Maximum softly-penalized likelihood for mixed effects logistic regression

Philipp Sterzinger

`phillip.sterzinger@warwick.ac.uk`
and
Ioannis Kosmidis
`ioannis.kosmidis@warwick.ac.uk`

Department of Statistics, University of Warwick
Coventry CV4 7AL, UK

June 2, 2022

S1 Supplementary Material

All labels for the sections, equations, tables, figures and so on in the current document have been prefixed by "S" (e.g. Section S1, equation (S1), Table S3, etc). The supplementary material for *Maximum softly-penalized likelihood for mixed effects logistic regression* contains:

i) ML, MSPL and `bglmer` (Chung et al., 2013) estimates for the full Culcita data set of McKeon et al. (2012) pertaining to Section 3 of the main text (Section S2)

ii) A summary of the simulation study of Section 8 of the main paper (Section S3), and

iii) Further simulations on synthetic data (Section S4)

This document and the R scripts and datasets to reproduce our results are available at `https://github.com/psterzinger/softpen_supplementary`.

S2 Culcita data

Table S1 MSPL, ML and `bglmer` estimates from the full Culcita dataset of (McKeon et al., 2012), using "none" as reference category. Asymptotic standard errors based on the negative Hessian of the approximate log-likelihood are given in parentheses

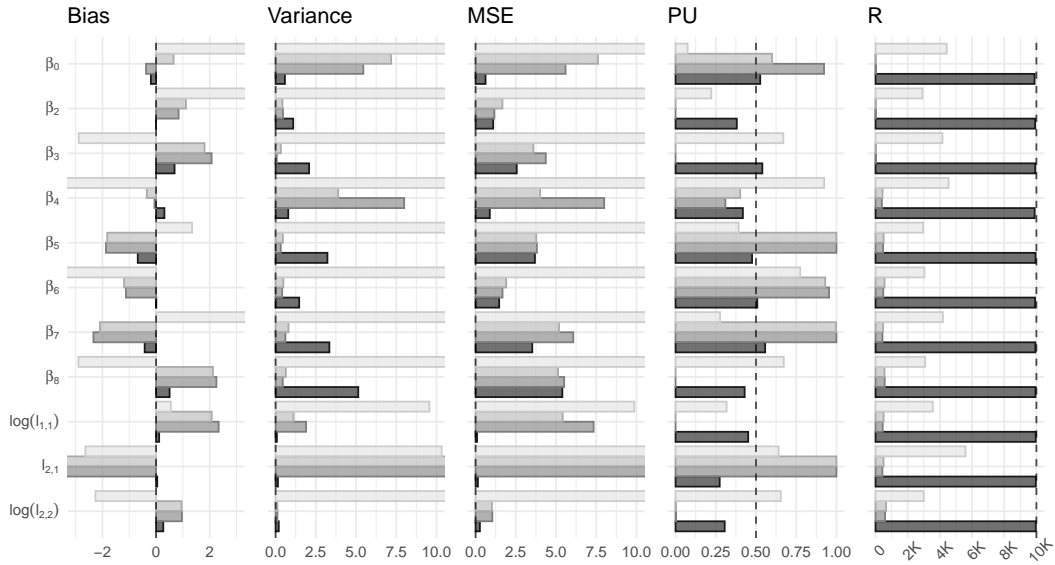|	ML[BFGS]	ML[CG]	bglmer[t]	bglmer[n]	MSPL
β_0	5.01	5.01	3.57	3.55	4.23
	(1.80)	(1.80)	(1.53)	(1.52)	(1.53)
β_2	-3.75	-3.75	-2.21	-2.19	-3.40
	(1.46)	(1.46)	(1.11)	(1.11)	(1.35)
β_3	-4.36	-4.36	-2.68	-2.66	-3.93
	(1.55)	(1.55)	(1.16)	(1.15)	(1.42)
β_4	-5.55	-5.55	-3.65	-3.62	-4.95
	(1.72)	(1.72)	(1.27)	(1.27)	(1.56)
$\log \sigma$	1.26	1.26	1.24	1.24	1.13
	(0.37)	(0.37)	(0.39)	(0.39)	(0.36)

S3 Conditional inference data

Fig. S1 Full simulation results from the simulation study in Section 8 of the main text

Table S2 Percentiles of centred parameter estimates from the simulation study in Section 8 of the main text

