Received: 20 December 2020 | Revised: 18 January 2022 | Accepted: 11 February 2022

DOI: 10.1002/sim.9361

RESEARCH ARTICLE

Statistics in Medicine WILEY

Improved estimation in negative binomial regression

Euloge Clovis Kenne Pagui | **Alessandra Salvan** | **Nicola Sartori**

Department of Statistical Sciences, University of Padova, Padova, Italy

Correspondence
Euloge Clovis Kenne Pagui, Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121 Padova, Italy.
Email: kenne@stat.unipd.it

Funding information
Università degli Studi di Padova, Grant/Award Numbers: BIRD185955, BIRD203991

Abstract

Negative binomial regression is commonly employed to analyze overdispersed count data. With small to moderate sample sizes, the maximum likelihood estimator of the dispersion parameter may be subject to a significant bias, that in turn affects inference on mean parameters. This article proposes inference for negative binomial regression based on adjustments of the score function aimed at mean or median bias reduction. The resulting estimating equations generalize those available for improved inference in generalized linear models and can be solved using a suitable extension of iterative weighted least squares. Simulation studies confirm the good properties of the new methods, which are also found to solve in many cases numerical problems of maximum likelihood estimation. The methods are illustrated and evaluated using two case studies: an Ames salmonella assay data set and data on epileptic seizures. Inference based on adjusted scores turns out to generally improve on maximum likelihood, and even on explicit bias correction, with median bias reduction being overall preferable.

KEYWORDS

adjusted score, iterative weighted least squares, maximum likelihood, mean and median bias reduction, parameterization invariance

1 | INTRODUCTION

Regression models for count data are employed in many contexts, especially in social sciences, economics, biology, and epidemiology. It is not uncommon that empirical counts display substantial overdispersion and a popular modeling approach is negative binomial regression, see for example, section 7.3 of the book by Agresti[1] and the monograph by Hilbe[2] for recent accounts.

Frequentist inference about mean and shape parameters in negative binomial regression is typically based on the likelihood and this is the method of choice for standard software, such as the `glm.nb` function of the R package MASS.[3] Maximum likelihood has been studied starting from Fisher[4] and Anscombe[5] for independent and identically distributed data and from Lawless[6] for the regression setting. With moderate sample sizes, the maximum likelihood estimator of the shape parameter may be subject to a substantial bias that can influence the inferential conclusions also about regression coefficients.

General improved estimation methods based on adjustments of the likelihood equations have been proposed starting from the contributions of Firth[7] and Kenne Pagui et al,[8] resulting in mean or median bias reduction, respectively. While mean bias reduction yields an estimator with reduced bias, median bias reduction is such that each component of the estimator is, with high accuracy, median unbiased, that is, it has the same probability of underestimating and overestimating the corresponding parameter component. Mean bias reduction is invariant under linear transformation of

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

the parameters, while median bias reduction is invariant under monotone component-wise parameter transformations. Unlike traditional bias correction, that subtracts an estimate of the bias from the maximum likelihood estimate, see for instance section 9.2 of Cox and Hinkley,[9] both mean and median bias reduction methods do not rely on finiteness of the maximum likelihood estimate and have the advantage of solving practical issues related to boundary estimates that can occur with positive probability in models for discrete data.[10] Obtaining the quantities required for mean and median bias reduction, as well as development of efficient software, is not always straightforward, but is necessary in order to make these methods available to practitioners. A major effort has been devoted to generalized linear models[11-13] leading to the `brglm2` package[14] for the software R.[15] Additional effort was needed for other specific models, such as beta regression[12,16] or cumulative link models.[17]

Negative binomial regression does not fall into the generalized linear models class when the shape parameter is unknown, as is the case in practical applications. Therefore, due to its widespread use, it is of interest to provide the quantities required for mean and median bias reduction, together with an efficient implementation, and to assess whether the general theoretical properties of the methods produce appreciable improvements over standard maximum likelihood. Previous work in this direction includes the paper by Saha and Paul,[18] who, for independent and identically distributed data, derived a bias corrected maximum likelihood estimator for the shape parameter and showed that it is preferable to other methods.[19,20] The authors also give an example of bias correction involving negative binomial regression, although the expression of the correction is not provided.

In this article, we derive the adjusted score equations for mean and median bias reduction for negative binomial regression and show that they can be solved by iterative weighted least squares after an appropriate adjustment of the ordinary working variate, or adjusted dependent variable, for maximum likelihood. Moreover, the method is developed for various link functions and parameterizations of the shape parameter. An R implementation is given in the `brnb` function which has been added to the current version of the R package `brglm2`.[14] Mean and median bias reduced estimators are compared with the maximum likelihood estimator through an extensive simulation experiment under different scenarios. Two case studies, the Ames salmonella reverse mutagenicity assay[21] and the epileptic seizures data,[22] are also considered and include comparison with other methods previously used for the same data, in particular bias correction.[18] The results indicate that, overall, mean and median bias reduction are both preferable to standard likelihood inference, even after bias correction, especially with moderate sample sizes. Median bias reduction provides the best empirical coverage of Wald-type confidence intervals for all parameters. Moreover, numerical problems that lead to unavailability of the maximum likelihood estimate, and therefore of its bias correction, occur more frequently than with mean or median bias reduction. In addition, traditional bias correction is seen to be less accurate than mean and median bias reduction when the number of covariates is large relative to the sample size.

The rest of the article is organized as follows. In Section 2, we introduce the notation for the negative binomial regression model. In Section 3, we give the adjusted score functions for mean and median bias reduction, together with computational details. Sections 4 and 5 contain simulation results and case studies, respectively. A brief discussion is given in Section 6. The Supplementary Material contains additional figures and simulation results together with R code to reproduce the analyses in the article.

2 | NEGATIVE BINOMIAL REGRESSION

Using Poisson regression when overdispersion is present typically leads to underestimation of standard errors of regression coefficients and therefore to potentially misleading inferential conclusions. Negative binomial regression[1,2] allows to model overdispersion introducing a shape parameter in the variance specification, so that, for count mean response μ_i, $i = 1, \dots, n$, the inflated variance has the form $\mu_i + \kappa\mu_i^2$, where $\kappa > 0$. Poisson regression is a limiting case as κ approaches zero.

Let y_i, $i = 1, \dots, n$, be realizations of independent negative binomial random variables Y_i with mean μ_i and variance $V(Y_i) = \mu_i + \kappa\mu_i^2$, where $\kappa > 0$. The probability mass function of Y_i is

$$f_{Y_i}(y_i; \mu_i, \kappa) = \frac{\Gamma(y_i + \kappa^{-1})}{y_i!\Gamma(\kappa^{-1})}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\left(\frac{1}{1+\kappa\mu_i}\right)^{1/\kappa}, \tag{1}$$

$y_i = 0, 1, \dots, \kappa > 0$ and $\mu_i > 0$. In a regression setting, we consider $\mu_i = g^{-1}(\eta_i)$, where $g^{-1}(\cdot)$ is the inverse of the link function, $\eta_i = x_i^\top\beta$ is the linear predictor, with $\beta = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ and $x_i = (x_{i1}, \dots, x_{ip})^\top$ a vector of covariates, with

$x_i^\top$ the ith row of the model matrix X. When an intercept is included in the linear predictor, $x_{i1} = 1$, $i = 1, \dots, n$. The usual choice for the link function is $g(\cdot) = \log(\cdot)$. For sake of generality, the derivation below is for a generic monotone reparameterization of κ, say $\phi = \phi(\kappa)$, with inverse $\kappa(\phi)$ and derivative $\kappa'(\phi)$. Common choices are $\phi = 1/\kappa$, $\phi = \log\kappa$ or $\phi = \sqrt{\kappa}$.

