Uniwersytet Ekonomiczny w Katowicach

blisko

międzynarodowo

przez całe życie

O kowariancji estymatorów Pathaka

Krzysztof Szymoniak-Książek
Kolegium Zarządzania
Katedra Statystyki, Ekonometrii i Matematyki

Plan wystąpienia

- Cel prezentacji
- Schemat Pathaka
- Twierdzenie o kowariancji
- Dowód twierdzenia
- Przykład empiryczny
- Podsumowanie

Cel prezentacji

Zaproponowanie estymatora kowariancji pomiędzy estymatorami wartości średniej w schemacie Pathaka.

Zamiar ten zrealizowano poprzez uogólnienie znanego estymatora wariancji estymatora średniej zaproponowanego przez Pathaka [1976].

Oznaczenia

- Populacja skończona:

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

- Badane cechy:

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

- Wartości średnie:

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

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

Schemat Pathaka

- Koszty:

$$\boldsymbol{c} = (c_1, \ldots, c_N)$$

- Budżet badania B:

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

Losujemy do momentu, gdy suma kosztów zbadania wartości cech wylosowanych elementów przekroczy lub osiągnie budżet.

[Pathak, 1976]

Schemat Pathaka

- Próba wylosowana schematem Pathaka:

$$s = (s_1, s_2, \ldots., s_{M+1}),\ s_i \in U$$

- Zdefiniujmy:

$$T_M = (M, \{s_1, s_2, \ldots., s_M\}, s_{M+1})$$

Odnotujmy, że T_M jest statystyką dostateczną [Pathak, 1976].

- Ponadto oznaczmy:

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

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

Twierdzenie Pathaka

Nieobciążone estymatory średnich oparte na próbie wylosowanej schematem Pathaka dane są wzorami:

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

$$\bar{y}_M = \frac{1}{M}\sum_{i=1}^{M} y_{(i)} \qquad (2)$$

[Pathak, 1976]

Dowód

Zauważmy, że $x_{(1)}$ jest nieobciążonym estymatorem $\bar{x}$. Oznaczmy $t_1 = x_{(1)}$. Niech $k > 1$ oraz niech $U_1, U_2, \dots, U_k, U_j \in U$. Załóżmy, że

$$T_M = \left(k, \{U_1, U_2, \dots, U_k\}, U_j\right).$$

Wówczas

$$P(s_1 = U_1 | T_M) = \cdots = P(s_1 = U_k | T_M) = \frac{1}{k} \text{ oraz } P\left(s_1 = U_j \middle| T_M\right) = 0.$$

Zatem

$$E_{t_1|T_M}(t_1|T_M) = \frac{1}{k}\sum_{i=1}^{k} x_{U_i} = \frac{1}{M}\sum_{i=1}^{M} x_{(i)} = \bar{x}_M .$$

Dowód

Korzystając z własności warunkowej wartości oczekiwanej

$$E_Y\left(E_{X|Y}(X|Y)\right) = E_X(X)$$

otrzymujemy, że

$$E(\bar{x}_M) = E_{T_M}\left(E_{t_1|T_M}(t_1|T_M)\right) = E_{t_1}(t_1) = \bar{x}.$$

Zatem $\bar{x}_M$ jest nieobciążonym estymatorem $\bar{x}$.

Analogicznie pokazujemy, że $\bar{y}_M$ jest nieobciążonym estymatorem $\bar{y}$.

Twierdzenie 2

a)

$$Cov(\bar{x}_M, \bar{y}_M) = \frac{1}{2N(N-1)} \sum_{j,k=1, j\neq k}^{N} (x_j - x_k)(y_j - y_k)\left(E\left(\frac{1}{M}\middle| s_1 = j, s_2 = k\right) - \frac{1}{N}\right) \quad (3)$$

Twierdzenie 2

b)

Nieobciążony estymator $Cov(\bar{x}_M, \bar{y}_M)$ dany jest wzorem:

$$\hat{C}_{xy} = \Delta \cdot \sum_{i=1}^{M} \left(x_{(i)} - \bar{x}_M\right)\left(y_{(i)} - \bar{y}_M\right), \qquad (4)$$

gdzie

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

Dowód

b)

