Bias reduction in generalized nonlinear models

Ioannis Kosmidis
and
David Firth

Department of Statistics

THE UNIVERSITY OF WARWICK

JSM 2009

Outline

Bias reduction in estimation

- In regular parametric models the maximum likelihood estimator $\hat{\beta}$ is consistent and the expansion of its bias has the form

$$E(\hat{\beta} - \beta_0) = \frac{b_1(\beta_0)}{n} + \frac{b_2(\beta_0)}{n^2} + \frac{b_3(\beta_0)}{n^3} + \ldots .$$

- Firth (1993): Adjust the score functions U_t to

$$U_t^* = U_t + A_t \quad (t = 1, \ldots, p) .$$

  For appropriate functions A_t, $U_t^* = 0$ $(t = 1, \ldots, p)$ results to estimators $\tilde{\beta}$ with no $O(n^{-1})$ bias term.

- Mehrabi & Mathhews (1995), Heinze & Schemper (2002;2005), Bull et al (2002;2007) and others.
  - → ML estimates are not required.
  - → Estimators with "better" properties.

Exponential family of distributions

- Random variable Y from the exponential family of distributions:

$$f(y\,;\theta) = \exp\left\{\frac{y^T\theta - b(\theta)}{\lambda} + c(y,\lambda)\right\} ,$$

where the dispersion λ is assumed known.

$$\mu = E(Y\,;\theta) = \frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta} ,$$

$$\sigma^2 = \mathrm{var}\,(Y\,;\theta) = \lambda\frac{\mathrm{d}^2 b(\theta)}{\mathrm{d}\theta^2} .$$

Generalized nonlinear model

- $y_1, \ldots, y_n$ realizations of independent random variables $Y_1, \ldots, Y_n$ from the exponential family.
- For a generalized nonlinear model (GNM)

$$g(\mu_r) = \eta_r(\beta) \quad (r = 1, \ldots, n)\,,$$

where g is the link function and $\eta_r : \Re^p \to \Re$.

- Score functions:

$$U_t = \sum_{r=1}^{n} \frac{w_r}{d_r}(y_r - \mu_r)x_{rt} \quad (t = 1, \ldots, p)\,,$$

where $w_r = d_r^2/\sigma^2$, $d_r = \mathrm{d}\mu_r/\mathrm{d}\eta_r$ and $x_{rt} = \partial\eta_r/\partial\beta_t$.

Adjusted score functions for GNMs

Bias-reducing adjusted score functions (Kosmidis & Firth, 2008)

$$U_t^* = \sum_{r=1}^{n} \frac{w_r}{d_r} \left[y_r + \frac{1}{2} h_r \frac{d'_r}{w_r} + d_r \, \mathrm{tr} \left\{ F^{-1} \mathcal{D}^2 \left(\eta_r ; \beta \right) \right\} - \mu_r \right] x_{rt} \, ,$$

→ $d'_r = \mathrm{d}^2 \mu_r / \mathrm{d}\eta_r^2$ and h_r is the r-th diagonal of $H = X F^{-1} X^T W$,

Adjusted score functions for GNMs

Bias-reducing adjusted score functions (Kosmidis & Firth, 2008)

$$U_t^* = \sum_{r=1}^{n} \frac{w_r}{d_r} \left[\overbrace{y_r + \frac{1}{2} h_r \frac{d_r'}{w_r} + d_r \,\mathrm{tr}\left\{ F^{-1} \mathcal{D}^2 \left(\eta_r ; \beta \right) \right\}}^{y_r^*} - \mu_r \right] x_{rt} \,,$$

→ $d_r' = \mathrm{d}^2\mu_r / \mathrm{d}\eta_r^2$ and h_r is the r-th diagonal of $H = X F^{-1} X^T W$,

Implementation

→ Replace y_r with the adjusted responses y_r^* in iterative reweighted least squares (IWLS).

- In terms of modified working observations

$$\zeta_r^* = \zeta_r - \xi_r \quad (r = 1, \ldots, n) \ ,$$

where
→ $\zeta_r = \sum_{t=1}^{p} \beta_t x_{rt} + (y_r - \mu_r)/d_r$ is the working observation for maximum likelihood, and
→ $\xi_r = -d'_r h_r/(2 w_r d_r) - \mathrm{tr}\left\{F^{-1}\mathcal{D}^2\left(\eta_r;\beta\right)\right\}/2.$

Modified working observations

Modified iterative re-weighted least squares

- Iteration

$$\tilde{\beta}_{(j+1)} = (X^T W_{(j)} X)^{-1} X^T W_{(j)} (\zeta_{(j)} - \xi_{(j)}) \,,$$

- The $O(n^{-1})$ bias of the maximum likelihood estimator for generalized nonlinear models is

$$b_1/n = (X^T W X)^{-1} X^T W \xi$$

(Cook et al. 1986; Cordeiro & McCullagh, 1991).

- Thus the iteration takes the form

$$\tilde{\beta}_{(j+1)} = \hat{\beta}_{(j)} - b_{1,(j)}/n \,.$$

Illustration: The RC(1) model

- Two-way cross-classification by factors X and Y with R and S levels, respectively. Entries are realizations of independent Poisson random variables.
- The RC(1) model (Goodman, 1979, 1985)

$$\log \mu_{rs} = \lambda + \lambda_r^X + \lambda_s^Y + \rho \gamma_r \delta_s \, .$$

- Modified working observation:

$$\zeta^*_{rs} = \zeta_{rs} + \frac{h_{rs}}{2\mu_{rs}} + \gamma_r C(\rho, \delta_s) + \delta_s C(\rho, \gamma_r) + \rho C(\gamma_r, \delta_s) \, ,$$

where for any given pair of unconstrainted parameters κ and ν, $C(\kappa, \nu)$ denotes the corresponding element of F^{-1}; if either of κ or ν is constrained, $C(\kappa, \nu) = 0$.