Let $\theta = (\beta^\top, \phi)^\top$. Noting that for any $a > 0$, $\Gamma(y + a)/\Gamma(a) = a(a + 1)\cdots(a + y - 1)$, the log likelihood is

$$\ell(\beta, \phi) = \sum_{i=1}^{n} m_i \left\{ \sum_{j=0}^{y_i^*} \log(1 + \kappa j) + y_i \log \frac{\mu_i}{1 + \kappa\mu_i} - \frac{1}{\kappa} \log(1 + \kappa\mu_i) \right\},$$

where m_i is a fixed weight for the ith observation, $y_i^* = y_i - 1$, $\sum_{j=0}^{y_i^*} \log(1 + \kappa j)$ is zero when $y_i^* < 0$ and $\kappa = \kappa(\phi)$. Weights m_i are typically equal to 1, but can be greater than 1 with grouped data. The score function $U = U(\theta) = (\partial/\partial\theta)\ell(\theta)$ has components $U_\beta = (\partial/\partial\beta)\,\ell(\beta, \kappa(\phi))$ and $U_\phi = (\partial/\partial\phi)\ell(\beta, \kappa(\phi))$ given by

$$U_\beta = X^\top W D^{-1}(y - \mu),$$
$$U_\phi = \kappa'(\phi) \sum_{i=1}^{n} m_i \left\{ S_{1i} - \frac{\mu_i y_i}{\kappa\mu_i + 1} + \frac{(\kappa\mu_i + 1)\log(\kappa\mu_i + 1) - \kappa\mu_i}{\kappa^3\mu_i + \kappa^2} \right\},$$

where D is a diagonal matrix with diagonal elements $d_i = d\mu_i/d\eta_i$, W is a diagonal matrix with diagonal elements $w_i = m_i d_i^2/V(Y_i)$, $y = (y_1, \dots, y_n)^\top$, $\mu = (\mu_1, \dots, \mu_n)^\top$ and $S_{1i} = \sum_{j=0}^{y_i^*} j/(\kappa j + 1)$. The expected information[6] is

$$i(\theta) = \begin{bmatrix} i_{\beta\beta} & 0_p \\ 0_p^\top & i_{\phi\phi} \end{bmatrix} = \begin{bmatrix} X^\top W X & 0_p \\ 0_p^\top & \kappa'(\phi)^2 i_{\kappa\kappa} \end{bmatrix},$$

where 0_p is a p-vector of zeros and

$$i_{\kappa\kappa} = \kappa^{-4} \sum_{i=1}^{n} m_i \left\{ \sum_{j=0}^{+\infty} \frac{\Pr(Y_i > j)}{(\kappa^{-1} + j)^2} - \frac{\kappa\mu_i}{\mu_i + \kappa^{-1}} \right\}.$$

The maximum likelihood estimate $\hat\theta^\top = (\hat\beta^\top, \hat\phi)$ is obtained as solution of the equations $U_\beta = 0$ and $U_\phi = 0$ that can be solved using a Fisher scoring algorithm. Exploiting the orthogonality between β and ϕ, the current iterate $\hat\phi^{(j)}$ is found by replacing $\hat\beta^{(j)}$ into the jth Fisher scoring iteration for $U_\phi = 0$. The procedure is alternated until convergence. With simple algebra, the jth iteration of Fisher scoring algorithm for $U_\beta = 0$ updates the current iterate $\hat\beta^{(j)}$ providing

$$\hat\beta^{(j+1)} = (X^\top W^{(j)} X)^{-1} X^\top W^{(j)} z^{(j)}, \tag{2}$$

where the superscript (j) indicates that the quantity is evaluated at $\hat\beta^{(j)}$ and z is the vector with elements $z_i = \eta_i + (y_i - \mu_i)/d_i$, $i = 1, \dots, n$, usually called the adjusted dependent variables or working variates. Equation (2) has the same form as the iterative weighted least squares (IWLS) iteration in generalized linear models.

3 | MEAN AND MEDIAN BIAS REDUCTION

Bias of maximum likelihood estimators in small samples or with sparse data can result in significant loss of accuracy of the related inferential procedures. An extensive amount of literature has focused on methods for reducing mean bias. A general classification separates explicit methods, also called bias correction, obtained subtracting from the maximum likelihood estimate an estimate of its first order bias, from implicit methods, also called bias reduction, obtained modifying the score function. A unified review is presented by Kosmidis.[23] See also Greenland et al[24] for an expository discussion of sparse data bias and available remedies. Such classification also holds for methods aiming at improving other centering properties of the estimator, such as the median centering. Generally, explicit methods are one-step approximations to the corresponding implicit methods, using the maximum likelihood estimate as a starting value. Therefore, they are less accurate and rely on existence of the latter. We recall below mean and median bias reduction and obtain the relevant expressions for negative binomial regression.

Consider a regular model with d-dimensional parameter θ, log likelihood $\ell(\theta)$, score function $U(\theta)$, and expected information $i(\theta)$, the latter assumed in the following to be of order n. We let $U_{\theta_r}(\theta)$ be a component of $U(\theta)$, $r = 1, \dots, d$, and $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^\top$ be the observed information. While the maximum likelihood estimate $\hat\theta$ is a solution of the equation $U(\theta) = 0$, the improved estimates proposed here are based on adjusted score equations having the general form $U(\theta) + A(\theta) = 0$, with $A(\theta)$ a model-dependent adjustment term of order $O(1)$ under repeated sampling. All the proposed adjustments involve the quantities (see Kosmidis and Firth[12] for their first use)

$$P_{\theta_r}(\theta) = E_\theta\{U(\theta)U(\theta)^\top U_{\theta_r}(\theta)\}, \quad Q_{\theta_r}(\theta) = E_\theta\{-j(\theta)U_{\theta_r}(\theta)\}, \quad r = 1, \dots, d.$$

In particular, Firth[7] showed that the leading term of order $O(n^{-1})$ of the bias of the maximum likelihood estimator is reduced to order $O(n^{-2})$ with $A(\theta) = A^*(\theta)$, where $A^*(\theta)$ has elements

$$A^*_{\theta_r}(\theta) = \frac{1}{2}\mathrm{tr}\left\{i(\theta)^{-1}(P_{\theta_r}(\theta) + Q_{\theta_r}(\theta))\right\}, \quad r = 1, \dots, d, \tag{3}$$

with $\mathrm{tr}(\cdot)$ the trace operator. We let $U^*(\theta) = U(\theta) + A^*(\theta)$ and we denote by θ^* the corresponding estimator, solution of $U^*(\theta) = 0$.

The bias corrected maximum likelihood estimator, see for example, section 9.2 of Cox and Hinkley[9] and section 5.3 of Barndorff-Nielsen and Cox,[25] is given by $\tilde\theta = \hat\theta - b(\hat\theta)$, where $b(\theta)$ is the term of order $O(n^{-1})$ of the bias of $\hat\theta$ and is equal to $-i(\theta)^{-1}A^*(\theta)$.[7] Also $\tilde\theta$ has bias of order $O(n^{-2})$, although the availability of $\tilde\theta$ relies on the existence of $\hat\theta$.

Both bias reduction and bias correction are tied to a specific parameterization. This means that if $\psi = \psi(\theta)$ is a nonlinear reparameterization of θ, the transformed estimator $\psi(\theta^*)$ or $\psi(\tilde\theta)$ will not have reduced bias of order $O(n^{-2})$. Equivariance under nonlinear componentwise reparameterizations is obtained with median bias reduction of Kenne Pagui et al,[8] leading to the estimator $\theta^\dagger$ satisfying, in the continuous case, the improved median centering property $\Pr_\theta(\theta^\dagger_r \le \theta_r) = 1/2 + O(n^{-3/2})$, $r = 1, \dots, d$, in contrast with the corresponding $O(n^{-1/2})$ order of error for the maximum likelihood estimator. More in detail, the adjusted score for median bias reduction, as given in formula (10) of the cited paper, has the form $U^\dagger(\theta) = U(\theta) + A^\dagger(\theta)$, with $A^\dagger(\theta) = A^*(\theta) - i(\theta)F(\theta)$. The vector $F(\theta)$ involves the quantities $P_{\theta_r}(\theta)$ and $Q_{\theta_r}(\theta)$ and its expression is given in the Appendix. The median bias reduced estimator $\theta^\dagger$ is obtained as a solution of $U^\dagger(\theta) = 0$.

Since both $A^*(\theta)$ and $A^\dagger(\theta)$ are of order $O(1)$, θ^* and $\theta^\dagger$ have the same asymptotic normal distribution as the maximum likelihood estimator.[7,8] In practice, standard errors are computed using diagonal elements of the inverse Fisher information, evaluated at the corresponding estimate, that is, $i(\theta^*)^{-1}$ and $i(\theta^\dagger)^{-1}$ respectively. The asymptotic coverage error of the associated Wald confidence intervals will be the same as with maximum likelihood, although empirical coverage error is typically better due to improved centering.

