Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2018, Vol. 27(5) 1365–1375
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216661213
journals.sagepub.com/home/smm
SAGE

Inferential tools in penalized logistic regression for small and sparse data: A comparative study

Marianna Siino, Salvatore Fasola and Vito MR Muggeo

Abstract

This paper focuses on inferential tools in the logistic regression model fitted by the Firth penalized likelihood. In this context, the Likelihood Ratio statistic is often reported to be the preferred choice as compared to the 'traditional' Wald statistic. In this work, we consider and discuss a wider range of test statistics, including the robust Wald, the Score, and the recently proposed Gradient statistic. We compare all these asymptotically equivalent statistics in terms of interval estimation and hypothesis testing via simulation experiments and analyses of two real datasets. We find out that the Likelihood Ratio statistic does not appear the best inferential device in the Firth penalized logistic regression.

Keywords

Logistic regression, Firth penalized likelihood, sandwich formula, score statistic, gradient statistic

1 Introduction

Logistic regression is undoubtedly one of the most popular statistical models routinely used in many areas of applied statistics, such as Biology, Ecology and especially Medicine.[1] Given a set of predictors, the model aims to assess the specific effect of one or more explanatory variables in the regression equation while providing estimates of the probability of the dichotomous outcome; excellent books describe it in detail.[2,3]

While standard logistic regression belongs to the main frame of statistical methods for applied research, its application in the presence of sparse data needs some attention and caution.[4] Sparseness is sometimes referred as 'separation' of data, as described by Albert and Anderson,[5] and it can be caused by occurrence of small sample size, and/or rare events, and/or unbalanced or highly predictive risk factors. Performance of the usual asymptotic procedures, point estimates, confidence intervals and hypothesis testing deteriorate in finite samples with sparseness of data.[5,6] More specifically, point estimates are not guaranteed to exist, and when obtained, they can suffer from important bias leading to inconsistent estimators; interval estimates and p-values are doubtful due to the non-Normal sampling distribution of the usual test statistics.

Different strategies have been discussed in the literature, both 'corrective', and 'preventive'. The former approach means that the maximum likelihood estimates (MLE) are adjusted ex-post,[7–11] whereas in the preventive approach the estimates solve modified estimating equations accounting for the bias of the MLEs. In the Firth's penalized approach,[12] proper estimating equations are defined and parameter estimates are obtained accordingly, leading to estimators that are unbiased to the order n^{-1}. Such approach turns out particularly useful with nearly separated data: besides removing the first-order bias, the approach guarantees the point estimates to be finite even in the presence of monotone likelihood[13] when the usual ML estimates do not exist. In the context of small samples and/or sparse data, some comparisons have been discussed[14,15] ending up with the Firth penalized framework as the final recommendation. The Firth approach in logistic regression has been extended to multinomial responses[16] and ordinal responses via cumulative link regression models;[17] more generally, the penalized approach has also emerged noteworthy in developing prediction models.[18]

Department of Scienze Economiche, Aziendali e Statistiche, University of Palermo, Italy

Corresponding author:
Vito MR Muggeo, Department of Scienze Economiche, Aziendali e Statistiche, University of Palermo, Italy.
Email: vito.muggeo@unipa.it

Within the Firth penalized logistic regression, several authors have discussed inferential tools to carry out inference on the model parameters.[13,19] Comparisons have concerned the Likelihood Ratio and the Wald statistics only, the latter using a covariance matrix for estimators valid only in very large samples. In the end, likelihood-based confidence intervals are generally recommended over the Wald ones according to simulation evidences.

In this work, we focus on both interval estimation and hypothesis testing in the Firth penalized logistic regression. We extend discussion and comparisons involving also the other asymptotically equivalent test statistics: the Score statistic that, curiously, has been neglected in the aforementioned literature, and the recent Gradient statistic.[20,21] We also argue about inappropriateness of the usual Wald statistic that is usually considered in the literature.

The remainder of this paper is organized as follows. Section 2 summarizes the different test statistics employed to make inference, namely to get confidence intervals and p-values; Section 3 presents the design and the results of the simulation studies and Section 4 deals with the analysis of two well-known datasets in literature. The last Section is devoted to conclusions and final discussion.

2 The four likelihood-based statistics

Let Y_i be the binary outcome variable for unit i, and $\boldsymbol{x}_i$ the K-dimensional covariate vector, possibly including exposures, confounders, predictors, and the 1 for model intercept. Interest lies on the conditional expected value $\pi_i = E[Y|\boldsymbol{x}_i]$ related to covariates via the logistic regression equation

$$\text{logit}(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^{K} x_{ij}\beta_j. \tag{1}$$

Maximum likelihood estimates (MLEs) of the regression parameters are usually obtained by equating to zero the score vector $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = U_j(\boldsymbol{\beta}), j = 0, 1, 2, \ldots, K$. As discussed in the previous section, sparseness poses serious issues in obtaining reliable point estimates and related quantities, and better estimating equations are obtained within the Firth penalized approach. Firth[12,22] suggests the following modification of the classical score function $U_j(\boldsymbol{\beta})$ that allows to remove the $O(n^{-1})$ bias of the ML estimates

$$U_j^*(\boldsymbol{\beta}) = U_j(\boldsymbol{\beta}) + \frac{1}{2}\ \text{trace}\left\{\boldsymbol{I}(\boldsymbol{\beta})^{-1}\frac{\partial \boldsymbol{I}(\boldsymbol{\beta})}{\partial \beta_j}\right\}, \quad j = 0, 1, 2, \ldots, K \tag{2}$$

where $\boldsymbol{I}(\boldsymbol{\beta})$ is the expected Fisher information. This is equivalent to use the penalized log-likelihood

$$\ell^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \log|\boldsymbol{I}(\boldsymbol{\beta})|^{\frac{1}{2}}, \tag{3}$$

where the penalty $|\boldsymbol{I}(\boldsymbol{\beta})|^{\frac{1}{2}}$ is the so-called Jeffrey's invariant prior in a Bayesian context.[4,12,23,24]

