beta_t^*}$, $\psi_t^\dagger = e^{\beta_t^\dagger}$ of the odds ratios $\psi_t = e^{\beta_t}$. Estimators of ψ_t with improved bias properties have also been recently investigated in Lyles et al. (2012). The invariance properties of ML and median BR guarantee that $\hat{\psi}$ and $\psi^\dagger$ are the ML and median BR

Table 7 Estimates and estimated standard errors (in parentheses) for the logistic regression model for the infant birthweight data in Sect. 5.3

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7
ML	− 8.496	− 0.067	0.690	− 0.560	− 1.603	− 1.211	2.262
	(5.826)	(0.053)	(0.566)	(0.576)	(0.697)	(0.924)	(1.252)
Mean BR	− 7.401	− 0.061	0.622	− 0.531	− 1.446	− 1.104	1.998
	(5.664)	(0.052)	(0.552)	(0.564)	(0.680)	(0.901)	(1.216)
Median BR	− 7.641	−0.062	0.638	− 0.538	− 1.481	− 1.134	2.059
	(5.717)	(0.053)	(0.557)	(0.568)	(0.681)	(0.906)	(1.228)

Table 8 Simulation results based on 10, 000 samples under the ML fit of the model for the birthweight data in Sect. 5.3

	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7
B	ML	− 1.42	− 0.01	0.09	− 0.03	− 0.20	− 0.12	0.34
	Mean BR	− 0.08	< 0.01	0.01	< 0.01	− 0.01	< 0.01	0.02
	Median BR	− 0.38	< 0.01	0.03	− 0.01	− 0.07	− 0.04	0.09
B_ψ	ML	183.50	< 0.01	0.75	0.12	0.02	0.18	57.50
	Mean BR	47.17	< 0.01	0.41	0.11	0.05	0.17	18.75
	Median BR	56.66	< 0.01	0.50	0.11	0.04	0.21	23.74
RMSE	ML	6.86	0.06	0.66	0.66	0.82	1.11	1.49
	Mean BR	5.94	0.05	0.58	0.59	0.72	0.94	1.28
	Median BR	6.11	0.06	0.60	0.61	0.78	1.01	1.32
PU	ML	56.1	53.3	46.4	51.4	57.8	53.5	43.1
	Mean BR	48.2	49.2	51.3	49.6	48.1	48.9	52.2
	Median BR	50.0	49.6	49.9	49.9	50.6	50.3	50.0
C	ML	94.8	94.8	94.5	94.7	96.4	96.6	94.5
	Mean BR	96.3	96.2	96.0	96.2	97.2	98.1	96.1
	Median BR	96.1	96.0	95.8	95.9	97.0	97.8	96.0

All reported summaries, described in the caption of Table 1, for ML are conditional to the finiteness of the estimates. B_ψ is the estimated bias in the ψ parameterization, and < 0.01 is used for a value that is less than 0.01 in absolute value

estimators of ψ, respectively. As a result, $\psi_t^\dagger$ preserves its improved median bias properties. On the other hand, ψ_t^* is not, formally, the mean BR estimator of ψ. Nevertheless, it behaves best in terms of bias. The improved estimation and inference provided by mean and median BR become even more evident in more extreme modelling settings, as shown by the example in the next section.

5.4 Logistic regression for the link between sterility and abortion

We consider data from a retrospective, matched case–control study on the role of induced and spontaneous abortions in the aetiology of secondary sterility (Trichopoulos et al. 1976). The data are available in the `infert` data frame from the `datasets` R package. The two healthy control subjects from the same hospital were matched to each of 83 patients according to their age, parity and level of education. One of the cases could be matched with only one control; thus, there are a total of 248 records. Each record also provides the number of induced and spontaneous abortions, taking values 0, 1 and 2 or more.

As is meaningful for retrospective case–control studies (see, for example, McCullagh and Nelder 1989, Sect. 4.3.3), we consider a logistic regression model with one fixed effect for each matched combination of cases and controls, and the number of induced and spontaneous abortions as the two categorical covariates of interest. In particular, the log odds of secondary sterility for the jth individual in the ith case–control combination are assumed to be

$$\lambda_i + \beta_1 x_{ij} + \beta_2 x'_{ij} + \beta_3 z_{ij} + \beta_4 z'_{ij} \quad (i = 1, \ldots, 83;\ j = 1, \ldots, n_i), \tag{12}$$

where $n_i \in \{2, 3\}$, x_{ij}, x'_{ij} are indicator variables of 1 and 2 or more spontaneous abortions, respectively, and z_{ij} and z'_{ij} are indicator variables of 1 and 2 or more induced abortions, respectively. The parameters $\lambda_1, \ldots, \lambda_{83}$ are the fixed effects for each matched combination of cases and controls, and the parameters of interest are $\beta_1, \ldots, \beta_4$.

