Modele Linowe

Lista 5

Pierwszy zbiór danych z pliku CH06PR15.txt zawiera dane dotyczące poziomu satysfakcji pacjentów. Kolejne kolumny zawierają: wiek pacjenta (pierwsza kolumna), punkty opisujące ciężkość choroby, poziom niepokoju i poziom satysfakcji pacjenta.

1. **a)** Użyj modelu regresji, aby opisać zależność poziomu satysfakcji pacjentów od wieku, ciężkości choroby i poziomu niepokoju. Podaj dopasowane równanie regresji i współczynnik R^2. Oblicz współczynniki regresji oraz R^2 za pomocą wzorów teoretycznych oraz poleceń wbudowanych w R.

   b) Przetestuj hipotezę, że poziom satysfakcji pacjentów nie zależy od trzech zmiennych objaśniających. Podaj testowaną hipotezę, statystykę testową z liczbą stopni swobody, p-wartość oraz własne wnioski. Oblicz statystykę testową oraz p-wartość za pomocą wzoru teoretycznego oraz poleceń wbudowanych w R.

   c) Przetestuj hipotezy, że poziom satysfakcji pacjentów nie zależy od każdej z trzech zmiennych objaśniających osobno. Podaj testowane hipotezy, statystyki testowe z liczbą stopni swobody, p-wartości oraz własne wnioski. Oblicz statystyki testowe oraz p-wartości za pomocą wzorów teoretycznych oraz poleceń wbudowanych w R.

   d) Podaj osobne 95% przedziały ufności dla współczynników regresji przy wieku, ciężkości choroby i poziomu niepokoju. Jaki jest związek między tymi wynikami a wynikami testów z punktu c)?

2. Przedstaw na wykresach residua w zależności od przewidywanej satysfakcji i każdej ze zmiennych objaśniających. Czy występują jakieś nietypowe wzory lub wartości odstające?

3. Czy rozkład residuów jest w przybliżeniu normalny? Aby potwierdzić swoją odpowiedź, użyj testu Shapiro-Wilka (funkcja R - shapiro.test) i wykresu qqplot.

Drugi zbiór danych zawiera dane studentów informatyki, które omawialiśmy na zajęciach. Nazwa pliku jest csdata.dat. Zmienne są następujące: id (numeryczny identyfikator każdego studenta), GPA (średnia ocen po trzech semestrach), HSM; HSS; HSE (średnia z ocen z matematyki, przedmiotów ścisłych i języka angielskiego ze szkóły średniej), SATM, SATV (wyniki ustandaryzowanego testu dla uczniów szkół średnich w USA z matematyki i ze zdolności językowych) i SEX (płeć, kodowana jako 1 *dla mężczyzn i* 2 *dla kobiet).*

4. **a)** Przeprowadź następujące dwie regresje:

   (i) przewiduj GPA za pomocą HSM, HSS and HSE;

   (ii) przewiduj GPA za pomocą SATM, SATV, HSM, HSS and HSE.

   Wyznacz różnicę między statystykami SSE dla dwóch modeli i skonstruuj statystykę F do testowania hipotezy zerowej, że współczynniki przy dwóch zmiennych SAT są równe zero (w modelu z pięcioma predyktorami).

   b) Użyj funkcji *anova* do obliczenia tej samej statystyki testowej. Podaj statystykę, liczbę stopni swobody, p-wartość i wnioski.

5. Użyj regresji, aby przewidywać GPA przy pomocy SATM, SATV, HSM, HSE i HSS. Umieść zmienne w modelu w kolejności podanej powyżej. Oblicz sumy I typu i II typu.

   a) Co można wywnioskować o tym modelu na podstawie wyników sum I typu i II typu?

   b) Sprawdź (uruchamiając dodatkowe regresje), że suma kwadratów typu I dla zmiennej HSM jest różnicą statystyk SSM dla dwóch modeli:

   (i) GPA vs. SATM, SATV and HSM;

   (ii) GPA vs. SATM, SATV.

   c) Czy istnieją predyktory, dla których sumy I typu i II typu są takie same? Wyjaśnij dlaczego.