In the following sections, we summarize some statistics on which inference can be based: these are all likelihood-based, and refer to the same standard normal null distribution for large samples. However, in small to moderate samples, they can perform quite differently, as discussed later. We will indicate the full penalized ML estimate via $\hat{\boldsymbol{\beta}}^*$, and the restricted one via $\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}^*$; namely $\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}^*$ is the ML estimate of $\boldsymbol{\beta}$ given $\beta_j = \beta_{0j}$ fixed under the null hypothesis.

2.1 The Likelihood Ratio statistic

The Penalized Likelihood Ratio statistic is defined as the usual Likelihood Ratio statistic but involving the penalized log-likelihood (equation (3)). Thus, it is

$$L = \text{sign}(\hat{\beta}_j^* - \beta_{0j})\sqrt{-2\left\{\ell^*(\hat{\boldsymbol{\beta}}^*) - \ell^*(\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}^*)\right\}}. \tag{4}$$

Several authors have claimed that the penalized Likelihood Ratio statistic is the preferred choice when using the Firth penalty in logistic regression[13,14,19]

2.2 The Wald statistic

The Wald statistic for the jth regression parameter is naturally defined as

$$W = \frac{\hat{\beta}_j^* - \beta_{0j}}{\text{var}(\hat{\beta}_j^*)^{1/2}} \tag{5}$$

The key point here is how to compute $\text{var}(\hat{\beta}_j^*)$, the jth element on the main diagonal of the full covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*)$. All the aforementioned papers in literature use the inverse of the Information, namely $\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*) \approx \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}^*)$. However, it should be acknowledged the $\boldsymbol{I}^{-1}$ is not the right approach to compute the variance of the estimator, at least in moderate samples. In fact, from basics of statistical inference, the asymptotic variance of the ML estimator comes from the linear expansion of the score, leading to the so-called *sandwich* formula,[25] which reduces to $\boldsymbol{I}^{-1}$ only if the second Bartlett identity holds. In the Firth penalized framework, the simple approximation $\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*) \approx \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}^*)$ holds only in very large samples when the penalty effect gets negligible. In particular, in small to moderate samples, $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}^*)$ overestimates the true estimates uncertainty. This crucial issue appears to have been neglected in literature,[13,16,19] causing a (pointless) 'bad reputation' of the Wald statistic. Hence, a more reliable asymptotic variance for $\hat{\boldsymbol{\beta}}^*$ is provided by the sandwich formula $\boldsymbol{H}^*(\hat{\boldsymbol{\beta}}^*)^{-1}\boldsymbol{I}(\hat{\boldsymbol{\beta}}^*)\boldsymbol{H}^*(\hat{\boldsymbol{\beta}}^*)^{-1}$. It should be noted that any adjective 'unpenalized' or 'penalized' (and corresponding asterisk) to be placed before 'Information' is meaningless. There is a unique Information, hereafter denoted by $\boldsymbol{I} = \text{var}(\boldsymbol{U}) = \text{var}(\boldsymbol{U}^*)$; in fact the penalized and unpenalized score have the same variance since penalty in equation (2) does not depend on data. Evaluating the hessian $\boldsymbol{H}^*(\boldsymbol{\beta})$ requires the second derivatives of the penalized log-likelihood; even if these could be obtained analytically or numerically, via finite differences for instance, we found out that computing the hessian using the idea of induced smoothing[26] works better in practice.

We use W_S to indicate the 'fair' Wald statistic using the sandwich variance, and W to mean the 'unfair' one using the simple $\boldsymbol{I}^{-1}$.

2.3 The Score statistic

The Score statistic is well known in the mainstream inference background, but its spread in applications is somewhat limited with respect to the Wald and Likelihood Ratio statistics. Such limited diffusion reflects in the Firth penalized logistic regression framework, where it appears to have been not yet discussed.

The penalized Score takes the form

$$S = U_j^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)\sqrt{\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)_{jj}} \tag{6}$$

where $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)_{jj}$ is the jth element on the main diagonal of the inverse of the variance of the conditional Score U_j^*. It should be noted that $\boldsymbol{U}^*$ is biased, namely its expected value is not zero due to the penalty in equation (2); however, such bias appears to have a negligible effect on the limit distribution of S, making the standard Normal a valid reference distribution to carry out inference on the regression parameters; simulation studies later bear this out. Moreover, as discussed elsewhere,[27] the Jeffrey prior (and the Firth penalty accordingly) is data dependent: this means that it would not be possible to discuss a general behaviour of the bias of the score.

2.4 The Gradient statistic

The Gradient statistic[20] is a relatively new statistic introduced as a possibly simpler alternative to the most popular Likelihood Ratio, Wald and Score statistics.[21] The Gradient statistic can be derived through the geometric mean between W and S, namely

$$G = \text{sign}(\hat{\beta}_j^* - \beta_{0j})\sqrt{(\hat{\beta}_j^* - \beta_{0j})U_j^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)} \tag{7}$$

However, due to lack of the second Bartlett identity, G may be affected by the same inaccuracies of W. Thus, a robust formulation is

$$G_S = \text{sign}(\hat{\beta}_j^* - \beta_{0j})\sqrt{(\hat{\beta}_j^* - \beta_{0j})U_j^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)\left[\boldsymbol{H}^*(\hat{\boldsymbol{\beta}}^*)\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}^*)\right]_{jj}} \tag{8}$$

which comes from a multidimensional characterization of the geometric mean between W_S and S.[28]

3 Simulation study

We carry out a simulation study to assess the properties of the aforementioned statistics in making inference on the regression parameters in Firth's penalized logistic regression: the penalized LR (equation (4)), the Wald statistic (equation (5)) and its robust counterpart W_S using the sandwich variance, the Score statistic (equation (7)), and the Gradient statistics (equations (7) and (8)).

