Uniwersytet Ekonomiczny w Katowicach

blisko

międzynarodowo

przez całe życie

O estymatorach złożonych dla schematu losowania Pathaka

On complex estimators for the Pathak sampling scheme

Krzysztof Szymoniak-Książek
College of Management
Department of Statistics, Econometrics, and Mathematics
szymoniak-ksiazek@ue.katowice.pl

Outline

- Purpose of the presentation
- Pathak sampling scheme
- The covariance theorem
- Taylor linearization
- Complex estimators
- Numerical study
- Summary

The purpose of the presentation

The aim of the study was to evaluate the properties of complex estimators of population parameters that are functions of population mean values under the Pathak sampling scheme.

Complex estimators:

- product mean estimator
- ratio mean estimator
- regression mean estimator

Notation

- Finite population:

$$U = \{1, \dots, N\}$$

- Study variables:

$$\boldsymbol{x} = (x_1, \dots, x_N)$$
$$\boldsymbol{y} = (y_1, \dots, y_N)$$

- Population means:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Pathak scheme

- Cost vector:

$$\boldsymbol{c} = [c_1, \ldots, c_N]$$

- Research budget B:

$$B > \max_{i \neq j \in U} \{c_i + c_j\}$$

Sampling until the sum of costs for the selected elements exceeds or reaches the research budget. The element for which this occurs is not included in the sample.

[Pathak, 1976]

Pathak scheme

- A sample selected using the Pathak sampling scheme

$$s = (s_1, s_2, \ldots., s_M),\ s_i \in U$$

- For notational convenience, we denote:

$$x_{(i)} = x_{s_i}$$
$$y_{(i)} = y_{s_i}$$

Pathak theorem

The unbiased estimators of means based on a sample selected using the Pathak sampling scheme:

$$\bar{x}_M = \frac{1}{M}\sum_{i=1}^{M} x_{(i)}$$

$$\bar{y}_M = \frac{1}{M}\sum_{i=1}^{M} y_{(i)}$$

[Pathak, 1976]

Covariance theorem

a)

$$Cov(\bar{x}_M, \bar{y}_M) = \frac{1}{2N(N-1)} \sum_{j,k=1, j\neq k}^{N} (x_j - x_k)(y_j - y_k)\left(E\left(\frac{1}{M}\middle| s_1 = j, s_2 = k\right) - \frac{1}{N}\right)$$

Covariance theorem

b)

An unbiased estimator of $Cov(\bar{x}_M, \bar{y}_M)$ is given by the formula:

$$\hat{C}_{xy} = \Delta \cdot \sum_{i=1}^{M} \left(x_{(i)} - \bar{x}_M\right)\left(y_{(i)} - \bar{y}_M\right),$$

where

$$\Delta = \left[\frac{1}{M} - \frac{1}{N}\right]\frac{1}{M-1}.$$

Proof: [Szymoniak-Książek & Gamrot, 2025]

Comparison with Pathak's results

$$Var(\bar{x}_M) = \frac{1}{2N(N-1)} \sum_{j,k=1,j\neq k}^{N} (x_j - x_k)^2 \left(E\left(\frac{1}{M}\middle| s_1 = j, s_2 = k\right) - \frac{1}{N} \right)$$

$$Cov(\bar{x}_M, \bar{y}_M) = \frac{1}{2N(N-1)} \sum_{j,k=1,j\neq k}^{N} (x_j - x_k)(y_j - y_k) \left(E\left(\frac{1}{M}\middle| s_1 = j, s_2 = k\right) - \frac{1}{N} \right)$$

Comparison with Pathak's results

$$\hat{V}_x = \Delta \cdot \sum_{i=1}^{M} \left(x_{(i)} - \bar{x}_M\right)^2$$

$$\hat{C}_{xy} = \Delta \cdot \sum_{i=1}^{M} \left(x_{(i)} - \bar{x}_M\right)\left(y_{(i)} - \bar{y}_M\right)$$

Taylor linearization – assumptions

Let $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$ be a vector of parameters in U, and let $\boldsymbol{t} = (t_1, t_2, \ldots, t_k)$ be an unbiased estimator for $\boldsymbol{T}$. Assume that the function $f: R^k \to R$ satisfies the following conditions:

- $f(\boldsymbol{T}) = \bar{y}$
- f is bounded in some neighborhood of the point $\boldsymbol{T}$ and has continuous and bounded partial derivatives up to at least the third order in this neighborhood

Additionally, assume that $P\left(\frac{|t_i - T_i|}{T_i} < 1\right) = 1$, for each i.

