Computational Statistics and Data Analysis 138 (2019) 126–142

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Location-adjusted Wald statistics for scalar parameters☆

Claudia Di Caterina [a,b,*], Ioannis Kosmidis [c,d]

[a] Department of Statistical Sciences, University of Padova, via Battisti 241, 35121 Padova, Italy
[b] Faculty of Economics and Management, University of Bolzano, piazza Università 1, 39100 Bolzano, Italy
[c] Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[d] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

ARTICLE INFO

Article history:
Received 10 September 2018
Received in revised form 11 March 2019
Accepted 5 April 2019
Available online 16 April 2019

Keywords:
Beta regression
Bias reduction
Data separation
Generalized linear models
Infinite estimates
Magnetic resonance imaging

ABSTRACT

Inference about a scalar parameter of interest is a core statistical task that has attracted immense research in statistics. The Wald statistic is a prime candidate for the task, on the grounds of the asymptotic validity of the standard normal approximation to its finite-sample distribution, simplicity and low computational cost. It is well known, though, that this normal approximation can be inadequate, especially when the sample size is small or moderate relative to the number of parameters. A novel, algebraic adjustment to the Wald statistic is proposed, delivering significant improvements in inferential performance with only small implementation and computational overhead, predominantly due to additional matrix multiplications. The Wald statistic is viewed as an estimate of a transformation of the model parameters and is appropriately adjusted, using either maximum likelihood or reduced-bias estimators, bringing its expectation asymptotically closer to zero. The location adjustment depends on the expected information, an approximation to the bias of the estimator, and the derivatives of the transformation, which are all either readily available or easily obtainable in standard software for a wealth of models. An algorithm for the implementation of the location-adjusted Wald statistics in general models is provided, as well as a bootstrap scheme for the further scale correction of the location-adjusted statistic. Ample analytical and numerical evidence is presented for the adoption of the location-adjusted statistic in prominent modelling settings, including inference about log-odds and binomial proportions, logistic regression in the presence of nuisance parameters, beta regression, and gamma regression. The location-adjusted Wald statistics are used for the construction of significance maps for the analysis of multiple sclerosis lesions from MRI data.

1. Introduction

Testing hypotheses and constructing confidence intervals for scalar parameters are key statistical tasks that are usually carried out relying on large-sample results about likelihood-based quantities. Under a model that is specified partially or fully by the null hypothesis, the signed root of the logarithm of the likelihood ratio (LR) statistic, the score statistic and the Wald statistic are equivalent to the first order (Pace and Salvan, 1997, § 3.4.1), yet the use of the latter for inference is more widespread. The Wald statistic involves a direct comparison between the estimated and the hypothetical value

☆ Supplementary material available at https://github.com/ikosmidis/waldi/tree/master/inst/supplementary_1710-11217.

* Corresponding author.
E-mail addresses: claudia.dicaterina@unibz.it (C. Di Caterina), ioannis.kosmidis@warwick.ac.uk (I. Kosmidis).

https://doi.org/10.1016/j.csda.2019.04.004
0167-9473/

Table 1
Maximum likelihood (ML) and reduced-bias (RB) estimates, corresponding standard errors (in parenthesis) and 95% Wald-type confidence intervals for the parameters of model (1).

	Estimates				95% Confidence intervals			
	ML		RB		ML		RB	
β_1	1.123	(0.143)	1.114	(0.148)	0.843	1.403	0.824	1.405
β_2	−0.742	(0.143)	−0.734	(0.148)	−1.021	−0.462	−1.024	−0.444
β_3	0.486	(0.133)	0.441	(0.141)	0.225	0.747	0.165	0.717
β_4	−0.581	(0.133)	−0.532	(0.140)	−0.841	−0.321	−0.807	−0.257
γ_1	3.304	(0.223)	3.092	(0.225)	2.868	3.741	2.652	3.533
γ_2	1.747	(0.262)	1.654	(0.264)	1.232	2.261	1.138	2.171
γ_3	1.229	(0.267)	1.048	(0.271)	0.705	1.753	0.518	1.578

of the parameter, accounting also for estimation uncertainty. As a result, and in contrast to its main competitors, its computation does not require fitting the model under the null hypothesis, which can be time-consuming for complex models or when there is a need to perform many tests.

The Wald test can, though, demonstrate anomalies in its power, mainly because of the use of parameter estimates in the variance part of the statistic (Mantel, 1987). Hauck and Donner (1977) and Fears et al. (1996) study such anomalies in logistic regression models and in one-way random effects analyses of variance, respectively, and Væth (1985) gives mathematical conditions under which Wald procedures can suffer from scarce power in the more general context of exponential family models. In addition, the performance of Wald-type inference depends directly and, sometimes, critically on the properties of the estimator used in the statistic. A common strategy in enhancing Wald inference is to replace the maximum likelihood (ML) estimator with another that has improved frequentist properties and a limiting normal distribution. One prominent example is the use of a moment-based estimator for the dispersion parameter in generalized linear models with unknown dispersion, as is recommended in McCullagh and Nelder (1989, § 8.3) and implemented in the `summary.glm` function of the `stats` R package (R Core Team, 2018). Another recent example is in Kosmidis and Firth (2010), who illustrate that the finite-sample bias of the ML estimator of the precision parameter in beta regression models results in excessively narrow Wald-type confidence intervals (CIs) and anti-conservative Wald tests, and propose the use of a reduced-bias (RB) estimator to alleviate those issues. Below is a working illustration of that proposal in a beta regression model involving precision covariates.

Example 1.1. Smithson and Verkuilen (2006) use beta regression to investigate the relative contribution of nonverbal IQ to the distribution of $n = 44$ children's scores on a reading accuracy test, controlling for the presence of diagnosed dyslexia. The score of the ith child is assumed to be from a beta random variable with mean μ_i and variance $\mu_i(1-\mu_i)/(1+\phi_i)$, with $\phi_i > 0$. The score mean μ_i and precision ϕ_i are linked with covariates through the relationships

$$\log \frac{\mu_i}{1-\mu_i} = \beta_1 + \sum_{j=2}^{4} \beta_j x_{ij} \quad \text{and} \quad \log \phi_i = \gamma_1 + \sum_{j=2}^{3} \gamma_j x_{ij} \quad (i = 1, \ldots, n), \tag{1}$$

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\top \in \Re^4$, $\gamma = (\gamma_1, \gamma_2, \gamma_3)^\top \in \Re^3$ are the vectors of regression coefficients, x_{i2} takes value -1 if the ith child is dyslexic and 1 if not, x_{i3} is the nonverbal IQ score, and $x_{i4} = x_{i2}x_{i3}$ is the interaction between dyslexia status and nonverbal IQ score.

The parameters β and γ in (1) are estimated using the ML estimator and its RB version, as these are implemented in the R package `betareg` (Grün et al., 2012). Table 1 shows the resulting estimates, the corresponding estimated standard errors based on the expected information matrix, and the nominally 95% individual Wald-type CIs. As is also noted in Kosmidis and Firth (2010), bias reduction inflates estimated standard errors, and results in CIs that are wider than those based on ML, better reflecting the uncertainty about the values of the parameter. To illustrate this, the coverage probabilities of individual Wald-type intervals are estimated at levels 90%, 95% and 99%, using 50 000 samples simulated under the ML fit. The results in Table 2 suggest that the use of the RB estimates in Wald-type CIs brings the empirical coverage probabilities closer to the nominal value.

In the current paper, we improve Wald inference about scalar parameters through a novel, more explicit approach than the one presented in Example 1.1, which exploits the direct dependence of the Wald statistic on the estimator. Location-adjusted (LA) Wald statistics are defined whose expectations are asymptotically closer to that of the limiting normal distribution under the null hypothesis. The developments in this paper also shed light on why the use of RB estimators in the Wald statistics can improve inferential performance in cases, like it does in Example 1.1.

Corresponding methods have been first proposed by Bartlett (1937) for enhancing first-order inference based on the LR statistic, and have later been applied to other test statistics (see, for example, Cordeiro and Ferrari, 1991). The location adjustment that is introduced here depends on quantities that are either readily available or easily obtainable, analytically or numerically, for many well-used model classes. In particular, the proposed adjustment involves the expected information, an approximation to the bias of the estimator (see, for example Cox and Snell, 1968, for the

Table 2
Empirical coverage probabilities of individual Wald-type confidence intervals for $\beta_2, \beta_3, \beta_4, \gamma_2, \gamma_3$ in (1), based on the maximum likelihood estimator (ML) and its reduced-bias version (RB) at nominal levels 90%, 95%, and 99%. Reported rates based on 50 000 samples simulated under the ML fit in Table 1.

	ML			RB		
	90%	95%	99%	90%	95%	99%
β_2	86.9	92.4	97.7	88.1	93.4	98.2
β_3	84.8	91.0	97.1	87.2	92.9	98.0
β_4	85.0	91.2	97.2	87.3	92.9	98.0
γ_2	82.4	89.1	96.1	83.8	90.2	96.7
γ_3	79.1	86.0	94.4	82.7	89.2	96.1

first-term in the bias expansion of the ML estimator), and the derivatives of an appropriate transformation of the model parameters, which can be computed either analytically or by numerical or automatic differentiation methods.

We demonstrate how the correction in the location of the Wald statistic strikes a balance in being sufficient to deliver significant improvements to finite-sample inferential performance over other status-quo methods in prominent modelling scenarios, only with a small sacrifice to the computational simplicity of classical Wald inference, mainly due to extra matrix multiplications. A bootstrap procedure that exploits the computational simplicity of LA Wald statistics is also presented to deliver location- and scale-adjusted Wald statistics.

Our recommendation of adopting LA Wald statistics in statistical practice is supported by ample analytical and numerical evidence, along with case studies under well-used modelling settings. These settings include inference about log-odds and binomial proportions, inference from logistic regression models in the presence of nuisance parameters, beta regression, gamma regression and random-effects meta-analysis (see supplementary material) models. The LA Wald statistics are successfully used within the mass univariate probit regression framework in Ge et al. (2014, § 4.1) for the construction of brain significance maps to visualize the strength of association of patient characteristics to the occurrence of multiple sclerosis lesions from MRI data. This is a scenario where standard statistics, like the Wald and likelihood-ratio ones, have sub-optimal properties or completely fail to apply due to the occurrence of infinite estimates in many of the thousands voxel-wise regressions.

The current paper is structured as follows. The location adjustment to the standard Wald statistic and the Wald statistic based on RB estimators are introduced in Sections 2 and 3. Section 4 derives the computational complexity of the LA Wald statistics and presents an algorithm for their computation in general models. Section 5 illustrates the effect that the location adjustment has on the normal approximation to the distribution of the Wald statistic. Procedures for the computation of CIs are detailed in Section 6, and their performance is assessed under a beta regression setting in Section 7. Section 8 gives evidence on the accuracy of the LA Wald statistics when making inference on log-odds and binomial proportions. Section 9 obtains the closed form of the quantities required to compute LA Wald statistics for inference from generalized linear models, and shows simulation results for gamma and logistic regressions. In Section 10, the LA statistics are used for the construction of significance maps for the analysis of multiple sclerosis lesions from MRI data. Finally, Section 11 presents a bootstrap procedure for the scale adjustment of LA Wald statistics, before closing with discussion and further work in Section 12.