We consider different scenarios to induce, to some extend, sparseness in the data: small to moderate sample sizes and highly predictive risk factors. Also both categorical and numerical covariates are considered. To provide recommendations with practical interest, we consider scenarios with one covariate only (and very small sample $n = 20$), and also three covariates as detailed below.

For each scenario, we generate $B = 5000$ samples of Bernoulli data $Y_i \sim Ber(\pi_i)$, where $\text{logit}(\pi_i) = \eta_i$. In the single-covariate case, we consider three different sample sizes $n \in \{20, 50, 100\}$, and $\eta_i = 1 + \beta x_i$ where the covariate is continuous or binary. To define the covariate, we extract values $z_i \sim \mathcal{N}(0, 1)$, and we simply set $x_i = z_i$ to get a continuous covariate, or $x_i = I(z_i > 0)$ to obtain a binary covariate.

In the multiple-covariate case, $\eta_i = 1 + \beta x_{1i} + 0.5x_{2i} - 0.5x_{3i}$, where to define the covariates, correlated multinormal values are drawn

$$\begin{bmatrix} z_{1i} \\ z_{2i} \\ z_{3i} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 & -0.6 \\ 0.6 & 1 & 0 \\ -0.6 & 0 & 1 \end{bmatrix}\right) \tag{9}$$

and $x_{2i} = I(z_{2i} > 0)$ (binary covariate), $x_{3i} = z_{3i}$ (continuous covariate). The covariate of interest x_1 is obtained via $x_{1i} = I(z_{1i} > 0)$ or $x_{1i} = z_{1i}$. Due to the presence of additional covariates, we focus only on medium to large sample sizes, namely $n \in \{50, 100\}$.

We first address interval estimation and use the statistics to derive confidence intervals for the regression coefficient of interest: we assess the performance via coverage levels and average widths. Then, we use the statistics for hypothesis testing by assessing empirical type I error and power rates. In the simulation runs, we do not distinguish between separated or nearly separated datasets, since the Firth penalized approach always returns finite and reliable estimates in practice.

3.1 Interval estimation

Let $T(\beta_{0j})$ be one of the pivot statistics discussed in Section 2 evaluated at the candidate value β_{0j}. A $(1 - \alpha)100\%$ confidence interval (CI) for the parameter of interest β_j is defined as

$$\text{CI} = \{\beta_{0j} \in \mathbb{R} : z_{\alpha/2} \leq T(\beta_{0j}) \leq z_{1-\alpha/2}\} \tag{10}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the quantiles of the standard normal distribution for a given α.

Simulated data are obtained as described above, with values for the interest parameter $\beta \in \{0.5, 1.5\}$, both in single or multiple covariate context. For the 95% CI of the interest parameter, Tables 1 and 2 report the empirical coverage levels (CL), the average widths ($\text{AW} = \sum_b \{\text{Upp}_b - \text{Low}_b\}/B$), and the average symmetry ratios given by $\text{AS} = \sum_b \{(\text{Upp}_b - \hat{\beta}_b)/(\hat{\beta}_b - \text{Low}_b)\}/B$, where in each replicate b, $\hat{\beta}_b$ is the estimate, and Low_b and Upp_b are the lower and upper confidence limits. This ratio will be one for symmetric CI (such as those returned by W and W_S) and will get more different from one as the asymmetry of the CI increases.

Despite that the Likelihood Ratio statistic performs rather fairly, it does not appear to be 'the best' inferential device in the Firth penalized logistic regression, as claimed in previous works. In general, the CI performances, in terms of coverage level, are similar using the other statistics with some light differences. The 'usual' Wald test

Table 1. Empirical coverage levels, average widths, and average symmetry ratios (based on 5000 runs) of 95% confidence intervals for β_1 according to the different statistics: Wald (W), Robust-Wald (W_S), Likelihood Ratio (L), Score (S), Gradient (G) and Robust-Gradient (G_S).

