Bias-reduction methods for Poisson regression

Luca Presicce [1] Tommaso Rigon [1] Emanuele Aliverti [2]

[1]Department of Economics, Management and Statistics, University of Milano - Bicocca
[2]Department of Statistics, University of Padova

CONTACT ME

1. Introduction

We address the Poisson GLMs issues by considering a penalty for the likelihood, obtaining the estimates by maximizing a genuine likelihood in which pseudo-counts substitute the response.

👉 GLMs have limitations, one above all is the presence of bias in the MLE. A variety of methods have been proposed in the literature, albeit theoretically appealing, these methods may face bottlenecks.

👉 In Rigon and Aliverti (2022) a fast alternative is introduced for the logistic regression, resulting in an approximation of Firth (1993)'s method, a milestone in the bias-reduction literature.

👉 We present a bias-reduction method for Poisson regression models that brings computational advantages. Penalizing the likelihood with the Diaconis-Ylvisaker (DY) conjugate prior of Diaconis and Ylvisaker (1979).

2. Generalized linear model - Poisson regression

👉 We will focus on count response data. Where is common to assume that the response variables $Y_1, \dots, Y_n$ are such that

$$Y_i \overset{\text{ind}}{\sim} \text{Pois}(\lambda_i), \tag{1}$$

$$g(\lambda_i) = x_i^\intercal \beta = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}, \tag{2}$$

independently for $i = 1, \dots, n$. The vector $x_i = (x_{i,1}, \dots, x_{i,p})^\intercal$ contains the covariate information of the ith observation, whereas the p-dimensional vector $\beta = (\beta_1, \dots, \beta_p)^\intercal$ corresponds to the regression coefficients.

👉 The function $g(\cdot)$ is the *link function*, which here is assumed to be $g(\cdot) = \log(\cdot)$, that is, the so-called canonical link for Poisson models McCullagh and Nelder (1989).

👉 The demand for bias-reduction methodologies in Poisson models descends from their broad range of applications. Moreover, the growing complexity of the contexts in which these models are employed justifies the necessity of faster alternatives.

3. Diaconis and Ylvisaker conjugate priors

👉 Poisson GLM likelihood: Let $y = (y_1, \dots, y_n)^\intercal$ be realizations of independent Poisson random variables with means $\lambda_1, \dots, \lambda_n$, as in (1). Thus, the likelihood function assumes the following form

$$\mathscr{L}(\beta; y) \propto \exp\left[\sum_{i=1}^{n} \{y_i x_i^\intercal \beta - \exp(x_i^\intercal \beta)\}\right]. \tag{3}$$

👉 Conjugate DY prior distribution: Following Chen and Ibrahim (2003), we consider the DY conjugate prior for the regression coefficients β, with density proportional to

$$\pi(\beta) \propto \exp\left[\tau \sum_{i=1}^{n} \{\psi_i x_i^\intercal \beta - \exp(x_i^\intercal \beta)\}\right], \tag{4}$$

where $\tau > 0$ and $\psi = (\psi_1, \dots, \psi_n)^\intercal$ are prior hyperparameters such that $\psi_i > 0$, for $i = 1, \dots, n$. The hyperparameters: τ controls the variability and the strength of the prior belief on ψ; and ψ is the mode of the prior distribution, thus defining scale and location parameters for $\pi(\beta)$.

👉 Penalized posterior distribution: The posterior distribution can be obtained combining prior information (4) with the likelihood function (3) as $\pi(\beta \mid y) \propto \pi(\beta)\mathscr{L}(\beta; y)$, leading to

$$\pi(\beta \mid y) \propto \exp\left[(\tau+1)\left\{\sum_{i=i}^{n}\left(\frac{\tau\psi_i + y_i}{\tau+1}\right) x_i^\intercal \beta - \exp(x_i^\intercal \beta)\right\}\right]. \tag{5}$$

Since the posterior distribution is in the same family as the prior, we have conjugacy, i.e. $\pi(\beta \mid y)$ is again in the DY family with updated parameters.

👉 Pseudo-counts: The prior parameters ψ are updated, obtaining the pseudo-counts $\tilde{y}_i = (y_i + \tau\psi_i)/(\tau+1)$ for $i = 1, \dots, n$. While τ is updated as simply $(\tau+1)$.

👉 Penalized genuine likelihood: The penalized likelihood is equivalent to the original one minus multiplicative constant terms depending only on the hyperparameter τ, that is

$$\pi(\beta \mid y) \propto \exp\{(\tau+1)\ell(\beta; \tilde{y})\}, \tag{6}$$

where $\ell(\beta; \tilde{y})$ is the log-likelihood and $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)^\intercal$.

👉 MAP estimates as bias-reducted MLE: From (6), the maximum a posteriori (MAP) of $\pi(\beta \mid y)$ corresponds to the MLE obtained from $\ell(\beta; \tilde{y})$.

4. Penalized score equations

👉 Comparison between score equations: the score function is defined as $U(\beta) = \partial\ell(\beta)/\partial\beta$, and MLE estimates are obtained as the solution of the system $U_r(\beta) = 0$, for $r = 1, \dots, p$. A penalized likelihood leads to penalized score equations:

$$U_r(\beta) = \sum_{i=1}^{n}(y_i - \mu_i)x_{i,r}\,, \quad U_{r,\text{DY}}(\beta) = C(\tau)\sum_{i=1}^{n}(\tilde{y}_i - \mu_i)x_{i,r}\,, \quad U_{r,\text{FI}}(\beta) = \sum_{i=1}^{n}(y_i^* - \mu_i)x_{i,r}, \tag{7}$$

with $r = 1, \dots, p$ and $\mu_i = \exp(x_i^\intercal\beta)$.