2. Location-adjusted Wald statistic

2.1. Bias of the Wald statistic

Consider a sample $y = (y_1, \ldots, y_n)^\top$ of observations assumed to be realizations of independent random variables $Y_1, \ldots, Y_n$, with Y_i having conditional density or probability mass function $f(y_i|x_i; \theta)$, where $\theta \in \Theta \subseteq \Re^p$, $p \leq n$, is the parameter vector and $x_i = (x_{i1}, \ldots, x_{ik})^\top$ is a k-vector of explanatory variables for the ith observation ($i = 1, \ldots, n$). We partition θ as $\theta = (\psi, \lambda^\top)^\top$, where $\psi \in \Psi \subset \Re$ is a scalar parameter of interest and $\lambda \in \Lambda \subset \Re^{p-1}$ is a $(p-1)$-vector of nuisance parameters.

Assuming that the log-likelihood function $l(\theta) = \sum_{i=1}^n \log f(y_i|x_i; \theta)$ satisfies the usual regularity conditions (see, e.g., Pace and Salvan, 1997, § 3.4), a typical way to construct inference about ψ is using a Wald statistic. For example, the Wald test for $H_0 : \psi = \psi_0$ with $\psi_0 \in \Re$ computes p-values using the standard normal distribution and the observed value of the signed Wald statistic

$$t = (\hat{\psi} - \psi_0)/\kappa(\hat{\theta}). \tag{2}$$

In the above expression, $\hat{\theta} = (\hat{\psi}, \hat{\lambda}^\top)^\top = \arg\max_{\theta \in \Theta} l(\theta)$ is the ML estimate of θ and $\kappa(\theta)$ is the square root of the (ψ, ψ)-element of the variance–covariance matrix $\{i(\theta)\}^{-1}$ of the exact or asymptotic distribution of the estimator $\hat{\theta}$, usually taken as the inverse of the expected information $E\{\nabla l(\theta) \nabla l(\theta)^\top\}$, where ∇ denotes the gradient with respect to θ. Without loss of generality, we assume that the element at the first row and first column of $\{i(\theta)\}^{-1}$ is the asymptotic variance of $\hat{\psi}$.

The standard normal distribution is not always a good approximation to the exact distribution of (2) under H_0. This is commonly the case when the model is highly non-linear in the parameters or n is small or moderate relative to p (see, for instance, McCullagh and Nelder, 1989, § 6.2.4).

We show how this approximation can be easily improved by bringing the first null moment of the asymptotic distribution for the adjusted Wald statistic closer to zero. The word "null" here is used to highlight the fact that the expectation is taken with respect to the model, assuming that the null hypothesis holds.

Consider the transformation

$$T(\theta; \psi_0) = (\psi - \psi_0)/\kappa(\theta) \tag{3}$$

of the parameter θ. We call $T(\theta; \psi_0)$ the Wald transform. The Wald transform is of order $O\left(n^{1/2}\right)$ because $\kappa(\theta) = O\left(n^{-1/2}\right)$. Then, t in (2) is the ML estimator of (3). As is the case for $\hat{\theta}$, t is also subject to finite-sample bias, which can be reduced by subtracting the first term in the asymptotic expansion of its bias, as is shown, for example, in Efron (1975, Remark 11, p. 1214).

Assume that (3) is at least three times differentiable with respect to θ and that $\hat{\theta}$ is consistent. Then, using the Einstein summation convention, $T(\hat{\theta}; \psi_0) - T(\theta; \psi_0)$ can be expanded as

$$\begin{aligned}(\hat{\theta}^u - \theta^u)T_u(\theta; \psi_0) + \frac{1}{2}(\hat{\theta}^u - \theta^u)(\hat{\theta}^v - \theta^v)T_{uv}(\theta; \psi_0) \\ + \frac{1}{6}(\hat{\theta}^u - \theta^u)(\hat{\theta}^v - \theta^v)(\hat{\theta}^w - \theta^w)T_{uvw}(\theta; \psi_0) + O_p\left(n^{-3/2}\right),\end{aligned} \tag{4}$$

where $T_u(\theta; \psi_0)$, $T_{uv}(\theta; \psi_0)$ and $T_{uvw}(\theta; \psi_0)$ are the gradient, hessian and third derivative, respectively, of (3) ($u, v, w = 1, \ldots, p$), all with order $O\left(n^{1/2}\right)$. Taking expectations in (4) gives that $E\{T(\hat{\theta}; \psi_0) - T(\theta; \psi_0)\} = B(\theta; \psi_0) + O\left(n^{-3/2}\right)$ with

$$B(\theta; \psi_0) = b^u(\theta)T_u(\theta; \psi_0) + \frac{1}{2}i^{u,v}(\theta)T_{uv}(\theta; \psi_0), \tag{5}$$

where the first-order bias $b^u(\theta)$ is such that $E_\theta(\hat{\theta}^u - \theta^u) = b^u(\theta) + o\left(n^{-1}\right)$ and $i^{u,v}(\theta)$ can be understood as the (u, v)th element of $\{i(\theta)\}^{-1}$ ($u, v = 1, \ldots, p$). The above expression can also be derived using Kosmidis and Firth (2010, § 4.3, Remark 3), which gives the first term in the bias expansion of the ML estimator for a transformation of a scalar parameter in terms of that of the ML estimator for the parameter. Expression (5) can be written in matrix notation as

$$B(\theta, \psi_0) = \{b(\theta)\}^\top \nabla T(\theta; \psi_0) + \frac{1}{2}\mathrm{tr}\left[\{i(\theta)\}^{-1}\nabla\nabla^\top T(\theta; \psi_0)\right], \tag{6}$$

where $\mathrm{tr}(A)$ denotes the trace of matrix A, and ∇ and $\nabla\nabla^\top$ denote the gradient and the matrix of second derivatives with respect to θ, respectively.

2.2. Location-adjusted Wald statistic

The LA Wald statistic for ψ is then

$$t^* = t - \widehat{B}, \tag{7}$$

where $\widehat{B}$ is a suitable estimator of $B(\theta; \psi_0)$ in (6). Natural candidates for $\widehat{B}$ are $B(\hat{\theta}_0; \psi_0)$ and $B(\hat{\theta}; \psi_0)$, where $\hat{\theta}_0 = (\psi_0, \hat{\lambda}_0^\top)^\top$ and $\hat{\lambda}_0 = \arg\max_{\lambda \in \Lambda} l(\psi_0, \lambda)$ is the constrained ML estimate of λ. In either case, a calculation along the lines of Pace and Salvan (1997, § 9.42) shows that the null expectation of t^* is $O(n^{-3/2})$, that is asymptotically closer to zero than the null expectation of t, which is $O(n^{-1/2})$.

We focus on LA statistics with $\widehat{B} = B(\hat{\theta}; \psi_0)$ because their computation does not require any additional constrained or otherwise optimization, than the already available fit for the full model.

2.3. Derivatives of the Wald transform

Expression (6) is convenient because it depends only on the derivatives of $T(\theta; \psi_0)$, the elements of $\{i(\theta)\}^{-1}$, and the first term in the expansion of the bias of the ML estimator, which has been derived in Cox and Snell (1968, expression (20)) for general parametric models, and is given in matrix form in Kosmidis and Firth (2010, § 2).

Both $i(\theta)$ and the first-order bias are readily available for a wide range of well-used model classes, especially those for which asymptotic bias reduction methods have been implemented (see, among others, Cook et al., 1986; Cordeiro and McCullagh, 1991; Cordeiro and Vasconcellos, 1997; Botter and Cordeiro, 1998; Simas et al., 2010; Cordeiro and Toyama Udo, 2008; Grün et al., 2012). In addition, the derivatives of the Wald transform can be written in terms of derivatives of $\kappa(\theta)$. Specifically, the gradient $\nabla T(\theta; \psi_0)$ and the hessian $\nabla\nabla^\top T(\theta; \psi_0)$ are

$$\left\{e_p - T(\theta; \psi_0)\nabla\kappa(\theta)\right\}/\kappa(\theta) \tag{8}$$

and

$$-\left[\nabla\kappa(\theta)\{\nabla T(\theta;\psi_0)\}^\top + \nabla T(\theta;\psi_0)\{\nabla\kappa(\theta)\}^\top + T(\theta;\psi_0)\nabla\nabla^\top\kappa(\theta)\right]/\kappa(\theta)\,, \tag{9}$$

respectively, where e_p is a p-vector with first element one and zeros everywhere else.

Expressions (8) and (9) can, in turn, be written in terms of the information matrix $i(\theta)$ and its derivatives. A straightforward application of matrix differentiation rules (see, e.g., Magnus and Neudecker, 1999) gives that the generic uth element of the p-dimensional gradient vector $\nabla\kappa(\theta)$ has the form

$$\frac{\partial\kappa(\theta)}{\partial\theta_u} = -\frac{1}{2\kappa(\theta)}\left[\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_u}\{i(\theta)\}^{-1}\right]_{\psi\psi} \qquad (u=1,\ldots,p)\,, \tag{10}$$

where $[\,\cdot\,]_{\psi\psi}$ denotes the (ψ,ψ) element of the square matrix in brackets. Further differentiation gives that the (u,v)th element in the $p\times p$ hessian $\nabla\nabla^\top\kappa(\theta)$ is

$$\begin{aligned}\frac{\partial^2\kappa(\theta)}{\partial\theta_u\partial\theta_v} = &-\frac{1}{4\{\kappa(\theta)\}^3}\left[\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_u}\{i(\theta)\}^{-1}\right]_{\psi\psi}\left[\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_v}\{i(\theta)\}^{-1}\right]_{\psi\psi}\\ &+\frac{1}{2\kappa(\theta)}\left[\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_v}\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_u}\{i(\theta)\}^{-1}\right]_{\psi\psi} - \frac{1}{2\kappa(\theta)}\left[\{i(\theta)\}^{-1}\frac{\partial^2 i(\theta)}{\partial\theta_u\partial\theta_v}\{i(\theta)\}^{-1}\right]_{\psi\psi}\\ &+\frac{1}{2\kappa(\theta)}\left[\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_u}\{i(\theta)\}^{-1}\frac{\partial i(\theta)}{\partial\theta_v}\{i(\theta)\}^{-1}\right]_{\psi\psi} \qquad (u,v=1,\ldots,p)\,.\end{aligned} \tag{11}$$

The information matrix $i(\theta)$ and its derivatives are model-dependent, but can be found directly for popular model classes, as done in Section 9.2 for generalized linear models.

3. Wald statistics and reduced-bias estimators

As is illustrated in Example 1.1 and the works cited therein, the use of RB estimators can deliver marked improvements in the inferential performance of Wald-type procedures. These performance gains can be partly explained by the fact that employing asymptotically efficient estimators with bias of order $o(n^{-1})$ in the Wald statistic results in the bias of the latter being as in (6) but without the term $\{b(\theta)\}^\top\nabla T(\theta;\psi_0)$. To see this, notice that $T(\tilde\theta;\psi_0)-T(\theta;\psi_0)$ admits the same expansion as in (4), with $\hat\theta$ replaced by the RB estimator $\tilde\theta$. Yet the expectation of the term $(\tilde\theta^u-\theta^u)T_u(\theta;\psi_0)$ is $o(n^{-1/2})$, giving $E\{T(\tilde\theta;\psi_0)-T(\theta;\psi_0)\}=\tilde B(\theta;\psi_0)+o(n^{-1/2})$, where,

$$\tilde B(\theta;\psi_0) = \frac{1}{2}\mathrm{tr}\left[\{i(\theta)\}^{-1}\nabla\nabla^\top T(\theta;\psi_0)\right]\,. \tag{12}$$

The above quantity is of the same order, $O(n^{-1/2})$, as (6). Given that the signs and the magnitude of terms in the right hand side of (6) are rarely known in general models, the elimination of $\{b(\theta)\}^\top\nabla T(\theta;\psi_0)$ through the use of RB estimators provides no guarantee of improvement, despite the positive results in the setting of Example 1.1. In fact, employment of RB estimators in the Wald statistics may even have a detrimental effect on inference.

