Penalized logistic regression for small or sparse data: interval estimators revisited

Marianna Siino[1], Salvatore Fasola[2], Vito M.R. Muggeo[1]

[1] Dip. Scienze Economiche, Aziendali e Statistiche, University of Palermo, ITALY

E-mail for correspondence: marianna.siino01@unipa.it

Abstract: This paper focuses on inferences in logistic regression models fitted by the Firth penalized log likelihood. In this context, many authors have claimed superiority of the likelihood ratio statistic with respect to the (wrong) Wald statistic via simulation evidence. We re-assess such findings by detailing the inferential tool and including in the comparisons the (right) Wald statistic and also other statistics neglected in previous literature. Simulation evidence and a real data set analysis withdraw previous findings by showing that the likelihood ratio statistic is not the best inferential device in Firth penalized logistic regression.

Keywords: Penalized likelihood; Logistic regression; Sandwich formula; Score-based CIs; Gradient-based CIs.

1 Introduction

The logistic regression equation reads as $\text{logit}(\pi_i) = \sum_{j=1}^{K} x_{ij}\beta_j$ where $E[Y|x_i] = \pi_i$, Y is the dichotomic response variable and x_i a K-dimensional covariate vector. To estimate the regression parameter β_j, Firth (1993) suggested to modify the classical score function $U_j(\boldsymbol{\beta})$ through $U_j^*(\boldsymbol{\beta}) = U_j(\boldsymbol{\beta}) + 0.5\text{tr}\left\{I(\boldsymbol{\beta})^{-1}\partial I(\boldsymbol{\beta})/\partial\beta_j\right\}$, $j = 1, 2, \ldots, K.$, corresponding to the penalized log-likelihood $\ell^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \log|I(\boldsymbol{\beta})|^{\frac{1}{2}}$ being the penalty $|I(\boldsymbol{\beta})|^{\frac{1}{2}}$ the so-called Jeffrey's invariant prior. Since the Firth penalized approach allows to remove the first order $O(n^{-1})$ bias of the MLEs, and it also guarantees finite estimates with sparse data where the classical ML estimates do not exist, such penalized approach is widespread in practice, especially in medical statistics involving small samples.

We focus on the construction of confidence intervals for the β_js, based on test statistics computed using penalized likelihood quantities. Existing approaches previously discussed consider CIs based on the penalized likelihood ratio and the Wald statistics (e.g., Heinze and Schemper, 2002; Bull et al., 2007). However we note two possible drawbacks in previous aforementioned studies: first they consider a meaningless and then wrong Wald statistic; second the other statistics - such as the well-known Score and the relatively more recent Gradient statistics are totally ignored. In the next

sections we revisit and compare them in terms of coverage levels for the interval estimators

2 Methods

Let β_j be the interest parameter in the logit regression equation. A relevant $(1-\alpha)100\%$ confidence interval is defined as $\{\beta_{0j} \in \mathbb{R} : z_{\frac{\alpha}{2}} \leq T(\beta_{0j}) \leq z_{1-\frac{\alpha}{2}}\}$, where $T(\beta_{0j})$ is any pivot statistic discussed nextly, and $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ are the appropriate quantiles of the standard normal distribution. The Likelihood ratio and the Wald statistics currently discussed in the literature are

$$L = \text{sign}(\hat{\beta}_j^* - \beta_{0j})\sqrt{-2\{\ell^*(\hat{\boldsymbol{\beta}}^*) - \ell^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)\}} \qquad W = \frac{\hat{\beta}_j^* - \beta_{0j}}{\sqrt{\text{I}^{-1}(\hat{\boldsymbol{\beta}}^*)_{jj}}}$$

where $\hat{\boldsymbol{\beta}}^*$ is the full (unrestricted) penalized ML estimate, and $\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*$ is the restricted penalized ML estimates, that is the penalized estimates obtained fixing β_j at β_{0j}. The penalized likelihood ratio statistic is quite straightforward, but the form of the Wald statistic deserves some discussion: it uses the jth element of the main diagonal of the (unpenalized) inverse information evaluated at the penalized ML estimate. Notice the Information, that is the variance of U^*, does not depend on the penalty and thus we write it as I, without asterisk. Now the question is if such quantity - employed in the aforementioned literature - represents the right formula to compute the variance of estimator $V[\hat{\boldsymbol{\beta}}^*]$. The answer is no. From basics of Inference, it should be remarked that variance of the ML estimator comes from the sandwich formula reducing to inverse of Information *only if the model is correctly specified* and the second Bartlett identity holds. In the Firth penalized likelihood, clearly the second Bartlett identity does not hold, namely $E[-H^*(\boldsymbol{\beta})] \neq \text{I}(\boldsymbol{\beta})$, where H^* is the hessian depending on the penalty. Thus the usual sandwich formula cannot be simplified, but a reliable variance for $\hat{\boldsymbol{\beta}}^*$ is provided by the sandwich estimate $V(\hat{\boldsymbol{\beta}}^*) \approx H^*(\hat{\boldsymbol{\beta}}^*)]^{-1}\text{I}(\hat{\boldsymbol{\beta}}^*)H^*(\hat{\boldsymbol{\beta}}^*)^{-1}$. For large samples the penalty effect vanishes, making valid the simple approximation $V(\hat{\boldsymbol{\beta}}^*) \approx \text{I}(\hat{\boldsymbol{\beta}}^*)^{-1}$; however in small to moderate samples $\text{I}(\hat{\boldsymbol{\beta}}^*)^{-1}$ typically *overestimates* the variance of $\hat{\boldsymbol{\beta}}^*$. This crucial issue appears to have been overlooked in literature, causing a (pointless) 'bad reputation' of the Wald statistic.
Likelihood ratio and Wald represent two, possibly the most famous, likelihood based statistics useful for interval estimators, but other options are available; see Muggeo and Lovison (2014) for a nice treatment about all four likelihood-based statistics. To complete our discussion about interval estimators in Firth penalized logistic regression, we write down 2 additional statistics, the Score and the Gradient statistic,

