Analiza regresji (Regression analysis)

Zadania:

1. Plik *realest.txt* zawiera następujące dane na temat domów na przedmieściach Chicago: cena domu (*Price*), liczba sypialni (*Bedroom*), powierzchnia w stopach kwadratowych (*Space*), liczba pokoi (*Room*), szerokość frontu działki w stopach (*Lot*), roczny podatek od nieruchomości (*Tax*), liczba łazienek (*Bathroom*), liczba miejsc parkingowych w garażu (*Garage*) i stan domu (*Condition*, 0-dobry, 1-wymaga remontu). Dopasować liniowy model regresji opisujący zależność ceny domu od pozostałych zmiennych w zbiorze.

   (a) Wyznacz współczynniki modelu i współczynnik dopasowania R^2. Które zmienne są istotne w modelu?

   (b) Jaki wpływ na cenę ma zwiększenie liczby sypialni o 1, kiedy wartości wszystkich pozostałych zmiennych objaśniających są ustalone? Znaleźć uzasadnienie tego pozornie błędnego wyniku. Porównać ten wynik z wynikiem otrzymanym dla modelu linowego opisującego zależność ceny domu jedynie od liczby sypialni.

   (c) Masz dom w tej okolicy, w dobrym stanie, z 3 sypialniami, o powierzchni 1500 stóp kwadratowych, z 8 pokojami, 40 stopami szerokości działki, 2 łazienkami, 1 miejscem w garażu i podatkiem w wysokości 1000 dolarów. Za ile spodziewasz się go sprzedać? Wykonaj predykcje korzystając z definicji oraz funkcji predict.

2. Zbiór *USPop* w bibliotece *car* zawiera informacje o liczbie ludności w USA w latach 1790-2000.

   (a) Używając metody nieliniowych najmniejszych kwadratów (*nls*) dopasować model wzrostu populacji

   $$y_i = \frac{\beta_1}{1 + e^{\beta_2 + \beta_3 x_i}} + \varepsilon_i,$$

   gdzie y_i jest wielkością populacji w roku x_i. Jako wartości początkowe przyjąć: $\beta_1 = 350, \beta_2 = 4.5, \beta_3 = -0.3$

   (b) Obliczyć wartości estymatorów $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$. Na podstawie otrzymanego modelu wrysować krzywą do wykresy rozproszenia zmiennej *population* względem zmiennej *years*.

   (c) Na podstawie modelu dopasowanego w punkcie (a) oszacować jaki będzie stan liczbowy populacji w USA w roku 2015?

3. (a) Wygenerować zbiór danych zgodnie z równaniem:

   $$y = g(x) + \varepsilon,$$

   gdzie: $g(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$, a ε ma rozkład normalny z parametrami $\mu = 0$, $\sigma = 0.1$.

   (b) Przetestować na tym zbiorze działanie trzech estymatorów regresji:

   - jądrowego (**ksmooth**),

- lokalnie wielomianowego (**loess**),
- spline'u kubicznego (**smooth.spline**),

ze swymi parametrami defaultowymi. Nanieść dopasowane krzywe na wykres rozproszenia.

(c) Obliczyć $ISE = n^{-1} \sum_{i=1}^{n} [g(x_i) - \hat{g}(x_i)]^2$ dla wszystkich trzech estymatorów. Zbadać jak ISE zalezy od n.