Bias corrected estimators simulation

1 Simulation description

The goal of the simulation was to compare different estimators when a part of the data is measured with errors. There were generated $N = 10000$ or $N = 100000$ observations $x \sim N(2, 1), z \sim Bernoulli(0.7)$ and $e \sim N(0, 1)$. The value of interests was computed by formula

$$Y_i = 1 + x_i - 2z_i + e_i.$$

Then in scenario I there were generated z^* such that they misclassify z with probability 0.07. In scenario II probability of misclassification when $z = 0$ was increased to 0.15 while probability of misclassification with $z = 1$ did not change. Then from whole population (with precisely measured z's) was drawn an audit sample of size n_ audit = 1000. Four estimators were used to predict coefficients of x, z and intercept. The simulation was then repeated 500 times.

2 Estimators description

In simulation there were considered four estimators: precise, naive, classic bca, and upgraded bca. Precise estimator was standard OLS estimator computed on true values of z measured without measurement error. Naive estimator was standard OLS estimator computed on z^* that did not correct potential bias. Classic bca was defined as in [1] even though its assumptions were violated

$$\hat{\psi}^{bca\_naive} = \left(I + \frac{\hat{\kappa}}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\xi}_i \hat{\xi}_i^T \right)^{-1} \begin{bmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{bmatrix} \right) \hat{\psi}. \tag{1}$$

Upgraded bca estimator was similar as the classic one but improved to deal with models with not equal probability of different types of errors and smaller asympthotic accuracy:

$$\hat{\psi}^{bca} = \left(I + \frac{1}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\xi}_i \hat{\xi}_i^T \right)^{-1} \begin{bmatrix} \hat{\kappa} & 0 \\ \hat{\lambda} & 0 \end{bmatrix} \right) \hat{\psi}, \tag{2}$$

where

$$\hat{\kappa} = \sqrt{n} \cdot FPR, \quad \hat{\lambda} = \sqrt{n} \cdot \left[\frac{\sum_{i=1}^{m} (\hat{\theta}_i - \theta_i)}{m}, \ \frac{\sum_{i=1}^{m} x_i (\hat{\theta}_i - \theta_i)}{m} \right]^T$$

$$\kappa = \lim_{n \to \infty} \sqrt{n} FPR, \quad \lambda = \lim_{n \to \infty} \frac{\sum_{i=1}^{n} q_i (\hat{\theta}_i - \theta_i)}{\sqrt{n}}$$

with FPR being the model *false positive ratio*, m being size of an used audit sample with both z and z^*, $\hat{\theta}_i, \theta_i$ are values predicted by ML model (measured with error) and their true values, $\hat{\psi}$ is naive OLS estimator, and $\hat{\xi} = [z^*, 1, x]^T$.

3 Simulation problem

For the estimator ψ^{bca_naive} to work it is necessary to satisfy conditions described in [1], especially $\mathbb{E}[(\hat{\theta}_i - \theta_i) q_i] = 0$ as in assumption 2 *(iii)*. However when q_i includes an intercept it gets form of $\mathbb{E}\hat{\theta}_i = \mathbb{E}\theta_i$ which is violated in scenario I because (in this case z^* is in place of $\hat{\theta}$)

$$\mathbb{E}z^* = P(z^* = 1) = 0.93 \cdot 0.7 + 0.07 \cdot 0.3 = 0.672 \neq 0.7 = \mathbb{E}z.$$

In scenario II even though the model has smaller accuracy, this condition is satisfied because

$$\mathbb{E}z^* = P(z^* = 1) = 0.93 \cdot 0.7 + 0.15 \cdot 0.3 = 0.696 \approx 0.7 = \mathbb{E}z$$

so better performance of the bca estimator can be expected.

4 Results

For each of the simulations there was computed difference between true and predicted coefficient. Their means are biases of estimated coefficients. Tables 1 and 2 presents their values for each of the estimators. Then differences between true and predicted values of coefficients were squared and added for each simulation. Mean value of these sums of squares are presented in tables 4-6. We can see that in scenario II classical bca estimator performed better than in scenario I even though the classifying model in this case has lower accuracy. After improving this estimator it started to perform well also in scenario I. The results of the upgraded bca estimator were slightly better in scenario I than in scenario II which is not a surprise due to better accuracy of the predictive model.

Table 1: Mean biases of estimated coefficients in scenario I

	N=10000			N=100000		
	Intercept	x	z	Intercept	x	z
Precise	−0.0007	0.0003	0.0004	0.0008	−0.0002	−0.0007
Naive	−0.2993	0.0001	0.3616	−0.2978	−0.0003	0.3607
Bca classic	−0.1934	0.0000	0.2041	−0.1931	−0.0003	0.2049
Bca upgraded	−0.0549	0.0018	0.0620	−0.0548	0.0010	0.0641

Table 2: Mean biases of estimated coefficients in scenario II

	N=10000			N=100000		
	Intercept	x	z	Intercept	x	z
Precise	−0.0007	0.0003	0.0004	0.0008	-0.0002	−0.0007
Naive	−0.3234	0.0002	0.4525	−0.3213	−0.0003	0.4511
Bca classic	−0.0932	0.0001	0.1220	−0.0927	−0.0003	0.1227
Bca upgraded	−0.0740	0.0018	0.0994	−0.0738	0.0010	0.1010

Table 3: MSE for scenario I for $N = 10000$

Statistic	Precise	Naive	Bca classic	Bca upgraded
Mean	0.001	0.222	0.082	0.011
St. Dev.	0.001	0.032	0.028	0.011
Min	0.000	0.129	0.019	0.0001
Max	0.011	0.327	0.167	0.061

Table 4: MSE for scenario I for $N = 100000$

Statistic	Precise	Naive	Bca classic	Bca upgraded
Mean	0.0001	0.219	0.081	0.011
St. Dev.	0.0001	0.010	0.023	0.009
Min	0.000	0.192	0.014	0.0001
Max	0.001	0.246	0.153	0.059

Table 5: MSE for scenario II for N=10000

Statistic	Precise	Naive	Bca classic	Bca upgraded
Mean	0.001	0.311	0.028	0.020
St. Dev.	0.001	0.039	0.020	0.016
Min	0.000	0.197	0.0002	0.000
Max	0.011	0.439	0.114	0.083

Table 6: MSE for scenario II for N=100000

Statistic	Precise	Naive	Bca classic	Bca upgraded
Mean	0.0001	0.307	0.027	0.019
St. Dev.	0.0001	0.012	0.019	0.014
Min	0.000	0.276	0.000	0.0002
Max	0.001	0.338	0.107	0.076

References

[1] Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher. Inference for regression with variables generated by AI or machine learning, 2025.