For the negative binomial regression model (1), we have $d = p + 1$ and the quantity $A^*(\theta) = (A^{*\top}_\beta, A^*_\phi)^\top$, whose derivation is in the Appendix, has

$$A^*_\beta = X^\top W\xi\,, \qquad A^*_\phi = \kappa'(\phi)\sum_{i=1}^n \frac{m_i h_i d_i^2 \mu_i^2}{2w_i V(Y_i)^2} + \frac{1}{2} i^{-1}_{\phi\phi} R_{\phi\phi}\,,$$

where $\xi = (\xi_1, \dots, \xi_n)^\top$, with $\xi_i = h_i d'_i/(2d_i w_i)$. The quantity h_i appearing in ξ_i and in A^*_ϕ 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 $d'_i = d^2\mu_i/d\eta_i^2$. The expression of $R_{\phi\phi}$ is given in the Appendix. For independent and identically distributed observations, with $\mu_i = \mu$, bias reduction for the shape parameter κ was considered by Zhang et al[26] in Example 4. The adjustment term $A^\dagger(\theta) = (A^{\dagger\top}_\beta, A^\dagger_\phi)^\top$ for median bias reduction has

$$A^\dagger_\beta = X^\top W(\xi + Xu)\,, \qquad A^\dagger_\phi = A^*_\phi + i^{-1}_{\phi\phi} S_{\phi\phi}\,, \tag{4}$$

where expressions for u and $S_{\phi\phi}$ are derived in the Appendix.

With simple algebra, the jth iteration of IWLS which updates the current iterate $\beta^{*(j)}$ leads to

$$\beta^{*(j+1)} = (X^\top W^{(j)}X)^{-1}X^\top W^{(j)} z^{*(j)}\,, \tag{5}$$

where $z^{*(j)} = z^{(j)} + \xi^{(j)}$ is the adjusted version of the working variate z defined in (2). The jth iteration of IWLS for $\beta^\dagger$ has the same expression as (5), with working variate $z^* + Xu$ in place of z^*.

All the improved methods for negative binomial regression, together with maximum likelihood fitting, are implemented in the `brnb` R function of the R package `brglm2`. Maximum likelihood fitting can also be performed using the `glm.nb` function of the MASS R package.

4 | SIMULATION STUDIES

In this section, the properties of the estimators are assessed through simulation under different scenarios corresponding to combinations of values of n, ϕ, and β. For each setting, we run 10 000 Monte Carlo replications. For all configurations, we used the logarithmic link function and the identity transformation for the shape parameter ($\phi = \kappa$). Maximum likelihood (ML), mean and median bias reduced (BR) estimates were computed using the `brnb` R function. Convergence is achieved when the absolute difference between the previous and current estimates is less than 10^{-8}. The default option sets to 100 the maximum number of iterations.

For a scalar parameter γ, let us denote by $\overline{\gamma}_r$, $r = 1, \ldots, R$, the rth Monte Carlo value of an estimator $\overline{\gamma}$ and by $se(\overline{\gamma}_r)$ its corresponding standard error, computed using the square root of a diagonal element of the inverse of Fisher information, evaluated at $\overline{\gamma}_r$. Let, in addition, $\text{MSE} = \sum_{r=1}^{R}(\overline{\gamma}_r - \gamma)^2/R$ be the empirical mean squared error, $\text{BIAS} = \overline{\gamma}_\bullet - \gamma$ the empirical mean bias, where $\overline{\gamma}_\bullet = \sum_{r=1}^{R}\overline{\gamma}_r/R$, and $\text{SD} = \sqrt{\sum_{r=1}^{R}(\overline{\gamma}_r - \overline{\gamma}_\bullet)^2/(R-1)}$ the empirical standard deviation. Moreover, let $I(A)$ be the indicator function of the set A and z_α the α-quantile of the standard normal distribution. Estimators are evaluated in terms of empirical probability of underestimation, $\text{PU} = \sum_{r=1}^{R} I(\overline{\gamma}_r \leq \gamma)/R$; estimated relative (mean) bias, $\text{RBIAS} = (\overline{\gamma}_\bullet - \gamma)/|\gamma|$; estimated root mean squared error, $\text{RMSE} = \sqrt{\text{MSE}}$; estimated coverage probability of 95% Wald-type confidence intervals, $\text{WALD} = \sum_{r=1}^{R} I(|\overline{\gamma}_r - \gamma| \leq se(\overline{\gamma}_r)z_{1-\alpha/2})/R$ and the relative increase in estimated mean squared error from its absolute minimum due to bias, $\text{IBMSE} = \{\text{MSE} - \text{SD}^2\}/\text{SD}^2 = \text{BIAS}^2/\text{SD}^2$. Except for RMSE, the performance measures are expressed in percentages.

We first conducted a simulation study with constant mean μ, that is, with intercept only, and shape parameter κ. In particular, for sample sizes $n = 20, 50, 150, 300$, $R = 10\,000$ Monte Carlo samples were drawn from the negative binomial with values of the parameters $\mu = 2, 5$ ($\beta = \log 2, \log 5$) and $\kappa = 0.5, 0.75, 1, 1.5, 2$. When the empirical variance is less than the mean, which happened in some simulated samples only with $n = 20$ and $n = 50$, ML, mean and median BR estimates do not exist. We denote by A_1 the number of simulated samples, out of 10 000, where this occurred. Nonconvergence of the ML algorithm was also observed in some simulated samples with empirical variance greater than the mean, especially with $\mu = 2$ and small values of κ. Few of these cases showed nonconvergence also for mean and median BR. In particular, in the $10\,000 - A_1$ samples with empirical variance greater than the mean, we found A_2 nonconvergence samples using ML. Among these A_2 samples, we found A_3 nonconvergence samples using mean BR and, finally, among these A_3 samples, we found A_4 nonconvergence samples using median BR. For each setting, with $n = 20, 50$, Table 1 gives the values of $A_j, j = 1, \ldots, 4$.

TABLE 1 Computational diagnostics in 10 000 replications

		$\mu = 2$					$\mu = 5$				
	κ	**0.5**	**0.75**	**1**	**1.5**	**2**	**0.5**	**0.75**	**1**	**1.5**	**2**
$n = 20$	A_1	535	214	108	43	17	18	3	0	0	0
	A_2	163	85	42	17	12	6	3	0	3	1
	A_3	2	1	0	0	1	0	0	0	2	1
	A_4	0	0	0	0	0	0	0	0	2	1
	10 000 - A_1 - A_2	9302	9701	9850	9940	9971	9976	9994	10 000	9997	9999
$n = 50$	A_1	36	6	0	0	0	0	0	0	0	0
	A_2	16	0	1	0	0	0	0	0	0	0
	A_3	2	0	0	0	0	0	0	0	0	0
	A_4	0	0	0	0	0	0	0	0	0	0
	10 000 - A_1 - A_2	9948	9994	9999	10 000	10 000	10 000	10 000	10 000	10 000	10 000

Note: A_1 indicates the number of samples with empirical variance less than the empirical mean. Of the remaining $10\,000 - A_1$ samples, A_2 is the number of nonconvergence samples using ML, which include A_3 nonconvergence samples using mean BR, which in turn include A_4 nonconvergence samples using median BR. The quantity 10 000 - A_1 - A_2 represents the number of samples with convergence for all methods out of 10 000 replications.