		Binary covariate						Normal covariate					
		$\beta = 0.5$			$\beta = 1.5$			$\beta = 0.5$			$\beta = 1.5$		
n	stats	CL	AW	AS	CL	AW	AS	CL	AW	AS	CL	AW	AS
20	W	**0.990**	4.75	1.00	**0.984**	5.60	1.00	**0.988**	2.44	1.00	0.951	3.30	1.00
	W_S	0.954	4.13	1.00	**0.962**	4.49	1.00	**0.962**	2.21	1.00	**0.901**	2.64	1.00
	L	0.957	4.91	1.14	**0.976**	6.10	1.43	**0.964**	2.52	1.23	**0.961**	3.62	1.64
	S	0.953	3.85	0.99	0.955	4.33	0.97	0.954	2.19	1.10	**0.929**	3.21	1.29
	G	0.948	5.63	1.33	**0.975**	7.57	1.96	0.954	2.74	1.40	**0.958**	4.04	2.14
	G_S	**0.934**	5.00	1.25	**0.967**	6.02	1.72	0.943	2.60	1.38	**0.933**	3.62	2.04
50	W	**0.968**	2.83	1.00	**0.973**	3.71	1.00	**0.972**	1.41	1.00	0.949	1.99	1.00
	W_S	0.953	2.68	1.00	0.952	3.18	1.00	**0.961**	1.36	1.00	**0.927**	1.81	1.00
	L	0.951	2.86	1.07	**0.963**	3.94	1.36	0.954	1.41	1.13	0.948	2.05	1.34
	S	0.950	2.60	1.00	0.957	3.23	0.98	0.953	1.31	1.03	**0.940**	1.95	1.07
	G	0.946	2.96	1.13	**0.960**	4.42	1.68	0.951	1.44	1.21	0.947	2.13	1.55
	G_S	**0.942**	2.85	1.11	**0.930**	3.86	1.58	0.947	1.42	1.20	**0.937**	2.03	1.52
100	W	0.953	1.95	1.00	**0.971**	2.56	1.00	**0.962**	0.96	1.00	0.954	1.37	1.00
	W_S	0.948	1.91	1.00	0.946	2.35	1.00	**0.958**	0.95	1.00	0.946	1.31	1.00
	L	0.946	1.95	1.04	0.957	2.63	1.22	0.955	0.96	1.09	0.954	1.39	1.22
	S	0.946	1.87	1.00	0.954	2.38	0.99	0.952	0.93	1.01	0.951	1.35	1.03
	G	0.943	1.98	1.07	0.952	2.76	1.39	0.952	0.97	1.13	0.954	1.41	1.35
	G_S	**0.940**	1.95	1.07	**0.940**	2.59	1.36	0.951	0.97	1.13	0.949	1.38	1.34

Note: Boldfaces refer to CL outside the 'plausible' range $.95 \pm 2.57\{(.95 \cdot .05)/B\}^{1/2}$. The logistic regression equation is $\text{logit}(\pi_i) = 1 + \beta x_i$ (see text for details).

Table 2. Empirical coverage levels, average widths, and average symmetry ratios (based on 5000 runs) of the 95% confidence intervals for β_1 according to the different statistics: Wald (W), Robust-Wald (W_S), Likelihood Ratio (L), Score (S), Gradient (G) and Robust-Gradient (G_S).

		Binary covariate						Normal covariate					
		$\beta = 0.5$			$\beta = 1.5$			$\beta = 0.5$			$\beta = 1.5$		
n	stats	CL	AW	AS	CL	AW	AS	CL	AW	AS	CL	AW	AS
50	W	**0.974**	4.10	1.00	**0.969**	4.93	1.00	**0.969**	2.40	1.00	**0.962**	3.03	1.00
	W_S	0.952	3.82	1.00	0.955	4.29	1.00	0.955	2.30	1.00	**0.940**	2.76	1.00
	L	0.952	4.23	1.10	**0.966**	5.50	1.37	0.950	2.42	1.10	0.951	3.16	1.30
	S	0.954	3.78	1.00	**0.959**	4.51	1.00	0.949	2.24	1.03	0.950	2.96	1.10
	G	0.946	4.50	1.19	**0.964**	6.34	1.76	0.943	2.50	1.16	0.947	3.31	1.49
	G_S	**0.932**	4.22	1.16	0.951	5.40	1.62	**0.936**	2.43	1.15	**0.932**	3.10	1.46
100	W	**0.963**	2.75	1.00	**0.967**	3.52	1.00	**0.960**	1.62	1.00	0.956	2.02	1.00
	W_S	0.956	2.66	1.00	0.945	3.17	1.00	0.951	1.59	1.00	0.948	1.93	1.00
	L	0.955	2.77	1.05	0.957	3.77	1.27	0.949	1.62	1.06	0.952	2.04	1.18
	S	0.953	2.64	1.00	0.957	3.31	1.00	0.949	1.55	1.01	0.952	1.98	1.03
	G	0.952	2.83	1.09	0.953	4.09	1.51	0.946	1.64	1.10	0.951	2.08	1.29
	G_S	0.944	2.75	1.08	**0.931**	3.68	1.44	0.943	1.62	1.10	0.945	2.02	1.28

Note: Boldfaces refer to CL outside the 'plausible' range $.95 \pm 2.57\{(.95 \cdot .05)/B\}^{1/2}$. The logistic regression equation is $\text{logit}(\pi_i) = 1 + \beta x_{1i} + 0.5x_{2i} - 0.5x_{3i}$ (see text for details).

heavily overestimates the estimator variance and thus provides wider CIs with coverage levels higher than the nominal one, even if $n = 100$. In the more difficult scenario with small samples and strong predictor (that is $n = 20$ and $\beta = 1.5$), L exhibits a conservative behaviour with somewhat large average widths: even with binary covariate, the average with of L is about 40% larger than the average width from S.

Overall, Score-based CIs appear to perform slightly better than the other competitors, even in the most difficult scenarios with small samples ($n = 20$) and strong predictors causing sparsity and sampling zeros. The coverage levels are quite close to the nominal 95% and the average width is always the lowest. The top performance of S is likely due to its simple structure: it is a sum of random variables and thus it is expected to converge quite fastly to the Normal distribution, and moreover, as discussed in the previous section, its *exact* variance is represented by the classical Fisher information matrix; curiously, its bias appears to have a negligible effect. Interestingly, the robust Wald statistic attains pretty good coverage levels with comparable average widths, sometimes even better than the Likelihood Ratio statistic. The good performance of W_S also gives evidence that the Normal distribution of the regression coefficients estimator is not an issue actually, as long as the Firth penalization is employed and a finite estimate is guaranteed. This result is coherent with discussion reported in Royall.[25] The Gradient statistics provide CIs with acceptable coverage levels but somewhat wider average widths; in particular the simple G, which does not require computation of the second derivatives, produces CIs with the largest width. Also the asymmetry amount of CIs based on G and G_S is fairly larger than the others.