				Percentiles				
		5%	10%	25%	50%	75%	90%	95%
MSPL	β_0	-2.17	-2.03	-0.01	-0.00	0.02	0.13	0.27
	β_2	-2.18	-2.17	-0.00	0.00	0.00	2.17	2.18
	β_3	-1.07	-0.87	-0.65	-0.00	2.17	2.18	2.18
	β_4	-0.60	-0.46	-0.21	0.11	0.52	1.88	2.31
	β_5	-4.35	-2.18	-2.17	0.00	0.66	1.08	2.18
	β_6	-2.30	-1.56	-0.51	-0.00	0.52	1.62	2.30
	β_7	-3.27	-2.73	-1.84	-0.23	0.78	1.68	2.82
	β_8	-3.24	-2.35	-0.96	0.41	2.05	3.39	4.31
	$\log l_{1,1}$	-0.06	-0.05	-0.02	0.01	0.12	0.44	0.66
	$l_{2,1}$	-0.66	-0.27	-0.02	0.12	0.21	0.35	0.45
	$\log l_{2,2}$	-0.42	-0.31	-0.08	0.26	0.58	0.87	1.02
bglmer[t]	β_0	-0.88	-0.86	-0.80	-0.50	0.06	3.26	4.33
	β_2	0.17	0.22	0.46	1.31	1.61	1.85	1.88
	β_3	1.17	1.20	1.41	1.74	2.09	2.57	2.73
	β_4	-4.37	-3.96	-0.65	0.24	1.01	1.51	1.79
	β_5	-3.11	-2.79	-2.17	-1.72	-1.37	-1.09	-0.87
	β_6	-2.15	-1.99	-1.69	-1.31	-0.78	-0.19	0.05
	β_7	-3.59	-3.19	-2.68	-2.09	-1.50	-0.97	-0.72
	β_8	0.92	1.18	1.55	2.07	2.58	3.07	3.46
	$\log l_{1,1}$	1.05	1.19	1.40	1.80	2.04	4.23	4.60
	$l_{2,1}$	-39.24	-34.03	-3.87	-2.86	-1.63	-1.06	-0.84
	$\log l_{2,2}$	0.43	0.55	0.73	0.94	1.16	1.36	1.51
bglmer[n]	β_0	-2.20	-1.71	-1.29	-0.72	-0.24	-0.22	2.65
	β_2	0.22	0.29	0.51	0.87	1.19	1.37	1.44
	β_3	1.77	1.81	1.88	2.03	2.19	2.41	2.51
	β_4	-5.50	-4.93	-1.17	1.04	1.99	2.51	2.69
	β_5	-2.82	-2.51	-2.20	-1.82	-1.49	-1.19	-1.01
	β_6	-2.12	-1.91	-1.59	-1.12	-0.69	-0.30	-0.05
	β_7	-3.65	-3.26	-2.83	-2.31	-1.84	-1.40	-1.09
	β_8	1.23	1.44	1.80	2.25	2.66	3.12	3.31
	$\log l_{1,1}$	0.95	1.04	1.21	1.91	3.81	4.55	4.92
	$l_{2,1}$	-38.54	-34.65	-4.03	-2.73	-1.29	-0.92	-0.78
	$\log l_{2,2}$	0.44	0.58	0.74	0.95	1.16	1.36	1.52
ML	β_0	-1.11	1.58	5.84	9.60	11.76	13.58	14.87
	β_2	-8.42	-7.29	0.76	4.96	7.59	10.71	12.75
	β_3	-10.06	-8.74	-7.04	-5.40	2.16	5.85	9.43
	β_4	-14.59	-13.43	-11.64	-9.51	-5.65	-1.42	1.28
	β_5	-10.07	-6.58	-1.34	0.74	2.89	8.63	16.11
	β_6	-12.92	-10.61	-7.41	-4.88	-1.04	7.29	8.40
	β_7	-6.77	-5.02	-0.91	5.61	7.60	10.38	13.80
	β_8	-19.81	-11.80	-5.30	-1.21	0.80	3.99	8.77
	$\log l_{1,1}$	-2.75	-2.05	-0.91	1.21	2.36	2.62	2.66
	$l_{2,1}$	-7.47	-7.19	-5.74	-1.51	0.54	0.81	1.09
	$\log l_{2,2}$	-6.62	-4.54	-3.22	-1.77	0.39	0.85	1.04

S4   Further Simulations

S4.1   Simulation 1: Extreme fixed effects

This section presents a simulation that seeks to provoke degenerate fixed effects estimates through a strong dependence of the responses on the particular fixed effects – a phenomenon known to occur in standard logistic regression (Albert and Anderson, 1984; Kosmidis and Firth, 2021).