Expression (12) proves useful, though, for adjusting the location of Wald statistics based on RB estimators in settings where the latter have been found to be particularly desirable, especially with reference to Wald-type inference. Such a scenario, which is also popular in recent applied work, involves models for categorical data where RB estimators are always finite, even in cases where the ML estimates have infinite components (see, for example, Heinze and Schemper, 2002; Kosmidis and Firth, 2011; Kosmidis, 2014, for logistic regression and regression models with nominal and ordinal responses, respectively). Despite that the Wald statistic and its LA version cannot be directly computed when $\hat\psi$ is infinite, the Wald statistic based on the RB estimator $\tilde t = T(\tilde\theta;\psi_0)$ and its LA version

$$\tilde t^* = \tilde t - \tilde B(\tilde\theta;\psi_0)\,,$$

are both well-defined. By the same arguments to those in Section 2 for the development of t^*, the null expectation of $\tilde t^*$ is $O(n^{-3/2})$ and, hence, asymptotically closer to zero than that of $\tilde t$.

4. Implementation and computational complexity

Assuming that the estimates for θ are available and that the information $i(\theta)$ has been computed at those estimates, the inversion of $i(\theta)$ involves, typically, $O(p^3)$ operations. Hence, the complexity for the computation of the Wald statistic (2) is $O(p^3)$.

After having computed the derivatives of $i(\theta)$ and $b(\theta)$ at the estimates, and without exploiting any model structures or sparsity in $i(\theta)$, the computation of $B(\hat\theta;\psi_0)$ requires $O(p^4)$ operations. This number of operations is due to the matrix multiplications for computing the gradient (8), the hessian (9) and, finally, (6). As a result, the LA Wald statistic (7) based on $B(\hat\theta;\psi_0)$ has, generally, computational complexity $O(p^4)$. Note here that the extra operations are just straightforward

matrix multiplications and, hence, the computing time can be significantly reduced by appropriate vectorization, pre-computing some of the quantities, and parallelizing others across the parameters, like, for instance, the products $\{i(\theta)\}^{-1}\partial i(\theta)/\partial\theta_u$ across $u \in \{1, \ldots, p\}$.

Algorithm 1 Location-adjusted Wald statistics

1: **procedure** LAWALD(θ^*, θ_0, i, b, c, e)
2: $SE(\theta, i) \leftarrow inverse(i(\theta))[k, k]$ ▷ $\kappa(\theta)$
3: $p \leftarrow length(\theta^*)$ ▷ length of the vector of estimates θ^*
4: **if** $e = 1$ **then** ▷ if θ^* is ML estimate
5: $B \leftarrow b(\theta^*)$
6: **end if** ▷ $b(\theta^*)$
7: $I \leftarrow i(\theta^*)$
8: $II \leftarrow inverse(I)$
9: $S \leftarrow vector(p)$ ▷ S as a p-vector
10: $T \leftarrow vector(p)$ ▷ T as a p-vector
11: **for** $j \in \{1, 2, \ldots, p\}$ **do**
12: $S[j] \leftarrow Sqrt(II[j, j])$ ▷ estimated standard error
13: $T[j] \leftarrow (\theta^*[j] - \theta_0[j])/S[j]$ ▷ Wald statistic
14: **if** $c = 1$ **then** ▷ if location-adjustment is requested
15: $U \leftarrow numericgradient(SE, \theta = \theta^*, k = j)$ ▷ $\nabla\kappa(\theta)$ at $\theta := \theta^*$
16: $V \leftarrow numerichessian(SE, \theta = \theta^*, k = j)$ ▷ $\nabla\nabla^\top\kappa(\theta)$ at $\theta := \theta^*$
17: $A \leftarrow -T[j] * U$
18: $A[j] \leftarrow 1 + A[j]$
19: **if** $e = 1$ **then** ▷ if θ^* is ML estimate
20: $W \leftarrow dotproduct(A, B)$ ▷ dot product of A and B
21: **end if**
22: **if** $e = 2$ **then** ▷ if θ^* is RB estimate
23: $W \leftarrow 0$
24: **end if**
25: $X \leftarrow wdotproduct(A, U, II)$ ▷ dot product of A and U with weight-matrix II
26: $Y \leftarrow sum(V * II)$ ▷ sum of elements from element-wise product of V and II
27: $Y \leftarrow T[j] * Y$
28: $T[j] \leftarrow T[j] - (W + X + Y/2)/S[j]$ ▷ location-adjusted Wald statistic
29: **end if**
30: **end for**
31: **return** T
32: **end procedure**

If p is not prohibitively large, and at the expense of additional computing cost due to matrix inversions, the derivatives of $\kappa(\theta)$ in (10) and (11) can be also calculated using numerical differentiation techniques at $\hat{\theta}$, provided there is an appropriate computer implementation of the standard errors as a function of the parameters. In this way, the location adjustment can be obtained for general models, requiring only the ML or RB estimates, along with ready implementations of the expected information matrix and an approximation of the bias function. Algorithm 1 details such a procedure in pseudo-code and the `waldi` R package (https://github.com/ikosmidis/waldi) implements the computations for generalized linear models and beta regression models. Notice that Steps 12–29 of Algorithm 1 for calculating t^* or $\tilde{t}^*$ can be performed in parallel across parameters, with significant savings in execution time when multiple computing units are available. The `waldi` R package provides the parallel computation of the LA Wald statistic.

5. Effect of location adjustment on distributional approximation

As shown in Section 2, the location adjustment of the Wald statistic delivers statistics with null expectations closer to zero than the standard Wald statistic. This correction often extends to higher-order moments than the mean, but it is not, overall, sufficient to deliver a drop in the order of the error of the normal approximation to the whole null distribution of the statistic, as Bartlett-type corrections for the LR statistics do.

To see this, suppose that Y_i has an exponential distribution with mean $\mu_i = e^{-\theta} > 0$, $\theta \in \Re$. The log-likelihood about θ is $l(\theta) = n\theta - n\bar{y}e^{\theta}$, where $\bar{y} = \sum_{i=1}^{n} y_i/n$ is the sample mean. The ML estimate, the expected information and the first-order bias are $\hat{\theta} = -\log\bar{y}$, $i(\theta) = n$, and $b(\theta) = (2n)^{-1}$, respectively. The derivatives of the Wald transform are $dT(\theta; \theta_0)/d\theta = n^{1/2}$ and $d^2T(\theta; \theta_0)/d\theta^2 = 0$, and so $B(\theta; \theta_0) = n^{-1/2}/2$. Hence, the Wald statistic (2) for $H_0 : \theta = \theta_0$ is $t = -n^{1/2}(\log\bar{y} + \theta_0)$ and the LA Wald statistic in (7) is $t^* = -n^{1/2}(\log\bar{y} + \theta_0) - n^{-1/2}/2$.

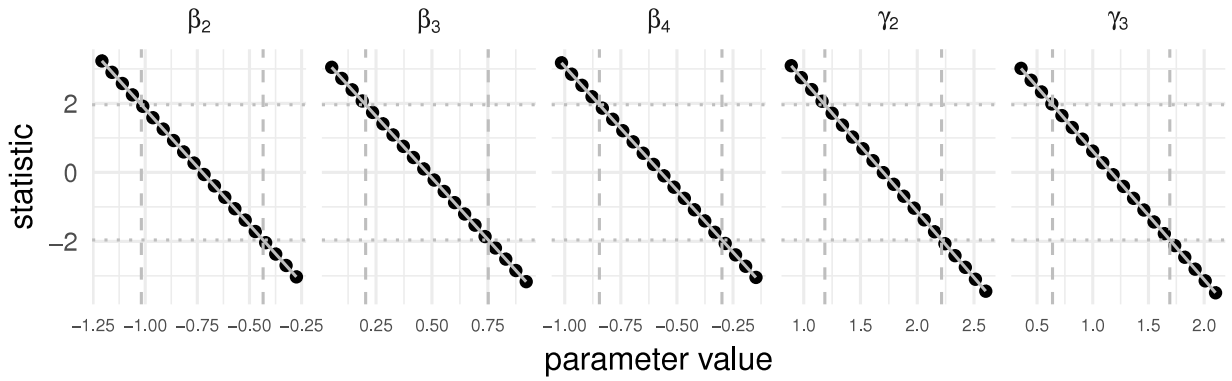

Fig. 1. Computation of the 95% confidence intervals based on t^* in Table 3, using linear interpolation of the values of $T(\hat{\theta};\psi) - B(\hat{\theta};\psi)$ on a equispaced grid of 20 values for ψ. The solid grey line is the linear interpolator, the dotted lines are at $\pm z_{0.975}$, and the vertical dashed lines are at the endpoints of the intervals.

The Edgeworth expansions (Hall, 1992, § 2.3) of the null distribution functions $F(z)$ and $F^*(z)$ of t and t^*, respectively, give

$$F(z) = \Phi(z) - n^{-1/2}\frac{z^2+2}{6}\phi(z) - n^{-1}\frac{2z^5 - 11z^3 + 57z}{144}\phi(z) + O\left(n^{-3/2}\right),$$
$$F^*(z) = \Phi(z) - n^{-1/2}\frac{z^2-1}{6}\phi(z) - n^{-1}\frac{2z^5 - 23z^3 + 75z}{144}\phi(z) + O\left(n^{-3/2}\right),$$

where $\Phi(z)$ and $\phi(z)$ are the distribution and density functions of the standard normal random variable. The corresponding Cornish–Fisher expansions (Hall, 1992, § 2.5) of the α-level quantiles q_α of $F(z)$ and q^*_α of $F^*(z)$ in terms of the α-level standard normal quantiles z_α are, then,

$$q_\alpha = z_\alpha + n^{-1/2}\frac{z_\alpha^2+2}{6} - n^{-1}\frac{11z_\alpha^3 - 65z_\alpha}{144} + O\left(n^{-3/2}\right), \tag{13}$$
$$q^*_\alpha = z_\alpha + n^{-1/2}\frac{z_\alpha^2-1}{6} - n^{-1}\frac{11z_\alpha^3 - 65z_\alpha}{144} + O\left(n^{-3/2}\right), \tag{14}$$

provided that $\epsilon < \alpha < 1-\epsilon$ for any $0 < \epsilon < 1/2$. Both q^*_α and q_α have non-zero $O(n^{-1/2})$ terms, and hence t^* does not deliver an improvement in the sense of a drop in the asymptotic order of the distributional approximation. Nevertheless, a careful term-by-term comparison reveals that $(q_\alpha - z_\alpha) \simeq (q^*_\alpha - z_\alpha) + n^{-1/2}/2$. As a result, the quantiles of the distribution of t^* are closer to the standard normal quantiles than those of t.