Data: Peridontal condition and calcium intake

Table: Periodontal condition and calcium intake (Goodman, 1981, Table 1.a.)

Periodontal condition	Calcium intake level			
	1	2	3	4
A	5	3	10	11
B	4	5	8	6
C	26	11	3	6
D	23	11	1	2

- For identifiability, set $\lambda_1^X = \lambda_1^Y = 0$, $\gamma_1 = \delta_1 = -2$ and $\gamma_4 = \delta_4 = 2$.
- Simulate 250000 data sets under the maximum likelihood fit.
- Estimate biases, mean squared errors and coverage of nominally 95% Wald-type confidence intervals.

Results

Table: Results for the dental health data. For the method of maximum likelihood, simulation results are all conditional upon finiteness of the estimates (about 3.5% of the simulated datasets resulted in infinite MLEs).

	Estimates		Simulation results					
	ML	BR	Bias ($\times 10^2$)		MSE ($\times 10$)		Coverage (%)	
			ML	BR	ML	BR	ML	BR
λ	2.31	2.35	−4.19	−0.25	2.28	1.49	96.9	96.6
λ_2^X	−0.13	−0.13	0.48	−0.01	1.45	1.16	95.8	96.2
λ_3^X	0.55	0.52	2.97	−0.22	1.50	1.18	95.7	96.0
λ_4^X	0.07	0.10	−5.00	0.02	3.34	1.87	97.1	97.3
λ_2^Y	−0.53	−0.53	−0.59	0.06	1.00	0.80	96.0	96.4
λ_3^Y	−1.17	−1.05	−16.81	1.19	6.55	2.80	97.1	96.1
λ_4^Y	−0.80	−0.75	−7.21	0.22	3.19	1.69	97.3	97.3
ρ	−0.20	−0.18	−1.76	−0.03	0.05	0.03	95.5	95.0
γ_2	−1.55	−1.48	−6.08	0.68	6.30	5.37	95.6	96.7
γ_3	0.90	0.91	1.88	1.43	6.94	5.34	93.8	95.2
δ_2	−1.16	−1.11	−7.00	−0.27	9.00	7.20	94.7	96.4
δ_3	3.11	2.84	37.42	−4.92	35.55	18.13	92.8	92.4

ML, maximum likelihood; BR, bias-reduced; MSE, mean squared error.

Penalized likelihood interpretation of bias reduction

- Firth (1993): for a generalized linear model with canonical link, the adjusted scores, correspond to penalization of the likelihood by the Jeffreys (1946) invariant prior.
- In models with non-canonical link and $p \geq 2$, there need not exist such a penalized likelihood interpretation.

Penalized likelihood interpretation of bias reduction

Theorem

Existence of penalized likelihoods

In the class of generalized linear models, there exists a penalized log-likelihood l^* such that $\nabla l^*(\beta) \equiv U^*(\beta)$, for all possible specifications of design matrix X, if and only if the inverse link derivatives $d_r = 1/g_r'(\mu_r)$ satisfy

$$d_r \equiv \alpha_r \sigma^{2\omega} \quad (r = 1, \ldots, n)\,,$$

where α_r $(r = 1, \ldots, n)$ and ω do not depend on the model parameters.

Penalized likelihood interpretation of bias reduction

The form of the penalized likelihoods for bias-reduction

When $d_r \equiv \alpha_r \sigma^{2\omega}$ $(r = 1, \ldots, n)$ for some ω and α,

$$l^*(\beta) = \begin{cases} l(\beta) + \dfrac{1}{4} \sum_r \log \kappa_{2,r}(\beta)^{h_r} & (\omega = 1/2) \\ l(\beta) + \dfrac{\omega}{4\omega - 2} \log |F(\beta)| & (\omega \neq 1/2) . \end{cases}$$

- → The canonical link is the special case $\omega = 1$.
- → With $\omega = 0$, the condition refers to models with identity-link.
- → For $\omega = 1/2$ the working weights, and hence F, H, do not depend on β.
- → If $\omega \notin [0, 1/2]$, bias-reduction also increases the value of $|F(\beta)|$. Thus, approximate confidence ellipsoids, based on asymptotic normality of the estimator, are reduced in volume.

Discussion

- A computational and conceptual framework for bias-reduction in generalized nonlinear models.
- λ was assumed known but this is not restricting the applicability of the results. The dispersion is usually estimated separately from the parameters β.
- Bias reduction can be beneficial in terms of the properties of the resultant estimators.
- Bias and point estimation are *not* strong statistical principles:
  $\rightarrow$ Bias relates to parameterization thus improving the bias violates exact equivariance under reparameterization.
  $\rightarrow$ Reduction in bias can be accompanied by inflation in variance.

Some references

Bull, S. B., Mak, C. and Greenwood, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* **39**, 57–74.

Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B: Methodological*, **53**, 629–643.

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

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

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics* **13**, 10–69.

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

Kosmidis, I. and D. Firth (2008). Bias reduction in exponential family nonlinear models. Technical Report 8-5, CRiSM working paper series, University of Warwick. Accepted for publication in *Biometrika*.

Wei, B. (1997). *Exponential Family Nonlinear Models*. New York: Springer-Verlag Inc.