1 Laboratoria 4, 5 i 6

Wczytaj do pakietu statystycznego dane ze zbioru 'regresja wielokrotna.xlsx'.

1. Dla każdej z $\binom{11}{2}$ par utworzonych ze zmiennych $Y, X_1, \ldots, X_{10}$ wykonaj wykres rozrzutu* i po przeanalizowaniu tych rysunków odpowiedz na następujące pytania:

   (a) Które ze zmiennych objaśniających $X_1, \ldots, X_{10}$ mogą mieć najmocniejszy liniowy wpływ na zmienną objaśnianą Y?

   (b) Czy pojawia się problem współliniowości, to znaczy, czy istnieje choć jedna para silnie ze sobą skorelowanych zmiennych objaśniających?

   (c) Czy pojawiają się obserwacje odstające?

   * to polecenie można wykonać za pomocą jednej komendy.

2. Wyznacz macierz korelacji próbkowych dla zmiennych $Y, X_1, \ldots, X_{10}$ i po przeanalizowaniu tak otrzymanych współczynników ponownie odpowiedz na pytania (a) i (b) z poprzedniego punktu.

3. Skonstruuj model regresji liniowej opisujący zależność między zmienną Y a zmiennymi objaśniającymi $X_1, \ldots, X_{10}$.

   (a) Wyznacz estymator najmniejszych kwadratów $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_{10})^T$.

   (b) Czy którakolwiek ze zmiennych objaśniających z tego (pełnego) modelu ma liniowy wpływ na zmienną objaśnianą? Odpowiedź uzasadnij podając p-wartość testu F.

   (c) Wyznacz współczynniki determinacji R^2 i $AdjR^2$.

4. Rozwiąż problem współliniowości.

   (a) Spośród zmiennych objaśniających, dla których VIF przekracza 10 (lub równoważnie $TOL := 1/VIF < 0,1$), wybierz tę z największą wartością VIF i usuń ją z modelu.

   (b) Oblicz *wskaźniki podbicia wariancji* w modelu regresji zawierającym **pozostałe** zmienne objaśniające. Jeśli któryś z tych wskaźników jest większy od 10 wróć do poprzedniego punktu.

5. Zidentyfikuj i ewentualnie usuń z próby obserwacje, które mogą być wpływowe. W tym celu przeanalizuj

   (a) *wpływy* (*leverages*) kolejnych obserwacji, czyli liczby $h_{11}, \ldots, h_{nn}$ tworzące główną przekątna macierzy $\boldsymbol{H}$,

   (b) *odległości Cooke'a* $D_1, \ldots, D_n$,

   (c) *studentyzowane rezydua* $r_1, \ldots, r_n$,

   (d) $DFFITS_1, \ldots, DFFITS_n$.

   By ułatwić identyfikację obserwacji wpływowych wykonaj wykres rozproszenia dla punktów $(1, D_1), (2, D_2), \ldots, (n, D_n)$, umieszczając na nim próg odcięcia (np. $y = \frac{4}{n-p}$).

6. Wykorzystując zmienne i obserwacje, które nie zostały usunięte, ponownie zbuduj model regresji liniowej opisujący zależność między zmienną Y a zmiennymi objaśniającymi.

   (a) Wyznacz estymator najmniejszych kwadratów $\widehat{\boldsymbol{\beta}}$.

   (b) Czy którakolwiek ze zmiennych objaśniających z tego (pełnego) modelu ma liniowy wpływ na zmienną objaśnianą? Odpowiedź uzasadnij podając p-wartość testu F.

   (c) Wyznacz współczynniki determinacji R^2 i $AdjR^2$. Czy po usunięciu niektórych zmiennych lub obserwacji polepszyło się dopasowanie modelu do danych?

7. Wykorzystaj **regresję krokową**, opcje *forward* i *backward* (w pakiecie Statistica opcje: *metoda wprowadzania postępującego* i *metoda eliminacji wstecz*) do wyboru podzbioru zmiennych objaśniających „najlepiej" opisującego liniowy wpływ zmiennych objaśniających na zmienną Y. Wykorzystaj także inne opcje regresji krokowej, dostępne w używanym przez Ciebie pakiecie (w Statistice opcje *metoda krokowa postępująca* i *metoda krokowa wstecz*). Oczywiście, przy tej analizie użyj zmodyfikowanych danych, powstałych po usunięciu niektórych zmiennych i niektórych obserwacji (punkty 4. i 5.).