👉 Comparison between pseudo-counts: The last two score equations in (7) provide a penalized system of score equations depending on a different set of pseudo-counts, respectively defined as

$$\tilde{y}_i = \frac{y_i + \tau\psi_i}{\tau+1}\,, \quad y_i^* = y_i + \frac{h_{i,i}}{2}\,, \quad i = 1, \dots, n. \tag{8}$$

Where y^* depend on the diagonal elements of the leverage matrix $H(\beta)$, while $\tilde{y}$ depends only on the prior hyperparameters τ and ψ. Firth (1993) removes the first-order term of the asymptotic bias, and we fix the values of τ and ψ introducing similarity between $\tilde{y}$ and y^*.

👉 Mean approximation for faster computation: One could approximate each $h_{i,i}$ for $i = 1, \dots, n$ with their mean, obtained as the ratio between the trace of $H(\beta)$ and the dimension of its diagonal n, that is p/n.
We call this computational expedient *mean approximation*, which avoids the burdens due to the computation of $H(\beta)$ at each iteration of the optimization, retaining the bias-reduction.

👉 Proposed hyperparameters choice: We allow $\tau = \tau(\alpha)$ and each $\psi_i = \psi_i(\alpha)$ to depend on a fixed constant $\alpha > 0$, chosen so that $\tau(\alpha)\psi_i(\alpha) = p/(2n)$, obtaining

$$\tilde{y}_i(\alpha) = \frac{y_i + \tau(\alpha)\psi_i(\alpha)}{\tau(\alpha)+1} = \frac{y_i + p/(2n)}{\tau(\alpha)+1}, \qquad i = 1, \dots, n. \tag{9}$$

For example, one could set $\tau(\alpha) = p/(\alpha n)$ and $\psi_i(\alpha) = \alpha/2$ for $i = 1, \dots, n$.

5. Illustration

👉 Dataset & model: We consider the *infert* data, on $n = 248$ women in a case-control study, where the number of spontaneous aborts is considered as a count outcome. The models consider a regression of the response variable on an intercept and six covariates, leading to $p = 7$.

👉 Simulation study: We simulate **10000** datasets from a Poisson regression model with parameter $\hat{\beta}_{\text{MLE}}$, allowing us to compare the performances in terms of bias and root mean squared error (RMSE).

		β_1	β_2	β_3	β_4	β_5	β_6	β_7
BIAS	$\hat{\beta}_{\text{MLE}}$	-0.20	0.21	0.21	0	0.00	-0.01	0.00
	$\hat{\beta}_{\text{DY}}$	-0.09	0.09	0.09	0	-0.01	0.02	-0.02
	$\hat{\beta}_{\text{FI}}$	0.00	0.00	0.00	0	0.00	0.00	0.00
RMSE	$\hat{\beta}_{\text{MLE}}$	1.57	1.49	1.49	0.02	0.07	0.16	0.17
	$\hat{\beta}_{\text{DY}}$	0.78	0.60	0.60	0.02	0.07	0.15	0.17
	$\hat{\beta}_{\text{FI}}$	0.73	0.53	0.53	0.02	0.07	0.16	0.17

Table: Simulation study results on the regression coefficients estimates obtained for spontaneous abort counts problem from the three models considered.

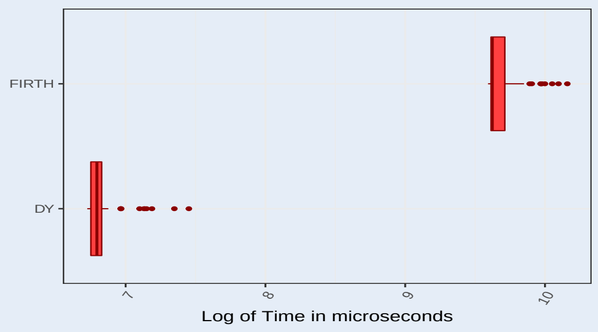

Figure: Timing boxplot of the estimation processes for the DY penalized and the Firth's models.

👉 Computational competition: We investigate the computational timing of the estimation process for both our and Firth (1993) model, considering **100** replications.

6. Discussion

👉 $\hat{\beta}_{\text{MLE}}$ performs significantly worse than all the reduced-bias methodologies in terms of bias and RMSE.

👉 Empirical similarity in terms of bias between the penalized estimator $\hat{\beta}_{\text{DY}}$ and $\hat{\beta}_{\text{FI}}$.

👉 Proposed estimation process takes $\approx$ 19 times less than the compared one, retaining comparable performance.

7. References

M.-H. Chen and J. G. Ibrahim, "Conjugate priors for generalized linear models," *Statist. Sinica*, vol. 13, no. 2, pp. 461–476, 2003.

P. Diaconis and D. Ylvisaker, "Conjugate Priors for Exponential Families," *Ann. Statist.*, vol. 7, no. 2, pp. 269 – 281, 1979.

D. Firth, "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, vol. 80, no. 1, pp. 27–38, 1993.

P. McCullagh and J. A. Nelder, *Generalized Linear Models*. London: Chapman & Hall / CRC, 1989.

T. Rigon and E. Aliverti, "Conjugate priors and bias reduction for logistic regression models," *arXiv*, 2022.