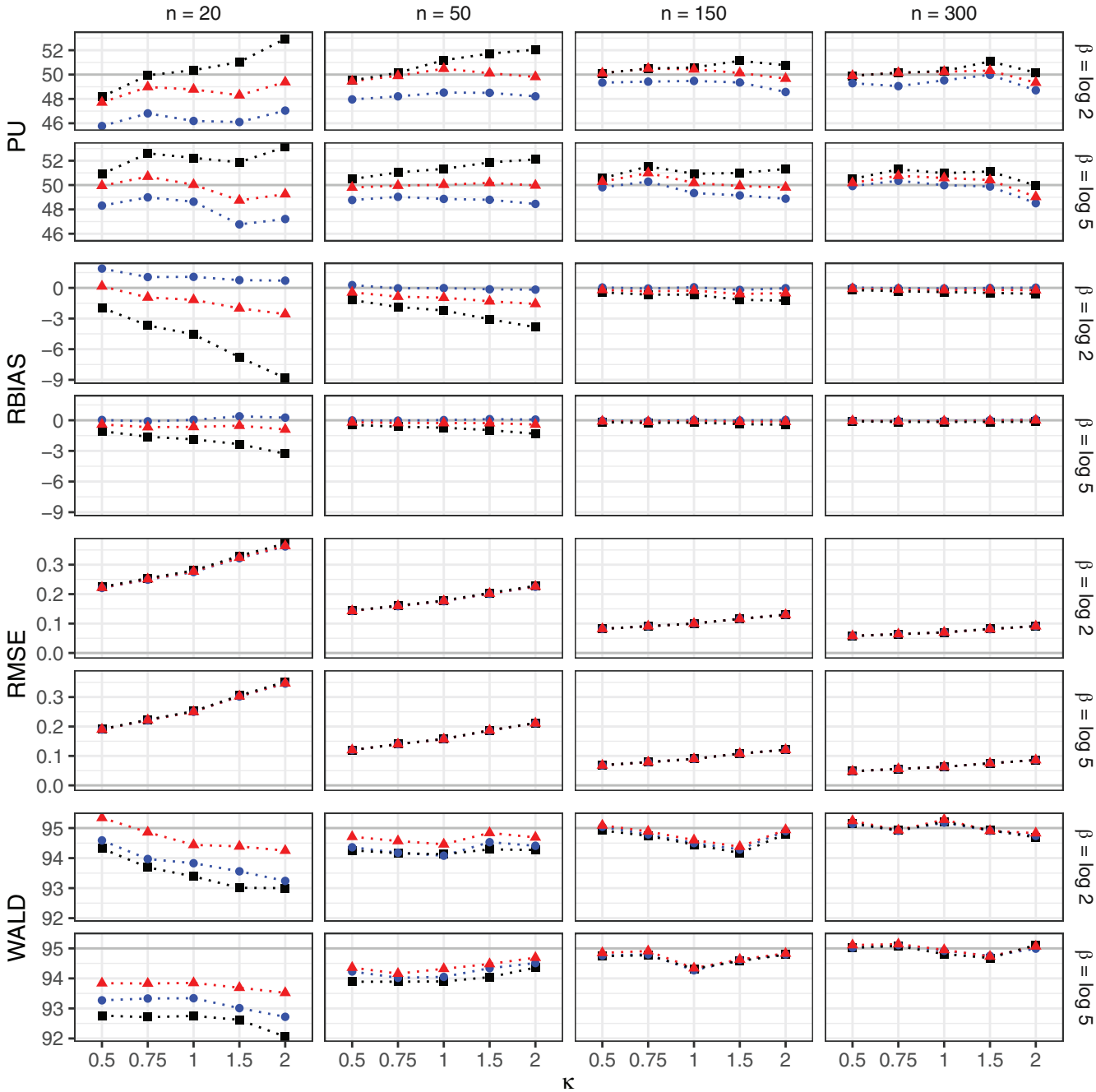

FIGURE 1 Estimated probability of under estimation (PU), relative bias (RBIAS), root mean squared error (RMSE) and coverage probability of 95% Wald-type confidence intervals (WALD) for the intercept $\beta = \log \mu$, with $\kappa = 0.5, 0.75, 1, 1.5, 2$ and $\beta = \log 2, \log 5$. Results for ML (black squares), mean BR (blue circles), and median BR (red triangles). Except for RMSE, the vertical axes represent percentages

In order to compare the methods on the same samples, the results reported in Figures 1 and 2 are based on the simulated samples in which all methods converged (10 000 - A_1 - A_2 samples in Table 1 for $n = 20, 50$, while all methods converge for $n = 150,300$). Moreover, as done in References 8,11, for $n = 20, 50$, the results based, for each method, on all the samples in which that method converged are displayed in Figures S1 and S2 in Section S.2 of the Supplementary Material. The qualitative conclusions of the simulation described below are unchanged.

Both mean and median BR achieve the desired goals, that is, are effective in mean and median centering, respectively, and are both preferable to ML. We recall, however, that mean centering is tied to a specific parameterization. Under this respect, the larger empirical relative bias of median BR of the shape parameter is not observed in other parameterizations, such as the inverse or the log parameterizations. Median BR provides empirical coverage of the 95% Wald-type confidence

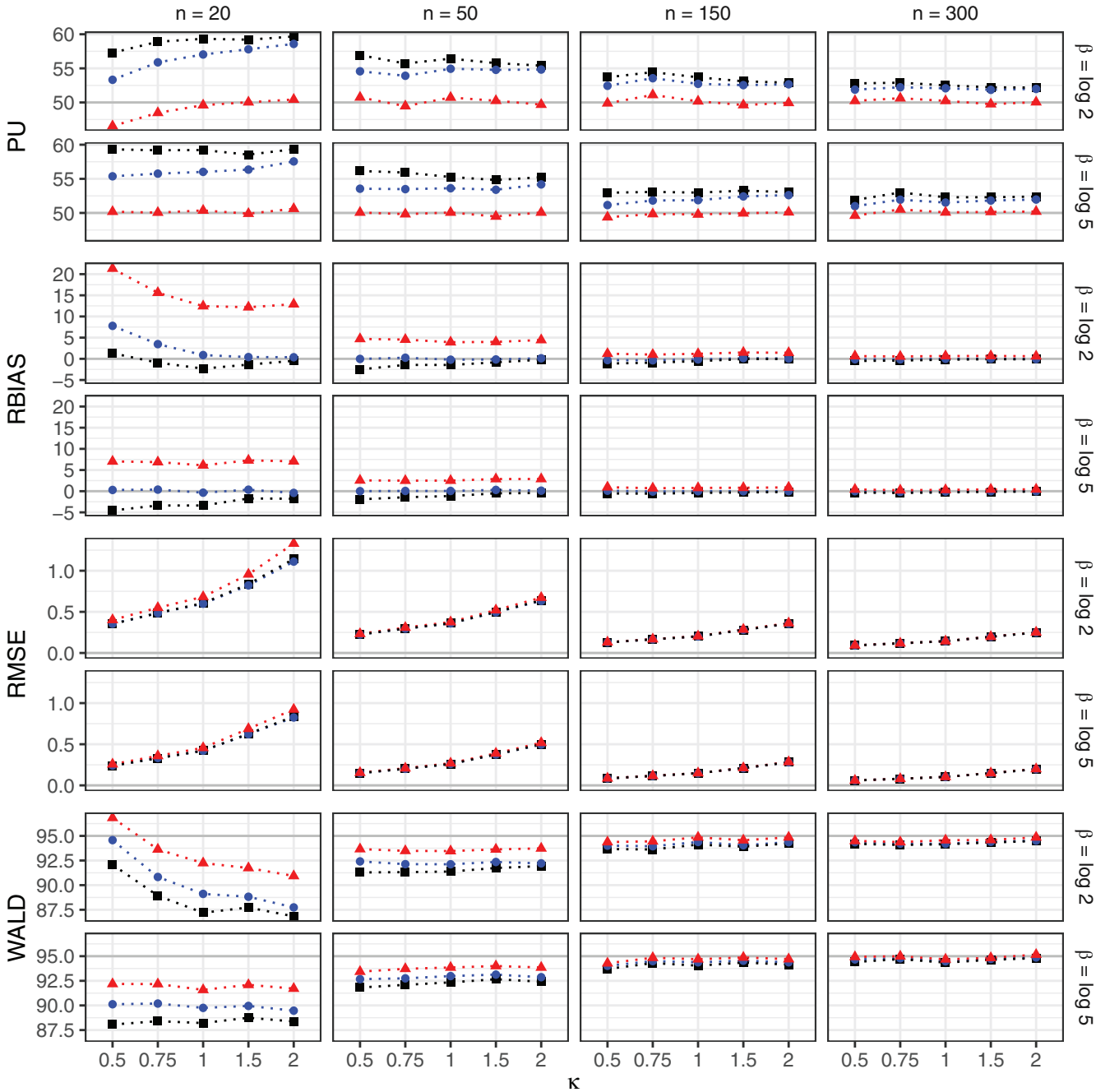

FIGURE 2 Estimated probability of under estimation (PU), relative bias (RBIAS), root mean squared error (RMSE) and coverage probability of 95% Wald-type confidence intervals (WALD) for the shape parameter κ, with $\kappa = 0.5, 0.75, 1, 1.5, 2$ and $\beta = \log 2, \log 5$. Results for ML (black squares), mean BR (blue circles), and median BR (red triangles). Except for RMSE, the vertical axes represent percentages

intervals better than its competitors, especially for the shape parameter and small sample sizes. As expected, all three estimators improve as the sample size and μ increase.