For this, we simulate from a mixed effects logistic regression with univariate random effects and logistic link function as follows. For five clusters $i = 1, \ldots, 5$ and within cluster observations $j = 1, \ldots, n$, $n \in \{50, 100, 200\}$, we draw an i.i.d. vector of fixed effects covariates $\boldsymbol{x}_{ij} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ where $X_{i1} = 1, X_{i2} \sim \mathrm{N}(0,1), X_{i3} \sim \mathrm{Ber}\left(\frac{1}{2}\right), X_{i4} \sim \mathrm{Ber}\left(\frac{1}{4}\right)$, and $X_{i5} \sim \exp(1)$. The fixed effect covariates are drawn once and held fixed over the simulation. To control the degree of dependence of the responses on a particular fixed effect covariate, the parameter of fixed effects is set as $\boldsymbol{\beta} = (1, -0.5, \lambda, 0.25, -1)$, where λ takes integer values from -10 to 10. For each specification of n, λ, we draw 100 samples from the model

$$
\begin{aligned}
Y_{ij} \mid u_i &\sim \mathrm{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i, \\
u_i &\sim \mathrm{N}(0,9) \quad (i = 1, \ldots, 5; j = 1, \ldots, n), \quad n \in \{50, 100, 200\}
\end{aligned}
\tag{S1}
$$

The random effects dispersion parameter $\sigma = 3$ was chosen as to avoid estimation issues associated with small random effects. We estimate the parameters using our proposed MSPL with the penalties given in Section 5 of the main text, ML and `bglmer` from the `blme` R (R Core Team, 2022) package (Chung et al., 2013) with a normal and t prior for the fixed effects and a gamma prior for the random effects variance. We approximate the log-likelihood with a 20-point adaptive Gauss-Hermite quadrature approximation. For ML and MSPL, we optimize the approximate log-likelihood using the optimization methods "CG", "BFGS", "nlminb" and "L-BFGS-B" from the `optimx` R package (Nash, 2014) and report the best fit. Both `bglmer` specifications use the default `bglmer` optimization settings.

