Statystyka z modelami liniowymi

Lista 5 - Regresja liniowa wieloraka

Zadanie 1: ***W tym zadaniu zostaną wykorzystane dane z pliku CH06PR15.txt. Kolejne kolumny zawierają: wiek pacjenta (pierwsza kolumna) oraz wyniki opisujące ciężkość choroby, poziom niepokoju pacjenta i poziom satysfakcji pacjenta.***

a) Wyznacz regresję liniową przyjmując wiek, ciężkość choroby i poziom niepokoju jako zmienne objaśniające, a satysfakcję jako zmienną odpowiedzi. Podaj równanie regresji liniowej prostej. Oblicz estymatory nieznanych parametrów za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*.

b) Podaj wartość R^2 oraz wynik testu istotności dla hipotezy zerowej, że trzy zmienne objaśniające nie są związane z zmienną odpowiedzi (podaj hipotezę zerową i alternatywną, statystykę testową ze stopniami swobody, p-wartość oraz krótką konkluzję). Wartości R^2, statystyki testowej oraz p-wartości oblicz za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*.

c) Podaj osobne 95% przedziały ufności dla współczynników regresji wieku, ciężkości choroby i poziomu niepokoju. Opisz wyniki testów istotności dla poszczególnych współczynników regresji (podaj hipotezę zerową i alternatywną, statystykę testową ze stopniami swobody, p-wartość i krótką konkluzję). Przedziały ufności, wartości statystyk testowych oraz p-wartości oblicz za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*. *Jaki jest związek między wynikami testów istotności a przedziałami ufności ?*

d) Zaznacz na wykresie reszty względem przewidywanej satysfakcji i każdej z zmiennych objaśniających. Czy zauważasz jakieś nietypowe tendencje lub wartości odstające? Czy reszty są w przybliżeniu mają rozkład normalny? Użyj testu Shapiro-Wilka i wykresów (histogram i qqplot).

e) Użyj ogólnego testu F, żeby porównać modele z jedną i trzema zmiennymi objaśniającymi (podaj hipotezę zerową i alternatywną, statystykę testową ze stopniami swobody, p-wartość i krótką konkluzję). Wartości statystyk testowych oraz p-wartości oblicz za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*.

Zadanie 2: *Wpływ wymiaru.*

a) Wygeneruj macierz planu $X_{1000\times 950}$ tak, aby jej elementy były niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma = 0.1)$. Następnie wygeneruj wektor zmiennej odpowiedzi zgodnie z modelem

$$Y = X\beta + \varepsilon,$$

gdzie $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$, $\varepsilon \sim N(0, I)$.

b) Wyestymuj wartości współczynników regresji i użyj testów T na poziomie istotności 0.05, aby zidentyfikować prawdziwe regresory, gdy model jest konstruowany przy użyciu pierwszych k kolumn macierzy planu dla $k \in \{1, 2, 5, 10, 50, 100, 500, 950\}$. Dla każdego z tych modeli podaj

- sumę kwadratów wartości resztowych $SSE = \left\|Y - \hat{Y}\right\|^2$;
- średni błąd kwadratowy estymatora wartości oczekiwanej Y: $MSE = \left\|X(\hat{\beta} - \beta)\right\|^2$;
- wartość statystyki AIC: $AIC = n \log(SSE/n) + 2k$;
- p-wartości odpowiadające dwóm pierwszym zmiennym objaśniającym;
- liczbę fałszywych odkryć.

Który model zostanie wybrany na podstawie AIC? (model, którego wartość statystyki AIC jest minimalna).

c) Powtórz punkt b), gdy modele są konstruowane przy użyciu zmiennych o największych (nie pierwszych) estymowanych współczynnikach regresji. Porównaj wartości obliczonych statystyk z wartościami uzyskanymi w punkcie b). Który model zostałby wybrany na podstawie AIC ?

d) Powtórz generowanie Y oraz kroki b) i c) 1000 razy. Dla każdego z podproblemów oszacuj moc identyfikacji X_1, X_2 i oczekiwaną liczbę fałszywych odkryć. Dodatkowo porównaj średnią rozmiar modelu wybranego przez AIC dla punktów b) i c).