Table 9 Estimates and estimated standard errors (in parentheses) for the parameters of interest in model (12) for the sterility data in Sect. 5.4

Method	β_1	β_2	β_3	β_4
ML	3.268 (0.592)	6.441 (0.955)	2.112 (0.587)	4.418 (0.948)
CL	2.044 (0.453)	3.935 (0.725)	1.386 (0.463)	2.819 (0.735)
Mean BR	2.055 (0.472)	3.954 (0.708)	1.305 (0.474)	2.714 (0.744)
Median BR	2.083 (0.478)	3.997 (0.713)	1.330 (0.482)	2.760 (0.754)

Due to the many nuisance parameters, the maximum likelihood estimators of $\beta_1, \ldots, \beta_4$ are highly biased leading to misleading inference. A solution that is specific to logistic regression is to eliminate the fixed effects by conditioning on their sufficient statistics and maximize the conditional likelihood (CL). This can be done, for example, using the `clogit` function in the `survival` R package. As shown in Table 9, both mean and median BR give estimates that are close to the maximum CL estimates, practically removing all the bias from the ML estimates, and resulting also in a correction for the estimated standard errors.

This desirable behaviour of mean BR and median BR is in line with published theoretical results in stratified settings with nuisance parameters. In particular, Lunardon (2018) has recently shown that inferences based on mean BR in stratified settings with strata-specific nuisance parameters are valid under the same conditions for the validity of inference (Sartori 2003) based on modified profile likelihoods (see, for example, Barndorff-Nielsen 1983; Cox and Reid 1987; McCullagh and Tibshirani 1990; Severini 1998). The same equivalence is shown for median BR in Kenne Pagui et al. (2017).

The advantage of mean and median BR over maximum CL is their generality of application. As is shown in Table 2 mean and median BR can be used in models where a sufficient statistic does not exist, and hence, direct elimination of the nuisance parameters is not possible. One such example is probit regression, which is typically the default choice in many econometric applications stemming out from prospective studies. The further algorithmic simplicity for mean and median BR makes them also competitive to the various modified profile likelihoods.

6 Multinomial logistic regression

6.1 The Poisson trick

Suppose that $y_1, \ldots, y_n$ are k-vectors of counts with $\sum_{j=1}^{k} y_{ij} = m_i$ and that $x_1, \ldots, x_n$ are corresponding p-vectors of explanatory variables. The multinomial logistic regression model assumes that conditionally on $x_1, \ldots, x_n$ the vectors of counts $y_1, \ldots, y_n$ are realizations of independent multinomial vectors, with $y_i = (y_{i1}, \ldots, y_{ik})$, where the probabilities for the ith multinomial vector satisfy

$$\log \frac{\pi_{ij}}{\pi_{ik}} = x_i^\top \gamma_j \quad (j = 1, \ldots, k-1), \tag{13}$$

with $\sum_{j=1}^{k} \pi_{ij} = 1$. Typically, $x_{i1} = 1$ for every $i \in \{1, \ldots, n\}$. The above model is also known as the baseline category logit (see, for example, Agresti 2002, Sect. 7.1) because it uses one of the multinomial categories as a baseline for the definition of the log odds. Expression (13) has the kth category as baseline, but this is without loss of generality since any other log odds can be computed using simple contrasts of the parameter vectors $\gamma_1, \ldots, \gamma_{k-1}$.

Maximum likelihood estimation can be done either by direct maximization of the multinomial log-likelihood for (13) or using maximum likelihood for an equivalent Poisson log-linear model. Specifically, if $y_{11}, \ldots, y_{nk}$ are realizations of independent Poisson random variables with means $\mu_{11}, \ldots, \mu_{nk}$, where

$$\begin{aligned} \log \mu_{ij} &= \lambda_i + x_i^\top \gamma_j \quad (j = 1, \ldots, k-1)\,, \\ \log \mu_{ik} &= \lambda_i\,, \end{aligned} \tag{14}$$

then the score equations for λ_i are $m_i = \sum_{j=1}^{k} \mu_{ij}$, forcing the Poisson means to add up to the multinomial totals and the maximum likelihood estimates for $\gamma_1, \ldots, \gamma_{k-1}$ to be exactly those that result from maximizing the multinomial likelihood for model (13) directly.

Kosmidis and Firth (2011) proved that the equivalence of the multinomial logistic regression model (13) and the Poisson log-linear model (14) extends to the mean BR estimates of $\gamma_1, \ldots, \gamma_{k-1}$, if at each step of the iterative procedure for solving the adjusted score equations, the current values of the Poisson expectations $\mu_{i1}, \ldots, \mu_{ik}$ are rescaled to sum up to the corresponding multinomial totals. Specifically, the results in Kosmidis and Firth (2011) suggest to prefix the IWLS update in (8) for the Poisson log-linear model (14) with the extra step

$$\bar{\mu}_{is}^{(j)} \leftarrow m_{is} \frac{\mu_{is}^{(j)}}{\sum_{t=1}^{k} \mu_{it}^{(j)}} \quad (i = 1, \ldots, n; s = 1, \ldots, k)$$

that rescales the Poisson means to sum to the multinomial totals. Then, W and the ML and mean BR quantities in the last row of Table 2 are computed using $\bar{\mu}_{is}^{(j)}$ instead of $\mu_{is}^{(j)}$.