[Bracha, 1996]

Taylor linearization

Then

$$f(\boldsymbol{t}) = f(\boldsymbol{T}) + \sum_{i=1}^{k}(t_i - T_i)\frac{\partial f}{\partial t_i}\bigg|_{\boldsymbol{t}=\boldsymbol{T}} +$$

$$\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(t_i - T_i)(t_j - T_j)\frac{\partial^2 f}{\partial t_i \partial t_j}\bigg|_{\boldsymbol{t}=\boldsymbol{T}} + \frac{1}{6}\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{k}(t_i - T_i)(t_j - T_j)(t_l - T_l)\frac{\partial^3 f}{\partial t_i \partial t_j \partial t_l}\bigg|_{\boldsymbol{t}=\tilde{\boldsymbol{T}}},$$

where $\boldsymbol{T} \leq \tilde{\boldsymbol{T}} \leq \mathbf{t}$.

$$AB(f(\boldsymbol{t})) = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} E\left((t_i - T_i)(t_j - T_j)\right)\frac{\partial^2 f}{\partial t_i \partial t_j}\bigg|_{\boldsymbol{t}=\boldsymbol{T}}$$

$$AMSE(f(\boldsymbol{t})) = \sum_{i=1}^{k}\sum_{j=1}^{k} E\left((t_i - T_i)(t_j - T_j)\right)\left(\frac{\partial f}{\partial t_i}\right)\left(\frac{\partial f}{\partial t_j}\right)\bigg|_{\boldsymbol{t}=\boldsymbol{T}}$$

[Bracha, 1996]

Product mean estimator

The statistic

$$\bar{y}_p = \frac{\bar{x}_M \bar{y}_M}{\bar{x}}, \qquad \bar{x} > 0$$

is called the product estimator of the mean $\bar{y}$.

(Murthy, 1964)

Approximate bias:

$$AB(\bar{y}_p) = \frac{Cov(\bar{x}_M, \bar{y}_M)}{\bar{x}}$$

Estimator:

$$\widehat{AB}(\bar{y}_p) = \frac{\hat{C}_{xy}}{\bar{x}_M}$$

Product mean estimator

Approximate MSE:

$$AMSE(\bar{y}_p) = D^2(\bar{y}_M) + 2\frac{\bar{y}}{\bar{x}}Cov(\bar{y}_M, \bar{x}_M) + \frac{\bar{y}^2}{\bar{x}^2}D^2(\bar{x}_M)$$

Estimator:

$$\widehat{AMSE}(\bar{y}_p) = \hat{V}_y + 2\frac{\bar{y}_M}{\bar{x}_M}\hat{C}_{xy} + \frac{\bar{y}_M^2}{\bar{x}_M^2}\hat{V}_x$$

Ratio mean estimator

The statistic

$$\bar{y}_q = \frac{\bar{y}_M}{\bar{x}_M}\bar{x}$$

is called the ratio estimator of the mean $\bar{y}$.

(Hansen et al., 1953)

Approximate bias:

$$AB(\bar{y}_q) = \frac{1}{\bar{x}}\left(\frac{\bar{y}}{\bar{x}}D^2(\bar{x}_M) - Cov(\bar{y}_M, \bar{x}_M)\right)$$

Estimator:

$$\widehat{AB}(\bar{y}_q) = \frac{1}{\bar{x}_M}\left(\frac{\bar{y}_M}{\bar{x}_M}\hat{V}_x - \hat{C}_{xy}\right)$$

Ratio mean estimator

Approximate MSE:

$$AMSE(\bar{y}_q) = D^2(\bar{y}_M) - 2\frac{\bar{y}}{\bar{x}}Cov(\bar{y}_M, \bar{x}_M) + \frac{\bar{y}^2}{\bar{x}^2}D^2(\bar{x}_M)$$

