

Ćwiczenia. Plik drukarki.txt zawiera dane na temat konserwacji drukarek. Druga kolumna zawiera liczbę kopiarek a kolumna pierwsza zawiera czas (w godzinach) potrzebny na konserwację tej liczby kopiarek.

- (a) Narysuj dane
- (b) Czy relacja między zmiennymi jest w przybliżeniu liniowa?
- (c) Narysuj histogram rozkładu czasu potrzebnego na konserwację drukarek.
- (d) Skonstruuj model regresji liniowej opisujący zależność pomiędzy liczbą kopiarek a czasem potrzebnym na konserwację:
 - i) Napisz wyestymowany model regresji.
 - ii) Przedstaw histogram wartości resztowych.
 - iii) Narysuj dane wraz z dopasowaną prostą regresji.
 - iv) Przedstaw i zinterpretuj wykres kwantyl-kwantyl.
 - v) Wyznacz graficznie 95% przedziały ufności i przedziały predykcyjne, bazując na wybranym podzbiorze obserwacji.
 - vi) Jaka jest dolna granica przedziału ufności dla trzeciej obserwacji?
 - vii) Wyznacz macierz korelacji i podaj wartość współczynnika korelacji Pearsona.
 - viii) Czy prawdziwa wartość współczynnika korelacji jest istotnie różna od zera?
 - ix) Podaj 95% przedział ufności dla współczynnika korelacji Pearsona.

Zadania

Z1. Dane kfm (pakiet ISwR): 50 wierszy i 7 kolumn. Zostały zebrane przez Kim Fleischer Michaelsen i zawierają dane dla 50 niemowląt w wieku w przybliżeniu 2 miesięcy. Dzieci były ważone tuż przed i tuż po karmieniu piersią. Zbadano też ilość spożytego mleka oraz inne dane:

<code>no</code>	numer identyfikacyjny
<code>dl.milk</code>	ilość spożytego mleka (dl/24h).
<code>sex</code>	płeć, zmienna czynnikowa o dwóch poziomach 'boy' i 'girl'.
<code>weight</code>	waga dziecka (kg)
<code>ml.suppl</code>	dodatkowy substytut mleka (ml/24h)
<code>mat.weight</code>	waga matki (kg)
<code>mat.height</code>	wzrost matki (cm)

(a) Przeanalizuj i zinterpretuj wpływ zmiennych objaśniających `sex`, `weight`, `ml.suppl`, `mat.weight`, `mat.height` na ilość spożytego mleka (`dl.milk`) przy użyciu modelu regresji wielokrotnej.

(b) Zauważ, że zmienna 'płeć' (sex) jest zmienną o dwóch poziomach: boy (chłopiec) i 'girl' (dziewczynka). Jakie ma to znaczenie dla przeprowadzonej analizy?

(c) Przedstaw i zinterpretuj wykres zależności parami oraz macierz korelacji między zmiennymi.

(d) Przyjmując poziom istotności równy 0.05 dokonaj krokowej procedury wyboru modelu startując z modelu zawierającego wszystkie zmienne i eliminując zmienne nieistotne o największej p-wartości. Napisz i zinterpretuj otrzymany końcowy model.

Z2. Dane iris (zawarte w podstawowym pakiecie danych w R) zawierają informacje o długości i szerokości kielicha (sepal length/ width) i długości i szerokości płatków (petal length/ width) w centymetrach, dla 50 kwiatów pochodzących z trzech gatunków iris: *Iris setosa*, *Iris versicolor*, *Iris virginica* (więcej: Anderson (1935)).

(a) Narysuj histogramy rozkładu długości i szerokości kielicha dla trzech analizowanych gatunków.

(b) Przedstaw wykresy korelacji parami dla długości i szerokości kielicha oraz długości i szerokości płatków (dla każdego gatunku osobno). Jakie widzisz zależności?

(c) Porównaj wartości współczynników korelacji dla długości kielicha i długości płatków. Dla którego z gatunków korelacja ta jest największa? Sprawdź istotność tych współczynników korelacji. Jakie dwie zmienne są najmocniej skorelowane dla poszczególnych gatunków?

(d) Dla gatunku *setosa* przeanalizuj i zinterpretuj wpływ zmiennych objaśniających Petal.Width (szerokość płatków), Sepal.Length (długość kielicha), Sepal.Width (szerokość kielicha) na zmienną Petal.Length (długość płatków) przy użyciu modelu regresji wielokrotnej. Przyjmując poziom istotności równy 0.05 dokonaj krokowej procedury wyboru modelu eliminując zmienne nieistotne o największej p-wartości. Napisz i zinterpretuj otrzymany końcowy model.

Wskazówka: Aby wyodrębnić gatunki można użyć procedury:

```
iris.versicolor <- iris[iris$Species=="versicolor",1:4]
```

```
iris.setosa <- iris[iris$Species=="setosa",1:4]  
iris.virginica <- iris[iris$Species=="virginica",1:4]
```