The same argument applies the case of median BR. Given that the extra term in the IWLS update for median bias reduction in (11) depends on the parameters only through the response means, the same extra step of rescaling the Poisson means before the IWLS update of the parameters will result in an iteration that delivers the median BR estimates of the multinomial logistic regression model using the equivalent Poisson log-linear model.

6.2 Invariance properties

The mean BR estimator is invariant under general affine transformations of the parameters, and hence, direct contrasts result in mean BR estimators for any other baseline category for the response and any reference category in the covariates, without refitting the model. This is a particularly useful guarantee when modelling with baseline category models. In contrast, a direct transformation of the median BR estimates with baseline category k or a specific set of contrasts for the covariates is not guaranteed to result in median BR estimates for other baseline categories or contrasts in general.

6.3 Primary food choices of alligators

In order to investigate the extent that non-invariance impacts estimation and inference, we consider the data on food choice of alligators analysed in Agresti (2002, Sect. 7.1.2). The data come from a study of factors influencing the primary food choice of alligators. The observations are 219 alligators captured in four lakes in Florida. The nominal response variable is the primary food type, in volume, found in an alligator's stomach, which has five categories (fish, invertebrate, reptile, bird and other). The data set classifies the primary food choice according to the lake of capture (Hancock, Oklawaha, Trafford, George), gender (male and female) and size of the alligator (≤ 2.3 m long, > 2.3 m long).

Let $s = 1$ for alligator size > 2.3 metres and 0 otherwise, and let z^H, z^O, z^T, z^G be indicator variables for the lakes; for instance, $z^H = 1$ for alligators on the lake Hancock and 0 otherwise. A possible model for the probabilities of food choice is

$$\log(\pi_{ic}/\pi_{i1}) = \gamma_{c1} + \gamma_{c2}s_i + \gamma_{c3}z_i^O + \gamma_{c4}z_i^T + \gamma_{c5}z_i^G \quad (c = 2, 3, 4, 5)\,, \qquad (15)$$

where π_{ic} is the probability for category c, with values corresponding to fish ($c = 1$), invertebrate ($c = 2$), reptile ($c = 3$), bird ($c = 4$) and other ($c = 5$). Model (15) is based on the choice of contrasts that would be selected by default in R. In order to investigate the effects of lack of invariance of median bias reduction, the set of contrasts used in Agresti (2002, Section 7.1.2) is considered where George is the reference lake and > 2.3 is the reference alligator size. These choices result in writing the food choice log odds as

$$\log(\pi_{ic}/\pi_{i1}) = \gamma'_{c1} + \gamma'_{c2}s'_i + \gamma'_{c3}z_i^H + \gamma'_{c4}z_i^O + \gamma'_{j5}z_i^T \quad (c = 2, 3, 4, 5), \qquad (16)$$

where $s' = 1$ for alligator size ≤ 2.3 metres and 0 otherwise. The coefficients in the linear predictors of (15) and (16) are related as $\gamma_{c1} = \gamma'_{c1}+\gamma'_{c2}+\gamma'_{c3}, \gamma_{c2} = -\gamma'_{c2}, \gamma_{c3} = \gamma'_{c4}-\gamma'_{c3}$, $\gamma_{c4} = \gamma'_{c5} - \gamma'_{c3}$ and $\gamma_{c5} = -\gamma'_{c3}$.

Table 10 gives the ML, mean BR and median BR estimates, along with the corresponding estimated standard errors of the coefficients of model (15). Table 10 shows also results of median $\text{BR}_{\gamma'}$, which correspond to the median BR estimates of γ' transformed in the γ parameterization. As in logistic regression, the mean and median BR estimates are shrunken relative to the maximum likelihood ones with a corresponding shrinkage effect on the estimated standard errors.

The median BR and median $\text{BR}_{\gamma'}$ estimates are almost the same, indicating that median BR, in this particular setting, is not affected by its lack of invariance under linear contrasts. The differences between the three methods are more notable when the observed counts are divided by two, as given in Table 11. In this case, data separation results in two of the ML estimates being infinite. This can generally happen with positive probability when data are sparse or when there are large covariate effects (Albert and Anderson 1984). As is the case for logistic regression (see Sect. 5.3), both mean and median BR deliver finite estimates for all parameters. The finiteness of the mean BR estimates has also been observed in Bull et al. (2002).