   Uwaga: W ten sposób można otrzymać różne modele, więc do dalszej analizy **wybierz jeden z nich** (za pomocą współczynnika C_p Mallowsa albo skorygowanego R^2) i nazwij go *modelem M*.

   (a) Wyznacz estymator najmniejszych kwadratów $\widehat{\boldsymbol{\beta}}$ w modelu *M*.

   (b) Czy którakolwiek ze zmiennych objaśniających z *modelu M* ma liniowy wpływ na zmienną objaśnianą? Odpowiedź uzasadnij podając p-value (p-wartość) odpowiedniego testu.

   (c) Dla każdej ze zmiennych objaśniających, które znalazły się w *modelu M*, sprawdź, czy ma ona liniowy wpływ na zmienną objaśnianą, gdy w modelu uwzględnione zostały pozostałe zmienne. Podaj p-wartość odpowiedniego testu i sformułuj wniosek.

   (d) Wyznacz przedział ufności na poziomie ufności 0.95 dla współczynników regresji, odpowiadających zmiennym z *modelu M*.

   (e) Wyznacz współczynniki determinacji R^2 i $AdjR^2$ w modelu *M*.

8. Przeanalizuj zachowanie reszt w *modelu M*, by sprawdzić czy spełnione są założenia występujące w modelu regresji liniowej (tzn. czy błędy pochodzą z rozkładu normalnego, mają średnią zero i tę samą wariancję). W tym celu należy wykonaj

   (a) wykresy kwantylowe dla reszt,

   (b) wykresy reszt względem każdej ze zmiennych objaśniających,

   (c) wykresy reszt względem wartości przewidywanych przez model,

9. Wyznacz przewidywaną przez *model M* wartość zmiennej objaśnianej Y, gdy zmienne objaśniające $X_1, X_2, \ldots, X_{10}$ mają wartość $1, 2, \ldots, 10$.

Zadania teoretyczne:

1. Udowodnij, że $\boldsymbol{P} := \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ jest macierzą symetryczną, taką że $\boldsymbol{P}^2 = \boldsymbol{P}$. Dla ustalonego wektora $\boldsymbol{y} \in \mathbb{R}^n$ wyznacz wektor $\boldsymbol{Py}$.

   Uwaga. $\mathbf{1}_n \in \mathbb{R}^n$ oznacza wektor złożony z samych jedynek, a $\boldsymbol{I}_n$ to macierz jednostkowa stopnia n.

2. Załóżmy, że w modelu regresji liniowej

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$$

   macierz eksperymentu $\boldsymbol{X}$ jest rzędu pełnego. Dla każdego wektora $\boldsymbol{l} \in \mathbb{R}^p$ znajdź wektor $\boldsymbol{a} \in \mathbb{R}^n$, taki że

$$\mathbb{E}(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{l}^T\boldsymbol{\beta} \quad \text{dla wszystkich} \quad \boldsymbol{\beta} \in \mathbb{R}^p. \quad (*)$$

   Czy istotne jest założenie, że $\boldsymbol{X}$ jest macierzą rzędu pełnego? Jeśli tak, to podaj przykład macierzy $\boldsymbol{X}$ i wektora $\boldsymbol{l}$, dla których (*) nie zachodzi.

   Uwaga. Dla macierzy $\boldsymbol{X}$ wymiaru $n \times p$, z $n \geq p$, określenie *macierz rzędu pełnego* oznacza, że rząd$(\boldsymbol{X}) = p$ (kolumny $\boldsymbol{X}$ są liniowo niezależne). Wówczas $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ istnieje, ale $\boldsymbol{X}$ jest odwracalna wtedy i tylko wtedy, gdy $n = p$.

3. Podaj postać każdej z macierzy $\boldsymbol{X}$, $\boldsymbol{X}^T\boldsymbol{X}$ i $\boldsymbol{H}$ dla modelu regresji liniowej z jedną zmienną objaśniającą i z danymi $(x_1, y_1), \ldots, (x_n, y_n)$.