Laboratorium statystyczne 2

Regresja wielokrotna, diagnostyka modelu

4. Dane: *pyłki*.

a. Zbuduj model liniowy pełny (wraz z interakcją): pyłek jako funkcja częstotliwości odwiedzin (czyli czas^(-1)) i rodzaju owada, który zbierał pyłki. Czy interakcja jest istotna?

b. Zbuduj taki sam model bez interakcji (model addytywny) (*model0*).

c. Uzasadnij (testem ANOVA) dlaczego model addytywny jest wystarczający.

d. Przeprowadź diagnostykę modelu *model0* (skrypt *wykres diagnostyczny.R*). Wskaż przypadki, które przekraczają graniczną wartość wpływu. Skomentuj wartości zmiennych, opisujących ten przypadek. Który z nich przekracza maksymalnie dopuszczalną odległość Cooka? Jakie są skutki zachowania tego przypadku do dalszej analizy?

e. Zbuduj model addytywny bez tego punktu (*model1*). Które współczynniki równania najbardziej się zmieniły? Jak zmieniły się p-wartości testu istotności współczynników regresji?

f. Przeprowadź diagnostykę modelu *model1*. Wskaż przypadki, które przekraczają graniczną wartość wpływu. Skomentuj wartości zmiennych, opisujących ten przypadek. Który z nich przekracza maksymalnie dopuszczalną odległość Cooka? Czy któryś z nich był wpływowy dla modelu *model0* ? (Uwaga na numerację punktów modelu *model1* – został zbudowany bez punktu 1!)

g. Zbuduj *model 2* po usunieciu wszystkich punktów wpływowych. Czy zauważasz jakąś istotną zmianę w stosulu do modeli *model0* i *model1*?

h. Narysuj wykres frakcji zebranego pyłku jako funkcji częstotliwości zbierania dla *wszystkich* danych i na tym tle narysuj równanie regresji (*model2*). Spróbuj zlokalizować trzy punkty wpływu. Czy wykres wizualnie pasuje do danych, z wyjątkiem tych punktów?

i. Wyznacz dla pełnego kompletu danych przybliżoną prostą regresji z metody strzałki Tukeya (skrypt *strzalka Tukeya.R*). Zauważ, że reguła stopu (|błąd|<0.1) daje przekształcenie o potędze -1. Dołącz wykres tej prostej do poprzedniego i oceń która z tych najlepiej pasuje do danych. Przybliżona prosta regresji z metody strzałki Tukeya nazywana jest prostą odporną (np. Jobson). Czy w tym przypadku zasługuje na swoją nazwę?

j. Przeprowadź te same obliczenia, wyrażając zmienną pyłek w skali logitowej (nowa zmienna lpyl). Czy jest model w skali logitowej jest lepszy?

6. Dane *nino*

Zbuduj model sztorm.modelF2 zależności logarytmu wskaźnika sztormu z interakcjami do drugiego rzędu (postać: y ~ (x1+x2+…)^2) dla zmiennych objaśniających el_nino, west.africa, sztormy, huragany.

a. Oblicz[1] dla tego modelu wskaźnik VIF[2]. Jaka jest interpretacja $\sqrt{VIF}$? Skomentuj wielkość VIF dla poszczególnych zmiennych przyjmując wartość progową 10 i skonfrontuj z p-wartością testu na istotność współczynniku w modelu zależności logarytmu wskaźnika sztormu z interakcjami do drugiego rzędu.

b. Przeprowadź metodą krokową AIC oraz BIC wybór zmiennych dla tego modelu. Porównaj ostateczne modele uzyskane w tych modelach. Który z nich jest bardziej korzystny przy prognozowaniu, a który przy opisie relacji wskaźnik sztormu – zmienne objaśniające?

c. Przeanalizuj wskaźnik VIF dla modelu ostatecznego, uzyskanego metodą BIC. Zauważ, że interakcja zmiennych *sztormy* i *huragany* ma bardzo wysoki wskaźnik VIF.

d. Zbuduj model addytywny, bez interakcji zmiennych *sztormy* i *huragany*. Oceń testem ANOVA który model wybrać: z interakcją czy bez interakcji?

---

[1] W R jest to funkcja vif{car}

[2] W R, gdy w równaniu występują czynniki, używany jest uogólniony wskaźnik GVIF. Szczególnie ważny gdy w modelu występują czynniki (jak w tym przykładzie)(p.też http://www.quantoid.net/lab2.pdf.) Jako wartość progową w tym przypadku przyjmuje się wartość 2, czasami 3

e. Oceń metodą Cp Mallowsa modele wskaźnika sztormu uzyskane metodą krokową AIC oraz BIC. Który z tych modeli jest lepszy względem tego kryterium? Wykorzystaj procedurę mle.cp{wle}

f. Znajdź najlepsze (wg kryterium AIC) podmodele modelu *sztorm.modelF2* (procedura regsubsets {leaps}) dla liczby zmiennych nie przekraczających 5

g. Oblicz dla każdego z nich współczynnik determinacji (rsq), resztową sumę kwadratów (rss), skorygowany współczynnik determinacji (adjr2), C_p Mallowsa (cp), współczynnik BIC (bic). Wybierz najlepszy model.

h. Oblicz współczynniki tego modelu.