In order to better assess the properties of the estimators considered in Tables 10 and 11, we designed a simulation study where the multinomial totals for each covariate setting in the alligator food choice data set are progressively increased as a fraction of their observed values. Specifically, we consider the sets of multinomial totals $\{rm_1, \ldots, rm_n\}$ for $r \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4, 5\}$, where m_i $(i = 1, \ldots, n)$ is the observed multinomial total for the ith combination of covariate values. For each value of r, we simulate 10,000 data sets from the ML fit of model (15) given in Table 10 and then compare the mean BR, median BR and median $\text{BR}_{\gamma'}$ estimators in terms of relative bias and percentage of underestimation. The ML estimator is not considered in the comparison because the probability of infinite estimates is very high, ranging from 1.3% for $r = 5$ up to 76.4% for $r = 0.5$. In contrast, mean BR and median BR produced finite estimates for all data sets and r values considered.

Figures 3 and 4 show the relative bias and the percentage of underestimation, respectively, for each parameter as a function of r. Overall, mean BR is preferable in terms of mean bias, while median BR achieves better median cen-

Table 10 Estimates and estimated standard errors (in parentheses) of the multinomial regression model (15) for the alligator data in Sect. 6

Method	c	γ_{c1}	γ_{c2}	γ_{c3}	γ_{c4}	γ_{c5}
ML	2	− 1.75 (0.54)	− 1.46 (0.40)	2.60 (0.66)	2.78 (0.67)	1.66 (0.61)
	3	− 2.42 (0.64)	0.35 (0.58)	1.22 (0.79)	1.69 (0.78)	− 1.24 (1.19)
	4	− 2.03 (0.56)	0.63 (0.64)	− 1.35 (1.16)	0.39 (0.78)	− 0.70 (0.78)
	5	− 0.75 (0.35)	− 0.33 (0.45)	− 0.82 (0.73)	0.69 (0.56)	− 0.83 (0.56)
Mean BR	2	− 1.65 (0.52)	− 1.40 (0.40)	2.46 (0.65)	2.64 (0.66)	1.56 (0.60)
	3	− 2.25 (0.61)	0.32 (0.56)	1.12 (0.76)	1.58 (0.75)	− 0.98 (1.02)
	4	− 1.90 (0.54)	0.58 (0.61)	− 1.04 (1.01)	0.40 (0.76)	− 0.62 (0.74)
	5	− 0.72 (0.35)	− 0.31 (0.44)	− 0.72 (0.71)	0.67 (0.56)	− 0.78 (0.55)
Median BR	2	− 1.71 (0.53)	− 1.41 (0.40)	2.51 (0.65)	2.69 (0.67)	1.61 (0.61)
	3	− 2.33 (0.62)	0.34 (0.57)	1.16 (0.77)	1.62 (0.76)	− 1.12 (1.10)
	4	− 1.96 (0.54)	0.60 (0.62)	− 1.20 (1.08)	0.39 (0.77)	− 0.66 (0.76)
	5	− 0.73 (0.35)	− 0.32 (0.44)	− 0.77 (0.71)	0.67 (0.56)	− 0.80 (0.55)
Median $BR_{\gamma'}$	2	− 1.70 (0.53)	− 1.41 (0.39)	2.52 (0.65)	2.70 (0.66)	1.61 (0.61)
	3	− 2.35 (0.63)	0.34 (0.57)	1.16 (0.77)	1.62 (0.77)	− 1.12 (1.11)
	4	− 1.97 (0.55)	0.60 (0.63)	− 1.21 (1.09)	0.39 (0.77)	−0.66 (0.76)
	5	− 0.73 (0.35)	− 0.32 (0.45)	− 0.78 (0.72)	0.67 (0.56)	− 0.80 (0.55)

Table 11 Estimates and estimated standard errors (in parentheses) of the multinomial regression model (15) for the alligator data in Sect. 6 after halving the frequencies, and rounding them to the closest integer

Method	c	γ_{c1}	γ_{c2}	γ_{c3}	γ_{c4}	γ_{c5}
ML	2	− 1.83 (0.76)	− 1.55 (0.59)	2.66 (0.94)	2.81 (0.95)	1.64 (0.87)
	3	− 3.39 (1.25)	1.40 (1.19)	1.13 (1.29)	1.44 (1.29)	$-\infty$ ($+\infty$)
	4	− 2.31 (0.86)	0.66 (1.03)	$-\infty$ ($+\infty$)	0.58 (1.16)	− 0.78 (1.29)
	5	− 0.82 (0.49)	− 0.04 (0.67)	− 1.35 (1.18)	0.28 (0.81)	− 1.25 (0.88)
Mean BR	2	− 1.64 (0.72)	− 1.43 (0.59)	2.40 (0.91)	2.54 (0.92)	1.46 (0.84)
	3	− 2.76 (1.00)	1.08 (0.96)	0.93 (1.15)	1.22 (1.15)	− 1.24 (1.71)
	4	− 2.02 (0.78)	0.55 (0.90)	− 1.30 (1.70)	0.57 (1.08)	− 0.57 (1.12)
	5	− 0.76 (0.49)	− 0.03 (0.66)	− 1.03 (1.06)	0.29 (0.81)	− 1.08 (0.84)
Median BR	2	−1.76 (0.74)	− 1.45 (0.59)	2.48 (0.93)	2.62 (0.93)	1.54 (0.86)
	3	− 3.00 (1.08)	1.23 (1.03)	1.02 (1.18)	1.31 (1.18)	− 2.04 (2.45)
	4	− 2.15 (0.81)	0.59 (0.95)	− 2.17 (2.49)	0.56 (1.11)	−0.67 (1.19)
	5	− 0.79 (0.49)	−0.04 (0.66)	−1.19 (1.11)	0.28 (0.81)	−1.16 (0.86)
Median $BR_{\gamma'}$	2	− 1.74 (0.74)	−1.45 (0.58)	2.50 (0.92)	2.64 (0.93)	1.54 (0.85)
	3	− 3.12 (1.14)	1.24 (1.08)	1.03 (1.24)	1.32 (1.24)	−2.05 (2.61)
	4	− 2.15 (0.81)	0.60 (0.95)	− 2.20 (2.51)	0.55 (1.11)	− 0.67 (1.19)
	5	− 0.79 (0.49)	− 0.03 (0.66)	− 1.20 (1.11)	0.27 (0.81)	− 1.16 (0.86)

