Wieloraka regresja liniowa

Regresję wielokrotną stosuje się do oszacowania związku między dwiema lub większą liczbą zmiennych niezależnych a jedną zmienną zależną. Wielokrotnej regresji liniowej możemy użyć, gdy chcemy sprawdzić jak silna jest zależność między dwiema lub więcej zmiennymi niezależnymi a jedną zmienną zależną lub w sytuacji gdy chcemy znać wartość zmiennej zależnej przy określonej wartości zmiennych niezależnych.

Model regresji wielorakiej (Multiple linear regression) :

$$\boldsymbol{Y = a_o + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + \varepsilon}$$

Aby znaleźć najlepiej dopasowaną linię regresji dla każdej zmiennej niezależnej, metodą wielokrotnej regresji liniowej obliczane są:

- współczynniki regresji
- statystyka **t** całego modelu.
- wartość **p**

1. Wczytaj plik 'choroby_serca.txt'. Będziemy sprawdzać, czy istnieje liniowa zależność między jazdą na rowerze do pracy, paleniem a chorobami serca na podstawie badań przeprowadzonych w 500 miastach. Wartości podane w pliku przedstawiają odsetek w ludzi w danym mieście jeżdżących do pracy rowerem, palących papierosy i zdiagnozowanych na choroby serca.
2. Pierwszym krokiem, który chcemy wykonać przed utworzeniem modelu regresji będzie sprawdzenie, czy nasze dane spełniają podstawowe założenia regresji liniowej.
    a. Oblicz współczynnik korelacji pomiędzy zmiennymi niezależnymi. Czy nie są one zbyt silnie skorelowane?
    b. Sprawdź, czy zmienna zależna ma rozkład normalny.
    c. Czy zależność pomiędzy zmiennymi niezależnymi a zmienną zależną ma charakter liniowy?
3. Utwórz model liniowy aby zbadać wpływ zmiennych niezależnych (procent ludzi jeżdżących rowerem oraz palących papierosy) na zmienną zależną (odsetek osób z chorobami serca).
4. Wyświetl wynik modelu i odpowiedz na pytania. Każdą odpowiedź uzasadnij.
    a. Czy możemy odrzucić hipotezę zerową mówiącą, że jazda na rowerze do pracy, jak i palenie tytoniu nie mają wpływu na częstość występowania chorób serca?
    b. Zinterpretuj obliczone współczynniki modelu. Jaka procentowa zmiana zachorowalności na serce wiąże się z każdym procentowym wzrostem liczby osób podróżujących rowerem do pracy? A jaka zmiana przypada na każdy procent wzrostu liczby palących?
    c. Ile wynoszą metryki dopasowania modelu- RSS, RSE oraz R^2 ?
5. Czy nasz model spełnia założenie homoscedastyczności (występowania stałej wariancji reszt dla wszystkich obserwacji) ?
6. Przedstaw na wykresie linię regresji.