Załóżmy, że $e(\bar{x}\bar{y})$ jest nieobciążonym estymatorem $\bar{x}\bar{y}$ oraz niech

$$\hat{C}_{xy} = \bar{x}_M \bar{y}_M - e(\bar{x}\bar{y}). \qquad (5)$$

Zauważmy, że

$$E\left(\hat{C}_{xy}\right) = E(\bar{x}_M \bar{y}_M) - E\left(e(\bar{x}\bar{y})\right) = E(\bar{x}_M \bar{y}_M) - \bar{x}\bar{y} =$$

$$E(\bar{x}_M \bar{y}_M) - E(\bar{x}_M)E(\bar{y}_M) = Cov(\bar{x}_M, \bar{y}_M)$$

Zatem $\hat{C}_{xy}$ jest nieobciążonym estymatorem $Cov(\bar{x}_M, \bar{y}_M)$.

Dowód

b)

Zauważmy, że

$$\bar{x}\bar{y} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} x_i y_i + \frac{1}{N^2}\sum_{\substack{j,k=1 \\ j\neq k}}^{N} x_j y_k$$

Rozważmy

$$t_2 = \frac{1}{N} x_{(1)} y_{(1)} + \frac{N-1}{N} x_{(1)} y_{(2)}$$

$$E_{t_2}(t_2) = E_{t_2}\left(\frac{1}{N} x_{(1)} y_{(1)} + \frac{N-1}{N} x_{(1)} y_{(2)}\right) = \frac{1}{N^2}\sum_{i=1}^{N} x_i y_i + \frac{1}{N^2}\sum_{\substack{j,k=1 \\ j\neq k}}^{N} x_j y_k = \bar{x}\bar{y}$$

Dowód

b)

Zatem t_2 jest nieobciążonym estymatorem $\bar{x}\bar{y}$ bazującym na dwóch pierwszych jednostkach. Niech $W = E_{t_2|T_M}(t_2|T_M)$.

Zauważmy, że

$$W = E_{t_2|T_M}(t_2|T_M) = E_{t_2|T_M}\left(\left(\frac{1}{N}x_{(1)}y_{(1)} + \frac{N-1}{N}x_{(1)}y_{(2)}\right)\middle| T_M\right) =$$

$$= \frac{1}{N}\frac{1}{M}\sum_{i=1}^{M} x_{(i)}y_{(i)} + \frac{N-1}{NM(M-1)}\sum_{\substack{j,k=1 \\ j\neq k}}^{M} x_{(j)}y_{(k)} =$$

$$= \bar{x}_M\bar{y}_M - \Delta \cdot \sum_{i=1}^{M}(x_{(i)} - \bar{x}_M)(y_{(i)} - \bar{y}_M)$$

Dowód

b)

$$\bar{x}_M\bar{y}_M - \Delta\cdot\sum_{i=1}^{M}(x_{(i)}-\bar{x}_M)(y_{(i)}-\bar{y}_M) =$$

$$= \frac{1}{M^2}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)} - \Delta\cdot\sum_{i=1}^{M}\left(x_{(i)}y_{(i)} - x_{(i)}\bar{y}_M - y_{(i)}\bar{x}_M + \bar{x}_M\cdot\bar{y}_M\right)$$

$$= \frac{1}{M^2}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)} - \Delta\cdot\sum_{i=1}^{M}\left(x_{(i)}y_{(i)} - x_{(i)}\bar{y}_M - y_{(i)}\bar{x}_M + \frac{1}{M^2}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}\right)$$

$$= \frac{1}{M^2}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)} - \Delta\sum_{i=1}^{M}\left(x_{(i)}y_{(i)} - x_{(i)}\frac{1}{M}\sum_{l=1}^{M} y_{(l)} - y_{(i)}\frac{1}{M}\sum_{l=1}^{M} x_{(l)} + \frac{1}{M^2}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}\right)$$

$$= \frac{1}{M^2}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)} - \Delta\cdot\left(\sum_{l=1}^{M} x_{(l)}y_{(l)} - \frac{2}{M}\left(\sum_{l=1}^{M} x_{(l)}y_{(l)} + \sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}\right) + \frac{1}{M^2}\sum_{i=1}^{M}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}\right)$$

$$= \left(\frac{1}{M^2} - \Delta\cdot\left(1 - \frac{2}{M} + \frac{1}{M}\right)\right)\sum_{l=1}^{M} x_{(l)}y_{(l)} + \left(\frac{1}{M^2} - \Delta\cdot\left(\frac{-2}{M} + \frac{1}{M}\right)\right)\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}$$