Table S3 shows the number of estimates per specification which resulted in an degenerate parameter estimate. We considered an estimate degenerate, if it is larger than 50 in absolute value, the gradient is larger than 0.001 in absolute value, or if the estimated asymptotic standard errors are larger than 40. Figure S2 shows the dispersion of the estimates β_3 around the true value (indicated by dashed horizontal line) per specification for all estimation methods. For presentability, the y-axis has been cropped to omit overly extreme estimates. For the ML, 672 estimates are cropped, while for all other methods, all estimates are shown. Note that the boxplots for the ML do not depict the empirical distribution of the ML estimate of β_3 over different samples as theoretically infinite estimates assume finite values in estimation due to numerical precision limitations and resulting premature declaration of convergence. We see from Table S3 that the MSPL gives the most stable parameter estimation, with one out of 6000 samples exhibiting estimation issues due to failed convergence. The ML on the other hand becomes highly problematic for large values of β_3 and even returns degenerate estimates for moderate values of β_3 with non-negligible frequency. The `bglmer` estimates, even though they penalize fixed effects more harshly, as can be seen by the shrinkage-induced bias of the `bglmer` estimates of β_3 in Figure S2, they encounter estimation issues substantially more often than the MSPL estimates.
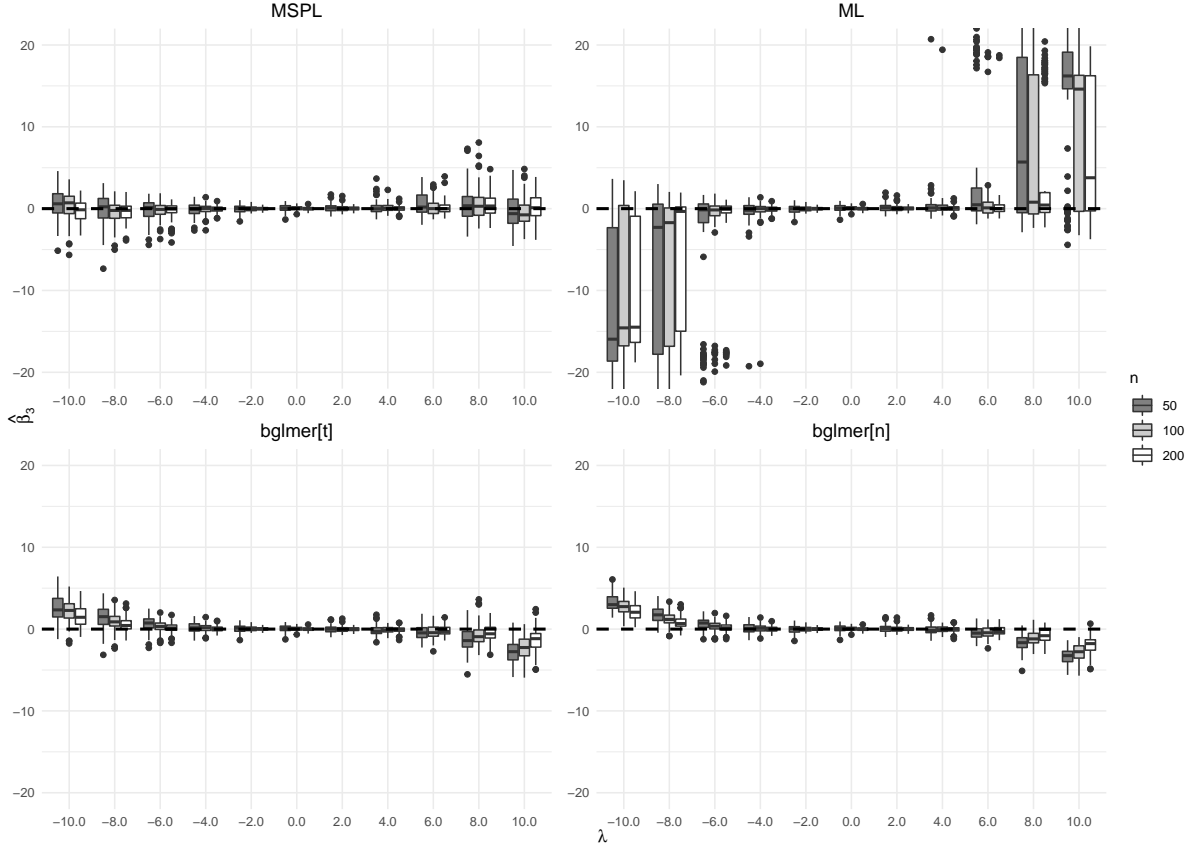
Fig. S2 Centered estimation output of β_3 from Simulation 1

Table S3 Percentage of degenerate estimates from Simulation 1

		λ										
		-10	-8	-6	-4	-2	0	2	4	6	8	10
MSPL	n=50	0	0	0	0	0	0	0	0	0	0	0
	n=100	0	0	0	0	0	0	0	0	0	**1**	0
	n=200	0	0	0	0	0	0	0	0	0	0	0
ML	n=50	**73**	**46**	**17**	**2**	0	0	0	**1**	**15**	**49**	**80**
	n=100	**63**	**45**	**10**	**1**	0	0	0	**1**	**5**	**31**	**61**
	n=200	**65**	**26**	**8**	0	0	0	0	0	**2**	**24**	**49**
bglmer[t]	n=50	0	0	0	0	0	0	0	0	0	0	0
	n=100	0	0	0	**1**	0	0	0	0	0	0	0
	n=200	0	0	**1**	0	**1**	0	0	0	0	0	**1**
bglmer[n]	n=50	**1**	0	0	0	0	**1**	0	0	0	0	0
	n=100	0	0	**1**	0	**1**	0	0	0	**1**	0	0
	n=200	0	0	0	0	0	0	0	0	0	0	0

S4.2 Simulation 2: Extreme random effects variance

In this simulation, we seek to provoke degenerate random effects variance estimates, that is random effects variance estimates that are either zero or infinite. One of the peculiarities of

Bernoulli-response (or Binomial-response) mixed effects models is that there can be separation of the observations with respect to the random effects covariates. Analogously to separation in logistic regression models (Albert and Anderson, 1984), where covariate constellations such that the responses can be separated by a hyperplane spanned by the covariate column vectors, lead to infinite parameter estimates, it is known that certain constellations of random effects covariates can lead to data separation and consequently degenerate random effects estimates (see for example Sauter and Held (2016) or the discussion on `https://stats.stackexchange.com/questions/44755`). We consider a simple simulation to provoke such data configurations by simulating from a mixed effects logistic regression with univariate random effects and vary the dependence of the responses on the grouping variable by controlling the random effects variance parameter.