Results for multiple (correlated) covariates do not differ with respect to the one-covariate models, and therefore they will not be further discussed.

3.2 Hypothesis testing

All the statistics presented in Section 2 can be clearly used for hypothesis testing. Moreover, in addition to the six likelihood-based statistics, we also consider a permutation-based test based on the Likelihood Ratio statistic.[29] Here the null distribution is obtained via 1000 permutations of residuals and the p-value is computed as the portion of permuted values exceeding the observed value in the sample. Note that such approach corresponds to the exact conditional logistic regression when there is a single covariate. Here we focus on testing for the effect of the interest covariate, namely the hypothesis $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. The power functions of all statistics under investigation are obtained at seven values of $\beta \in \{\pm 1.5, \pm 1.0, \pm 0.5, 0\}$. Tables 3 and 4 portray the empirical rejection rates at nominal level 0.05 for the single and multiple covariate cases, respectively. Clearly, the entries in the middle columns represent the size, i.e. the type I error probability.

As for interval estimation, results again suggest that the Likelihood Ratio statistic does not perform the best. Under the null hypothesis, all the considered test statistics exhibit acceptable empirical rejection rates. W is affected by the wrong variance formula and thus it exhibits the lowest rejection rates especially in small to moderate samples. As expected, there is no uniformly most powerful test, and the behaviour of the test statistics depends on the true value of the regression coefficient. Both the Gradient statistics, and in particular the robust one G_S, are featured by high power, especially on the 'right side' when the parameter takes positive values; however, G_S tends to return high rejection rates also under the null hypothesis, making it rather useless, at least in theory. In large samples, the penalty effect vanishes and, to some extent, differences disappear.

When the model involves multiple covariates, findings are substantially unchanged for the six likelihood-based statistics, but the permutational statistic exhibits comparatively better performance.

4 Examples

We illustrate the different statistics and relevant findings on two real datasets previously analysed in the literature.

4.1 Osteogenic sarcoma data

The first dataset is taken from Metha and Patel.[30] The data refer to a study on $n = 46$ patients with osteogenic sarcoma. The three-year disease-free interval (DFI3) is the response variable, while the categorical explanatory variables are gender (SEX), the presence of any osteoid pathology (AOP) and lymphocytic infiltration (LI). The main interest lies on the effect of LI. The classical MLEs do not exist finite, because of separation caused by the variable LI: there are no disease-free individuals among subjects without infiltration. We fit a penalized logistic regression model with additive linear effects and compute the 95% confidence intervals for the regression coefficient of covariate LI; results are reported in Table 5.

Table 3. Empirical rejection rates (based on 5000 runs) for hypothesis testing $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ for different test statistics: Wald (W), Robust-Wald (W_S), Likelihood Ratio (L), Permutation Likelihood Ratio (L_p), Score (S), Gradient (G) and Robust-Gradient (G_S).

		True β values						
n	Test	-1.5	-1.0	-0.5	0	0.5	1.0	1.5
					Binary covariate			
20	W	0.208	0.084	0.021	**0.009**	0.013	0.014	0.028
	W_S	0.312	0.155	0.056	**0.038**	0.059	0.097	0.167
	L	0.339	0.176	0.063	**0.039**	0.047	0.076	0.122
	L_p	0.270	0.128	0.049	**0.026**	0.039	0.059	0.099
	S	0.361	0.194	0.074	0.044	0.054	0.083	0.127
	G	0.353	0.190	0.075	0.048	0.061	0.098	0.164
	G_S	0.354	0.193	0.078	**0.059**	0.102	0.165	0.273
50	W	0.669	0.351	0.103	**0.034**	0.074	0.163	0.272
	W_S	0.674	0.364	0.116	0.046	0.100	0.231	0.432
	L	0.701	0.390	0.128	0.050	0.102	0.231	0.429
	L_p	0.664	0.358	0.106	**0.040**	0.089	0.210	0.391
	S	0.725	0.410	0.132	0.051	0.105	0.234	0.424
	G	0.710	0.406	0.131	0.053	0.118	0.256	0.460
	G_S	0.714	0.408	0.133	0.055	0.130	0.283	0.511
100	W	0.948	0.632	0.200	**0.042**	0.158	0.421	0.690
	W_S	0.948	0.636	0.204	0.045	0.168	0.455	0.733
	L	0.955	0.647	0.217	0.049	0.174	0.469	0.739
	L_p	0.941	0.615	0.199	0.043	0.160	0.437	0.716
	S	0.956	0.655	0.219	0.050	0.175	0.467	0.732
	G	0.955	0.651	0.219	0.051	0.178	0.475	0.753
	G_S	0.955	0.651	0.220	0.053	0.187	0.491	0.771
					Normal covariate			
20	W	0.323	0.143	0.036	**0.006**	0.028	0.143	0.335
	W_S	0.587	0.341	0.113	**0.029**	0.097	0.328	0.607
	L	0.653	0.392	0.133	**0.040**	0.125	0.370	0.665
	L_p	0.673	0.417	0.148	0.046	0.140	0.405	0.692
	S	0.666	0.417	0.144	0.046	0.131	0.392	0.673
	G	0.697	0.437	0.162	0.054	0.149	0.423	0.708
	G_S	0.732	0.478	0.183	**0.066**	0.170	0.456	0.735
50	W	0.962	0.749	0.238	**0.027**	0.247	0.747	0.971
	W_S	0.970	0.780	0.277	**0.034**	0.280	0.779	0.977
	L	0.973	0.812	0.313	0.043	0.313	0.807	0.981
	L_p	0.976	0.821	0.322	0.047	0.324	0.814	0.982
	S	0.975	0.820	0.324	0.043	0.318	0.812	0.983
	G	0.976	0.823	0.328	0.047	0.329	0.816	0.983
	G_S	0.978	0.834	0.339	0.049	0.340	0.823	0.984
100	W	1.000	0.978	0.532	**0.039**	0.519	0.979	1.000
	W_S	1.000	0.980	0.546	0.043	0.530	0.981	1.000
	L	1.000	0.983	0.574	0.050	0.556	0.984	1.000
	L_p	1.000	0.983	0.575	0.051	0.561	0.984	1.000
	S	1.000	0.984	0.576	0.052	0.560	0.985	1.000
	G	1.000	0.985	0.581	0.052	0.565	0.985	1.000
	G_S	1.000	0.985	0.584	0.053	0.569	0.985	1.000