Estimator:

$$\widehat{AMSE}(\bar{y}_q) = \hat{V}_y - 2\frac{\bar{y}_M}{\bar{x}_M}\hat{C}_{xy} + \frac{\bar{y}_M^2}{\bar{x}_M^2}\hat{V}_x$$

Regression mean estimator

The statistic

$$\bar{y}_r = \bar{y}_M + \frac{\hat{C}_{xy}}{\hat{V}_x}(\bar{x} - \bar{x}_M)$$

is called the regression estimator of the mean $\bar{y}$.

(Neyman, 1934)

Regression mean estimator

Approximate bias:

...

Estimator:

$$\widehat{AB}(\bar{y}_r) = \frac{N-M}{MN}\left(\frac{\bar{x}\hat{C}_{xy} - \frac{1}{2}\bar{y}\hat{V}_x}{\hat{V}_x}\right)$$

Regression mean estimator

Approximate MSE:

$$AMSE(\bar{y}_r) = \frac{Cov(\boldsymbol{x},\boldsymbol{y})}{Var(\boldsymbol{x})}\left(\frac{Cov(\boldsymbol{x},\boldsymbol{y})D^2(\bar{x}_M)}{Var(\boldsymbol{x})} - 2Cov(\bar{x}_M,\bar{y}_M)\right) + D^2(\bar{y}_M)$$

Estimator:

$$\widehat{AMSE}(\bar{y}_r) = \frac{\hat{C}_{xy}}{\hat{V}_x}\left(\frac{\hat{C}_{xy}\hat{V}_x}{\hat{V}_x} - 2\hat{C}_{xy}\right) + \hat{V}_y = \frac{\hat{C}_{xy}^2}{\hat{V}_x} - 2\frac{\hat{C}_{xy}^2}{\hat{V}_x} + \hat{V}_y = \hat{V}_y - \frac{\hat{C}_{xy}^2\hat{V}_y}{\hat{V}_x\hat{V}_y}$$

$$= \hat{V}_y\left(1 - \frac{\hat{C}_{xy}^2}{\hat{V}_x\hat{V}_y}\right) = \hat{V}_y(1-\hat{\rho}^2)$$

Numerical study

Consider a population U of size $N = 10000$ such that the vector of study variables is a realization from a bivariate normal distribution:

$$(y_i, x_i) \sim N_2\left([5,5], \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{bmatrix}\right).$$

Furthermore, assume that the cost vector is a realization from a uniform distribution: $c_i \sim U(0{,}10)$.

The subsequent analysis will be conducted separately for the three complex estimators of population mean: the product estimator, the ratio estimator, and the regression estimator.

