Bias reduction in generalized nonlinear models

Ioannis Kosmidis
and
David Firth

Department of Statistics

The University of Warwick

JSM 2009
Outline

1. Reduction of the bias
2. Generalized nonlinear models
3. Illustration
4. Generalized linear models
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.


→ ML estimates are not required.

→ Estimators with “better” properties.
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 $\lambda$ is assumed known.

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

$$\sigma^2 = \text{var} (Y ; \theta) = \lambda \frac{d^2 b(\theta)}{d\theta^2}.$$
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 \( \eta_r : \mathbb{R}^p \rightarrow \mathbb{R} \).

- Score functions:

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

  where \( w_r = d_r^2 / \sigma^2 \), \( d_r = d\mu_r / d\eta_r \) and \( x_{rt} = \partial \eta_r / \partial \beta_t \).
Adjusted score functions for GNMs

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

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

\[ \rightarrow d'_r = \frac{d^2 \mu_r}{d\eta_r^2} \text{ and } h_r \text{ is the } r\text{-th diagonal of } H = X F^{-1} X^T W, \]
Adjusted score functions for GNMs

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

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

\[ \rightarrow d_r' = \frac{d^2 \mu_r}{d \eta_r^2} \text{ and } h_r \text{ is the } r\text{-th diagonal of } H = XF^{-1}X^TW, \]
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

→ $\xi_r = -d'_r h_r/(2w_r d_r) - \text{tr} \left\{ F^{-1} D^2 (\eta_r; \beta) \right\} /2$. 

Kosmidis, I. Bias reduction in generalized nonlinear models
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 unconstrained parameters $\kappa$ and $\nu$, $C(\kappa, \nu)$ denotes the corresponding element of $F^{-1}$; if either of $\kappa$ or $\nu$ is constrained, $C(\kappa, \nu) = 0$. 
Data: Peridontal condition and calcium intake

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

<table>
<thead>
<tr>
<th>Periodontal condition</th>
<th>Calcium intake level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>26</td>
</tr>
<tr>
<td>D</td>
<td>23</td>
</tr>
</tbody>
</table>

- 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).

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Simulation results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
</tr>
<tr>
<td></td>
<td>ML</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>2.31</td>
</tr>
<tr>
<td>$\lambda_2^X$</td>
<td>-0.13</td>
</tr>
<tr>
<td>$\lambda_3^X$</td>
<td>0.55</td>
</tr>
<tr>
<td>$\lambda_4^X$</td>
<td>0.07</td>
</tr>
<tr>
<td>$\lambda_2^Y$</td>
<td>-0.53</td>
</tr>
<tr>
<td>$\lambda_3^Y$</td>
<td>-1.17</td>
</tr>
<tr>
<td>$\lambda_4^Y$</td>
<td>-0.80</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.20</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>-1.55</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.90</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>-1.16</td>
</tr>
<tr>
<td>$\delta_3$</td>
<td>3.11</td>
</tr>
</tbody>
</table>

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 $\alpha_r$ ($r = 1, \ldots, n$) and $\omega$ 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 \( \omega \) and \( \alpha \),

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

\( \rightarrow \) The canonical link is the special case \( \omega = 1 \).

\( \rightarrow \) With \( \omega = 0 \), the condition refers to models with identity-link.

\( \rightarrow \) For \( \omega = 1/2 \) the working weights, and hence \( F, H \), do not depend on \( \beta \).

\( \rightarrow \) 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.
- $\lambda$ was assumed known but this is not restricting the applicability of the results. The dispersion is usually estimated separately from the parameters $\beta$.
- Bias reduction can be beneficial in terms of the properties of the resultant estimators.
- Bias and point estimation are not strong statistical principles:
  - Bias relates to parameterization thus improving the bias violates exact equivariance under reparameterization.
  - Reduction in bias can be accompanied by inflation in variance.