For five clusters $i = 1, \ldots, 5$ and within cluster observations $j = 1, \ldots, n$, $n \in \{50, 100, 200\}$, we draw an i.i.d. vector of fixed effects covariates $\boldsymbol{x}_{ij} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ where $X_{i1} = 1$, $X_{i2} \sim \mathrm{N}(0, 1)$, $X_{i3} \sim \mathrm{Ber}\left(\frac{1}{2}\right)$, $X_{i4} \sim \mathrm{Ber}\left(\frac{1}{4}\right)$, and $X_{i5} \sim \exp(1)$. The fixed effect covariates are drawn once and held fixed over the simulation. Likewise, the fixed effects $\boldsymbol{\beta} = (1, -0.5, 0.5, 0.25, -1)$ are held fixed over the simulation, while $\lambda = \log \sigma$ is varied over the integer values from -5 to 2. For each specification of n, λ, we draw 100 samples from the model

$$
\begin{aligned}
Y_{ij} \mid u_i &\sim \mathrm{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i, \\
u_i &\sim \mathrm{N}(0, \exp(\lambda)^2) \quad (i = 1, \ldots, 5; j = 1, \ldots, n), \quad n \in \{50, 100, 200\}
\end{aligned}
\tag{S2}
$$

We estimate the parameters using our proposed MSPL with the penalties given in Section 6 of the main text, ML and `bglmer` from the `blme` R package (Chung et al., 2013) with a normal and t prior for the fixed effects and a gamma prior for the random effects variance. We approximate the log-likelihood with a 20-point adaptive Gauss-Hermite quadrature approximation. For ML and MSPL, we optimize the approximate log-likelihood using the optimization methods "CG", "BFGS", "nlminb" and "L-BFGS-B" from the `optimx` R package (Nash, 2014) and report the best fit. Both `bglmer` specifications use the default `bglmer` optimization settings. Figure S3 shows the dispersion of the estimates for $\log \sigma$ around the true value (indicated by dashed horizontal line), for each estimation method and specification of λ, n. For the ML and `bglmer` estimates, these boxplots do not approximate the distribution of the maximum likelihood estimator as, owed to numerical precision limitations, parameter estimates which ought to be infinite or zero are not estimated as such so that the point masses at the boundaries of the parameter space are missing. For bglmer[t], 6 estimates are not shown and for bglmer[n], 4 estimates are not shown due to failed estimation. While both the MSPL and `bglmer` shrink negative estimates of $\log \sigma$ towards zero, the shrinkage induced by `bglmer` is considerably stronger, as can be seen by the absolute amount of shrinkage and the smaller dispersion of the estimates. Moreover, we see that for larger values of $\log \sigma$ `bglmer` is unable to guard against infinite estimates of the random effects variance.