With the aim of checking the improvement in the order of bias for mean BR, and of the distance from 0.5 of the probability of underestimation for median BR, we simulated $N_r = 2^r N_0$ samples of size $n_r = 2^r n_0$ for $r \in \{0, \dots, 5\}$ with $n_0 = 20$ and $N_0 = 4000$. This simulation design guarantees that the simulation standard error is asymptotically bounded for any r. The results are given in Section S.3 of the Supplementary Material for $\mu = 5$ and $\kappa = 1$ and are in line with the theory. Indeed, mean BR provides a reduction in the order of the bias from $O(n^{-1})$ to $O(n^{-2})$. Similarly, the order of error in the probability of underestimation is seen to decrease from $O(n^{-1/2})$ to $O(n^{-3/2})$ for median BR, even though theoretically the result is only guaranteed for continuous models.

We now consider a second simulation study involving covariates. The linear predictor is connected to the mean with log link and the identity transformation for the shape parameter is considered ($\phi = \kappa$). In particular, we let

$$\log \mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}, \tag{6}$$

where x_{i2} and x_{i3} are independent realizations of Bernoulli random variables with probabilities 0.8 and 0.5, respectively; x_{i4} are independent realizations of a uniform on $(1, 2)$; x_{i5} are independent realizations of a Poisson with mean 2.5, $i = 1, \ldots, n$. The true parameter values are $\beta_1 = 1, \beta_2 = -0.75, \beta_3 = -1.5, \beta_4 = 1$, and $\beta_5 = -0.5$. Four values are considered for the shape parameter, $\kappa = 0.5, 0.75, 1, 1.5$. The sample sizes considered are $n = 40, 80, 150, 300$. For each combination of β, κ and n, we run 10 000 Monte Carlo replications, where the values of the explanatory variables x_{i2}, x_{i3}, x_{i4}, and x_{i5} were held constant throughout the simulations.

The summaries of the simulation results for the regression coefficients are presented in Figure 3 with $\kappa = 0.75$. Other values of κ gave similar results, which are summarized in Figures S4 to S6 in Section S.4 of the Supplementary Material. Figure 4 summarizes results for κ. With $\kappa = 0.75$ and $n = 40$, we found 39, 11 and 9 samples out of 10 000 where the IWLS algorithm did not reach convergence for ML, mean BR, and median BR, respectively, while no convergence problems arose for $n = 80, 150, 300$. A sufficient condition for existence of the ML estimate[27] is satisfied in most of the nonconvergence cases. Therefore, nonconvergence is mostly due to numerical problems. Moreover, the percentage of samples showing nonconvergence decreases as κ increases. As for the previous simulation study, the reported results are based only on samples which led to convergence for all methods. Looking at the four performance measures, it appears

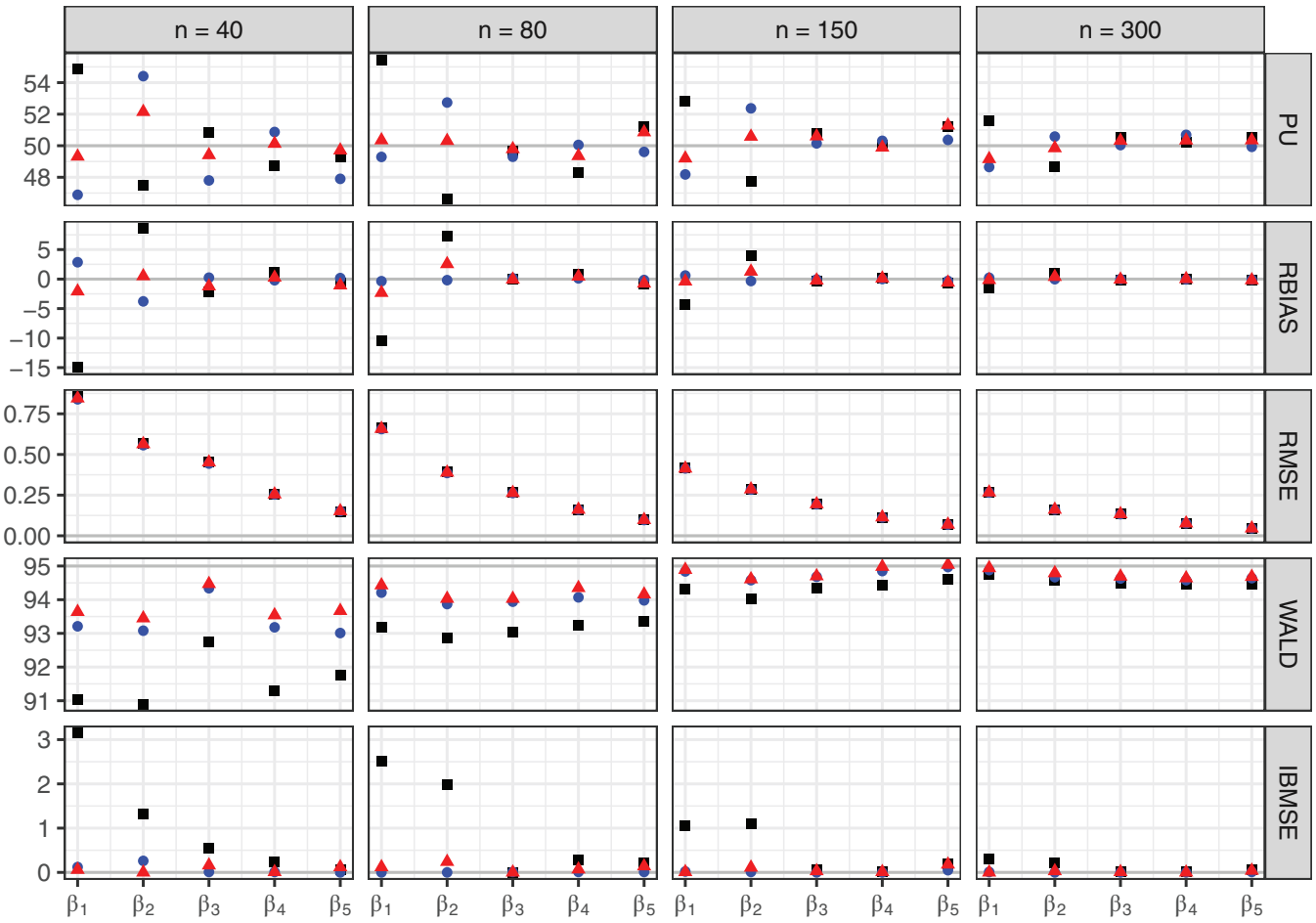

FIGURE 3 Estimated probability of under estimation (PU), relative bias (RBIAS), root mean squared error (RMSE), coverage probability of 95% Wald-type confidence intervals (WALD) and increase in estimated mean squared error (IBMSE) for estimation of regression parameters $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ with $\kappa = 0.75, n = 40, 80, 150, 300$. Simulation results for ML (black squares), mean BR (blue circles), and median BR (red triangles). Except for RMSE, the vertical axes represent percentages