tring for all the parameters. We note that even median $BR_{\gamma'}$ has bias and probabilities of underestimation very close to those obtained directly under the γ parameterization. This confirms the indications from the observed data that, even if not granted by the theory, median BR is close to invariant under contrasts in the current model setting. As expected, the frequency properties of the three estimators converge to what we expect from standard ML asymptotics as r increases. In particular, the bias converges to 0 and the percentage of underestimation to 50%.

7 Discussion

Fisher orthogonality (Cox and Reid 1987) of the mean and dispersion parameters dictates that the mixed approach to bias reduction is valid also for generalized linear models with dispersion covariates in Smyth (1989), and that estimation can be done by direct generalization of the IWLS iterations in (5) and (11), for mean and median bias reduction, respectively.

Inference and model comparison has been based on Wald-type statistics. For special models, it is possible to form penalized likelihood ratio statistics based on the penalized

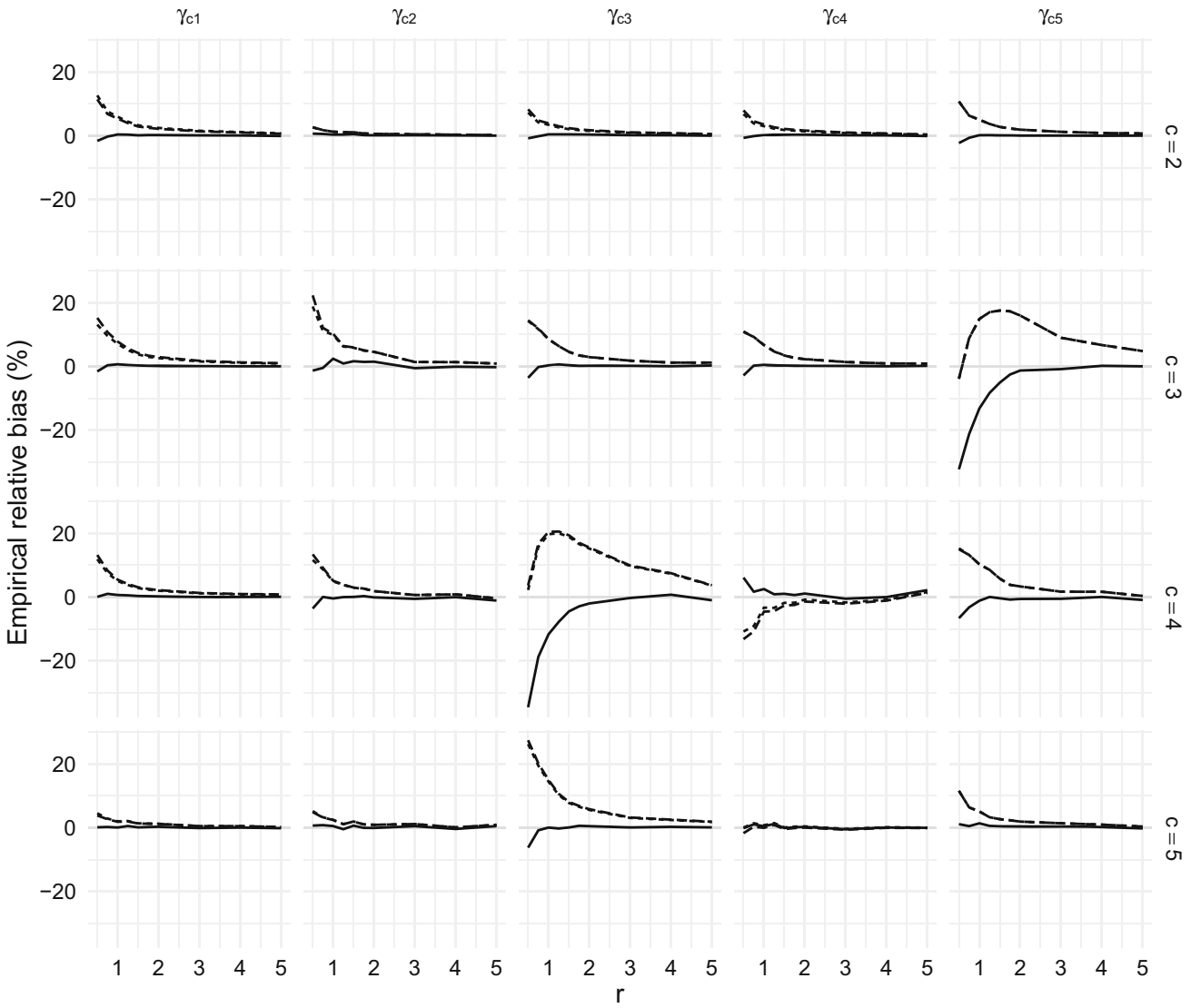

Fig. 3 Empirical relative bias based on 10,000 simulated samples from the ML fit of model (15) given in Table 10, for each $r \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4, 5\}$. The curves correspond to the mean BR (solid), median BR (dashed) and median $\text{BR}_{\gamma'}$ (long dashed) estimators. The grey horizontal line is at zero

log-likelihood that corresponds to the adjusted scores. A prominent example is logistic regression where the mean bias-reducing adjusted score is the gradient of the log-likelihood penalized by the logarithm of the Jeffreys' prior (see Heinze and Schemper 2002, where the profiles of the penalized log-likelihood are used for inference). In that case, the estimator from mean BR coincides with the mode of the posterior distribution obtained using the Jeffreys' prior (see also Ibrahim and Laud 1991). The same happens for Poisson log-linear models and for multinomial baseline category models. Even when a penalized log-likelihood corresponding to adjusted scores is not available (see Theorem 1 in Kosmidis and Firth 2009, for necessary and sufficient conditions for the existence of mean bias-reducing penalized likelihoods for generalized linear models), the adjustments to the score can, however, be seen as model-based penalties to the inferential quantities for maximum likelihood. In this sense, the adjustments introduce some implicit regularization to the estimation problem, which is just enough to achieve mean or median BR.

In this framework, a general alternative to Wald-type statistics is score-type statistics with known asymptotic distributions, which can be readily defined as in Lindsay and Qu (2003). Let $(\beta^\top, \phi)^\top = (\psi^\top, \lambda^\top)^\top$, with $\dim(\psi) = p_1$ and $\dim(\lambda) = p - p_1$, $i^{\psi\psi}(\psi, \lambda)$ be a $p_1 \times p_1$ matrix collecting the rows and columns of $\{i(\psi, \lambda)\}^{-1}$ corresponding to ψ, and λ^*_ψ the estimator of λ resulting from the solution of the mean bias-reducing adjusted score equations on λ for fixed