$$S = U_j^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)\sqrt{\text{I}^{-1}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)_{jj}}, \qquad G = \text{sign}(\hat{\beta}_j^* - \beta_{0j})\sqrt{(\hat{\beta}_j^* - \beta_{0j})U_j^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)}.$$

where $\mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^{*})_{jj}$ is the inverse of the variance of the conditional Score U_j^* given the remaining components. Interestingly the Score statistic is well known in the mainstream inference background, but its use appears to be quite limited in practical application. The Gradient statistic is relatively new, but takes advantage of its computational simplicity as it just needs estimates and first derivatives.

3 Simulation study

To compare the four likelihood based statistics, we generate Bernoulli data $Y_i \sim Ber(\pi_i)$ where $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$. We set $\beta_0 = 1$ while $\beta_1 \in \{0.5, 1.5\}$, and three different sample sizes $n \in \{20, 50, 100\}$. For each scenario we consider a balanced binary variable $x_i = I(i > n/2)$ and a continuous covariate with equally spaced values, i.e. $x_i = i/n$.

TABLE 1. Simulation results (based in 1000 runs): empirical coverage level (CL) and average width (AW) of the CIs based on the incorrect Wald (W), the correct Wald with the sandwich variance (W_S), Likelihood Ratio (L), Score (S) and Gradient (G) statistics.

		$x_i = i/n$				$x_i = I(i > n/2)$			
		$\beta_1 = 0.5$		$\beta_1 = 1.5$		$\beta_1 = 0.5$		$\beta_1 = 1.5$	
n	Test	CL	AW	CL	AW	CL	AW	CL	AW
20	W	0.991	4.69	0.984	5.52	0.994	7.25	0.991	8.00
	W_S	0.970	4.09	0.934	4.43	0.982	6.70	0.981	7.16
	L	0.957	4.87	0.973	6.05	0.968	7.40	0.981	8.40
	S	0.951	3.83	0.934	4.31	0.968	6.40	0.972	7.20
	G	0.957	5.59	0.972	7.55	0.962	7.97	0.977	9.29
50	W	0.968	2.81	0.970	3.65	0.968	4.73	0.971	5.63
	W_S	0.958	2.66	0.957	3.15	0.962	4.55	0.943	5.22
	L	0.959	2.83	0.965	3.87	0.959	4.75	0.950	5.80
	S	0.961	2.58	0.947	3.19	0.955	4.44	0.953	5.38
	G	0.950	2.93	0.967	4.41	0.955	4.86	0.947	6.05
100	W	0.955	1.94	0.959	2.59	0.965	3.31	0.966	3.96
	W_S	0.946	1.90	0.938	2.36	0.959	3.25	0.951	3.80
	L	0.946	1.94	0.940	2.67	0.954	3.31	0.954	4.01
	S	0.948	1.86	0.945	2.40	0.951	3.19	0.957	3.85
	G	0.943	1.96	0.937	2.84	0.953	3.34	0.950	4.08

Table 1 reports the empirical coverage level (CL) and the relevant average width (AW) of the 95% CIs for β_1. Broadly speaking the likelihood ratio statistic does not perform the best, and the *fair* W_S behaviour reasonably well, even better than L for $n \geq 50$. Overall, score-based CIs appear to outperform the competitors in terms of CL and AW, also in the most difficult

scenarios with small samples ($n = 20$) and strong predictor causing sampling zeroes and sparse data ($\beta_1 = 1.5$). As expected, differences attenuate at large samples.

4 Example: osteogenic sarcoma data

We consider an example from Metha and Patel (1995) on $n = 46$ patients with osteogenic sarcoma. A three year disease-free interval (DFI3) is the response, while the explanatory variables are gender (SEX) and the presence of any osteoid pathology (AOP) and lymphocytic infiltration (LI). Notice the classical MLEs cannot be computed since there is problem of separation caused by the variable LI. We estimate a penalized logistic regression model with additive linear effects and compute the 95% confidence intervals for β_{LI}, see Table 2. It is worth emphasizing the variable LI turns out to be significant only according to W_S, L, S and G but the 95% CI based on S is the narrowest.

TABLE 2. 95% confidence intervals for β_{LI} based on the five pivot statistics and exact inference (E).

95%CI	W	W_S	L	S	G	E
Inf	-5.504	-4.637	-7.363	-4.805	-10.147	$-\infty$
Sup	0.582	-0.286	-0.188	-0.104	-0.356	0.160

References

Bull, S.B., Lewinger, J.P., Lee, S.S.F. (2007). Confidence intervals for multinomial logistic regression in sparse data, **26**, 903-918.

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

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

Muggeo V., Lovison G. (2014). The 'three plus one' likelihood-based test statistics: unified geometrical and graphical interpretations. *The American Statistician*, **68**, 302–306.