GLM bca estimators

1 Simulation description

The goal of the simulation was to check efficiency of the additive bias corrected estimator adapted for generalised linear models. There were generated $N = 100000$ observations $x \sim N(1, 1)$, $z \sim Bernoulli(0.7)$. Then there were generated Y from Bernoulli distribution with probability $\sigma(1 + x - 2z)$, where σ is the sigmoid function.

Then there were generated labels z^* with probability of misclassification equal to 0.07 when $z = 1$ and 0.1 for $z = 0$. Then from the whole population (with precisely measured z's) was drawn an audit sample of size n_ audit = 1000. The simulation was then repeated 500 times.

2 Estimators description

In simulation there were used three estimators – precise, naive and bca. The precise estimator uses precisely measured data and is impossible in real world. However it was used as a benchmark to estimate the best possible precision of estimator the used model. The naive estimator uses available data with measurement error and does not correct them in any way. The bca estimator uses incorrectly measured data but also has available an audit sample which contains both precisely and imprecisely measured data. Then uses this information to correct bias in final estimations. The exact formula for it is

$$\hat{\psi}_{bca} = \hat{\psi} - \frac{1}{\sqrt{n}} \hat{A}^{-1} \hat{b}$$

where $\hat{\psi}$ is the naive estimator,

$$\hat{A} = \frac{1}{n} \sum_{i=1}^{n} \mu'(\hat{\psi}^T \hat{\xi}_i) \hat{\xi}_i \hat{\xi}_i^T,$$

$$\hat{b} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\xi}_i \left(\mu(\hat{\psi}^T \xi_i) - \mu(\hat{\psi}^T \hat{\xi}_i) \right),$$

μ is a linking function (in our case a sigmoid function) and $\hat{\xi}$ is a vector containing observed values.

3 Results

As expected the best was the precise estimator and naive one was the worst. What is worth to notice is a fact that the bca estimator has decreased absolute value of biases in all coefficients (when compared to naive one) more than thrice. Its average sum square error was about ten times smaller that for naive estimator. Even though the results are still not perfect, the improvement is significant.

Table 1: Mean bias of estimating coefficients

	Precise	Naive	Bca
z	0.0005	0.504	0.140
Intercept	−0.001	−0.382	−0.108
x	0.0003	−0.049	−0.016

Table 2: SSE

Statistic	Precise	Naive	Bca
Mean	0.001	0.403	0.036
St. Dev.	0.001	0.028	0.023
Min	0.000	0.307	0.0001
Max	0.006	0.501	0.123