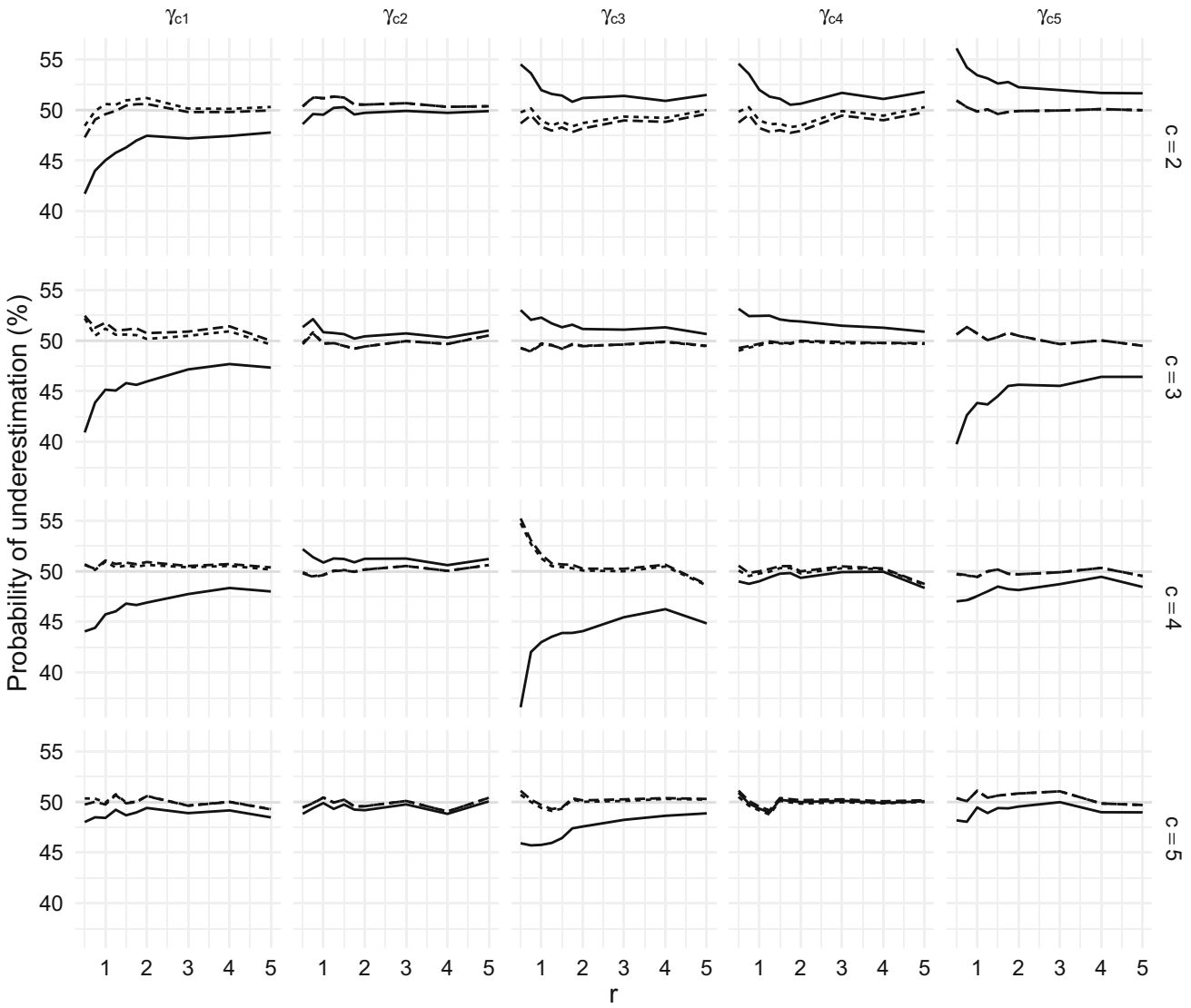

Fig. 4 Empirical probability of underestimation based on 10,000 simulated samples from the ML fit of model (15) given in Table 10, for each $r \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4, 5\}$. The curves correspond to the mean BR (solid), median BR (dashed) and median BR$_{\gamma'}$ (long dashed) estimators. The grey horizontal line is at 50

ψ. Since the scores have an asymptotic normal distribution with mean zero and variance–covariance matrix $i(\psi, \lambda)$ and the mean bias-reducing adjustment is of order $O(1)$,

$$\{s_\psi(\psi, \lambda^*_\psi) + A^*_\psi(\psi, \lambda^*_\psi)\}^\top i^{\psi\psi}(\psi, \lambda^*_\psi) \left\{s_\psi(\psi, \lambda^*_\psi) + A^*_\psi(\psi, \lambda^*_\psi)\right\} \tag{17}$$

has an asymptotic null $\chi^2_{p_1}$ distribution. The same result holds for median BR, by replacing λ^*_ψ and A^*_ψ with $\lambda^\dagger_\psi$ and $A^\dagger_\psi$. The adjusted score statistic can then be used for constructing confidence intervals and regions and testing hypotheses on any set of parameters of the generalized linear models, including constructing tables similar to analysis of deviance tables for maximum likelihood.

Finally, as is illustrated in the example of Sect. 5.4 and shown in Lunardon (2018) and Kenne Pagui et al. (2017), mean BR and median BR can be particularly effective for inference about a low-dimensional parameter of interest in the presence of high-dimensional nuisance parameters, while providing, at the same time, improved estimates of the nuisance parameters.