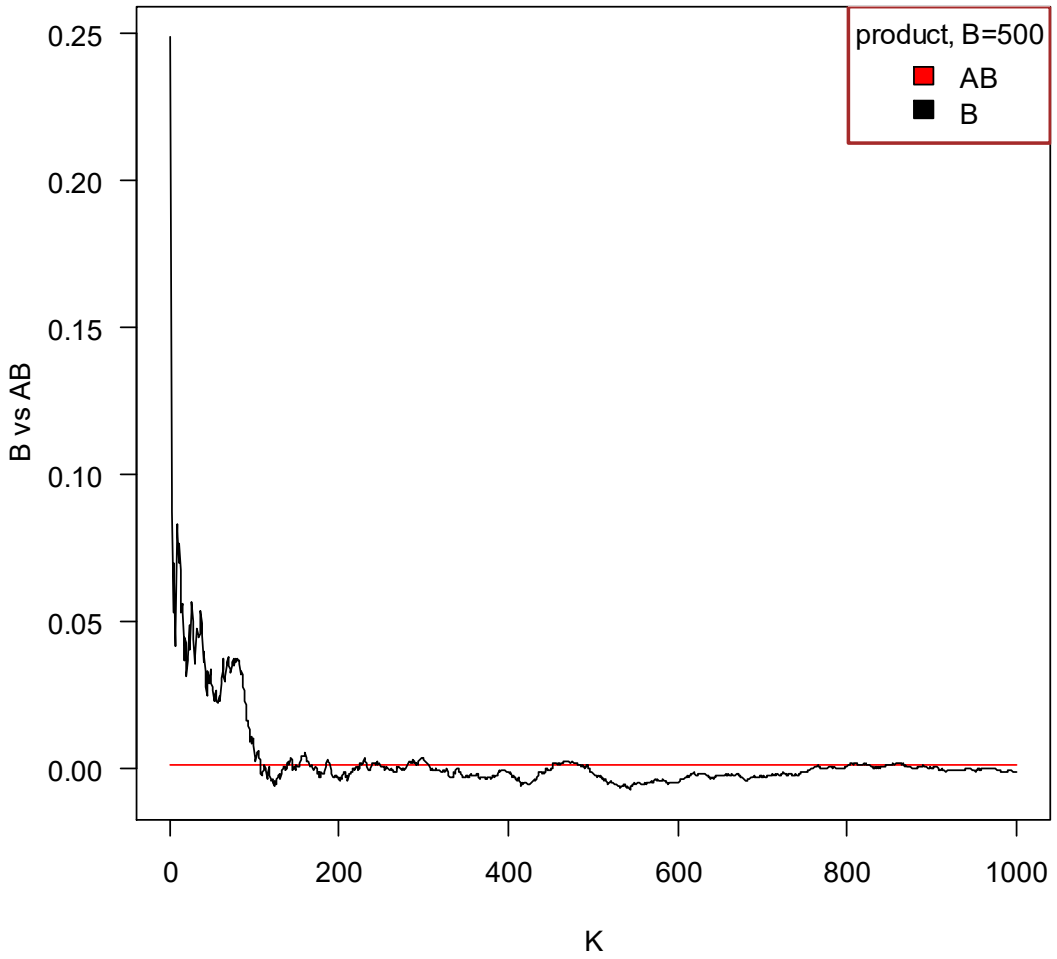
Numerical study

A sample is drawn K times using the Pathak sampling scheme under a budget constraint, and a complex estimator of the population mean $\bar{y}$ is computed.