Table S4 shows the number of estimates for $\log \sigma$ per specification which resulted in an degenerate random effects variance estimate. We considered an estimate degenerate, if it is larger than $\log(50)$ or smaller than $\log(0.01)$, the gradient is larger than 0.001 in absolute value, or if the estimated asymptotic standard errors are larger than 40. We see that MSPL is the most stable estimation routine and that both `bglmer` and ML exhibit estimation degeneracies frequently for both small and large values of the random effects variance.
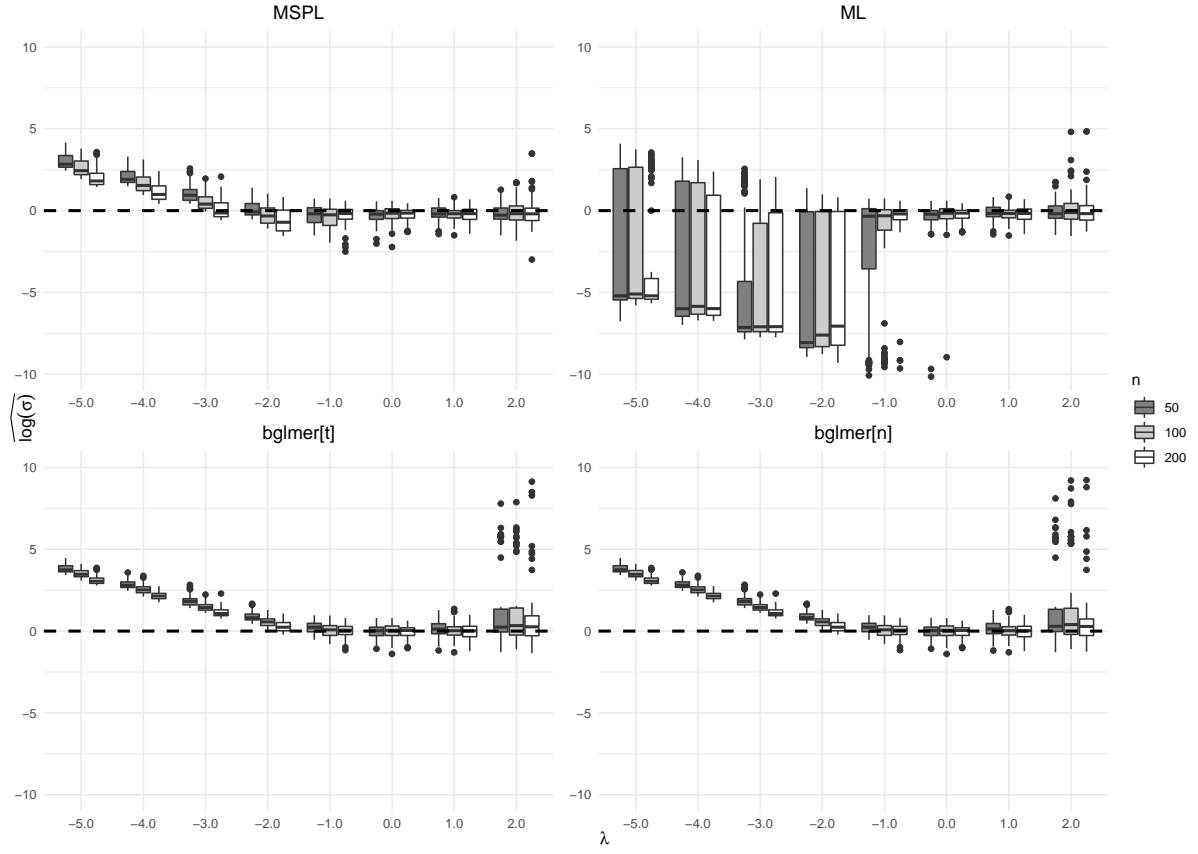
Fig. S3 Centered estimation output of $\log \sigma$ from Simulation 2

Table S4 Percentage of degenerate estimates from Simulation 2

		λ							
		-5	-4	-3	-2	-1	0	1	2
MSPL	n=50	0	0	0	0	0	0	0	**1**
	n=100	0	0	0	0	0	0	0	0
	n=200	0	0	0	0	0	0	0	**2**
ML	n=50	**73**	**70**	**76**	**70**	**25**	**2**	0	0
	n=100	**68**	**68**	**75**	**58**	**16**	**1**	0	**7**
	n=200	**78**	**68**	**69**	**52**	**4**	0	0	**5**
bglmer[t]	n=50	0	0	0	0	0	0	0	**12**
	n=100	**2**	**2**	**6**	**3**	**3**	0	0	**17**
	n=200	**26**	**35**	**23**	**15**	**4**	0	0	**11**
bglmer[n]	n=50	0	0	0	0	0	0	0	**12**
	n=100	**2**	**1**	**2**	0	0	0	0	**17**
	n=200	**20**	**22**	**20**	**15**	**3**	0	0	**11**

References

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.

Chung, Y., S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika 78*(4), 685–709.

Kosmidis, I. and D. Firth (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika 108*(1), 71–82.

McKeon, C. S., A. C. Stier, S. E. McIlroy, and B. M. Bolker (2012). Multiple defender effects: synergistic coral defense by mutualist crustaceans. *Oecologia 169*(4), 1095–1103.

Nash, J. C. (2014). On best practice optimization methods in R. *Journal of Statistical Software 60*(2), 1–14.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sauter, R. and L. Held (2016). Quasi-complete separation in random effects of binary response mixed models. *Journal of Statistical Computation and Simulation 86*(14), 2781–2796.