Modele Linowe

Lista 3

W zadaniach $1-2$ ***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.***

1. **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 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 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?

2. **a)** 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. Co można wywnioskować o danych na podstawie statystyki R^2?

   b) Przetestuj hipotezę, że GPA nie jest skorelowane z PH na podstawie testu F. Podaj testowaną hipotezę, statystykę testową z liczbą stopni swobody, p-wartość oraz własne wnioski.

   c) Przewiduj wyniki GPA dla uczniów, wyniki testu PH których wynoszą $25, 55, 85$. Podaj 90% przedziały predykcyjne.

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

   e) Która z dwóch zmiennych: wynik testu IQ czy wynik testu PH, jest lepszym predyktorem GPA?

W zadaniach $3-4$ ***zostaną wykorzystane dane z pliku ch01pr20.txt. Druga kolumna zawiera liczbę kopiarek, a pierwsza kolumna zawiera czas (w godzinach) potrzebny na utrzymanie tych kopiarek.***

3. **a)** Sprawdź, czy suma residuów wynosi zero.

   b) Przedstaw wykres residuów względem zmiennej objaśniającej i krótko omów wykres, zwracając uwagę na jakiekolwiek nietypowe wzory lub punkty.

   c) Przedstaw wykres residuów względem kolejności, w jakiej dane pojawiają się w pliku danych i krótko omów wykres, zwracając uwagę na jakiekolwiek nietypowe wzory lub punkty.

   d) Sprawdź rozkład residuów za pomocą histogramu i wykresu kwantylowo-kwantylowego. Jakie są wnioski?

4. Zmodyfikuj dane pliku **ch01pr20.txt**, dodając dodatkową obserwację $(1000; 2)$.

   a) Przeprowadź regresję ze zmienionymi danymi i utwórz tabelę porównującą wyniki tej analizy z wynikami analizy oryginalnych danych. W tabeli należy podać: dopasowane równanie regresji, t-test dla slope'a z p-wartością, R^2 i estymator σ^2. Krótko omów różnice.

   b) Powtórz punkty (b), (c) i (d) z zadania 3 powyżej na zmodyfikowanym zbiorze danych i pokaż nietypową obserwację na każdym z tych wykresów.

   c) Zmodyfikuj początkowy plik **ch01pr20.txt** w inny sposób, dodając obserwację $(1000, 6)$ i powtórz analizę z kroków a) i b) tego zadania. Porównaj wyniki.

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

5. **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).

   c) Oblicz współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu.

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

7. **a)** Utwórz nową zmienną odpowiedzi, biorąc logarytm stężenia roztworu (zdefiniuj $\tilde{Y} = \log(Y)$).

   b) Powtórz zadanie 5 z $\tilde{Y}$ jako zmienną odpowiedzi i czasem jako zmienną objaśniającą ($\tilde{Y} \sim time$). Podsumuj swoje wyniki.

   c) Na wykresie przedstaw stężenie roztworu względem czasu. Dodaj krzywą regresji i pasmo dla 95% przedziałów predykcji na podstawie wyników uzyskanych w punkcie b). Porównaj z wykresem uzyskanym w zadaniu 5.

   d) Oblicz współczynnik korelacji między obserwowanym a przewidywanym stężeniem roztworu opartym na modelu z punktu b) i porównaj z odpowiednim wynikiem z zadania 5.

8. 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?

Zadania teoretyczne $(+1pkt)$

1. **a)** Za pomocą R oblicz wartość krytyczną dla dwukierunkowego t-testu istotności z r stopniami swobody, $r \in \{5, 10, 50\}$, $\alpha = 0.05$. Oznacz tą wartość jako t_c.

   b) Za pomocą R oblicz wartość krytyczną dla testu istotności F z jednym stopniem swobody w liczniku i r stopniami swobody w mianowniku, $r \in \{5, 10, 50\}$, $\alpha = 0.05$. Oznacz tą wartość jako F_c.

   c) Sprawdź, czy kwadrat wartości t_c wynosi F_c i wyjaśnij dlaczego.

2. Poniżej zamieszczono część tablicy ANOVA:

	df	SS
Model	1	100
Error	20	400

   a) Ile obserwacji znajduje się w pliku?

   b) Oblicz estymator σ.

   c) Sprawdź, czy slope jest równy zero (podaj statystykę testową z liczbą stopni swobody i wniosek).

   d) Jaką część zmienności zmiennej odpowiedzi wyjaśnia model?

   e) Jaki jest próbkowy współczynnik korelacji między zmienną odpowiedzi a zmienną objaśniającą?

(*Małgorzata Bogdan, Liudmyla Zaitseva*)