6. Confidence intervals based on location-adjusted statistics

The LA Wald statistics t^* and $\tilde{t}^*$ can be used to obtain $100(1-\alpha)\%$ CIs by the numerical inversion of the approximate probability statements for each, that is by finding all ψ such that

$$z_{\alpha/2} \leq T(\hat{\theta};\psi) - B(\hat{\theta};\psi) \leq z_{1-\alpha/2} \quad \text{and} \quad z_{\alpha/2} \leq T(\tilde{\theta};\psi) - \tilde{B}(\tilde{\theta};\psi) \leq z_{1-\alpha/2}\,, \tag{15}$$

respectively. The numerical inversion can be performed by evaluation of the LA statistics on a grid of values for ψ and linear interpolation. Fig. 1 illustrates this process for the t^* intervals in Table 3.

7. Location-adjusted statistics in beta regression

7.1. Implementation

The LA statistics $\tilde{t}$ and $\tilde{t}^*$ involve the derivatives of the Wald transform, the expected information matrix and the first-order bias term of the ML estimator for beta regression models, whose general expressions, despite being obtainable, are complicated to either write down or implement. Grün et al. (2012, § 2) provide closed-form expressions for the latter two, and the `enrichwith` R package (Kosmidis, 2017) can generate the corresponding R functions of the parameters for their evaluation, which can then be used as input in Algorithm 1. In particular, the derivatives of the Wald transform can be computed using expressions (8) and (9), where $\nabla\kappa(\theta)$ and $\nabla\nabla^\top\kappa(\theta)$ are approximated, to high-accuracy, for each parameter and at the estimates using Richardson's extrapolation (as implemented, for example, in the `numDeriv` R package; Gilbert and Varadhan, 2016).

Table 3
Empirical coverage probabilities of individual confidence intervals at nominal levels 90%, 95% and 99% for the regression parameters in (1), based on the inversion of the location-adjusted Wald statistics $\tilde{t}$ and $\tilde{t}^*$, using standard normal quantiles and the quantiles of the bootstrap distribution of the statistics from a parametric bootstrap of size 500. Reported rates are based on 50 000 samples simulated under the ML fit in Table 1.

	95% Confidence intervals				Empirical coverage probability					
	t^*		$\tilde{t}^*$		t^*			$\tilde{t}^*$		
	Lower	Upper	Lower	Upper	90%	95%	99%	90%	95%	99%
Intervals using $N(0, 1)$ quantiles										
β_2	−1.019	−0.435	−1.031	−0.446	88.5	93.7	98.4	88.3	93.5	98.3
β_3	0.204	0.752	0.165	0.719	87.1	92.8	98.0	87.3	93.0	98.0
β_4	−0.845	−0.299	−0.809	−0.257	87.2	92.8	98.0	87.5	93.0	98.0
γ_2	1.186	2.214	1.134	2.169	83.5	90.0	96.6	83.9	90.3	96.8
γ_3	0.639	1.691	0.513	1.574	81.8	88.6	95.7	82.7	89.2	96.2
Studentized bootstrap intervals										
β_2	−1.059	−0.442	−1.091	−0.440	89.5	94.5	98.7	89.4	94.6	98.6
β_3	0.171	0.792	0.159	0.758	89.2	94.3	98.5	89.4	94.5	98.5
β_4	−0.871	−0.268	−0.853	−0.264	89.3	94.3	98.5	89.5	94.4	98.6
γ_2	1.112	2.303	1.040	2.241	89.9	94.7	98.7	90.1	94.9	98.8
γ_3	0.565	1.835	0.394	1.769	90.1	94.9	98.7	90.5	95.1	98.8

7.2. Reading skills

For the regression parameters in model (1) of Example 1.1, Table 3 reports individual 95% CIs derived from the inversion of t^* and $\tilde{t}^*$, and their empirical coverage probabilities. The empirical coverage probabilities are markedly closer to the nominal level than those of standard Wald CIs based on the ML estimates in Table 1. The intervals based on $\tilde{t}^*$ are also quite similar to the ones based on $\tilde{t}$ in Table 1, with only a slight improvement in terms of empirical coverage. This indicates that the extra term (12) in the bias of the Wald statistic based on RB estimators is of no consequence here; most of the correction to the location of the Wald statistic is achieved by using the RB estimator in place of the ML one.

Table 3 also shows results about the studentized bootstrap CIs (see Davison and Hinkley, 1997, § 2.4) for the model parameters, based on a parametric bootstrap of size 500 to estimate the quantiles of the distribution of t^* and $\tilde{t}^*$. The computation of the studentized bootstrap CIs is done, again, by numerically inverting the approximate probability statements for each parameter as in (15), after replacing the standard normal quantiles with the corresponding ones estimated via bootstrap. It should be noted here that in cases like beta regression an estimator of the cumulative distribution function of the test statistic being studentized is generally not available in closed form. As a result, the calculation of studentized bootstrap intervals is a computationally intensive process. For example, on a MacBook Pro laptop with 3.5 GHz Intel Core i7 processor and 16 GB of RAM, computing all CIs based on $\tilde{t}^*$ using the default arguments in the generic implementation of the `waldi_confint` function in the `waldi` R package takes about 1.4 s. On the other hand, the computation of the corresponding studentized bootstrap intervals takes about 140 s, because of the need to refit the model and calculate the LA Wald statistics 500 times.

Tipically, using the quantiles of the bootstrap distribution of t^* and $\tilde{t}^*$ instead of $N(0, 1)$ ones brings the empirical coverage probabilities of intervals based on the LA Wald statistics even closer to the respective nominal levels, most notably for the precision effects γ_2 and γ_3.

8. Inference about log-odds and binomial proportions

Hypothesis tests and confidence intervals about log-odds and binomial proportions are amongst the most common statistical tasks in applied data analysis. In the present section, we investigate the performance of the LA Wald statistics in these contexts and contrast it to that of classical proposals in the literature.

Suppose that Y_i has a Bernoulli distribution with mean $\mu_i = e^\theta/\left(1+e^\theta\right) \in (0, 1)$, $\theta \in \Re$. The log-likelihood function about θ, the ML estimate, its first-order bias and the expected information are $l(\theta) = n\bar{y}\theta - n\log\left(1+e^\theta\right)$, $\hat{\theta} = \log\left\{\bar{y}/(1-\bar{y})\right\}$, $b(\theta) = -\left(1+e^\theta\right)\left(1-e^\theta\right)\left(2ne^\theta\right)^{-1}$ and $i(\theta) = ne^\theta/\left(1+e^\theta\right)^2$, respectively. The well-known RB estimator of the log-odds θ is $\tilde{\theta} = \log\left\{(\bar{y}+a)/(1-\bar{y}+a)\right\}$, where $a = n^{-1}/2$ (Haldane, 1955; Anscombe, 1956). The first summand in the right-hand side of (6) equals

$$b(\theta)\frac{dT(\theta;\theta_0)}{d\theta} = \frac{e^{2\theta}(\theta_0-\theta+2)+2e^\theta(\theta-\theta_0)+\theta_0-\theta-2}{4n^{1/2}e^{\theta/2}(1+e^\theta)},$$

and the second is

$$\tilde{B}(\theta;\theta_0) = \frac{1}{2i(\theta)}\frac{d^2T(\theta;\theta_0)}{d\theta^2} = \frac{e^{2\theta}(\theta-\theta_0-4)+6e^\theta(\theta_0-\theta)+\theta-\theta_0+4}{8n^{1/2}e^{\theta/2}(1+e^\theta)}.$$

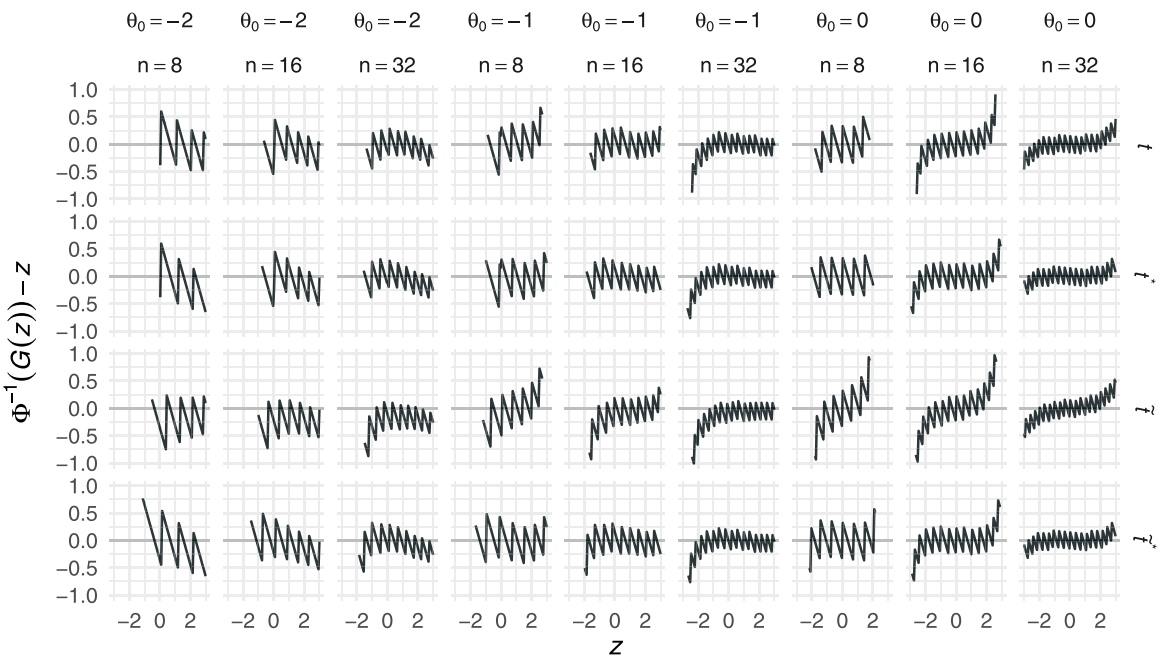

Fig. 2. The function $\Phi^{-1}(G(z)) - z$ when $G(z)$ is the null distribution function for t, t^*, $\tilde{t}$ and $\tilde{t}^*$ under the Bernoulli model of Section 8, for $\theta_0 \in \{-2, -1, 0\}$ and $n \in \{8, 16, 32\}$. The reference zero line (grey) corresponds to $G(z) = \Phi(z)$.

Hence, we obtain $B(\theta; \theta_0) = (8n^{1/2})^{-1}(\theta_0 - \theta)(e^{-\theta/2} + e^{\theta/2})$. The Wald statistics t and $\tilde{t}$ and their LA versions are

$$
\begin{aligned}
t &= \left\{n(\bar{y} - \bar{y}^2)\right\}^{1/2}\left(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\right),\\
\tilde{t} &= \frac{\{n(\bar{y}+a)(1-\bar{y}+a)\}^{1/2}}{1+2a}\left(\log\frac{\bar{y}+a}{1-\bar{y}+a} - \theta_0\right),\\
t^* &= \left\{(n\bar{y} - n\bar{y}^2)^{1/2} + (n\bar{y} - n\bar{y}^2)^{-1/2}/8\right\}\left(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\right),\\
\tilde{t}^* &= \tilde{t} - \frac{(1-\bar{y}+a)^{3/2}}{8n^{1/2}(\bar{y}+a)^{1/2}(1+2a)}\left\{\left(\frac{\bar{y}+a}{1-\bar{y}+a}\right)^2\left(\log\frac{\bar{y}+a}{1-\bar{y}+a} - \theta_0 - 4\right)\right.\\
&\quad \left. + \frac{6(\bar{y}+a)}{1-\bar{y}+a}\left(\theta_0 - \log\frac{\bar{y}+a}{1-\bar{y}+a}\right) + \log\frac{\bar{y}+a}{1-\bar{y}+a} - \theta_0 + 4\right\},
\end{aligned}
$$

respectively. All statistics depend on values of $n\overline{Y}$, which has a null binomial distribution with index n and probability $e^\theta/(1+e^\theta)$. If $\bar{y} = 0$, then $\hat{\theta} = -\infty$ and if $\bar{y} = 1$, then $\hat{\theta} = +\infty$. In these cases, $t = 0$ and $t^* = \pm\infty$, respectively. So, regardless of the value θ_0, when all observations are equal to zero or one the Wald test always accepts the hypothesis $\theta = \theta_0$, and its LA version always rejects it. For this reason and to enable comparison, the convention that $t^* = t = 0$ whenever $\bar{y} = 0$ or $\bar{y} = 1$ is adopted for every $\theta_0 \in \Re$. No such convention is necessary for $\tilde{t}$ and $\tilde{t}^*$ because the finiteness of $\tilde{\theta}$ guarantees their finiteness.