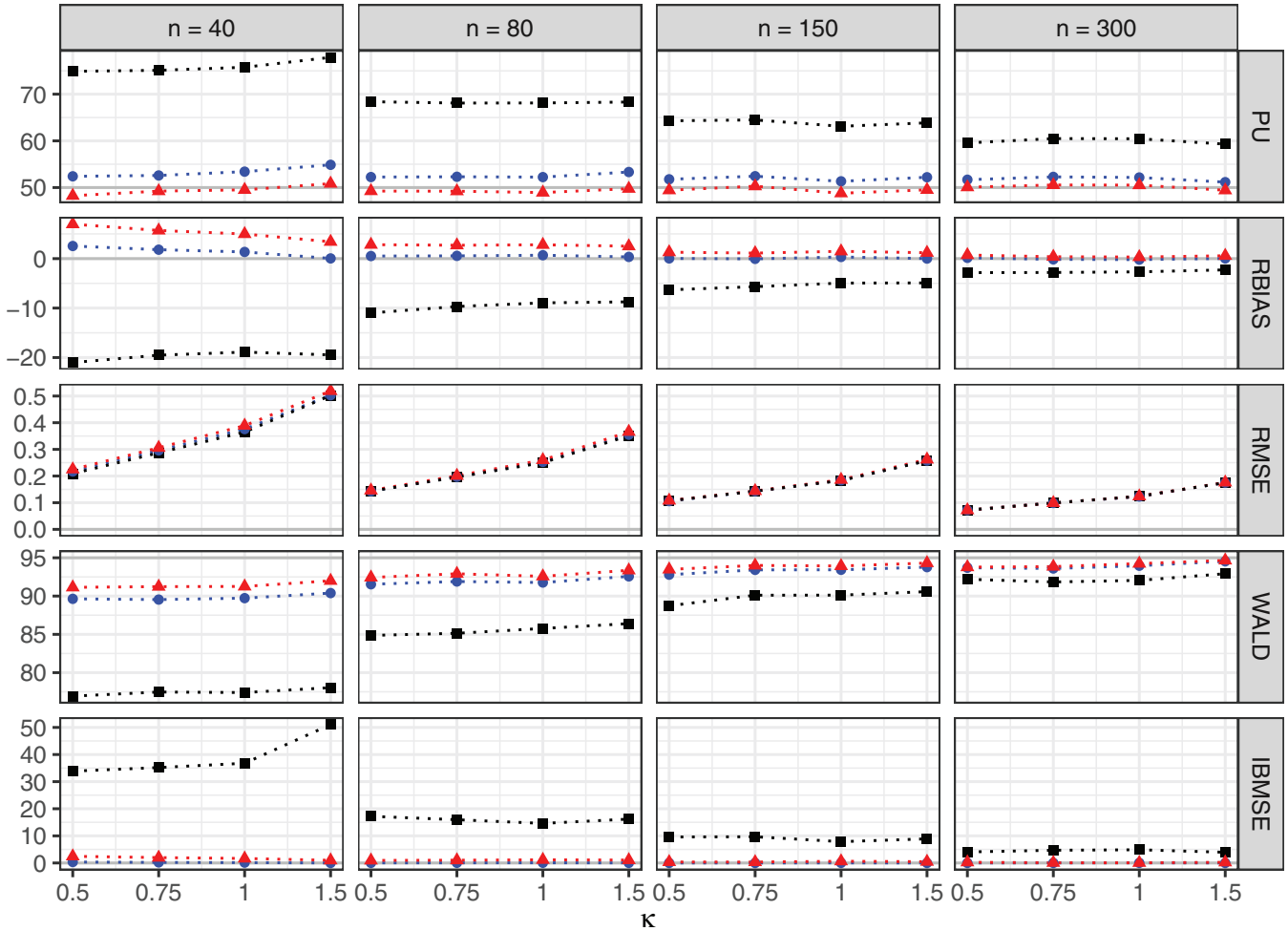

FIGURE 4 Estimated probability of under estimation (PU), relative bias (RBIAS), root mean squared error (RMSE), coverage probability of 95% Wald-type confidence intervals (WALD) and increase in estimated mean squared error (IBMSE) for estimation of shape parameter κ with $n = 40, 80, 150, 300$. Simulation results for ML (black squares), mean BR (blue circles), and median BR (red triangles). Except for RMSE, the vertical axes represent percentages

that mean and median BR are preferable to ML for small n. On the other hand, the results improve for all three methods as n increases. As κ increases, for estimation of regression coefficients, median BR is comparable to mean BR in terms of estimated relative (mean) bias, while it proves to be more accurate in achieving median centering. Moreover, in all scenarios, median BR provides the empirical coverages of Wald-type confidence intervals closest to the 95% nominal value. The results in Figure 4 show that the improvement given by both mean and median BR over ML is substantial in all scenarios and more pronounced than in the previous case with the intercept parameter only.

Finally, in Section S.5 of the Supplementary Material, we investigated by simulation the behavior of the methods when the generating model has no overdispersion. In particular, samples of size $n = 40$ are generated from a Poisson regression model with mean satisfying (6). As expected, in many samples the negative binomial fitting procedures did not converge. In such cases, the corresponding procedure for Poisson regression[14] was used. The results in Table S1 indicate that mean and median BR are essentially equivalent to, and sometimes better than, ML for inference about β.

5 | CASE STUDIES

We consider two case studies, namely the Ames salmonella assay data and the epileptic seizures data. The first data set has one explanatory variable with 6 levels and 3 observations each. The second data set has counts of epileptic seizures for 59 matched pairs.

TABLE 2 Ames salmonella assay: Parameter estimates and corresponding standard errors in parenthesis

	ML	Mean BC	Mean BR	Median BR
β_0	2.198 (0.325)	2.210 (0.348)	2.216 (0.352)	2.211 (0.359)
β_1	−0.001 (0.00039)	−0.001 (0.00042)	−0.001 (0.00042)	−0.001 (0.00043)
β_2	0.313 (0.088)	0.311 (0.095)	0.309 (0.096)	0.309 (0.098)
κ	0.049 (0.028)	0.063 (0.033)	0.065 (0.033)	0.069 (0.035)

5.1 | Ames salmonella data

We consider data from an Ames salmonella reverse mutagenicity assay, previously analyzed using negative binomial regression by several authors[6,18,21,28] in order to account for the observed overdispersion. The response variable Y corresponds to the number of revertant colonies observed on a plate, while covariate x is the dose level of quinoline on the plate. Three observations were taken at each of six dose levels leading to a total of 18 observations. We focus on the analysis based on the log-linear model[28]

$$\log \mu_i = \beta_0 + \beta_1 x_i + \beta_2 \log(x_i + 10). \tag{7}$$

In the above expression, the constant 10 represents the smallest non-zero dose level. The main interest is focused on testing significance of mutagenic effect, that is, the null hypothesis $H_0 : \beta_2 = 0$. The presence of overdispersion was confirmed by the Pearson statistic based on the residuals of the Poisson model. We also compared Poisson and negative binomial models using parametric bootstrap. The code is available in Section S.6 of the Supplementary Material. The results of both tests support the choice of a negative binomial model.

Table 2 shows the estimates obtained with ML, mean bias correction (BC), mean BR, and median BR using the identity transformation for the shape parameter ($\phi = \kappa$). Estimates of the regression coefficients have the same interpretation as in Poisson log linear models and the values here turn out to be comparable across methods. Mean and median bias reduced estimates of the shape parameter are comparable, but slightly different from the maximum likelihood estimate. This in turn reflects on the standard errors of the regression parameter estimates.

A simulation study, with covariates fixed at the observed values and true parameter value equal to the observed mean BR estimate is included in Section S.6 of the Supplementary Material and confirms the findings of the previous section, with mean and median BR showing an improved repeated sampling behavior with respect to ML.