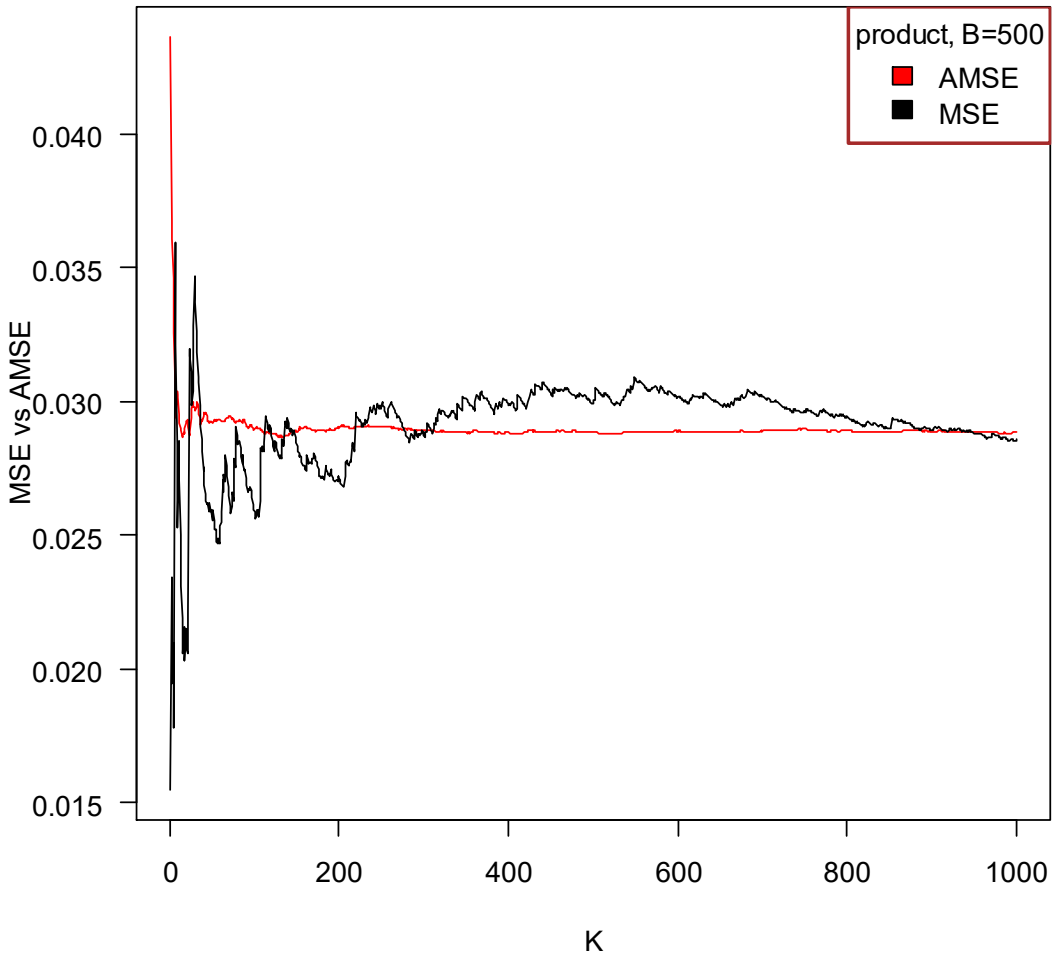
The first figure presents a comparison between the average (over $K = 1,2,\dots,1000$) bias of the complex estimator (black line) and the average estimator of its approximate bias (red line).

