Statystyka z modelami liniowymi

Lista 4 - Regresja liniowa prosta

Zadanie 1: ***W tym zadaniu zostaną wykorzystane dane z pliku ch01pr20.txt. Druga kolumna zawiera liczbę kopiarek, a pierwsza kolumna zawiera czas (w godzinach) potrzebny na utrzymanie tych kopiarek.***

a) Przedstaw dane na wykresie. Czy zależność jest w przybliżeniu liniowa? Wyznacz regresję liniową z y = *czas obsługi* i x = *liczba obsługiwanych maszyn.* Podaj równanie regresji liniowej prostej. Oblicz slope i intersept za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*. Do wykresu dodaj prostą regresji liniowej.

b) Wyznacz 95% przedział ufności dla slope'a i intersepta: za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*.

c) Przedstaw wyniki testu istotności slope'a i intersepta. Podaj testowane hipotezy, statystyki testowe z liczbą stopni swobody, p-wartości i własne wnioski. Oblicz statystyki testowe oraz p-wartości za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*.

d) Podaj estymator wartości oczekiwanej czasu obsługi, której można oczekiwać, gdyby serwisowanych było k maszyn oraz 95% przedział ufności dla tej wartości, $k \in \{1, 5, 8, 11, 25, 100\}$. Oblicz przedział ufności za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*. Jak długość przedziału ufności zależy od miejsca punktu względem danych?

e) Podaj przewidywany czas obsługi, którego można oczekiwać, jeśli k maszyn było serwisowanych oraz 95% przedział predykcyjny dla tego czasu, $k \in \{1, 5, 8, 11, 25, 100\}$. Oblicz przedział predykcyjny za pomocą wzorów teoretycznych oraz poleceń wbudowanych w *Python*/*R*. Jak długość przedziału predykcyjnego zależy od miejsca punktu względem danych?

f) Dodaj do wykresu z danymi 95% przedziały ufności oraz przedziały predykcyjne dla poszczególnych obserwacji. Wyjaśnij, dlaczego przedziały ufności są zawsze mniejsze niż przedziały predykcyjne.

Zadanie 2: ***W tym zadaniu zostaną wykorzystane dane z pliku*** **tabela1_6.txt*****, zawierający średnią ocen (GPA, druga kolumna), wynik w standardowym teście IQ (trzecia kolumna), płeć (czwarta kolumna) oraz punktację na teście psychologicznym Piers-Harris Children's Self-Concept Scale (PH, piąta kolumna) dla 78 uczniów siódmej klasy.***

a) Użyj prostego modelu regresji, aby opisać zależność GPA od wyników testu IQ. Podaj odpowiednie równanie regresji. Przedstaw dane i prostą regresji na wykresie. Oblicz współczynnik determinacji R^2 za pomocą wzorów teoretycznych oraz poleceń, wbudowanych w *Python*/*R*. Co można wywnioskować o danych na podstawie statystyki R^2?

b) Przetestuj hipotezę, że GPA nie jest skorelowane z IQ na podstawie testu F. Podaj testowaną hipotezę, statystykę testową z liczbą stopni swobody, p-wartość oraz własne wnioski. Oblicz statystykę F oraz p-wartość za pomocą wzoru teoretycznego oraz poleceń wbudowanych w *Python*/*R*.

c) Przewiduj GPA dla uczniów, IQ których wynosi $75, 100, 140$. Podaj 90% przedziały predykcyjne.

d) Dodaj do wykresu z danymi 90% przedziały predykcyjne. Ile obserwacji znajduje się poza tymi przedziałami?

e) Użyj prostego modelu regresji, aby opisać zależność GPA od wyników testu PH. Podaj odpowiednie równanie regresji. Przedstaw dane i prostą regresji na wykresie. Oblicz współczynnik determinacji R^2. Która z dwóch zmiennych: wynik testu IQ czy wynik testu PH, jest lepszym predyktorem GPA?