5.2 | Epileptic seizures data

We consider here the epileptic seizures data on 2-week seizure counts for 59 epileptics.[22] The data were analyzed by several Authors.[3,29] The number of seizures was recorded for a baseline period of 8 weeks, and then patients were randomly assigned to a treatment group or a control group. Counts were then recorded for four successive 2-weeks periods. The response was the number of observed seizures. We analyzed the data by comparing the response before and after the treatment, hence obtaining a set of 59 matched pairs. The only covariates in the linear predictor are then given by the two treatment indicators. As in the previous example, Poisson overdispersion was confirmed both by Pearson statistic and parametric bootstrap. The code is available in Section S.7 of the Supplementary Material. Therefore, we assume a negative binomial model for the response Y_{ij}, $i = 1, \dots, 59$, $j = 1, 2$, with mean and variance

$$\mu_{ij} = \exp(\lambda_i + x_{ij}^\top \beta), \qquad V(Y_{ij}) = \mu_{ij} + \kappa \mu_{ij}^2 ,$$

where intercepts λ_i determine the stratified structure corresponding to each subject, $x_{i1} = (0, 0)^\top$, while $x_{i2} = (1, 0)^\top$ if subject i received the placebo and $x_{i2} = (0, 1)^\top$ if subject i received the treatment. We focus on inference about $\beta = (\beta_1, \beta_2)^\top$ and κ, while the intercepts $\lambda_1, \dots, \lambda_{59}$ are treated as incidental nuisance parameters. This is a rather extreme case of fixed effects model for clustered data where it is well known that ML inference for the parameters of interest is problematic (see p. 292 of Cox and Hinkley[9]). Therefore, it is of particular interest to assess the behavior of BR methods, which provide improved estimates also for the nuisance parameters. Negative binomial regression for clustered data is considered in section 7.5.5 of Demidenko.[30]

Figure 5 displays the parameter estimates and the corresponding confidence intervals obtained with different methods. In addition to ML, mean BC, mean BR and median BR, modified profile likelihood[31] (MPL) for the three-dimensional parameter of interest $(\beta_1, \beta_2, \kappa)^\top$ has also been included for comparison. MPL for this model was previously proposed by Bellio and Sartori[29] due to its higher-order accuracy in models with many nuisance parameters.[32] In this setting, the same higher order accuracy is guaranteed by BR methods.[8,33] The difference between mean and median BR and MPL with respect to ML is particularly pronounced for κ. This reflects on the lengths of confidence intervals, with those from ML being inaccurately too short, as illustrated by the simulation results below.

We run 10 000 replications with covariates fixed at the observed value and true parameters set to the observed mean BR estimates. We found only 13 samples out of 10 000 where the IWLS algorithm did not reach convergence for ML, of these, 4 showed nonconvergence also for mean BR and median BR. The results are reported in Table 3 for the 9987

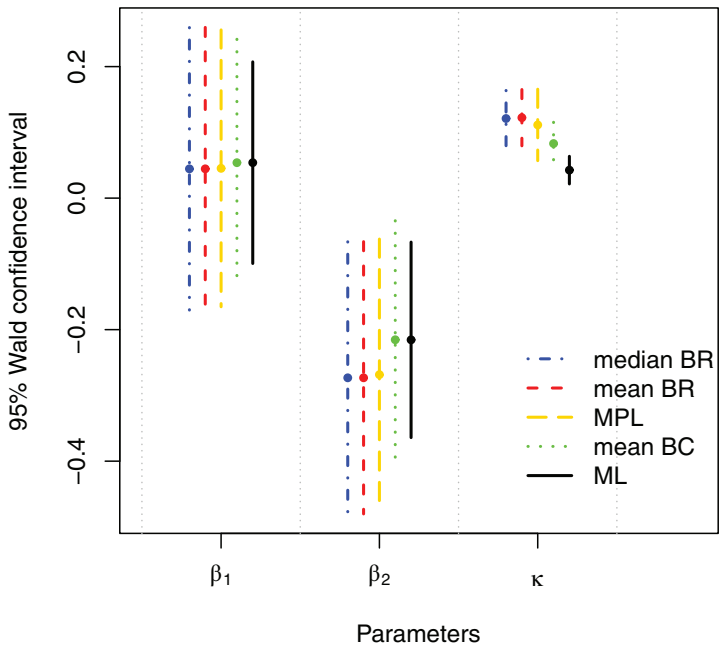

FIGURE 5 Epileptic seizures data: Points represent the parameter estimates while the vertical lines represent 95% Wald-type confidence intervals

TABLE 3 Epileptic seizures data: Simulation results for ML (hat), mean BC (tilde), mean BR (star), and median BR (dagger) of the parameters of interest

	PU	RBIAS	RMSE	WALD	IBMSE
$\hat{\beta}_1$	50.21	−0.05	0.11	80.90	0.00
$\tilde{\beta}_1$	50.28	−0.30	0.11	89.44	0.00
β_1^*	50.38	−0.12	0.11	94.28	0.00
$\beta_1^\dagger$	50.37	−0.10	0.11	94.25	0.00
$\hat{\beta}_2$	49.54	0.92	0.11	81.39	0.05
$\tilde{\beta}_2$	49.43	0.99	0.11	89.18	0.06
β_2^*	50.34	0.23	0.11	93.98	0.00
$\beta_2^\dagger$	50.36	0.22	0.11	93.96	0.00
$\hat{\kappa}$	100.00	−67.31	0.08	0.25	3331.25
$\tilde{\kappa}$	96.21	−36.26	0.05	31.22	360.50
κ^*	46.09	3.80	0.03	82.50	2.02
$\kappa^\dagger$	48.14	2.44	0.03	82.88	0.88

samples in which the IWLS algorithm achieved convergence for all the approaches. For the regression coefficients, all the approaches are almost equivalent in terms of PU, RBIAS, and RMSE, while coverage of confidence intervals based on BR methods substantially improves upon ML. Mean BC, while providing a noticeable improvement upon ML, is not as good as BR methods. This is mainly related to estimation of the shape parameter. Indeed, in this extreme scenario, only BR methods are seen to provide reasonable inference about κ.

Although not of direct interest in the present example, both mean and median BR provide improved estimates also of the nuisance parameters. The simulation results for these are presented in Figure S7 of the Supplementary Material. Once again, we can appreciate the improved performance of mean and median BR by looking at the coverages of 95% Wald-type confidence intervals which are closest to the nominal value.

6 | DISCUSSION

For negative binomial regression, we developed inference based on adjusted score equations for mean and median bias reduction.[7,8] Simulation results confirm the theoretical properties of the methods and indicate that they are both effective in improving over standard likelihood inference and even over traditional mean bias correction. This is especially notable when the number of parameters is large compared to the sample size, as is illustrated by the simulation results for the case study in Section 5.2. These methods also solve in most cases numerical problems that may occur with ML, and consequently with mean bias correction. Even though mean and median bias reduction aim at different centering properties, in practice they lead to similar conclusions. On the other hand, median bias reduction seems slightly preferable in terms of coverage accuracy of Wald confidence intervals. Other types of confidence intervals such as those based on the asymptotic distribution of the likelihood ratio or score statistics could be preferable to Wald intervals. However, these are not available for mean and median bias reduced estimators. Construction of adjusted score intervals could be the object of future research.

ACKNOWLEDGEMENTS

Open Access Funding provided by Universita degli Studi di Padova within the CRUI-CARE Agreement. [Correction added on 26 May 2022, after first online publication: CRUI funding statement has been added.]

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Euloge Clovis Kenne Pagui https://orcid.org/0000-0002-8998-9251
Nicola Sartori https://orcid.org/0000-0002-3063-8385

REFERENCES

1. Agresti A. *Foundations of Linear and Generalized Linear Models*. Hoboken, NJ: John Wiley & Sons; 2015.
2. Hilbe J. *Negative Binomial Regression*. 2nd ed. Cambridge, UK: Cambridge University Press; 2011.
3. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer; 2002.
4. Fisher RA. The negative binomial distribution. *Ann Eugen*. 1941;11:182-187.
5. Anscombe FJ. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*. 1950;37:358-382.
6. Lawless JF. Negative binomial and mixed Poisson regression. *Can J Stat*. 1987;15:209-225.
7. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80:27-38.
8. Kenne Pagui EC, Salvan A, Sartori N. Median bias reduction of maximum likelihood estimates. *Biometrika*. 2017;104:923-938.
9. Cox DR, Hinkley DV. *Theoretical Statistics*. London, UK: Chapman & Hall; 1974.
10. Kosmidis I, Firth D. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*. 2021;108:71-82.
11. Kosmidis I, Firth D. Bias reduction in exponential family nonlinear models. *Biometrika*. 2009;96:793-804.
12. Kosmidis I, Firth D. A generic algorithm for reducing bias in parametric estimation. *Electron J Stat*. 2010;4:1097-1112.
13. Kosmidis I, Kenne Pagui EC, Sartori N. Mean and median bias reduction in generalized linear models. *Stat Comput*. 2020;30:43-59.
14. Kosmidis I. brglm2: bias reduction in generalized linear models; 2021. R package version 0.8.2.
15. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021.
16. Grün B, Kosmidis I, Zeileis A. Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *J Stat Softw*. 2012;48:1-25.
17. Kosmidis I. Improved estimation in cumulative link models. *J Royal Stat Soc Ser B Stat Methodol*. 2014;76:169-196.
18. Saha K, Paul S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*. 2005;61:179-185.