$$= \left(\frac{1}{M^2} - \left[\frac{1}{M} - \frac{1}{N}\right]\frac{1}{M-1}\frac{M-1}{M}\right)\sum_{l=1}^{M} x_{(l)}y_{(l)} + \left(\frac{1}{M^2} + \left[\frac{1}{M} - \frac{1}{N}\right]\frac{1}{M-1}\frac{1}{M}\right)\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}$$

$$= \left(\frac{1}{MN}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \left(\frac{1}{M^2} + \frac{N-M}{MN}\frac{1}{M-1}\frac{1}{M}\right)\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}\right.$$

$$= \frac{1}{MN}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{N(M-1)+N-M}{NM^2(M-1)}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)} = \frac{1}{MN}\sum_{l=1}^{M} x_{(l)}y_{(l)} + \frac{NM-M}{NM^2(M-1)}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}$$

$$= \frac{1}{MN}\sum_{i=1}^{M} x_{(i)}y_{(i)} + \frac{N-1}{NM(M-1)}\sum_{j,k=1,j\neq k}^{M} x_{(j)}y_{(k)}$$

Dowód

b)

Ponieważ W jest funkcją T_M, to $E(W) = E_{T_M}(W)$. Zatem korzystając z własności warunkowej wartości oczekiwanej otrzymujemy

$$E(W) = E_{T_M}\left(E_{t_2|T_M}(t_2|T_M)\right) = E_{t_2}(t_2) = \bar{x}\bar{y}.$$

Zatem W jest nieobciążonym estymatorem $\bar{x}\bar{y}$.

Podstawiając W za $e(\bar{x}\bar{y})$ do (5)

$$\hat{C}_{xy} = \bar{x}_M\bar{y}_M - e(\bar{x}\bar{y})$$

otrzymujemy (4)

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

Dowód

a)

Odnotujmy, że

$$Cov(\bar{x}_M, \bar{y}_M) = E\left(\hat{C}_{xy}\right) = E\left(E\left(\hat{C}_{xy}\middle|T_M\right)\right).$$

Ponadto

$$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{2N}\sum_{\substack{j,k=1 \\ j\neq k}}^{N}(x_j - x_k)(y_j - y_k)$$

[Yuli Zhang, Huaiyu Wu, Lei Cheng, 2012]

Dowód

a)

Zatem dla ustalonej próby s zachodzi

$$\sum_{i=1}^{M}\left(x_{(i)}-\bar{x}_M\right)\left(y_{(i)}-\bar{y}_M\right)=\frac{1}{2M}\sum_{\substack{j,k=1\\ j\neq k}}^{M}\left(x_{(j)}-x_{(k)}\right)\left(y_{(j)}-y_{(k)}\right)$$

W szczególności

$$\Delta\cdot\sum_{i=1}^{M}\left(x_{(i)}-\bar{x}_M\right)\left(y_{(i)}-\bar{y}_M\right)=\Delta\cdot\frac{1}{2M}\sum_{\substack{j,k=1\\ j\neq k}}^{M}\left(x_{(j)}-x_{(k)}\right)\left(y_{(j)}-y_{(k)}\right)$$

Dowód

a)

Zauważmy, że

$$Cov(\bar{x}_M, \bar{y}_M) = E(\hat{C}_{xy}) = E\big(E(\hat{C}_{xy}|T_M)\big) =$$

$$= E\left(E\left(\Delta \cdot \sum_{i=1}^{M}(x_{(i)} - \bar{x}_M)(y_{(i)} - \bar{y}_M)\middle| T_M\right)\right) =$$

$$= E\left(E\left(\Delta \cdot \frac{1}{2M}\sum_{\substack{j,k=1 \\ j\neq k}}^{M}(x_{(j)} - x_{(k)})(y_{(j)} - y_{(k)})\middle| T_M\right)\right)$$

Dowód

a)

$$E_{T_M}\left(E_{\hat{C}|T_M}\left(\Delta\cdot\frac{1}{2M}\sum_{\substack{j,k=1\\ j\neq k}}^{M}\left(x_{(j)}-x_{(k)}\right)\left(y_{(j)}-y_{(k)}\right)\middle|T_M\right)\right)=$$

$$=E\left(E\left(\left[\frac{1}{M}-\frac{1}{N}\right]\frac{1}{2M(M-1)}M(M-1)\left(x_{(1)}-x_{(2)}\right)\left(y_{(1)}-y_{(2)}\right)\middle|T_M\right)\right)=$$

$$=E\left(\left[\frac{1}{M}-\frac{1}{N}\right]\frac{1}{2}\left(x_{(1)}-x_{(2)}\right)\left(y_{(1)}-y_{(2)}\right)\right)$$