8 Supplementary material

The supplementary material includes R code and a report to fully reproduce all numerical results and figures in the paper.

Acknowledgements Ioannis Kosmidis was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 (Turing award

number TU/B/000082). Euloge Clovis Kenne Pagui and Nicola Sartori were supported by the Italian Ministry of Education under the PRIN 2015 grant 2015EASZFS_003 and by the University of Padova (PRAT 2015 CPDA153257).

Appendix

Proof of Theorem 3.1

Proof Since $\hat{\phi} < \phi^* < \phi^\dagger$ and $z_{1-\alpha/2} < t_{n-p;1-\alpha/2}$, we have $\hat{I}_{1-\alpha} \subset I^*_{1-\alpha} \subset I^E_{1-\alpha}$ and $I^*_{1-\alpha} \subset I^\dagger_{1-\alpha}$ for any $n - p \geq 1$ and $\alpha \in (0, 1)$. We also have $I^\dagger_{1-\alpha} \subset I^E_{1-\alpha}$ if $g(\nu, \alpha) = \{(\nu - 2/3)/\nu\}^{1/2} t_{\nu;1-\alpha/2} - z_{1-\alpha/2} > 0$. For fixed natural $\nu \geq 1$, the function $g(\nu, \alpha)$ is positive when $\alpha \to 0^+$ and has only one zero in $\tilde{\alpha}(\nu)$. Hence, the condition is satisfied for $\alpha < \tilde{\alpha}(\nu)$. Moreover, it can be seen numerically that $\tilde{\alpha}(\nu)$ increases with ν, having a minimum in $\tilde{\alpha}(1) = 0.35562$.