6. Utwórz nową zmienną (nazwij ją SAT), która będzie sumą dwóch testów SAT. Użyj regresji, aby przewidzieć GPA przy pomocy trzech zmiennych: SATM, SATV i SAT. Opisz uzyskane wyniki i wytłumacz je.

7. Użyj regresji GPA na zmiennych objaśniających HSM, HSS, HSE, SATM, SATV i SEX. Dla tego modelu:

   a) Przeanalizuj wykresy *partial regression plots*. Podaj krótkie wyjaśnienie, czym one są i jakiego rodzaju informacje zawierają. Czy są jakieś nietypowe wzory lub obserwacje?

   b) Przeanalizuj residua studentyzowane wewnetrznie (studentized residuals) i zewnetrznie (studentized deleted residuals). Podaj krótkie wyjaśnienie, czym one są, jaka jest różnica między nimi i jakiego rodzaju informacje zawierają. Czy występują jakieś wartości odstające?

   c) Przeanalizuj miarę DFFITS dla wszystkich obserwacji. Podaj krótkie wyjaśnienie, czym ona jest i jakiego rodzaju informacje zawiera. Co można wywnioskować o tym modelu na podstawie analizy tej statystyki?

   d) Przeanalizuj odległosc Cook'a (Cook's distance) dla wszystkich obserwacji. Podaj krótkie wyjaśnienie, czym ona jest i jakiego rodzaju informacje zawiera. Co można wywnioskować o tym modelu na podstawie analizy tej statystyki?

   e) Przeanalizuj miarę DFBETA dla wszystkich obserwacji. Podaj krótkie wyjaśnienie, czym ona jest i jakiego rodzaju informacje zawiera. Co można wywnioskować o tym modelu na podstawie analizy tej statystyki?

   f) Przeanalizuj miarę Tolerancja (Tolerance). Podaj krótkie wyjaśnienie, czym ona jest i jakiego rodzaju informacje zawiera. Co można wywnioskować o tym modelu na podstawie analizy tej statystyki? Jaka wartość tej statystyki jest dla modelu z zadania 6?

   g) Czym są kryteria wyboru modeli AIC, BIC, Cp Mallows'a? Wybierz najlepszy model regresji przy użyciu BIC, kriterium Cp Mallows'a, modyfikowanego współczynnika determinacji.

Zadania teoretyczne $(+1pkt)$

1. Na podstawie danych estymowane są parametry modelu

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Estymatory są następujące $b_0 = 1$, $b_1 = 4$, $b_2 = 3$, $s = 3$.

a) Przewiduj wartość Y dla $X_1 = 2$ i $X_2 = 6$.

b) Estymowane odchylenie standardowe estymatora wartości oczekiwanej Y dla $X_1 = 2$ i $X_2 = 6$ wynosi 2. Estymuj wariancję błędu predykcji $\sigma^2(pred)$.

c) Powyższy model został dopasowany przy użyciu 20 obserwacji, a estymowane odchylenie standardowe b_1, $s(b_1)$, jest równe 1. Skonstruuj 95% przedział ufności dla β_1.

2. Dokonujemy analizy danych przy użyciu następującego modelu regresji wielorakiej:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Poniżej znajduje się tabela z sumami kwadratów typu I i II:

	Type I	Type II
X_1	300	30
X_2	40	25
X_3	20	?

Statystyka SST wynosi 760, a $n = 24$.

a) Ile wynosi suma kwadratów typu II dla X_3?

b) Przetestuj hipotezę, że $\beta_1 = 0$ (w pełnym modelu).

c) Przetestuj hipotezę, że $\beta_2 = \beta_3 = 0$.

d) Przetestuj hipotezę, że $\beta_1 = \beta_2 = \beta_3 = 0$.

e) Z modelu zostały usunięte wartości X_2 i X_3. Przetestuj hipotezę, że $\beta_1 = 0$ w modelu regresji liniowej prostej

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

(Uwaga - SSE i dfE są inne niż dla modelu z trzema zmiennymi).

f) Oblicz próbkowy współczynnik korelacji między Y i X_1.