Dowód

a)

Zatem

$$Cov(\bar{x}_M, \bar{y}_M) = E\left(\frac{1}{2}\sum_{\substack{j,k=1 \\ j\neq k}}^{N}(x_j - x_k)(y_j - y_k)\left[\frac{1}{M} - \frac{1}{N}\right]\alpha_{jk}\right), \qquad (6)$$

gdzie $\alpha_{jk} = 1$, jeśli $s_1 = j$ i $s_2 = k$ oraz $\alpha_{jk} = 0$ w przeciwnym przypadku.

Odnotujmy, że

$$E(\alpha_{jk}) = P(s_1 = j, s_2 = k) = \frac{1}{N(N-1)} \qquad (7)$$

Dowód

a)

Ponadto

$$E\left(\frac{1}{M}\alpha_{jk}\right) = E\left(E\left(\frac{1}{M}\alpha_{jk}\middle| s_1, s_2\right)\right) = \frac{1}{N(N-1)}E\left(\frac{1}{M}\middle| s_1 = j, s_2 = k\right) \quad (8)$$

Podstawiając (7) i (8) do (6) otrzymujemy

$$Cov(\bar{x}_M, \bar{y}_M) = E\left(\frac{1}{2}\sum_{\substack{j,k=1 \\ j\neq k}}^{N}(x_j - x_k)(y_j - y_k)\left[\frac{1}{M} - \frac{1}{N}\right]\alpha_{jk}\right) =$$

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

Porównanie z wynikami Pathaka

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

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

Porównanie z wynikami Pathaka

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

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

Eksperyment symulacyjny

$N = 100$

$\boldsymbol{x} = (x_1, x_2, \dots, x_N)$

$\boldsymbol{y} = (y_1, y_2, \dots, y_N)$

$\boldsymbol{c} = (c_1, c_2, \dots, c_N)$

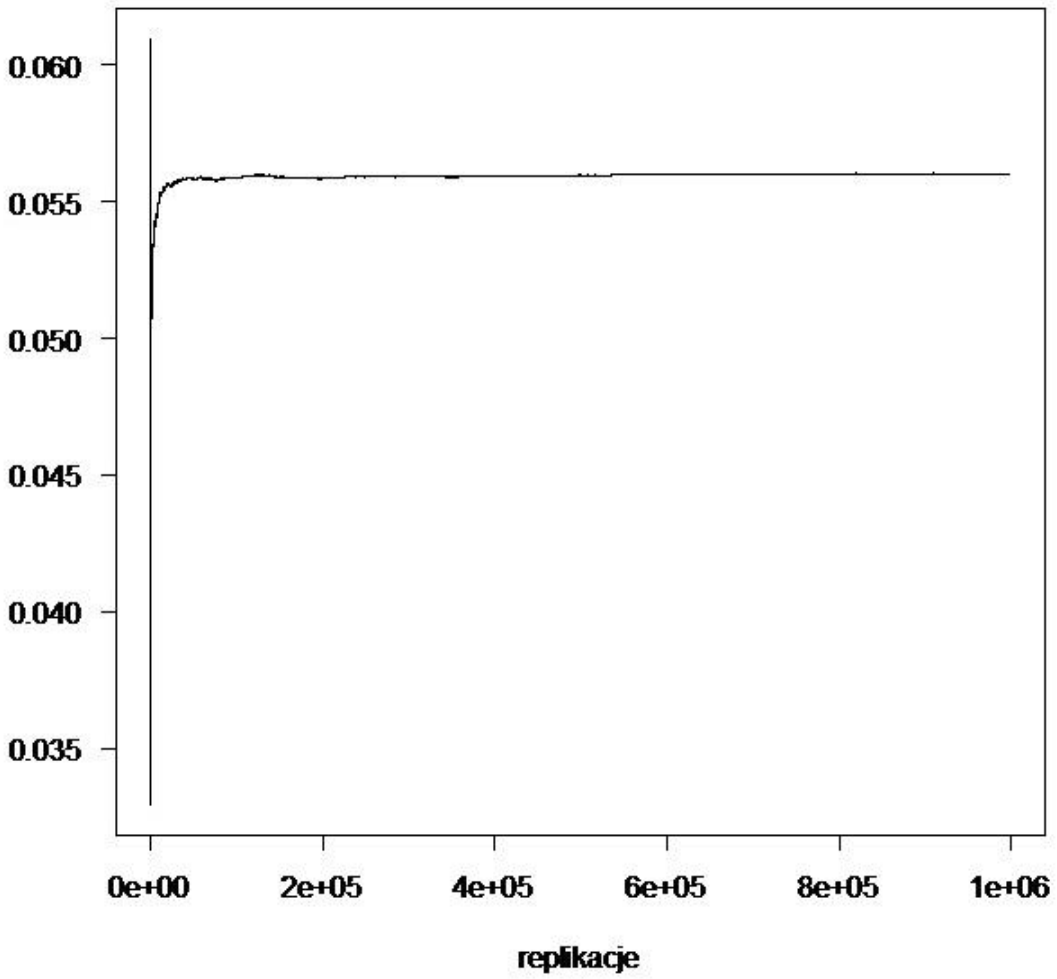
(x_i, y_i) – realizacje z rozkładu $N_2\left([0,0], \begin{bmatrix} 1 & \lambda \\ \lambda & 1 \end{bmatrix}\right)$ takie, że $r(\boldsymbol{x}, \boldsymbol{y}) \approx \lambda$