19. Clark SJ, Perry JN. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*. 1989;45:309-316.
20. Piegorsch WW. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*. 1990;46:863-867.
21. Margolin BH, Kim BS, Risko KJ. The Ames Salmonella/microsome mutagenicity assay: Issues of inference and validation. *J Am Stat Assoc*. 1989;84:651-661.
22. Thall PF, Vail SC. Some covariance models for longitudinal count data with overdispersion. *Biometrics*. 1990;46:657-671.
23. Kosmidis I. Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdiscipl Rev Comput Stat*. 2014;6:185-196.
24. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ*. 2016;352:i1981.
25. Barndorff-Nielsen OE, Cox DR. *Inference and Asymptotics*. London, UK: Chapman & Hall; 1994.
26. Zhang X, Paul S, Wang YG. Small sample bias correction or bias reduction? *Commun Stat Simul Comput*. 2021;50:1165-1177.
27. Gning LD, Pierre-Loti-Viaud D. On the existence of maximum likelihood estimators in Poisson-gamma HGLM and negative binomial regression model. *Electron J Stat*. 2013;7:2577-2594.
28. Breslow NE. Extra-Poisson variation in log-linear models. *J Royal Stat Soc Ser C (Appl Stat)*. 1984;33:38-44.
29. Bellio R, Sartori N. Practical use of modified maximum likelihoods for stratified data. *Biometr J*. 2006;48:876-886.
30. Demidenko E. *Mixed Models - Theory and Applications with R*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2013.
31. Barndorff-Nielsen OE. On a formula for the distribution of the maximum likelihood estimator. *Biometrika*. 1983;70:343-365.
32. Sartori N. Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*. 2003;90:533-549.
33. Lunardon N. On bias reduction and incidental parameters. *Biometrika*. 2018;105:233-238.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kenne Pagui EC, Salvan A, Sartori N. Improved estimation in negative binomial regression. *Statistics in Medicine*. 2022;41(13):2403-2416. doi: 10.1002/sim.9361

APPENDIX A

Quantities involved in $A^{\dagger}(\theta)$ for a general parametric model

As given in Section 2, the general expression of the median BR adjustment is $A^{\dagger}(\theta) = A^{*}(\theta) - i(\theta)F(\theta)$. The vector $F(\theta)$ has components $F_r = [i(\theta)^{-1}]_r^{\top}\tilde{F}_r$, where $\tilde{F}_r$ has elements

$$\tilde{F}_{r,t} = \mathrm{tr}[g_r\{(1/3)P_{\theta_t}(\theta) + (1/2)Q_{\theta_t}(\theta)\}], \qquad r,\ t = 1, \dots, d,$$

with the matrix g_r given by

$$g_r = (i^{rr}(\theta))^{-1}[i(\theta)^{-1}]_r[i(\theta)^{-1}]_r^{\top}, \qquad r = 1, \dots, d.$$

Above and elsewhere, $[C]_r$ denotes the rth column of a matrix C, while $i^{rr}(\theta)$ is the (r, r) element of $i(\theta)^{-1}$.

Quantities involved in $A^{*}(\theta)$ and $A^{\dagger}(\theta)$ for negative binomial regression

Let $\mathrm{diag}\{e_1, \dots, e_n\}$ denote a diagonal matrix having $(e_1, \dots, e_n)$ as its main diagonal. Let, in addition, 1_n be a n-vector of ones and I_n the identity matrix of order n.

In order to give the expressions of matrix quantities appearing in (3), we use the index s, $s = 1, \dots, p$, for elements of β and the subscript ϕ for the ϕ parameter. For simplicity, the argument θ will be omitted. We get

$$P_{\beta_s} + Q_{\beta_s} = \begin{bmatrix} X^{\top}X_s^D D^{-1}D'WX & 0_p \\ 0_p^{\top} & 0 \end{bmatrix},$$

where X_s^D denotes the diagonal matrix with elements of the sth column of the matrix X as its main diagonal and $D' = \mathrm{diag}\{d_1', \dots, d_n'\}$. Moreover, letting $R = P_{\phi} + Q_{\phi}$, we have

$$R = \begin{bmatrix} R_{\beta\beta} & R_{\beta\phi} \\ R_{\phi\beta} & R_{\phi\phi} \end{bmatrix},$$

with

$$
\begin{aligned}
R_{\beta\beta} &= \kappa'(\phi)X^\top D^2\Omega M^2\mathcal{V}^{-2}X,\\
R_{\beta\phi} &= R_{\phi\beta}^\top\\
&= \kappa'(\phi)^2X^\top D\Omega\{M(\kappa M+I_n)\}^{-1}\{E_1-ME_2-M^3(\kappa M+I_n)^{-1}\}1_n,\\
R_{\phi\phi} &= \kappa'(\phi)^3\sum_{i=1}^n m_i\left\{-2E(S_{3i})+\frac{2\kappa^2\mu_i^3+9\kappa\mu_i^2+6\mu_i}{\kappa^3(\kappa\mu_i+1)^2}-\frac{6}{\kappa^4}\log(\kappa\mu_i+1)\right.\\
&\quad\left.+2E(S_{1i}S_{2i})-\frac{2\mu_i}{\kappa\mu_i+1}E(S_{2i}Y_i)-\frac{2\{\kappa\mu_i-(\kappa\mu_i+1)\log(\kappa\mu_i+1)\}}{\kappa^2(\kappa\mu_i+1)}E(S_{2i})\right\}\\
&\quad+i_{\kappa\kappa}\kappa'(\phi)\kappa''(\phi),
\end{aligned}
$$

where $\Omega=\mathrm{diag}\{m_1,\ \dots\ ,m_n\}$, $M=\mathrm{diag}\{\mu_1,\ \dots\ ,\mu_n\}$, $\mathcal{V}=\mathrm{diag}\{v_1,\ \dots\ ,v_n\}$, with $v_i=V(Y_i)$, $S_{ai}=\sum_{j=0}^{y_i^*}j^a/(\kappa j+1)^a$, $a=1,2,3$, $E_1=\mathrm{diag}\{E(S_{21}Y_1),\ \dots\ ,E(S_{2n}Y_n)\}$ and $E_2=\mathrm{diag}\{E(S_{21}),\ \dots\ ,E(S_{2n})\}$.

In order to give the expressions for the additional quantities u and $S_{\phi\phi}$ appearing in (4), we denote by $i_{\beta\beta}^{ss}$ the (s,s) element of $i_{\beta\beta}^{-1}$ and we let $v_i'=dv_i/d\mu_i=1+2\kappa\mu_i$. Then, $u=(u_1,\ \dots\ ,u_p)^\top$ with

$$
u_s=[(X^\top WX)^{-1}]_s^\top X^\top\begin{bmatrix} h_{s,1}\left\{d_1v_1'/(6v_1)-d_1'/(2d_1)\right\}\\ \vdots\\ h_{s,n}\left\{d_nv_n'/(6v_n)-d_n'/(2d_n)\right\}\end{bmatrix}.
$$

In the above expression, $h_{s,i}$ is the ith diagonal element of XG_sX^TW, with $G_s=(i_{\beta\beta}^{ss})^{-1}[i_{\beta\beta}^{-1}]_s[i_{\beta\beta}^{-1}]_s^\top$.

Finally,

$$
\begin{aligned}
S_{\phi\phi} &= \kappa'(\phi)^3\sum_{i=1}^n m_i\left\{-\frac{2}{3}E(S_{3i})+\frac{1}{3}\frac{2\kappa^2\mu_i^3+9\kappa\mu_i^2+6\mu_i}{\kappa^3(\kappa\mu_i+1)^2}-\frac{2}{\kappa^4}\log(\kappa\mu_i+1)\right.\\
&\quad\left.+\frac{1}{2}E(S_{1i}S_{2i})-\frac{1}{2}\frac{\mu_i}{\kappa\mu_i+1}E(S_{2i}Y_i)-\frac{\kappa\mu_i-(\kappa\mu_i+1)\log(\kappa\mu_i+1)}{2\kappa^2(\kappa\mu_i+1)}E(S_{2i})\right\}\\
&\quad+\frac{1}{2}i_{\kappa\kappa}\kappa'(\phi)\kappa''(\phi)\ .
\end{aligned}
$$