Zadanie 3: Wygeneruj wektor $X = (X_1, \ldots X_{200})^T$ z wielowymiarowego rozkładu normalnego $\mathcal{N}(0, \frac{1}{500}I)$. Następnie wygeneruj 1000 wektorów Y z modelu $Y = 5 + \beta_1 X + \varepsilon$, gdzie

a) $\beta_1 = 0$, $\varepsilon \sim \mathcal{N}(0, I)$.

b) $\beta_1 = 0$, $\varepsilon_1, \ldots, \varepsilon_{200}$ są iid z rozkładu wykładniczego Exp(1).

c) $\beta_1 = 0$, $\varepsilon_1, \ldots, \varepsilon_{200}$ są iid z rozkładu logistycznego L$(0, 1)$.

d) $\beta_1 = 2$, $\varepsilon \sim \mathcal{N}(0, I)$.

e) $\beta_1 = 2$, $\varepsilon_1, \ldots, \varepsilon_{200}$ są iid z rozkładu wykładniczego z parametrem $\lambda = 1$.

f) $\beta_1 = 2$, $\varepsilon_1, \ldots, \varepsilon_{200}$ są iid z rozkładu logistycznego L$(0, 1)$.

Dla każdego powtórzenia eksperymentu przetestuj hipotezę $H_0 : \beta_1 = 0$ i estymuj prawdopodobieństwo odrzucenia H_0 na podstawie częstości odrzuceń w próbie (osobno dla każdego z punktów (a)-(f)). Porównaj te estymatory prawdopodobieństwa z teoretycznym prawdopodobieństwem błędu I rodzaju (a, b, c) oraz teoretyczną mocą (d, e, f) obliczoną przy założeniu, że szum ma rozkład normalny. Podsumuj wyniki.

Zadanie 4:

W następnych zadaniach zostaną wykorzystane dane z pliku ch03pr15.txt dotyczące stężenia roztworu. Pierwsza kolumna zawiera wartości stężenia roztworu, a druga - czas.

a) Przeprowadź regresję liniową z czasem jako zmienną objaśniającą i stężeniem roztworu jako zmienną odpowiedzi. Podaj odpowiednie równanie regresji. Przedstaw dane i prostą regresji na wykresie. Dodaj do wykresu 95% przedziały predykcyjne dla poszczególnych obserwacji. Jakie są wnioski?

b) Podsumuj wyniki regresji, podając wartość R^2 i wyniki testu istotności dla hipotezy zerowej, że stężenie roztworu nie zależy od czasu (podaj hipotezy zerową i alternatywną w odniesieniu do parametrów modelu, statystykę testową ze stopniami swobody, p-wartość i krótki wniosek). Oblicz współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu.

c) Użyj procedury Box'a-Cox'a, aby znaleźć odpowiednią transformację dla stężenia roztworu.

d) Utwórz nową zmienną odpowiedzi, biorąc logarytm stężenia roztworu (zdefiniuj $\tilde{Y} = \log(Y)$). Powtórz zadanie $a), b)$ z $\tilde{Y}$ jako zmienną odpowiedzi i czasem jako zmienną objaśniającą ($\tilde{Y} \sim time$). Podsumuj swoje wyniki.

e) Na wykresie przedstaw stężenie roztworu względem czasu. Dodaj krzywą regresji i pasmo dla 95% przedziałów predykcji na podstawie wyników uzyskanych w punkcie d). Porównaj z wykresem uzyskanym w zadaniu a. Oblicz współczynnik korelacji między obserwowanym a przewidywanym stężeniem roztworu opartym na danym modelu i porównaj z odpowiednim wynikiem z zadania b)).

e) Skonstruuj nową zmienną objaśniającą $\tilde{t} = time^{-1/2}$. Powtórz zadanie 7, używając modelu regresji ze stężeniem roztworu jako zmienną odpowiedzi i $\tilde{t}$ jako zmienną objaśniającą ($Y \sim \tilde{t}$). Podsumuj wyniki. Który model wydaje się najlepszy i dlaczego?