c_i - realizacje z rozkładu $U(0,100)$

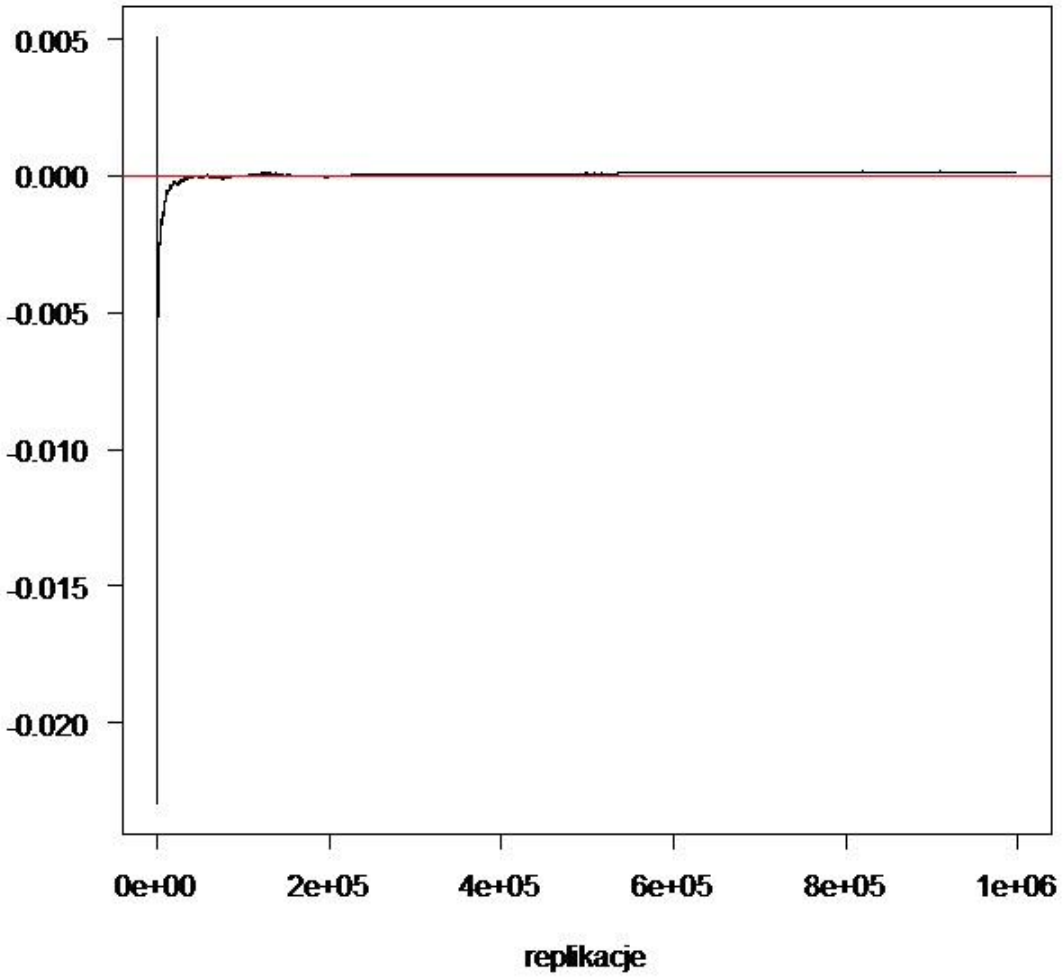
$B = 500$ – budżet badania