Even when $I^E_{1-\alpha} \subset I^\dagger_{1-\alpha}$, when $\nu > 1$, the absolute difference between the length of the intervals $I^\dagger_{1-\alpha}$ and $I^E_{1-\alpha}$ is smaller than the corresponding difference for $I^*_{1-\alpha}$ and $I^E_{1-\alpha}$, for any $\alpha > 0$. Indeed, this is true provided that the function $h(\nu, \alpha) = 2t_{\nu;1-\alpha/2}/\sqrt{\nu} - z_{1-\alpha/2}/\sqrt{\nu - 2/3} - z_{1-\alpha/2}/\sqrt{\nu}$ is positive. This is verified because, for fixed $\nu > 1$, $h(\nu, \alpha)$ is a monotonic decreasing function in α, converging to 0^+ as $\alpha \to 1^-$. On the other hand, if $\nu = 1$, $h(\nu, \alpha)$ is positive for $\alpha < 0.62647$ and negative otherwise. □

References

Agresti, A.: Categorical Data Analysis. Wiley, New York (2002)

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

Barndorff-Nielsen, O.: On a formula for the distribution of the maximum likelihood estimator. Biometrika **70**(2), 343–365 (1983)

Brazzale, A., Davison, A., Reid, N.: Applied Asymptotics: Case Studies in Small-Sample Statistics. Cambridge University Press, Cambridge (2007)

Bull, S.B., Mak, C., Greenwood, C.M.: A modified score function estimator for multinomial logistic regression in small samples. Comput. Stat. Data Anal. **39**(1), 57–74 (2002)

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

Cox, D.R., Reid, N.: Parameter orthogonality and approximate conditional inference (with discussion). J. R. Stat. Soc. Ser. B Methodol. **49**, 1–39 (1987)

Cox, D.R., Snell, E.J.: A general definition of residuals (with discussion). J. R. Stat. Soc. Ser. B Methodol. **30**, 248–275 (1968)

Efron, B.: Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). Ann. Stat. **3**, 1189–1242 (1975)

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

Green, P.J.: Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J. R. Stat. Soc. Ser. B Methodol. **46**(2), 149–192 (1984)

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

Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. Wiley, New York (2000)

Ibrahim, J.G., Laud, P.W.: On Bayesian analysis of generalized linear models using Jeffreys's prior. J. Am. Stat. Assoc. **86**(416), 981–986 (1991)

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

Konis, K.: Linear programming algorithms for detecting separated data in binary logistic regression models. Ph.D. Thesis, University of Oxford (2007)

Kosmidis, I.: Bias in parametric estimation: reduction and useful side-effects. Wiley Interdiscip. Rev: Comput. Stat. **6**(3), 185–196 (2014a)

Kosmidis, I.: Improved estimation in cumulative link models. J. R. Stat. Soc. Ser. B Methodol. **76**(1), 169–196 (2014b)

Kosmidis, I.: brglm2: bias reduction in generalized linear models. R package version 0.1.8 (2018)

Kosmidis, I., Firth, D.: Bias reduction in exponential family nonlinear models. Biometrika **96**(4), 793–804 (2009)

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

Kosmidis, I., Firth, D.: Multinomial logit bias reduction via the poisson log-linear model. Biometrika **98**(3), 755–759 (2011)

Kosmidis, I., Firth, D.: Jeffreys' prior, finiteness and shrinkage in binomial-response generalized linear models. (2018) arXiv:1812.01938v1

Lindsay, B.G., Qu, A.: Inference functions and quadratic score tests. Stat. Sci. **18**(3), 394–410 (2003)

Lunardon, N.: On bias reduction and incidental parameters. Biometrika **105**(1), 233–238 (2018)

Lyles, R.H., Guo, Y., Greenland, S.: Reducing bias and mean squared error associated with regression-based odds ratio estimators. J. Stat. Plan. Inference **142**(12), 3235–3241 (2012)

McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, London (1989)

McCullagh, P., Tibshirani, R.: A simple method for the adjustment of profile likelihoods. J. R. Stat. Soc. Ser. B Methodol. **52**(2), 325–344 (1990)

R Core Team.: R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing (2018)

Sartori, N.: Modified profile likelihoods in models with stratum nuisance parameters. Biometrika **90**(3), 533–549 (2003)

Severini, T.A.: An approximation to the modified profile likelihood function. Biometrika **85**(2), 403–411 (1998)

Smyth, G.K.: Generalized linear models with varying dispersion. J. R. Stat. Soc. Ser. B Methodol. **51**(1), 47–60 (1989)

Trichopoulos, D., Handanos, N., Danezis, J., Kalandidi, A., Kalapothaki, V.: Induced abortion and secondary infertility. Br. J. Obstet. Gynaecol. **83**(8), 645–650 (1976)

Wedderburn, R.W.M.: On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika **63**(1), 27–32 (1976)