AZ 2021/2022 **Laboratorium 2 - testy permutacyjne, korelacja częściowa, Metoda MNK**

2.1 Zbiór *airpollution.txt* zawiera dane dotyczące związku pomiędzy zanieczyszczeniem powietrza i śmiertelnością w 60 miastach amerykańskich. Między zmiennymi są:

Mortality - skorygowana wiekiem liczba zgonów na 100 000 mieszkańców,

Education - mediana liczby lat kształcenia,

JanTemp, JulTemp - średnie temperatury w styczniu i lipcu (w stopniach Fahrenheita),

SO2Pot - stężenie dwutlenku siarki.

(a) Oblicz współczynnik korelacji próbkowej pomiędzy zmienną *Mortality* a zmienną *Education* .

(b) Dokonaj permutacji zmiennej *Mortality* i oblicz współczynnik korelacji między spermutowaną zmienną, a zmienną *Education* . Powtórz punkt $k = 100000$ razy.

(c) Narysuj histogram uzyskanych korelacji. Nanieś na uzyskany wykres korelację z punktu **(a)**.

(d) Przeprowadź test permutacyjny. Czy zmienne *Mortality* i *Education* są skorelowane?

(e) Powtórz wcześniejsze podpunkty dla zmiennych *JulyTemp* i *S02Pot* .

2.2 Rozważmy trzy zmienne losowe:

- $Z \sim N(0, 1)$
- $X = 2Z + N_x, \quad N_x \sim N(0, 1)$
- $Y = -5Z + N_y, \quad N_y \sim N(0, 1)$

(a) Oblicz współczynnik korelacji i korelacji częściowej dla zmiennych X i Y.

(b) Wygeneruj próbkę z rozkładu zmiennej Z, X i Y. Oblicz współczynnik korelacji próbkowej i częściowej korelacji próbkowej między zmiennymi X i Y, można skorzystać z funkcji `pcor` z pakietu `ppcor`.

(c) Jaka jest różnica między współczynnikiem korelacji próbkowej i częściowej korealcji próbkowej dla zmiennych Y i $V = Z^2 + N_v, \quad N_v \sim N(0, 1)$. Wyjaśnij uzyskane wyniki.

2.3 Wygeneruj chmurę punktów w następujący sposób. Niech x_i będą punktami z zakresu $[0, 10]$ odległymi o 0.1, natomiast $y_i = x_i + \epsilon_i$, gdzie ϵ_i to zmienne losowe z rozkładu normalnego o średniej 0 i odchyleniu standardowym $\sigma = 3$. Wykonaj wykres rozproszenia (x_i, y_i).

(a) Oblicz współczynnik korelacji próbkowej korzystając z definicji oraz funkcji `cor()` w pakiecie R.

(b) Oblicz współczynniki prostej MNK korzystając z definicji oraz funkcji `lm()` w pakiecie R.

(c) Nanieś otrzymaną prostą MNK na wykres rozproszenia.

(d) Powtórz procedurę z punktów (a)-(c) dla $\sigma = 0.5$, $\sigma = 5$. Co możemy powiedzieć o wartości współczynnika korelacji próbkowej oraz współczynnikach prostej MNK?

2.4 W pakiecie R w bibliotece MASS znajduje się zbiór danych *hills* dotyczących biegów przełajowych, które odbyły się w Szkocji w 1984 roku. Zawiera on trzy zmienne:

time - rekordowy czas pokonania trasy (w minutach),

dist - długość trasy w milach (na mapie),

climb - całkowita różnica poziomów do pokonania na trasie (w stopach).

(a) Porównaj (wyświetlając w jednym oknie) wykresy rozproszenia *time* od *dist* i *time* od *climb*. Oblicz współczynniki korelacji *time* i *dist* oraz *time* i *climb*.

(b) Dopasuj proste MNK opisujące zależność *time* od *dist* oraz *time* od *climb*. Nanieś dopasowane proste na wykresy rozproszenia. Oblicz R^2 dla otrzymanych modeli posługując się dwiema metodami:

- z definicji, wyznaczając wartości SST, SSR i SSE,
- korzystając z funkcji `summary()`.

(c) Jaki rekordowy czas pokonania trasy o długości 15 mil przewidzimy posługując się prostą MNK?

2.5 Wczytaj i wyświetl na ekranie zbiór *anscombe_quartet.txt*.

(a) Do każdej z czterech par zmiennych dopasuj prostą MNK.

(b) Porównaj otrzymane współczynniki dopasowanych prostych MNK, współczynniki R^2 i współczynniki korelacji.

(c) W jednym oknie narysuj 4 wykresy rozrzutu Y_i od X_i, $i = 1, 2, 3, 4$. W którym przypadku możemy mówić o przybliżonej zależności liniowej y od x?