Note: The nominal level is 0.05 and boldfaces refer to rates outside the 'plausible' range $.05 \pm 2.57\{(.95 \cdot .05)/B\}^{1/2}$. The logistic regression equation is $\text{logit}(\pi_i) = 1 + \beta x_i$ (see text for details).

Table 4. Empirical rejection rates (based on 5000 runs) for hypothesis testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ for different test statistics: Wald (W), Robust-Wald (W_S), Likelihood Ratio (L), Permutation Likelihood Ratio (L_p), Score (S), Gradient (G) and Robust-Gradient (G_S).

		True β values						
n	Test	−1.5	−1	−0.5	0	0.5	1	1.5
					Binary covariate			
50	W	0.376	0.170	0.061	**0.018**	0.036	0.069	0.107
	W_S	0.411	0.199	0.075	**0.036**	0.077	0.148	0.227
	L	0.460	0.228	0.093	**0.040**	0.072	0.134	0.201
	L_p	0.468	0.234	0.099	0.043	0.083	0.162	0.245
	S	0.487	0.239	0.098	**0.040**	0.069	0.128	0.182
	G	0.477	0.236	0.100	0.046	0.085	0.151	0.225
	G_S	0.480	0.236	0.100	0.050	0.103	0.201	0.295
100	W	0.730	0.368	0.123	**0.041**	0.089	0.207	0.347
	W_S	0.736	0.380	0.129	0.046	0.108	0.261	0.447
	L	0.763	0.410	0.140	0.052	0.109	0.255	0.434
	L_p	0.766	0.413	0.146	0.053	0.115	0.269	0.460
	S	0.773	0.421	0.144	0.052	0.107	0.243	0.411
	G	0.768	0.419	0.145	0.054	0.113	0.272	0.454
	G_S	0.768	0.418	0.145	0.056	0.129	0.307	0.503
					Normal covariate			
50	W	0.962	0.749	0.238	**0.027**	0.247	0.747	0.971
	W_S	0.970	0.780	0.277	**0.034**	0.280	0.779	0.977
	L	0.973	0.812	0.313	0.043	0.313	0.807	0.981
	L_p	0.976	0.821	0.322	0.047	0.324	0.814	0.982
	S	0.975	0.820	0.324	0.043	0.318	0.812	0.983
	G	0.976	0.823	0.328	0.047	0.329	0.816	0.983
	G_S	0.978	0.834	0.339	0.049	0.340	0.823	0.984
100	W	1.000	0.978	0.532	**0.039**	0.519	0.979	1.000
	W_S	1.000	0.980	0.546	0.043	0.530	0.981	1.000
	L	1.000	0.983	0.574	0.050	0.556	0.984	1.000
	L_p	1.000	0.983	0.575	0.051	0.561	0.984	1.000
	S	1.000	0.984	0.576	0.052	0.560	0.985	1.000
	G	1.000	0.985	0.581	0.052	0.565	0.985	1.000
	G_S	1.000	0.985	0.584	0.053	0.569	0.985	1.000

Note: The nominal level is 0.05 and boldfaces refer to rates outside the 'plausible' range $.05 \pm 2.57\{(.95 \cdot .05)/B\}^{1/2}$. The logistic regression equation is $\text{logit}(\pi_i) = 1 + \beta x_{1i} + 0.5x_{2i} - 0.5x_{3i}$ (see text for details).

Table 5. Confidence intervals (95%) (and relevant width) for β_{LI} based on the Wald (W), Robust-Wald (W_S), Likelihood Ratio (L), Score (S), Gradient (G) and Robust-Gradient (G_S) statistics.

CI	W	W_S	L	S	G	G_S
Low	−5.504	−4.637	−7.363	−4.804	−10.148	−10.161
Upp	0.582	−0.286	−0.188	−0.104	−0.356	−0.355
Width	6.09	4.35	7.17	4.70	9.79	9.81

The results are coherent with the previous simulation findings. The CIs based on the Gradient statistics are the widest and the most asymmetric, although they end up with a significant effect of LI. The CI coming from W is also pretty wide but it is symmetric and includes the zero, leading to a (likely misleading) non-significant effect of the variable LI. On the other hand, the Score and the robust Wald CIs are the narrowest ones and also return a significant result for LI.