Fig. 2 compares $\Phi(z)$ to the null distribution of t and t^* for $\theta_0 \in \{-2, -1, 0\}$ and $n \in \{8, 16, 32\}$, plotting only over the range of z where each distribution takes values in $(0, 1)$. The adjustment of the Wald statistic appears to be effective in terms of bringing the corresponding distributions closer to standard normal, especially for the smaller sample sizes. The statistic $\tilde{t}^*$ performs best having a distribution that is similar to that of t^*, but taking values in $(0, 1)$ for a wider range of z. Note here that, in the majority of cases, the standard normal approximation to the distribution of the Wald statistic based on the RB estimator is worse than that of t.

The statistic $\tilde{t}^*$ can be readily inverted numerically as in (15) to produce CIs for the log-odds θ with excellent coverage properties. Applying the transformation $e^\theta/(1+e^\theta)$ to the endpoints of those intervals results in CIs for the binomial probability. As is demonstrated in the supplementary material, the latter have competitive coverage properties and shorter expected lengths, for most values in $(0, 1)$, than the widely accepted "add 2 successes and 2 failures" intervals proposed in Agresti and Coull (1998) and Agresti and Caffo (2000), with the added benefit of no overshooting outside $(0, 1)$. Based on these findings, we propose the transformation of the endpoints of the CIs based on $\tilde{t}^*$ for routine use when inference about a binomial proportion or log-odds is of interest.

Finally, the tests conducted via the LA Wald statistic are found to be more robust to the undesirable behaviour of Wald tests in binomial settings observed in Hauck and Donner (1977). To illustrate this, consider $n = 32$ and $\theta_0 = 0$. The values of the Wald statistic for $\bar{y} = 0.875$, $\bar{y} = 0.906$, $\bar{y} = 0.938$, $\bar{y} = 0.969$ and $\bar{y} = 1$ are 3.640, 3.741 3.708 3.380 and 0 (in 3 significant decimal places), respectively. As a result, the evidence against $H_0 : \theta = 0$ from the Wald test decreases when $\bar{y} > 0.906$, despite the fact that the sample mean is getting further away from the null probability of 0.5. In contrast, the corresponding values for the LA Wald statistic are 3.770, 3.913, 3.955, 3.816 and 0, respectively, thus the evidence against H_0 decreases only after $y > 0.938$. The same behaviour is observed for $\tilde{t}^*$.

9. Generalized linear models

9.1. Wald statistics

One of the key summaries in the output of standard statistical software when fitting generalized linear models are Wald statistics for the regression coefficients.

In generalized linear models (McCullagh and Nelder, 1989), the conditional distribution of Y_i given x_i is assumed to be from the exponential dispersion family (Jørgensen, 1987), with density or probability mass function

$$f(y_i|x_i;\theta_i) = \exp\left\{\frac{y\theta_i - h(\theta_i) - c_1(y)}{\phi/m_i} - \frac{1}{2}a\left(-\frac{m_i}{\phi}\right) + c_2(y)\right\}.$$

In the latter expression, $h(\cdot)$, $c_1(\cdot)$, $a(\cdot)$ and $c_2(\cdot)$ are sufficiently smooth functions, and $m_1, \ldots, m_n$ are known, non-negative observation weights. Special distributions with density or probability mass function of the above form are the normal, gamma, Poisson and binomial.

The conditional expectation $\mu_i = h'(\theta_i)$ of Y_i is linked to x_i as $g(\mu_i) = \eta_i = \beta^\top x_i$, where $g(\cdot)$ is an at least three times differentiable link function, and $h'(u) = dh(u)/du$. The variance of Y_i is $\phi V(\mu_i)/m_i$, where $V(\mu_i) = h''(\theta_i)$ is the variance function, with $h''(u) = d^2h(u)/du^2$, and ϕ is a dispersion parameter that allows shrinking or inflating the contribution of the mean to the sample variance.

The gradient of the log-likelihood with respect to β is inversely proportional to ϕ and, hence, the ML estimate for β can be obtained without knowing ϕ, through iteratively reweighted least squares (IRLS; Green, 1984). The expected information matrix on β and ϕ is

$$i(\beta,\phi) = \begin{bmatrix} \frac{1}{\phi}X^\top W(\beta)X & 0_k \\ 0_k^\top & \frac{1}{2\phi^4}\sum_{i=1}^{n} m_i^2 a''(-m_i/\phi) \end{bmatrix}, \tag{16}$$

where 0_k is a k-dimensional vector of zeros, $a''(u) = d^2a(u)/du^2$, X is the $n \times k$ model matrix with rows $x_1, \ldots, x_n$ and $W = \text{diag}\{w_1, \ldots, w_n\}$ with $w_i = m_i d_i^2/V(\mu_i)$, $d_i = d\mu_i/d\eta_i$.

The Wald statistic in (3) that is typically reported for $H_0 : \beta_j = \beta_{j0}$ $(j = 1, \ldots, k)$ in the output of software for fitting generalized linear models has the form

$$t_j = (\hat{\beta}_j - \beta_{j0})/\kappa_j(\hat{\beta}, \phi^*) \tag{17}$$

(see, for example, the `summary.glm` method in R), and is the estimate of the Wald transform $T_j(\beta,\phi;\beta_{j0}) = (\beta_j - \beta_{j0})/\kappa_j(\beta,\phi)$, where $\kappa_j(\beta,\phi)$ denotes the (j,j)th element of the matrix $[\phi\{X^\top W(\beta)X\}^{-1}]^{1/2}$, and ϕ^* is an estimator of ϕ.

As discussed by McCullagh and Nelder (1989), the ML estimator $\hat{\phi}$ is severely biased and not robust under misspecification of the conditional distribution of Y_i given x_i. For this reason, when ϕ is unknown, McCullagh and Nelder (1989) recommend to replace ϕ^* in (17) with the moment estimator $\tilde{\phi} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/\{(n-p)V(\hat{\mu}_i)\}$ that is based on the Pearson residuals.

9.2. Implementation

The bias terms $B_j(\beta,\phi;\beta_{j0})$ and $\tilde{B}_j(\beta,\phi;\beta_{j0})$ can be readily computed for all generalized linear models using the expression for the first term in the bias expansion of the ML estimators in Cordeiro and McCullagh (1991), the derivatives of $T_j(\beta,\phi;\beta_{j0})$ in (8) and (9), the expected information matrix (16), and its derivatives. The derivatives of $i(\beta,\phi)$ with respect to β can be written in closed form as

$$\frac{\partial i(\beta,\phi)}{\partial \beta_u} = \begin{bmatrix} \frac{1}{\phi}X^\top W_u'(\beta)X & 0_k \\ 0_k^\top & 0 \end{bmatrix} \quad \text{and} \quad \frac{\partial^2 i(\beta,\phi)}{\partial \beta_u \partial \beta_v} = \begin{bmatrix} \frac{1}{\phi}X^\top W_{uv}''(\beta)X & 0_k \\ 0_k^\top & 0 \end{bmatrix}, \tag{18}$$

where $W_u' = W(2R - L)T_u$, with $R = \text{diag}\{r_1, \ldots, r_n\}$, $r_i = d\log d_i/d\eta_i$, $L = \text{diag}\{l_1, \ldots, l_n\}$, $l_i = d\log V(\mu_i)/d\eta_i$, and $T_u = \text{diag}\{x_{1u}, \ldots, x_{nu}\}$. Furthermore, $W_{uv}'' = W(2R-L)^2T_uT_v + W(2R'-L')T_uT_v$, with $R' = \text{diag}\{r_1', \ldots, r_n'\}$, $r_i' = d^2\log d_i/d\eta_i^2$

Table 4
The ML estimates of the regression parameters β_1, β_2, β_3 and β_4 of the gamma regression model for the blood clotting dataset, and the ML ($\star$) and moment-based ($\dagger$) estimate of ϕ. The various versions of the Wald statistic are for the individual hypotheses $H_0 : \beta_j = 0$ ($j = 1, \ldots, 4$).

	Estimate	Wald		t^*
		$\hat{\phi}$	$\tilde{\phi}$	
β_1	5.503	34.126	29.282	28.953
β_2	−0.602	−12.842	−11.020	−10.896
β_3	−0.584	−2.563	−2.199	−2.173
β_4	0.034	0.520	0.446	0.441
ϕ	*0.017			
	†0.024			

and $L' = \text{diag}\{l'_1, \ldots, l'_n\}$, $l'_i = d^2 \log V(\mu_i)/d\eta_i^2$ ($u, v = 1, \ldots, k$). For generalized linear models with unknown dispersion parameter, the derivatives of $i(\beta, \phi)$ with respect to ϕ are

$$\frac{\partial i(\beta,\phi)}{\partial \phi} = \begin{bmatrix} -\dfrac{1}{\phi^2} X^\top W(\beta) X & 0_k \\ 0_k^\top & \dfrac{1}{2\phi^6}\displaystyle\sum_{i=1}^n m_i^2 a'''(-m_i/\phi) - \dfrac{2}{\phi^5}\sum_{i=1}^n m_i^2 a''(-m_i/\phi) \end{bmatrix} \tag{19}$$

and

$$\frac{\partial^2 i(\beta,\phi)}{\partial \phi^2} = \begin{bmatrix} \dfrac{2}{\phi^3} X^\top W(\beta) X & 0_k \\ 0_k^\top & \dfrac{10}{\phi^6}\displaystyle\sum_{i=1}^n m_i^2 a''(-m_i/\phi) - \dfrac{5}{\phi^7}\sum_{i=1}^n m_i^2 a'''(-m_i/\phi) \\ & \quad + \dfrac{1}{2\phi^8}\displaystyle\sum_{i=1}^n m_i^2 a_i^{iv}(-m_i/\phi) \end{bmatrix}, \tag{20}$$

and the mixed second derivative is

$$\frac{\partial^2 i(\beta,\phi)}{\partial \beta_u \partial \phi} = \begin{bmatrix} -\dfrac{1}{\phi^2} X^\top W'_u(\beta) X & 0_k \\ 0_k^\top & 0 \end{bmatrix}. \tag{21}$$

Algorithm 1 for computing the LA statistics can also be implemented by replacing steps 15 and 16 with the evaluation of derivatives (18)–(21). The `waldi` R package can compute the LA statistics using either the analytical derivatives or numerical differentiation of $\kappa(\beta, \phi)$, as in Section 7.