$R = 10^6$ - liczba replikacji próby

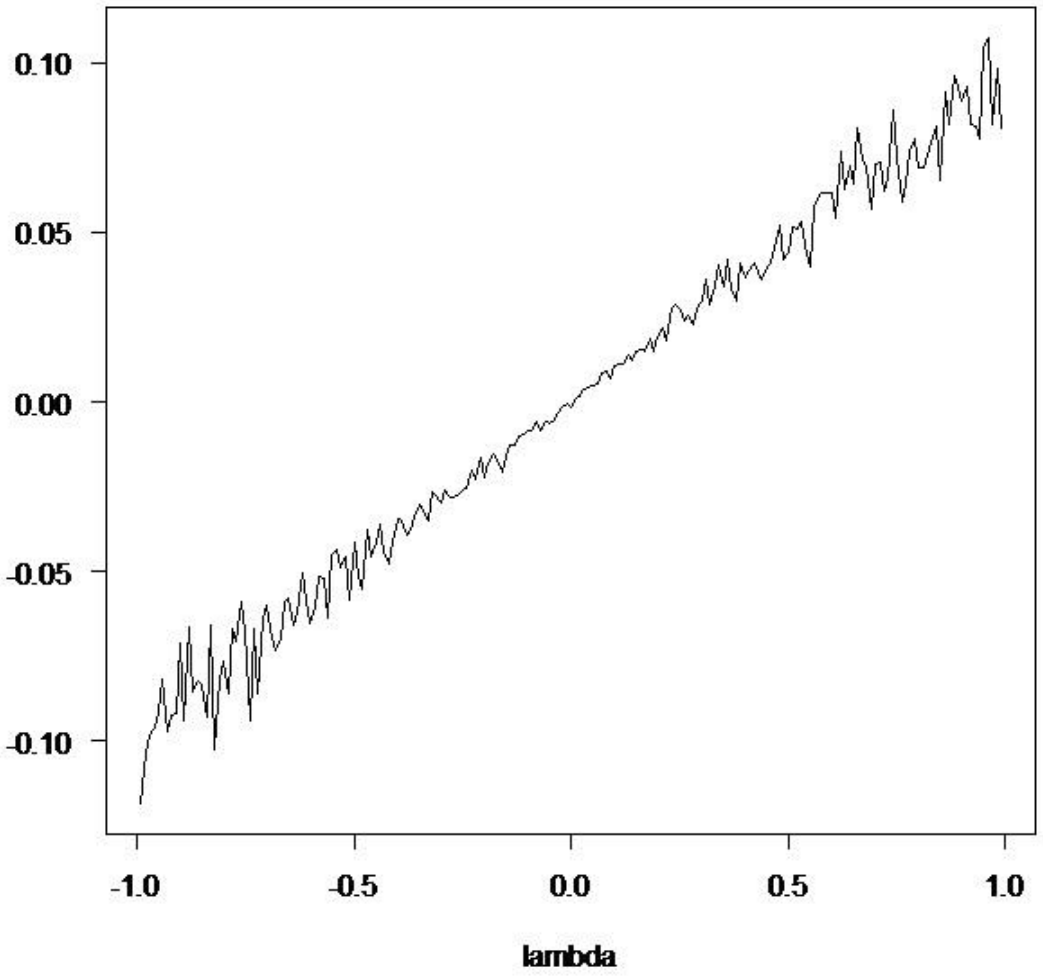
Uśredniony $\hat{C}_{xy}$ dla $\lambda = 0.8$