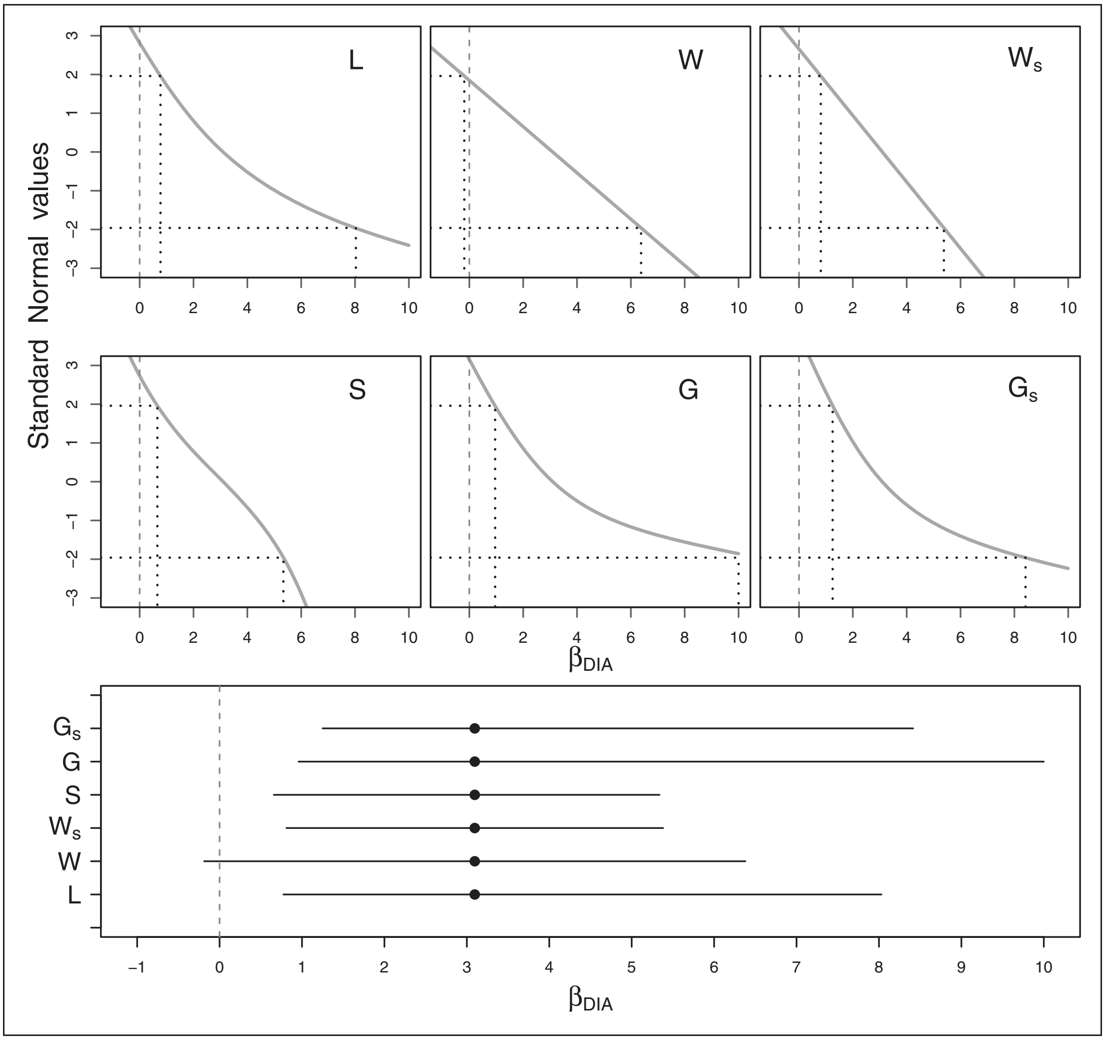

Figure 1. The six likelihood-based statistics (L = Likelihood Ratio, W = Wald, W_S = robust Wald, S = Score, G = Gradient, G_S = robust Gradient) profiled and corresponding 95% confidence intervals for the parameter β_{DIA} in the Urinary tract infection example.

4.2 Urinary tract infection

The second example concerns a retrospective case-control study on possible determinants of urinary infections on a sample of sexually active college women carried out at the University of Michigan.[31,32]

The 130 cases and 109 controls were classified according to six binary covariates, namely age (lower or greater than 23 years) and some indicators of sexual behaviour: use of oral contraceptive (OC), condom (VIC), lubricated condom (VICL), spermicide (VIS) or diaphragm (DIA). Since all the women using a diaphragm were cases, separation occurs and we estimate a penalized logistic regression model to obtain finite estimates. We focus on the effect of covariate ‘using the diaphragm’ (DIA) on the probability of infection, and compute the 95% confidence interval for corresponding regression coefficient β_{DIA} using the different statistics. The intervals are represented in Figure 1, together with the six curves corresponding to the pivotal statistics $T(\beta_{DIA})$ profiled. Again, the CIs based on S and W_S are the narrowest, while the CI based on W leads to different, and possibly misleading, conclusions. Again the gradient statistics lead to quite large and asymmetric CI but not including the zero.

5 Conclusions

Small datasets are rather frequent in medical research, and it is therefore quite imperative from a statistical perspective to set up appropriate methods to obtain valid and reliable inferential procedures. At this aim, many authors have carried out comparisons among the different 'classical' options, but interest was limited to the 'usual' Wald and the Likelihood Ratio statistics. The usual Wald uses a simple but inappropriate formula for the estimates variance, i.e. the inverse of Information. However, as previously discussed, the inverse of Information is appropriate only in large samples. In small to moderate samples, when the penalty is not negligible, a sandwich formula turns out to be more appropriate, and the resulting robust Wald-based CIs have the correct coverage levels. Moreover, the satisfactory performance of the robust Wald suggests that Normality of the estimator is not actually an issue as claimed previously in the literature; instead, as discussed by Royall,[25] using the appropriate variance appears to be enough to guarantee a large sample standard Normal distribution. Besides Wald and Likelihood Ratio, in this paper we have also discussed the Score, the Gradient statistics.

