Zadanie domowe 1 – badania interentowe

Zadanie domowe do wykonania w grupach projektowych

- Termin: 12.04, 23:59
- Forma: Skrypt R, R Notebook, JupyterNotebook, link do COLAB
- Nazwa pliku: **grupa_numer_zadanie_domowe**.(rozszerzenie)

Treść

1. Zakładamy, że badana populacja ma wielkość N=10000.
2. Tworzymy macierz zmiennych $\mathbf{X}$ o wymiarze $N \times (1+p)$ gdzie p=50 (czyli macierz z wyrazem wolnym, pierwsza kolumna składająca się z samych 1, oraz 50 zmiennych). Zakładamy, że 50 zmiennych generujemy *niezależnie* z rozkładu normalnego standaryzowanego.
3. Zmienna celu Y jest utworzona w następujący sposób:

$$Y_i = 1 + \exp\{3\sin(\beta^T \mathbf{X}_i)\} + X_{i5} + X_{i6} + \epsilon_i, \quad \epsilon_i \sim N(0,1)$$

gdzie X_{i5} i X_{i6} to 5 i 6 zmienna z macierzy wektor $\mathbf{X}$ nie licząc wyrazu wolnego, $\beta = (1,0,0,1,1,1,1,0,...,0)^T$ o wymiarach $p+1$ (pierwszy element to β_0 czyli wykraz wolny). Oznacza to, że w wektorze β tylko wyraz wolny oraz 4-7 elementy są niezerowe (od 3 do 6 zmiennej). Uwaga wektor $\mathbf{X}_i$ ma wymiar $(p+1) \times 1$ dzięki temu w wyniku $\beta^T \mathbf{X}_i$) otrzymujemy skalar.

4. Tworzymy dwie próby:

Próba A - próba losowa o wielkości $n_A = 500$, w której jednostki badanej populacji są losowane proporcjonalnie do prawdopodobieństwa określonego następująco:

$$\pi_{iA} \propto \frac{0.25 + |X_{i1}| + 0.03|Y_i|}{\sum_{i=1}^{N} 0.25 + |X_{i1}| + 0.03|Y_i|}$$

Próba B - przynależność do próby nielosowej o wielkości $n_B = 2000$ wylosowana z rozkładu Bernoulliego gdzie π_B określono następującym modelem nieliniowej regresji logistycznej

$$\pi_{iB} = \frac{\exp\left\{3.5 + \alpha^T \log(\mathbf{X}_i^2) - \sin(X_{i3} + X_{i4}) - X_{i5} - X_{i6}\right\}}{1 + \exp\left\{3.5 + \alpha^T \log(\mathbf{X}_i^2) - \sin(X_{i3} + X_{i4}) - X_{i5} - X_{i6}\right\}}$$

gdzie $\alpha = (0,0,0,3,3,3,3,...,0)^T$ oznacza parametry stojące przy macierzy $\log(\mathbf{X}^2)$.

5. Celem jest oszacowanie średniej w populacji $y = \sum_{i=1}^{N} y_i/N$.
6. Proszę dokonać losowania próby A i B 500 razy oraz wyznaczyć wartość oczekiwaną, obciążenie, wariancję oraz MSE następujących estymatorów:

- $\mu_{A1} = \sum_i w_i y_i / n_A$,
- $\mu_B = \sum_i y_i / n_B$,

gdzie $w_i = 1/\pi_{iA}$, n_A to wielkość próby losowej, a n_B to wielkość próby nielosowej.

Źródło: Yang, S., Kim, J. K., & Rui, S. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 82(2), 445–465. https://doi.org/10.1111/rssb.12354