Modele Linowe

Lista 4

1. *Wpływ korelacji.*

   a) Wygeneruj macierz $X_{100\times 2}$ taką, że jej wiersze są niezależnymi wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma/100)$, gdzie

   $$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

   Następnie wygeneruj wektor zmiennej odpowiedzi postaci $Y = \beta_1 X_1 + \varepsilon$, gdzie $\beta_1 = 3$, X_1 to pierwsza kolumna X, a $\varepsilon \sim N(0, I)$.

   b) Wyznacz 95% przedział ufności dla wartości β_1 i przeprowadź $t-$test na poziomie istotności $0,05$ dla hipotezy $\beta_1 = 0$ przy użyciu

   - modelu prostej regresji liniowej $Y = \beta_0 + \beta_1 X_1 + \varepsilon$;
   - modelu z dwiema zmiennymi objaśniającymi $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

   Porównaj i omów wyniki.

   c) Oblicz ręcznie odchylenie standardowe estymatora β_1 i moc identyfikacji X_1 w obu modelach.

   d) Wygeneruj 1000 niezależnych kopii wektora błędów losowych ε i 1000 odpowiednich kopii wektora zmiennej odpowiedzi. Dla każdego z tak wygenerowanych zbiorów danych wyznacz estymator β_1 i wykonaj test istotności dla β_1 w obu modelach (z jedną i dwoma zmiennymi objaśniającymi). Wyestymuj odchylenie standardowe β_1 oraz moc testu i porównaj te wartości z teoretycznymi wynikami uzyskanymi w punkcie c).

2. *Wpływ wymiaru.*

   a) Wygeneruj macierz planu $X_{1000\times 950}$ tak, że jej elementy są niezależnymi zmiennymi losowymi z rozkładu normalnego $N(0, \sigma = 0.1)$. Następnie wygeneruj wektor zmiennej odpowiedzi według modelu

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

   gdzie $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$.

   b) Wyestymuj wartości współczynników regresji i wykonaj $t-$ testy na poziomie istotności 0.05, aby zidentyfikować istotne regresory, jeżeli model jest zbudowany przy użyciu pierwszych k kolumn macierzy planu dla $k \in \{1, 2, 5, 10, 50, 100, 500, 950\}$.

   Dla każdego z tych modeli należy podać

- sumę kwadratów residiów $SSE = \left\| Y - \hat{Y} \right\|^2$;
- błąd średniokwadratowy estymatora wartości oczekiwanej Y: $MSE = \left\| X(\hat{\beta} - \beta) \right\|^2$;
- wartość kryterium AIC: $AIC = n \log\left(SSE/n\right) + 2k$;
- p-wartości dla dwóch pierwszych zmiennych objaśniających;
- liczbę fałszywych odkryć.

Który model należy wybrać na podstawie AIC? (model, który ma minimalną wartość AIC)

c) (+1 pkt) Powtórz punkt b), gdy modele są konstruowane przy pomocy zmiennych o największych (niekoniecznie pierwszych) estymowanych współczynnikach regresji. Porównaj wartości obliczonych statystyk z otrzymanymi w punkcie b). Który model należy wybrać na podstawie AIC?

d) Powtórz generowanie ε i Y oraz punkty b) i c) (jeśli zostało zrobione) 1000 razy. Dla każdego z zadań oblicz moc identyfikacji X_1, X_2 i średnią liczbę fałszywych odkryć. Dodatkowo oszacuj średni rozmiar modelu wybranego przez AIC dla punktów b) i c) (jeśli zostało zrobione).