Laboratorium 2 - Metoda MNK

2.1 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?

2.2 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?

2.3 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?

2.4 Wygeneruj dane w następujący sposób:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon,$$

gdzie $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_3 = 0.75$. Niech $x_{i,1}, x_{i,2}, x_{i,3} \sim N(0, 1)$ i $\varepsilon_i \sim N(0, 1)$.

a) Wygeneruj zbiór danych dla $n = 100$ obserwacji.

b) Wyznacz macierz eksperymentu.

c) Wyznacz rozkład QR macierzy eksperymentu.

d) Udowodnij, że $R\hat{\beta} = Q'y$.

e) Używając powyższego równania wyznacz wartość estymatora wektora współczynników $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]$