9.3. Gamma regression with unknown dispersion

The data in McCullagh and Nelder (1989, § 8.4.2) consist of 18 observations of mean blood clotting times in seconds for nine percentage concentrations of normal plasma and two lots of clotting agent. The clotting times are assumed here to be realizations of independent random variables $Y_1, \ldots, Y_{18}$, each Y_i having a Gamma$(\phi^{-1}, (\phi\mu_i)^{-1})$ distribution with

$$\log \mu_i = \beta_1 + \sum_{j=2}^4 \beta_j x_{ij} \quad (i = 1, \ldots, 18),$$

where x_{i2} denotes the normal plasma concentration, x_{i3} is a dummy variable encoding which of the two lots of clotting agent was employed for the ith observation, and $x_{i4} = x_{i2}x_{i3}$ is the interaction between the plasma concentration and the clotting agent.

Table 4 reports the ML estimates of the regression parameters and the Wald statistics for the individual hypotheses $H_0 : \beta_j = 0$ ($j = 1, \ldots, 4$) using either $\hat{\phi}$ or $\tilde{\phi}$ in (17), along with the corresponding LA Wald statistics. The values of the Wald statistic with $\tilde{\phi}$ are closer to t_j^* than those with $\hat{\phi}$, confirming that the recommendation of McCullagh and Nelder (1989) to use $\tilde{\phi}$ in (17) delivers some location correction. Fig. 3 shows empirical null rejection probabilities at several nominal levels when testing $H_0 : \beta_j = \beta_{j0}$ against $H_1 : \beta_j \neq \beta_{j0}$ and $H_1 : \beta_j < \beta_{j0}$ ($j = 2, 3, 4$), where β_{j0} is set at the estimate of β_j in Table 4. The test based on the LA Wald statistic performs better than that based on the Wald statistic with $\hat{\phi}$ and $\tilde{\phi}$. As expected, the latter is the least reliable. The statistics $\tilde{t}_j$ and $\tilde{t}_j^*$, based on RB estimators of β and ϕ, result in the same, to plotting accuracy, rejection probabilities as the Wald statistic using $\tilde{\phi}$ and t_j^*, respectively (see supplementary material). For this reason, we have not reported their values in Table 4 and empirical rejection probabilities in Fig. 3.

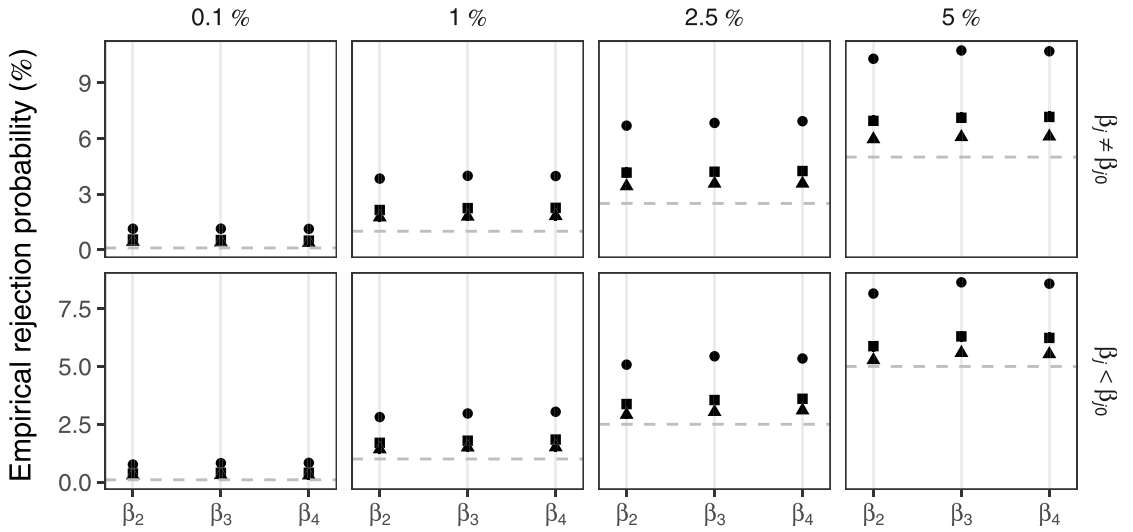

Fig. 3. Empirical null rejection probabilities when testing $H_0 : \beta_j = \beta_{j0}$ against $H_1 : \beta_j \neq \beta_{j0}$ (top row) and $H_1 : \beta_j < \beta_{j0}$ (bottom row) ($j = 2, 3, 4$) based on the normal approximation to the distribution of t^* (triangles) and the Wald statistic using $\tilde{\phi}$ (squares) and $\hat{\phi}$ (circles). The null value β_{j0} is set at the estimate of β_j in Table 4. Reported rates obtained using 50 000 simulated samples from the ML fit shown in Table 4, and for nominal levels (dashed grey line) 0.1% (left-most column), 1% (second column from left), 2.5% (third column from left), and 5% (right column).

Table 5
Maximum likelihood ($\hat{\gamma}$), maximum conditional likelihood ($\hat{\gamma}_c$) and reduced-bias ($\tilde{\gamma}$) estimates for γ in (22), with corresponding estimated standard errors (in parenthesis). The statistics are for $H_0 : \gamma = 0$ and involve the Wald statistic using the ML, maximum conditional likelihood and reduced-bias estimates (t, t_c and $\tilde{t}$, respectively), the signed roots of the logarithms of the likelihood and conditional likelihood ratio statistics (r and r_c, respectively), and the location-adjusted Wald statistics based on the ML and reduced-bias estimates (t^* and $\tilde{t}^*$, respectively). Approximate p-values based on the normal distribution are given in square brackets.

$\hat{\gamma}$	$\hat{\gamma}_c$	$\tilde{\gamma}$	t	t_c	$\tilde{t}$	r	r_c	t^*	$\tilde{t}^*$
1.4324	1.2561	1.1562	1.9511	1.8307	1.7362	2.1596	2.0214	1.9257	1.9064
(0.7341)	(0.6861)	(0.6659)	[0.0510]	[0.0671]	[0.0825]	[0.0308]	[0.0432]	[0.0541]	[0.0566]

9.4. Logistic regression with many nuisance parameters

The dataset in Davison (1988, Table 1) records 18 matched pairs of binomial observations from a study that investigates the effect of lulling on the crying of babies. Matching is per day and each day pair consists of the number of babies not crying out of a fixed number of control babies, and the outcome of "lulling" on a single child. A total of 143 babies are involved in the experiment. Interest is in testing the effect of lulling on the crying of children. Suppose that the crying status of baby j in day i is a Bernoulli random variable with probability μ_{ij} of not crying such that

$$\log \frac{\mu_{ij}}{1 - \mu_{ij}} = \beta_i + \gamma z_{ij} \quad (i = 1, \ldots, 18;\ j = 1, \ldots, n_i), \tag{22}$$

where z_{ij} takes value one if the jth baby out of the n_i children observed on day i was lulled, and zero otherwise. We assume independence between babies and across days. This generalized linear model has $\phi = 1$, and 18 nuisance parameters to allow for different probabilities of not crying across days and account for between-day variations in the experimental design. Estimation and inference about γ can be performed via either the likelihood or the conditional likelihood after eliminating $\beta_1, \ldots, \beta_{18}$ by conditioning on their sufficient statistics (see, for example, Agresti, 2002, § 6.7.1).

Table 5 reports the ML and maximum conditional likelihood (MCL) estimates of γ, the corresponding standard errors, and the values of typical statistics for testing $\gamma = 0$. The statistics $\tilde{t}$ and $\tilde{t}^*$ use the bias-reducing adjusted score estimators in Firth (1993) as implemented in the `brglm2` R package (Kosmidis, 2018). A conventional parametric bootstrap based on 1000 samples under the ML fit gives a p-value of 0.0230, which is similar to what a LR test (r in Table 5) gives. As is apparent, the values of t^* and $\tilde{t}^*$ are closer to that of the Wald statistic based on the MCL estimator $\hat{\gamma}_c$.

Fig. 4 shows the empirical null p-value distribution for the various statistics in Table 5 when testing $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$ and $H_1 : \gamma < 0$. These distributions are computed using 50 000 samples from model (22) with $\beta_1, \ldots, \beta_{18}$ set to their ML estimates from the observed data and $\gamma = 0$. There were 13 samples where $\hat{\gamma}$ and $\hat{\gamma}_c$ had infinite value and the LA Wald statistic for γ was given value zero, in line with the discussion in Section 8. No such conventions are necessary for the computation of $\tilde{t}$ and $\tilde{t}^*$, because the RB estimator is always finite (see, Kosmidis and Firth, 2018, for proof). The detection of infinite estimates was done prior to fitting the models using the linear programming algorithms

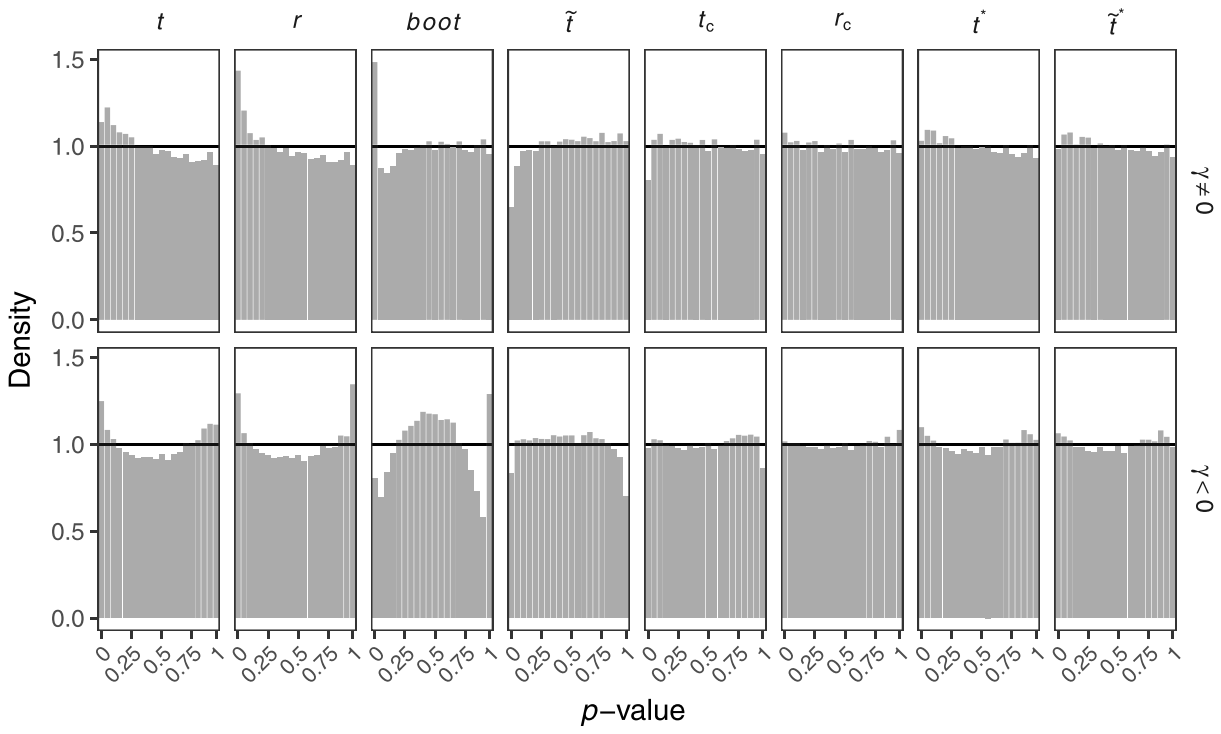

Fig. 4. Empirical null p-value distributions when testing $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$ (top row) and $H_1 : \gamma < 0$ (bottom row) using the normal approximation to the distribution of t, $\tilde{t}$, t_c, r, r_c, t^*, $\tilde{t}^*$ and parametric bootstrap based on 1000 samples under the ML fit (*boot*). The solid horizontal line at one is the uniform density. Results obtained from a simulation study with 50 000 replications.

in the unpublished PhD thesis of Konis (2007), as implemented in the `detect_separation` method of the `brglm2` R package.