The second figure shows an analogous comparison between the simulation mean squared error of the complex estimator of the mean (black line) and the average estimator of its approximate MSE (red line).

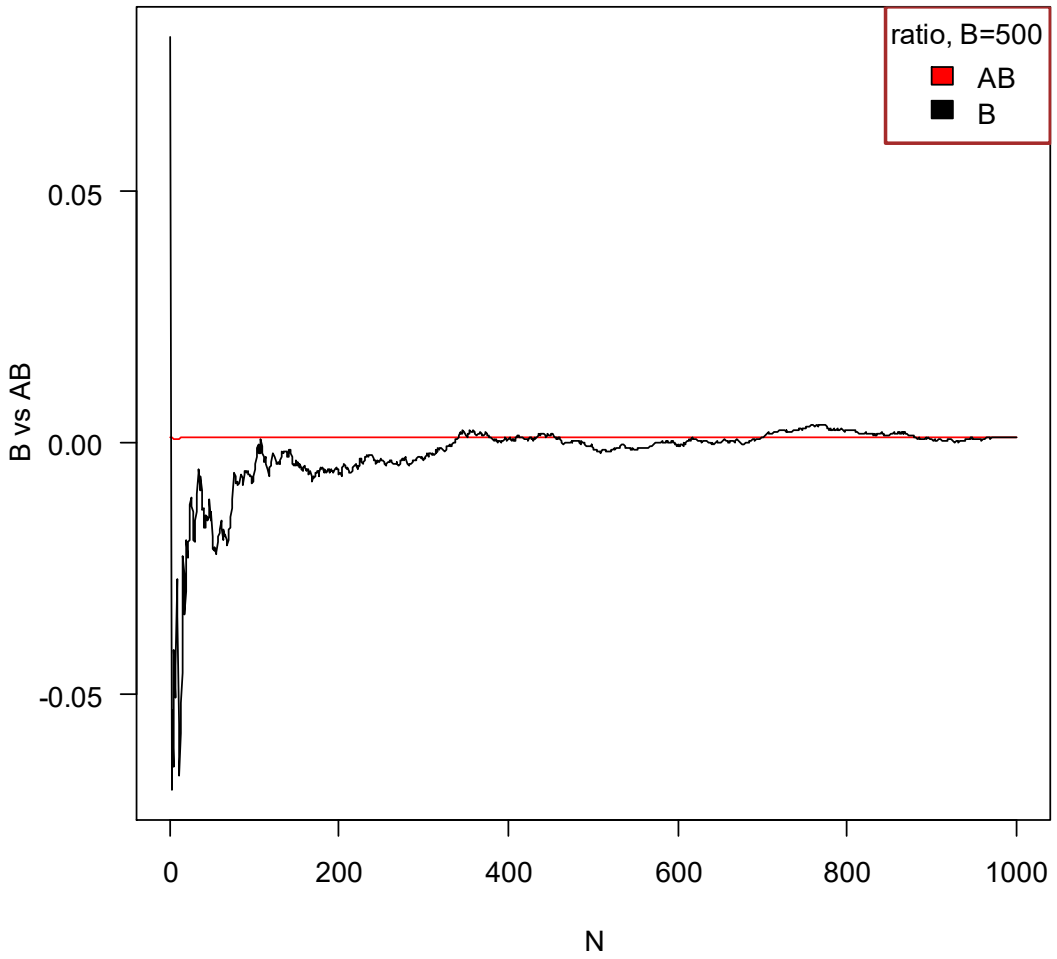
Product mean estimator – bias



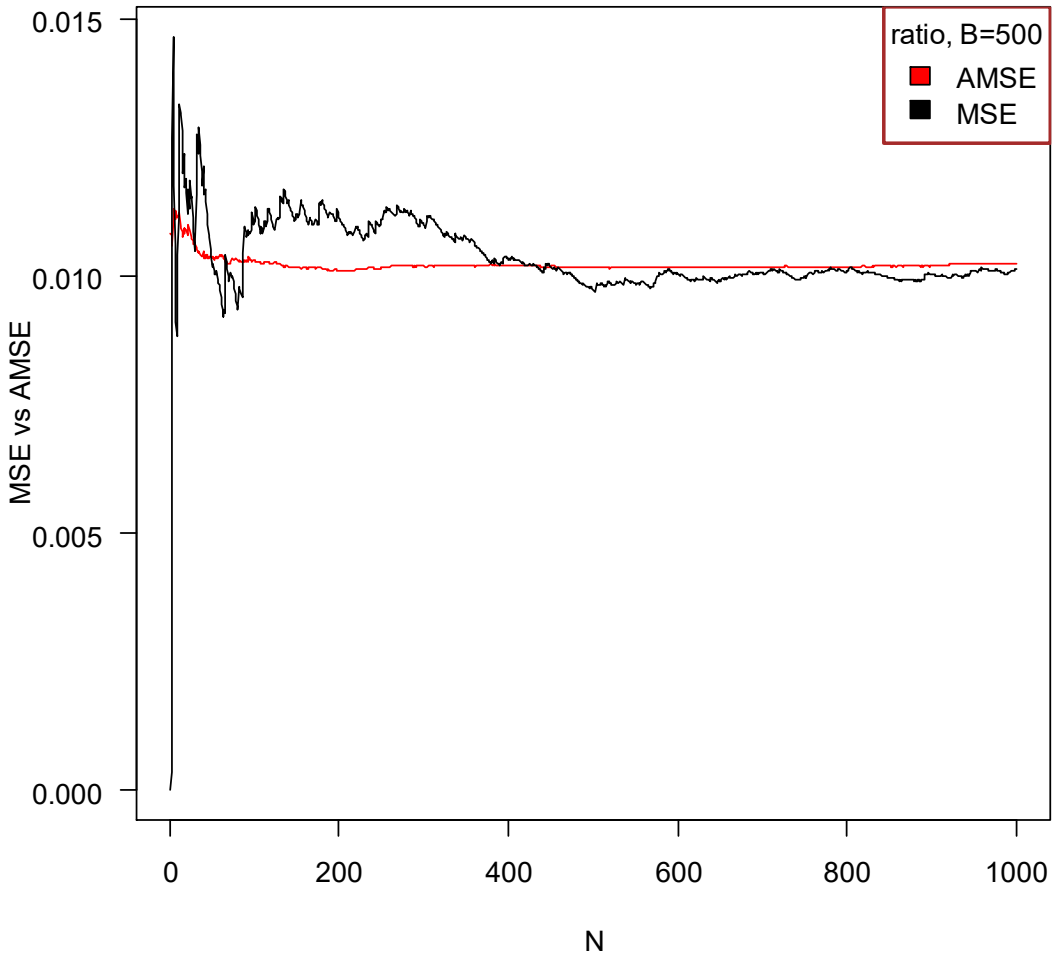
Product mean estimator – MSE



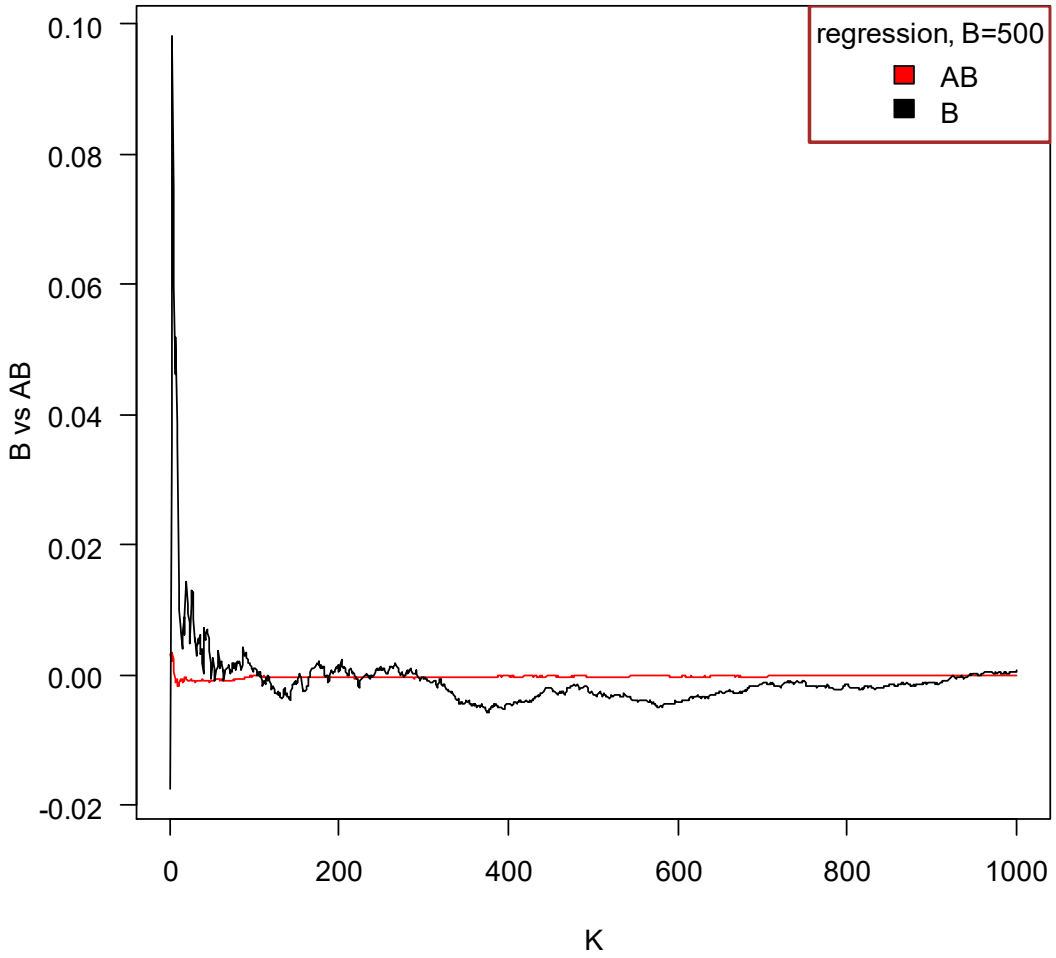
Ratio mean estimator – bias



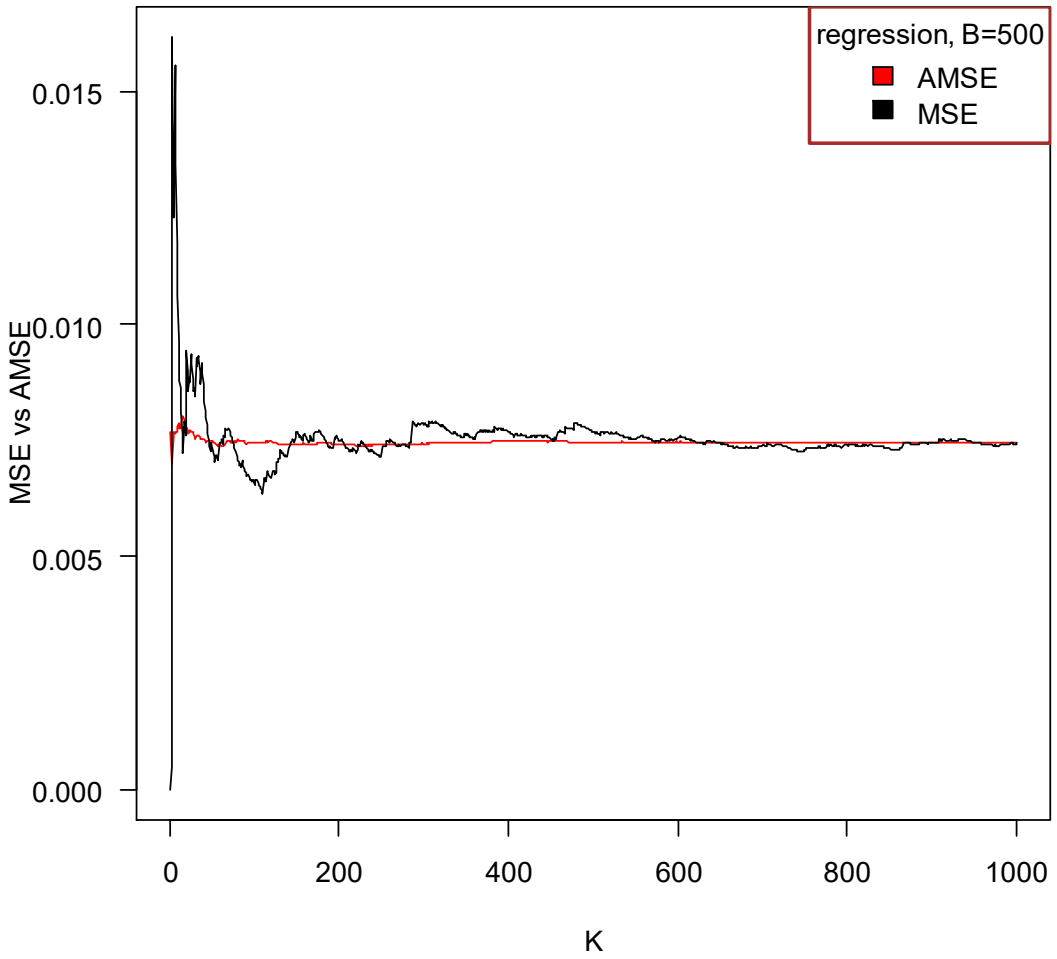
Ratio mean estimator – MSE



Regression mean estimator – bias



Regression mean estimator – MSE



Summary

For all considered complex estimators of the mean, even with a small number of repeated samples drawn from the population U, both the average sample biases and the sample mean squared errors of the complex estimators are very close to the values of the corresponding averaged approximate estimators of bias and mean squared error for the complex estimators of the population mean.

Bibliography

- Bracha, Cz. (1996). *Teoretyczne podstawy metody reprezentacyjnej*. PWN, Warszawa.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory.*
- Murthy, M. N. (1964). Product Method of Estimation. *Sankhyā: The Indian Journal of Statistics, Series A* (1961-2002), 26(1), 69–74.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection*. Journal of the Royal Statistical Society*, 97(4), 558–625.
- Pathak K. (1976): Unbiased estimation in fixed cost sequential sampling scheme, *Annals of Statistics*, 4(5), 1012-1017
- Szymoniak-Książek, K., & Gamrot, W. (2025). Controlling for survey costs when estimating covariances between Pathak estimators in fixed-budget sequential sampling. *Journal of Economics & Management*/University of Economics in Katowice, (47), 211-228.

Uniwersytet
Ekonomiczny
w Katowicach

www.ue.katowice.pl