In the context of interval estimation, the Score-based approach appears to return CIs with correct coverage levels across the scenarios and regardless of the number of covariates in the model. For hypothesis testing problems, differences among the approaches are minor, with a noticeable good performance of the simple Gradient statistic; its ease of computation represents a noteworthy advantage, since it just needs the point estimate and the first derivative. From a practical viewpoint, the different statistics exhibit approximately the same computational burden: profiling is a necessary step to obtain the endpoints of the CI in all but one statistic: in fact the robust Wald statistic does not need profiling, but the (penalized) hessian has to be computed to obtain reliable standard errors via the sandwich formula. The permutational approach is the heaviest as the model has to be fitted at each permuted dataset.

For applications, currently the Firth procedure is implemented in most statistical softwares/languages: for instance, the `brglm` or `pmlr` packages in R, the `firthlogit` program in Stata, and the `firth` option to the model statement in `proc logistic` in SAS. Currently it seems that only the Likelihood Ratio or the 'unfair' Wald statistics have been implemented for confidence intervals or hypothesis testing; it is hoped that some of the 'alternative' statistics, such as the Score, are implemented to obtain results with better statistical properties. The supplementary material available online includes R code to compute Score-based CI and p-value (smm.sagepub.com).

Finally, application and comparisons of the aforementioned statistics in more general logistic regressions, such as the cumulative link models for ordinal responses,[17] represent a noteworthy point to be investigated.

Acknowledgement

The authors would like to thank the reviewers for their insightful comments and suggestions which greatly improved the paper.

Declaration of conflicting interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplementary material

Supplementary material is available for this article online.

References

1. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York, NY: Springer, 2008.
2. Harrell F. *Regression modeling strategies*. New York, NY: Springer, 2001.
3. Hosmerm D, Lemeshowm S and Sturdivant R. *Applied logistic regression*, 3rd ed. New York, NY: Wiley, 2013.
4. Hamra G, MacLehose R and Cole S. Sensitivity analyses for sparse-data problems using weakly informative bayesian priors. *Epidemiology* 2013; **24**: 233–239.
5. Albert A and Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; **71**: 1–10.
6. Jennings D. Judging inference adequacy in logistic regression. *J Am Statis Assoc* 1986; **81**: 471–476.

7. Schaefer RL. Bias correction in maximum likelihood logistic regression. *Stat Med* 1983; **2**: 71–78.
8. Cordeiro GM and McCullagh P. Bias correction in generalized linear models. *J Royal Stat Soc Ser B (Methodological)* 1991; **53**: 629–643.
9. Bull SB, Greenwood CM and Hauck WW. Jackknife bias reduction for polychotomous logistic regression. *Stat Med* 1997; **16**: 545–560.
10. Cordeiro GM and Cribari-Neto F. On bias reduction in exponential and non-exponential family regression models. *Commun Stat-Simul Computat* 1998; **27**: 485–500.
11. Leung DHY and Wang YG. Bias reduction using stochastic approximation. *Australian & New Zealand J Stat* 1998; **40**: 43–52.
12. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**: 27–38.
13. Heinze G and Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409–2419.
14. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med* 2006; **25**: 4216–4226.
15. Maiti T and Pradhan V. A comparative study of the bias corrected estimates in logistic regression. *Stat Meth Med Res* 2008; **17**: 621–634.
16. Bull SB, Mak C and Greenwood CM. A modified score function estimator for multinomial logistic regression in small samples. *Computat Stat Data Analysis* 2002; **39**: 57–74.
17. Kosmidis I. Improved estimation in cumulative link models. *J Royal Stat Soc: Ser B (Statistical Methodology)* 2014; **76**: 169–196.
18. Moons K, Donders AR, Steyerberg E, et al. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* 2004; **57**: 1262–1270.
19. Bull SB, Lewinger JP and Lee SS. Confidence intervals for multinomial logistic regression in sparse data. *Stat Med* 2007; **26**: 903–918.
20. Terrell GR. The gradient statistic. *Comput Sci Stat* 2002; **34**: 206–215.
21. Muggeo VMR and Lovison G. The 'three plus one' likelihood-based test statistics: unified geometrical and graphical interpretations. *Am Statistician* 2014; **68**: 302–306.
22. Firth D. Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. In: Dodge Y and Whittaker J (eds) *Computational statistics*, vol 1. Neuchatel Switzerland, Springer, 1992, pp.553–557.
23. Jeffrey H. An invariant form for the prior probability in estimation problems. *Proc Royal Soc London Series A, Math Phys Sci* 1946; **186**: 453–461.
24. Chen M, Ibrahim J and Kim S. Properties and implementation of Jeffreys's prior in binomial regression models. *J Am Stat Assoc* 2008; **103**: 1659–1664.
25. Royall RM. Model robust confidence intervals using maximum likelihood estimators. *Int Stat Rev/Revue Internationale de Statistique* 1986; **54**: 221–226.
26. Jin Z, Shao Y and Ying Z. A monte carlo method for variance estimation for estimators based on induced smoothing. *Biostatistics* 2015; **16**: 179–188.
27. Greenland S and Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med* 2015; **34**: 3133–3143.
28. Lemonte AJ. On the gradient statistic under model misspecification. *Stat Probabil Lett* 2013; **83**: 390–398.
29. Potter DM. A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Stat Med* 2005; **24**: 693–708.
30. Mehta CR and Patel NR. Exact logistic regression: theory and examples. *Stat Med* 1995; **14**: 2143–2160.
31. Foxman B, Marsh J, Gillespie B, et al. Condom use and first-time urinary tract infection. *Epidemiology* 1997; **8**: 637–641.
32. Corcoran C, Mehta C, Patel N, et al. Computational tools for exact conditional logistic regression. *Stat Med* 2001; **20**: 2723–2739.