The empirical null p-value distributions for t and for the signed root of the logarithm of the LR statistic r are far from uniform. This is a well-studied issue when testing in the presence of nuisance parameters (Davison, 1988), which can be remedied by their elimination via conditioning on their sufficient statistics and carrying out inference based on the conditional likelihood; Fig. 4 illustrates that the distribution of p-values based on r_c is much closer to uniform than that of t, r and parametric bootstrap based on 1000 samples under the ML fit. Despite of its simplicity and of being based on a single fit of the model that involves all nuisance parameters, the LA statistic t^* and, particularly, $\tilde{t}^*$ deliver a dramatic improvement over all t, r, and bootstrap, having p-value distributions that are close to uniform and significantly lower computational cost than bootstrap, r, and r_c.

10. Significance maps from brain lesion data

Ge et al. (2014) propose a spatial probit model and develop the associated Bayesian machinery for the analysis of multiple sclerosis lesion maps, accounting for spatial dependence between lesion location and subject-specific covariates. After warping the brain data for each individual in the sample to a common atlas, the resulting lesion maps comprise one binary observation in each voxel, indicating the presence or absence of a lesion. The alternative approach, which is also explored in Ge et al. (2014, § 4.1), is mass univariate modelling of lesion occurrence, where a binary-response generalized linear model is fitted independently on each voxel. A key inferential output from either approach are significance maps, which highlight voxels according to the evidence against the null hypothesis of no covariate effect.

Spatial probit regression properly accounts for spatial dependence using a probit model with multivariate conditional auto-regressive priors for the regression parameters. However, as the resolution of the brain scans and the number of individuals increases, posterior sampling becomes a computationally tedious task due to the large dimension of the parameter space. For example, the moderately sized application in Ge et al. (2014) has 250 individuals with 274 596 voxels each, and requires sampling about 2 750 000 parameters from the posterior. Ge et al. (2014) achieved this by developing a partly-parallelizable Gibbs sampling procedure and distributing the computation using GPUs. On the other hand, mass univariate regression does not formally account for spatial variation, but is, nevertheless, computationally attractive. It can be carried out on a regular laptop in a matter of minutes, because of the concavity of the log-likelihood and the ability to parallelize computations over the voxels. Significance maps can then be produced by relying on the voxel value

of standard statistics, like the Wald and the signed root of the logarithm of the LR statistic. Ge et al. (2014) used Wald statistics from voxel-wise logistic regressions based on the RB estimator in Firth (1993).

Here, we use the sample of 50 patients from the supplementary material of Ge et al. (2014) to construct significance maps using both $\tilde{t}$ and its LA version $\tilde{t}^*$ for the parameters of a binary-response generalized linear model with probit link. The available covariates are multiple sclerosis type (0 if relapsing–remitting multiple sclerosis, 1 if secondary progressive multiple sclerosis), and standardized versions of age, sex, disease duration and two disease severity measures (PASAT and EDSS; see Ge et al., 2014, for details) for each individual. The lesion maps for the 50 individuals have 902 629 voxels each (resolution $91 \times 109 \times 91$). Even with such a small sample of patients, there are 4304 unique configurations of lesion occurrence across the voxels in the sample. One of those is the trivial configuration where there is no lesion occurrence for all patients, and which happens for 879 237 voxels. So, for mass univariate regression we need to fit 4304 probit regression models, each with an intercept and 6 parameters, one for every covariate.

The ML estimates for the parameters of age, disease duration, EDSS, PASAT, sex and type, were infinite for 63.7%, 63.7%, 63.2%, 63.6%, 78.3% and 75.5% of the non-trivial voxels, respectively, making t and t^* not very useful to produce significance maps. On the other hand, the computation of r (the signed root of the logarithm of the LR statistic) requires 6 times more fits – one extra fit of the nested model which results by omitting each covariate – than for Wald statistics. In addition, that computation fails in a considerable amount of cases due to separation either in the full or the nested models. This is because standard numerical algorithms (like the IRLS implementation in the `glm` R function) reach the maximum number of iterations without achieving the maximum of the log-likelihood, resulting in a negative value of the logarithm of the LR statistic. For example, the calculation of r failed in 18.1% of non-trivial voxels, while testing for disease duration.

Fig. 5 shows the results for disease duration accounting for all other covariates, using a single sagittal slice of the average white matter map supplied in the supplementary material of Ge et al. (2014). As expected, larger disease duration corresponds to more damage from multiple sclerosis, which is apparent by the positive association between disease duration scores and lesion occurrence, especially along the minor forceps. Use of $\tilde{t}$ results in 18.9% of the voxels being greater than 1 in absolute value. The corresponding percentage for the LA $\tilde{t}^*$ is 24.8%. As can be seen in the top right plot of Fig. 5, the location adjustment typically results in an inflation of the value of the Wald statistic, leading in turn to stronger signal detection.

11. Bootstrap for location- and scale-adjusted statistics

A further enhancement to inferential procedures using the Wald statistic can be achieved by also correcting for its variance. As in (7), we can define a location- and scale-adjusted statistic $t^{**} = (t - \hat{B})\hat{V}^{-1/2}$, where $\hat{V}$ is an estimator of the variance $V(\theta, \psi_0)$ of t in (2) or t^* in (7). One way to obtain an estimator for $V(\theta, \psi_0)$ is to derive the series expansion

$$V(\theta, \psi_0) = 1 + V_2(\theta, \psi_0) + O(n^{-2}),$$

where $V_2(\theta, \psi_0) = O(n^{-1})$, and use $\hat{V} = 1 + V_2(\hat{\theta}, \psi_0)$. Section 2.2 of the unpublished PhD thesis by Claudia Di Caterina (Di Caterina, 2017) evaluates the performance of such a scale adjustment in one-parameter problems. It is therein found that, while the location- and scale-adjusted Wald statistic has, in theory, a distribution that is asymptotically closer to $N(0, 1)$, $\hat{V}$ can take extreme, or even negative, values in prominent modelling settings. This renders the evaluation of t^{**} unstable or impossible.

A stable scale adjustment results by estimating $V(\theta, \psi_0)$ from the bootstrap distribution of t^*. The computation required is, practically, the same as the one used for getting studentized CIs in Section 7, with the difference that the bootstrap samples are used to estimate a variance instead of quantiles at tail probabilities. The bootstrap scale adjustment applies directly to the LA Wald statistic $\tilde{t}^*$ based on RB estimators in Section 3, to give the location- and scale-adjusted statistic $\tilde{t}^{**}$.

As an illustration, for the beta regression model of Example 1.1, the empirical coverage probabilities of 90%, 95% and 99% CIs for the interaction parameter β_4 derived by the inversion of t^{**} and $\tilde{t}^{**}$ using 500 bootstrap samples are 89.8%, 94.6%, 98.7%, and 90.2%, 94.8%, 98.8%, respectively. For the gamma regression model in Section 9.3, the empirical null rejection probabilities when testing $H_0 : \beta_4 = \beta_{40}$ at levels 0.1%, 1%, 2.5%, 5% via t^{**} based on 500 bootstrap replicates are, respectively, 0.32%, 1.50%, 3.02%, 5.33% for $H_1 : \beta_4 \neq \beta_{40}$, and 0.24%, 1.25%, 2.69%, 4.95% for $H_1 : \beta_3 < \beta_{30}$ (see supplementary material for more extensive results).

While the inversion of t^{**} or $\tilde{t}^{**}$ to get CIs is possible, it is clearly a laborious process because a separate lot of bootstrap samples is necessary at each point on the grid of values considered for ψ (see Section Section 6). Nevertheless, the computation of p-values based on t^{**} or $\tilde{t}^{**}$ requires only the generation of a single lot of bootstrap samples.

12. Concluding remarks

Correcting the first moment of the distribution of the Wald statistic has been found to be particularly effective in scenarios where the bias of the ML estimator impacts inferential conclusions on the parameter of interest. In contrast to its main competitors, such as the bootstrap and the LR and score statistics, the LA Wald statistic has computational complexity similar to that of the standard Wald one and does not need any extra model fits. As is done in the algorithm

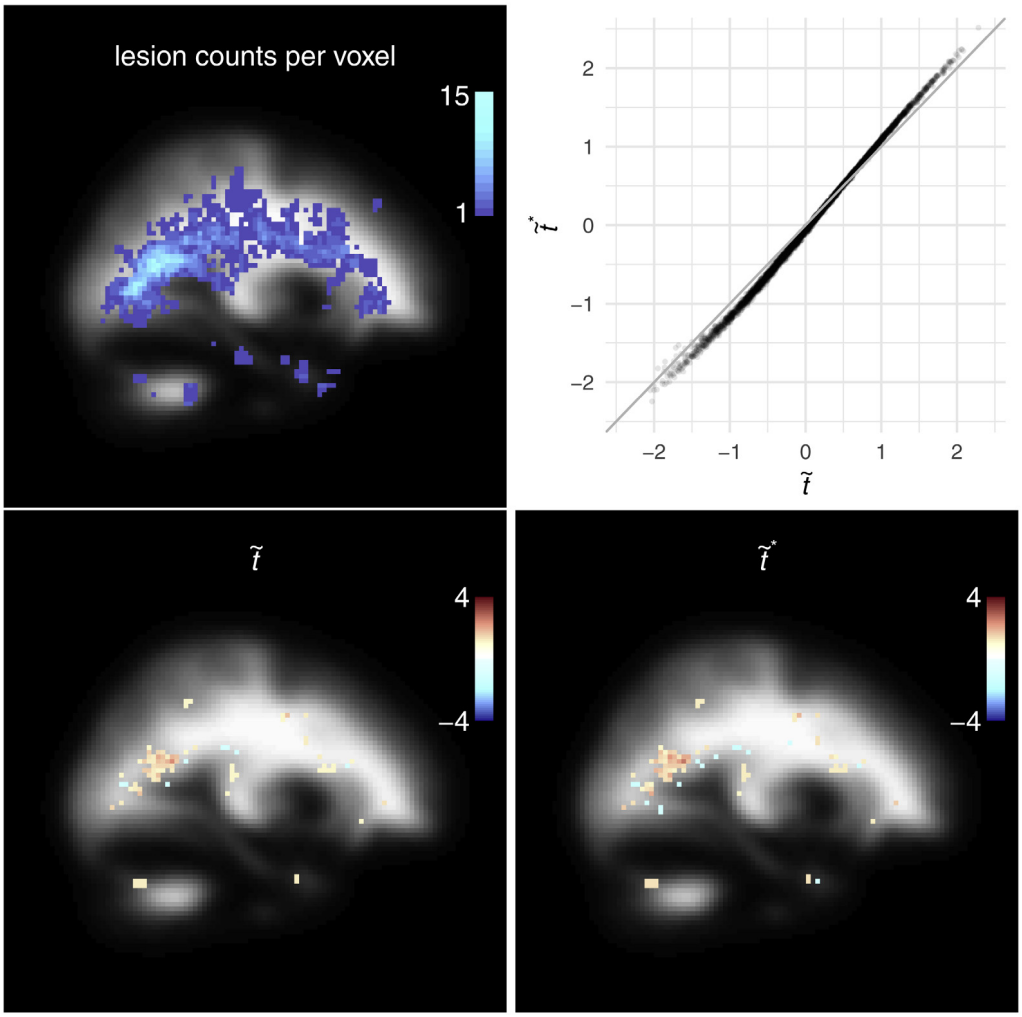

Fig. 5. A single sagittal slice of the average white matter map overlayed with non-zero counts of lesions per voxel amongst the 50 individuals (top left), significance maps for disease duration based on $\tilde{t}$ and $\tilde{t}^*$ (bottom left and bottom right, respectively), and $\tilde{t}$ versus $\tilde{t}^*$ (top right). Voxels with statistic values between -1 and 1 are not shown.

of Section 4, numerical differentiation can be used to approximate the gradient and hessian of $\kappa(\theta)$, in order to deliver general and modular implementations of the location adjustment that depend only on a ready implementation of the expected information matrix and the bias function. The `waldi` R package in the supplementary material provides such implementations. This approach, though general, does not scale as well as analytical derivatives do with p, due to the matrix inversions required during numerical differentiation. For instance, using numerical derivatives, the calculation of the LA statistic t^* in Table 5 takes approximately 6 times the computing time the analytical calculation does (using the `waldi` R package on a MacBook Pro laptop with 3.5 GHz Intel Core i7 processor and 16 GB of RAM; see supplementary material).