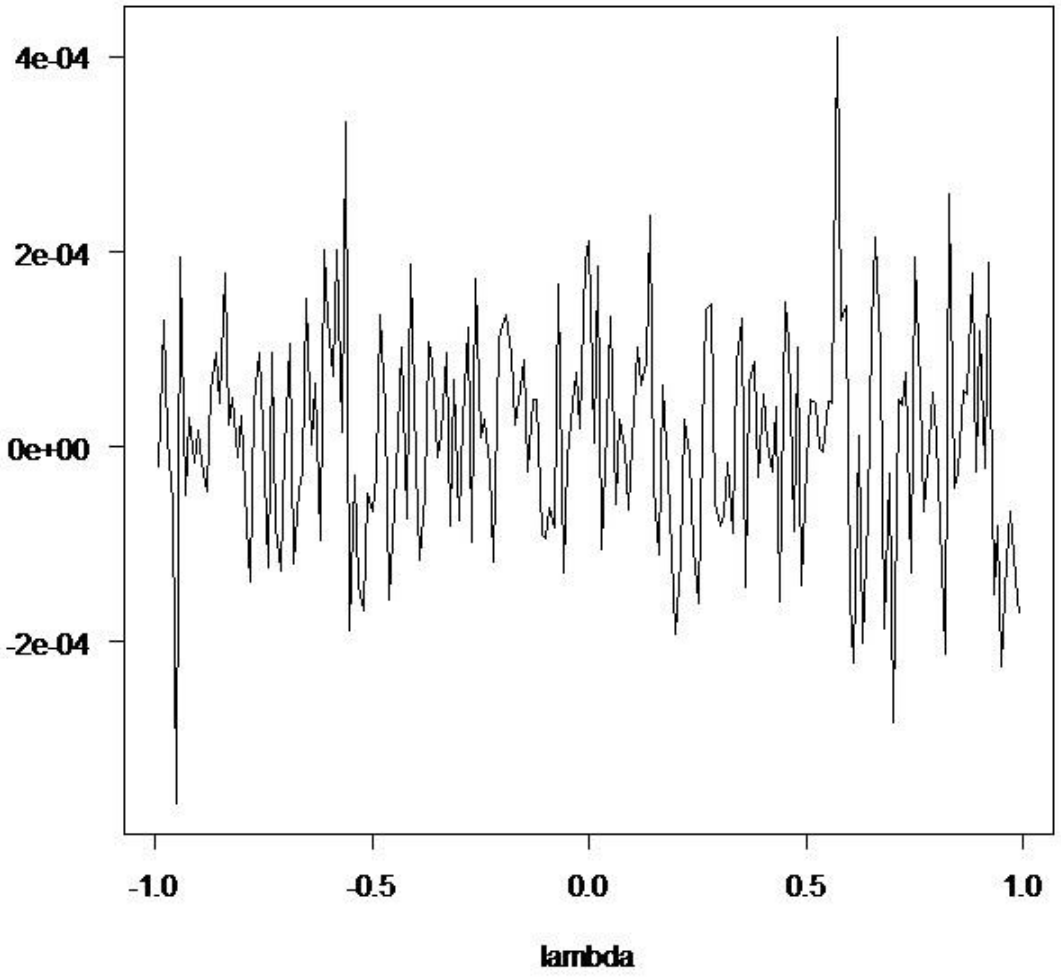
Uśredniony błąd dla $\lambda \approx 0.8$



Uśredniony $\hat{C}_{xy}$ w zależności od λ



Uśredniony błąd w zależności od λ



Podsumowanie

1. Rozważane estymatory kowariancji są nieobciążone.

2. Estymatory kowariancji mogą być przydatne do różnych celów np. ocena własności złożonych estymatorów parametrów populacji będących funkcjami wartości średnich lub globalnych.

Literatura

- Basu, D. (1958): On sampling with and without replacement, „Sankhya”, 20, 287-294
- Basu, D. (1969): Role of sufficiency and likelihood principle in sample survey theory, „Sankhya”, 31, 441-454
- Gamrot W. (2014): Estymacja wartości przeciętnej uwzględniająca koszt pozyskania danych, „Wydawnictwo UE”, Katowice
- Hajek, J. (1959): Optimum strategy and other problems in probability sampling, „Casopis Pest. Mat.”, 84, 387-423
- Pathak K. (1976): Unbiased estimation in fixed cost sequential sampling scheme, „Annals of Statistics”, 4(5), 1012-1017
- Zhang Y., Wu H., Cheng L. (2012): Some new deformation formulas about variance and covariance, „2012 Proceedings of International Conference on Modelling, Identification and Control, Wuhan, China”, 987-992

Uniwersytet
Ekonomiczny
w Katowicach

www.ue.katowice.pl