1 Laboratorium 1 i 2.

Wczytaj do jakiegoś pakietu statystycznego dane $(x_1, y_1), \ldots, (x_n, y_n)$ z pliku lab1.txt.

1. Dla każdej ze zmiennych y i x pojawiających się w pliku wyznacz podstawowe wskaźniki numeryczne charakteryzujące próbę: średnią, wariancję, odchylenie standardowe, medianę, pierwszy i trzeci kwartyl, minimum i maksimum. Skonstruuj też histogramy i box-ploty.

2. Wykonaj wykres rozproszenia zmiennych y i x i oblicz współczynnik korelacji próbkowej tych zmiennych. Czy chmura punktów na tym wykresie ma (w przybliżeniu) charakter liniowy? Uzasadnij dlaczego można wykorzystać model regresji liniowej

   $$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

   do opisu zależności między zmiennymi y i x?

3. Wyznacz wartości estymatorów najmniejszych kwadratów $\widehat{\beta_0}$ i $\widehat{\beta_1}$ parametrów β_0 i β_1.

4. Znajdź wartość estymatora $\widehat{\sigma}^2 := \dfrac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}$ parametru σ^2.

   Uwaga. Symbol $\widehat{y}_i$ oznacza prognozowaną przez model wartość zmiennej objaśnianej y, odpowiadającą wartości x_i zmiennej objaśniającej x, czyli $\widehat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_i$.

5. Na poziomie istotności $\alpha = 0,05$ zweryfikuj hipotezę $H_0 : \beta_1 = 0$ przy hipotezie alternatywnej $H_1 : \beta_1 \neq 0$. Jaka jest p-wartość dla tego testu? Czy na podstawie tych wyników można stwierdzić, że rozpatrywany w tym przykładzie model regresji liniowej ma sens?

   Uwaga. Niech $SE^2_{\beta_1} := \dfrac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Jeśli prawdziwa jest hipoteza zerowa, a błędy $\varepsilon_1, \ldots, \varepsilon_n$ są niezależne i mają ten sam rozkład $N(0, \sigma^2)$, to statystyka $T := \dfrac{\widehat{\beta_1}}{SE_{\beta_1}}$ ma rozkład t_{n-2} (t-Studenta z $n-2$ stopniami swobody). Na poziomie istotności α test odrzuca H_0, gdy $|T| \geq t_{\alpha/2,n-2}$ ($t_{\alpha/2,n-2}$ to kwantyl rzędu $1 - \alpha/2$ rozkładu t_{n-2}).

6. Skonstruuj przedział ufności dla β_1 na poziomie ufności 0.99.

   Uwaga. Jeśli $\varepsilon_1, \ldots, \varepsilon_n$ *i.i.d.* $N(0, \sigma^2)$ to przedział ufności na poziomie ufności $1 - \alpha$ dla współczynnika β_1 ma postać $\widehat{\beta_1} \pm t_{\alpha/2,n-2} \times SE_{\beta_1}$.

7. Dla $x_0 = 1$ oblicz prognozowaną przez model wartość $\widehat{Y}(x_0)$. Następnie wyznacz przedziały ufności na poziomie ufności $0,99$ dla $Y(x_0)$.

   Uwaga. Jeśli $\varepsilon_1, \ldots, \varepsilon_n$ *i.i.d.* $N(0, \sigma^2)$, to przedział ufności na poziomie ufności $1-\alpha$ dla prognozowanej przez model wartości $Y(x_0)$ ma postać

   $$\widehat{Y}(x_0) \pm t_{\alpha/2,n-2} \times SE_{\widehat{Y}(x_0)-Y(x_0)},$$

   gdzie $\widehat{Y}(x_0) = \widehat{\beta_0} + \widehat{\beta_1} x_0$ i

   $$SE^2_{\widehat{Y}(x_0)-Y(x_0)} = \widehat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

8. Narysuj histogram i wykres kwantylowy dla rezyduów. Narysuj też wykresy rozproszenia dla prób $(\widehat{y}_1, e_1), \ldots, (\widehat{y}_n, e_n)$ i $(x_1, e_1), \ldots, (x_n, e_n)$. Czy na podstawie tych rysunków można stwierdzić, że któreś założeń modelu regresji nie jest spełnione (patrz zadanie teoretyczne 3.)?

9. Zmodyfikuj dane z pliku lab1.txt, przyjmując, że ostatnią obserwacją jest $6,9347$ i $148,6400$, a nie $6,9347$ i $14,8640$. Wyznacz estymatory najmniejszych kwadratów parametrów β_0 i β_1 i porównaj je z estymatorami otrzymanymi poprzednio. Co zauważasz? Czy obserwacja $(6,9347; 148,640)$ jest a) odstająca, b) wpływowa?

10. Niech $n = 100$, $x_1, x_2, \ldots, x_n$ iid $U(0,1)$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ iid $N(0, \sigma^2)$, $\sigma = 0.1$, $\beta_0 = 1$, $\beta_1 = 2$.

    (a) Wygeneruj obserwacje $y_1, y_2, \ldots, y_n$, takie że

    $$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

    (b) Wykonaj wykres rozproszenia dla próby $(x_1, y_1), \ldots, (x_n, y_n)$. Czy chmura punktów ma (w przybliżeniu) charakter liniowy?

    (c) W oparciu o próbę $(x_1, y_1), \ldots, (x_n, y_n)$ wyznacz estymatory najmniejszych kwadratów $(\widehat{\beta}_0, \widehat{\beta}_1)$ parametrów (β_0, β_1). Porównaj wartości tych estymatorów z wartościami (β_0, β_1)?

    (d) Powtórz obliczenia dla $\sigma = 0.5$ i $\sigma = 1$. Jak zmieniają się precyzja estymatorów i współczynnik R^2, gdy rośnie σ^2?

Zadania teoretyczne. Rozważmy model regresji liniowej z jedną zmienną objaśniającą $y = \beta_0 + \beta_1 \cdot x + \varepsilon$, tworzony na podstawie próby $(x_1, y_1), \ldots, (x_n, y_n)$.

1. Wyprowadź wzory na estymatory najmniejszych kwadratów $\widehat{\beta}_0$ i $\widehat{\beta}_1$ parametrów β_0 i β_1, wiedząc że $\widehat{\beta}_0$ i $\widehat{\beta}_1$ są rozwiązaniem następującego problemu minimalizacyjnego:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \left[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right].$$

2. Uzasadnij, przeprowadzając odpowiednie rachunki, że $\widehat{\beta}_1$ można zapisać w równoważnej postaci $\widehat{\beta}_1 = r \dfrac{s_y}{s_x}$, gdzie r **jest współczynnikiem korelacji próbkowej Pearsona**, a s_y^2 i s_x^2 oznaczają wariancje próbkowe w próbach y-ów i x-ów.

3. Niech $e_1, \ldots, e_n$ oznaczają kolejne rezydua, tzn. niech $e_i := y_i - \widehat{y}_i$, $i = 1, \ldots, n$. Uzasadnij, przeprowadzając odpowiednie rachunki (!), następujący fakt: Jeśli model regresji liniowej poprawnie opisuje zależność między zmiennymi y i x, to

    - współczynnik korelacji próbkowej dla próby $(\widehat{y}_1, e_1), \ldots, (\widehat{y}_n, e_n)$ jest równy zero;
    - współczynnik korelacji próbkowej dla próby $(x_1, e_1), \ldots, (x_n, e_n)$ jest równy zero;

    Jak powinny wyglądać wykresy rozproszenia dla tych prób?