The theoretical framework for LA Wald statistics is modular, allowing to utilize alternative variance–covariance matrices or bias estimates. The derivation of the bias of the Wald statistic in (2) depends only on an explicit expression for the variance–covariance matrix of the estimator. Hence, the location adjustment can be performed with more general forms for the variance–covariance matrix of the asymptotic distribution of the estimator, including robust versions of it, like the ones of MacKinnon and White (1985). Moreover, for $b(\theta)$ in (6), we used the first term in the expansion of the bias of the ML estimator. In fact, the theory in Section 2 applies unaltered to any approximation of $b(\theta)$ that satisfies $E_\theta(\hat{\theta} - \theta) = b(\theta) + o\left(n^{-1}\right)$. This includes the bias itself, if that is available, and simulation-based estimators of it, obtained by jackknifing or parametric and non-parametric bootstrap, as described in Efron and Tibshirani (1993, Chapter 10) and Davison and Hinkley (1997, Chapter 2). The expression of the bias of the Wald statistic in (6) is also the same for Wald statistics based on any estimator that is $\sqrt{n}$-consistent, asymptotically normal and has a known variance–covariance matrix.

A bootstrap procedure that exploits the computational simplicity of LA Wald statistics has also been introduced to deliver location- and scale-adjusted Wald statistics. Note here that the variance of t^* can be reliably estimated with less bootstrap samples than those needed for accurate estimation of quantiles or tail probabilities of its bootstrap distribution. The latter are what is required, for instance, to construct studentized CIs and carry out the bootstrap hypothesis tests.

Further research can surely deal with the theoretical evaluation of the LA Wald statistic in the stratified settings considered by Sartori (2003). In addition, the quality of the normal approximation to the distribution of the location- and scale-adjusted statistics, which were here only briefly outlined, deserves more detailed investigations (see, e.g., Lee and Young, 2005, for relevant discussions in the context of bootstrap prepivoting). Another direction for future work involves the more systematic comparison of the various statistics for the construction of significance maps, including permutation-based approaches (see, among others, Winkler et al., 2014).

In closing, the main criticism about Wald procedures is their lack of invariance under non-linear transformations of the parameter. Specifically, the conclusions from Wald inferences depend on the parameterization of the model (see, for instance, Larsen and Jupp, 2003, who tackle this issue by introducing a geometric invariant Wald statistic). The LA Wald statistics improve significantly over standard ones, but their direct dependence on the Wald statistic and its bias renders them also not invariant. If parameterization invariance is critical to the inferential problem at hand, then both Wald statistics and their LA versions should be used with care.

Acknowledgements

Ioannis Kosmidis was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 (Turing award number TU/B/000082). Part of this work was completed when Ioannis Kosmidis was a Senior Lecturer at University College London, where Claudia Di Caterina spent a year as a visiting Ph.D. student. Claudia Di Caterina was partially funded by the Italian Ministry of Education under the PRIN 2015 grant 2015EASZFS_003. The authors are grateful to Thomas Nichols for helpful discussions and pointers on mass univariate regression for the occurrence of brain lesions. They also thank Nicola Sartori and Alessandra Salvan for useful talks about the methodology.

Appendix A. Supplementary data

The supplementary material (also available at https://github.com/ikosmidis/waldi/tree/master/inst/supplementary_1710-11217) provides code to fully reproduce all the numerical results and outputs in the paper, including additional and enriched outputs from the numerical computations and simulation experiments.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2019.04.004.

References

Agresti, A., 2002. Categorical Data Analysis, second ed. John Wiley & Sons, Inc., New York.

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Amer. Statist. 54, 280–288. http://dx.doi.org/10.1080/00031305.2000.10474560.

Agresti, A., Coull, B.A., 1998. Approximate is better than exact for interval estimation of binomial proportions. Amer. Statist. 52, 119–126. http://dx.doi.org/10.1080/00031305.1998.10480550.

Anscombe, 1956. On estimating binomial response relations. Biometrika 43, 461–464.

Bartlett, M.S., 1937. Properties of sufficiency and statistical tests. Proc. Roy. Soc. Lond. Ser. A 160, 268–282.

Botter, D.A., Cordeiro, G., 1998. Improved estimators for generalized linear models with dispersion covariates. J. Stat. Comput. Simul. 62, 91–104. http://dx.doi.org/10.1080/00949659808811926, URL: http://www.tandfonline.com/doi/abs/10.1080/00949659808811926.

Cook, R.D., Tsai, C.-L., Wei, B.C., 1986. Bias in nonlinear regression. Biometrika 73, 615–623.

Cordeiro, G.M., Ferrari, S.L.d.P., 1991. A modified score test statistic having chi-squared distribution to order n^{-1}. Biometrika 78, 573–582.

Cordeiro, G.M., McCullagh, P., 1991. Bias correction in generalized linear models. J. Roy. Stat. Soc. Ser. B 53, 629–643.

Cordeiro, G., Toyama Udo, M., 2008. Bias correction in generalized nonlinear models with dispersion covariates. Comm. Statist. Theory Methods 37, 2219–2225.

Cordeiro, G.M., Vasconcellos, K.L.P., 1997. Bias correction for a class of multivariate nonlinear regression models. Statist. Probab. Lett. 35, 155–164.

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

Davison, A.C., 1988. Approximate conditional inference in generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 50, 445–461, URL: http://www.jstor.org/stable/2345707.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge, UK.

Di Caterina, C., 2017. Reducing the Impact of Bias in Likelihood Inference for Prominent Model Settings (Ph.D. thesis). University of Padova.

Efron, B., 1975. Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). Ann. Statist. 3, 1189–1217.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall Ltd, New York.

Fears, T.R., Benichou, J., Gail, M.H., 1996. A reminder of the fallibility of the Wald statistic. Amer. Statist. 50, 226–227. http://dx.doi.org/10.1080/00031305.1996.10474384.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T.E., Johnson, T.D., 2014. Analysis of multiple sclerosis lesions via spatially varying coefficients. Ann. Appl. Stat. 8, 1095–1118. http://dx.doi.org/10.1214/14-AOAS718.

Gilbert, P., Varadhan, R., 2016. numDeriv: Accurate Numerical Derivatives. URL: https://CRAN.R-project.org/package=numDeriv, R package version 2016.8-1.

Green, P.J., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J. Roy. Stat. Soc. Ser. B 46, 149–192, URL: http://www.jstor.org/stable/2345503.

Grün, B., Kosmidis, I., Zeileis, A., 2012. Extended beta regression in R: Shaken, stirred, mixed, and partitioned. J. Stat. Softw. 48, 1–25, URL: http://www.jstatsoft.org/v48/i11/.
Haldane, J., 1955. The estimation of the logarithm of a ratio of frequencies. Ann. Hum. Genet. 20, 309–311.
Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer, New York.
Hauck, W.W., Donner, A., 1977. Wald's test as applied to hypotheses in logit analysis. J. Amer. Statist. Assoc. 72, 851–853, URL: http://www.jstor.org/stable/2286473.
Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21, 2409–2419.
Jørgensen, B., 1987. Exponential dispersion models (with discussion). J. Roy. Stat. Soc. Ser. B 49, 127–162.
Konis, K., 2007. Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models (Ph.D. thesis). University of Oxford, URL: https://ora.ox.ac.uk/objects/uuid:8f9ee0d0-d78e-4101-9ab4-f9cbceed2a2a.
Kosmidis, I., 2014. Improved estimation in cumulative link models. J. Roy. Stat. Soc. Ser. B 76, 169–196. http://dx.doi.org/10.1111/rssb.12025.
Kosmidis, I., 2017. enrichwith: Methods to enrich list-like R objects with extra components. URL: https://github.com/ikosmidis/enrichwith R package version 0.1.
Kosmidis, I., 2018. brglm2: Bias Reduction in Generalized Linear Models. URL: https://CRAN.R-project.org/package=brglm2 R package version 0.1.8.
Kosmidis, I., Firth, D., 2010. A generic algorithm for reducing bias in parametric estimation. Electron. J. Stat. 4, 1097–1112. http://dx.doi.org/10.1214/10-EJS579.
Kosmidis, I., Firth, D., 2011. Multinomial logit bias reduction via the poisson log-linear model. Biometrika 98, 755–759.
Kosmidis, I., Firth, D., 2018. Jeffreys' prior, finiteness and shrinkage in binomial-response generalized linear models. arXiv e-prints, (p. arXiv:1812.01938).
Larsen, P., Jupp, P., 2003. Parametrization-invariant Wald tests. Bernoulli 9, 167–182. http://dx.doi.org/10.3150/bj/1068129014.
Lee, S.M., Young, A.G., 2005. Parametric bootstrapping with nuisance parameters. Statist. Probab. Lett. 71, 143–153.
MacKinnon, J.G., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J. Econometrics 29, 305–325.
Magnus, J.R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester.
Mantel, N., 1987. Understanding Wald's test for exponential families. Amer. Statist. 41, 147–148, URL: http://www.jstor.org/stable/2684232.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.
Pace, L., Salvan, A., 1997. Principles of Statistical Inference: From a Neo-Fisherian Perspective. World Scientific, London.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.
Sartori, N., 2003. Modified profile likelihoods in models with stratum nuisance parameters. Biometrika 90, 533–549.
Simas, A.B., Barreto-Souza, W., Rocha, A.V., 2010. Improved estimators for a general class of beta regression models. Comput. Statist. Data Anal. 54, 348–366. http://dx.doi.org/10.1016/j.csda.2009.08.017, URL: http://www.sciencedirect.com/science/article/pii/S0167947309003107.
Smithson, M., Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol. Methods 11, 54–71.
Væth, M., 1985. On the use of Wald's test in exponential families. Internat. Statist. Rev. 53, 199–214.
Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397. http://dx.doi.org/10.1016/j.neuroimage.2014.01.060.