

Wybrane zadania z Kolokwium komputerowego
ze Statystycznych modeli liniowych i nieliniowych
rok ak. 2025/26

ZADANIE 1. Zbiór danych `datasets::iris` zawiera pomiary w centymetrach długości i szerokości działki kielicha (sepal) oraz długości i szerokości płatk (petal) 50 kwiatów trzech odmian gatunku irys. Wykonaj poniższe polecenia i odpowiedz na zadane pytania.

(A) Zbuduj model wyjaśniający zmienność długości działki kielicha od szerokości działki kielicha i szerokości płatk rozważanego gatunku kwiatów. (3 pkt)

(B) Sprawdź, czy istnieje interakcja między zmiennymi objaśniającymi w modelu zbudowanym w punkcie (A). (4 pkt)

(C) Wyznacz przedziały ufności dla wszystkich współczynników modelu (A). (5 pkt)

(D) Zmodyfikuj model otrzymany w punkcie (A) logarytmując zmienne objaśniające. Podaj interpretację współczynników w otrzymanym modelu. (7 pkt)

(E) Zbadaj, czy w zbudowanym modelu w punkcie (A) błędy są homoskedastyczne. W przypadku odpowiedzi negatywnej sprawdź, czy zlogarytmowanie zmiennych objaśniających w punkcie (D) wymusiło homoskedastyczność błędów. (8 pkt)

(F) Przeprowadź analizę kontrastów Helmerta, aby sprawdzić, czy rozważane odmiany kwiatów irys różnią się długością działki kielicha. (10 pkt)

ZADANIE 2. W zbiorze `Mroz` biblioteki `car` znajdują się dane dotyczące zatrudnienia zamężnych kobiet. Zbiór zawiera 8 zmiennych

`lfp` – zatrudnienie, czynnik z poziomami `no` (kobieta nie jest zatrudniona) i `yes` (kobieta jest zatrudniona),

`k5` – liczba dzieci do piątego roku życia,

`k618` – liczba dzieci w wieku od 6 do 18 lat,

`age` – wiek kobiety w latach,

`wc` – uczęszczanie na studia, czynnik z poziomami `no` (kobieta nie uczęszcza na studia) i `yes` (kobieta uczęszcza na studia),

`hc` – uczęszczanie męża na studia, czynnik z poziomami `no` (mąż nie uczęszcza na studia) i `yes` (mąż uczęszcza na studia),

`lwg` – logarytm oczekiwanej stawki wynagrodzenia,

`inc` – dochód rodziny bez dochodu kobiety.

(A) Stosując selekcję w przód zbuduj model za pomocą, którego można określić szanse na to, że losowo wybrana kobieta jest zatrudniona. (5 pkt)

(B) Na podstawie zbudowanego modelu w punkcie (A) odpowiedz na pytanie: jaki wpływ na zatrudnienie kobiety ma to, że uczęszcza na studia, a jaki wpływ na zatrudnienie kobiety ma jej wiek? (8 pkt)

(C) Przyjmując próg klasyfikacyjny 0.5 stwórz zmienną zero-jedynkową `pred`, za pomocą której można określić, czy na podstawie zbudowanego modelu zamężna kobieta jest zatrudniona, a następnie oblicz miary oceny jakości prognoz i na ich podstawie dokonaj oceny zbudowanego modelu. (10 pkt)

ZADANIE 3. Oceniono typ osobowości 3154 zdrowych mężczyzn w wieku 39-59 lat zamieszkałych w San Francisco. Obserwacje zapisano w zbiorze `faraway::wcgs`. Na podstawie tych obserwacji chcemy odpowiedzieć na pytanie, jaki jest wpływ masy ciała (`weight`), stężenia cholesterolu we krwi (`chol`) i typu osobowości (`dibep`) na to, czy u danego mężczyzny stwierdzono chorobę wieńcową (`chd`). W tym celu

zapoznaj się z informacjami zawartymi w pomocy RStudio, a następnie zbuduj odpowiedni model włączając do niego **tylko wymienione wyżej zmienne**.

(A) Podaj interpretacje współczynników zbudowanego modelu. (10 pkt.)

(B) Przyjmując próg klasyfikacyjny 0.6 stwórz zmienną zero-jedynkową **pred**, za pomocą której można określić, czy na podstawie zbudowanego modelu stwierdzono chorobę wieńcową, a następnie oblicz miary oceny jakości prognoz i na ich podstawie dokonaj oceny zbudowanego modelu. (10 pkt.)

UWAGA. W zbiorze `faraway::wgs` brakuje danych. Możesz je usunąć np. za pomocą funkcji `na.omit()`.

ZADANIE 4. Na podstawie zbioru `faraway::wgs` zbuduj model liniowy zależności ciśnienia skurczowego (`sdp`), od wieku (`age`), poziomu cholesterolu (`chol`), masy ciała (`weight`) oraz liczby wypalanych papierosów (`cigs`).

(A) Sprawdź, czy w zbudowanym modelu zostało naruszone założenie o jednorodności wariancji składnika losowego. (8 pkt.)

(B) Czy zlogarytmowanie zmiennych objaśniających wymusi homoskedastyczność błędów zbudowanego modelu. Zauważ, że zmiennej `cigs` nie możemy logarytmować, albowiem może przyjmować wartości równej zero. (8 pkt.)

(C) Podaj interpretacje współczynników stojących przy zmiennej `cigs` i `weight` w modelu zbudowanym w punkcie (B). (6 pkt.)

ZADANIE 5. Firma farmaceutyczna przetestowała trzy różne sposoby leczenia osób wysokim poziomem cholesterolu. W eksperymencie wzięło udział 72 uczestników. Celem badania było sprawdzenie działania nowej klasy leków (`drug`) na obniżenie poziomu cholesterolu (`cholesterol`). Zebrane dane znajdują się w zbiorze `datarium::heartattack`.

(A) Za pomocą **odpowiedniego** testu zweryfikuj hipotezę, że średni poziom cholesterolu różni się w zależności od podanego leku. (8 pkt.)

(B) Przeprowadź analizę kontrastów sum aby sprawdzić, czy średni poziom cholesterolu różni się w zależności od podanego leku. (10 pkt.)

Zbiór danych `pima` biblioteki `faraway` zawiera wyniki badania przeprowadzonego na 768 dorosłych Indiankach Pima mieszkających w pobliżu Phoenix przez Narodowy Instytut Cukrzycy oraz Chorób Trawiennych i Nerek. Zbiór ten zawiera następujące zmienne

`pregnant` – liczba przebytych ciąż,

`glucose` – stężenie glukozy w osoczu po 2 godzinach w doustnym teście tolerancji glukozy,

`diastolic` – ciśnienie rozkurczowe krwi (mm Hg),

`triceps` – grubość fałdu skórniego tricepsa (mm), antropometryczny wskaźnik oceny zasobów tkanki tłuszczowej,

`insulin` – ilość insuliny w surowicy (μ U/ml),

`bmi` – wskaźnik masy ciała (waga w kg/(wzrost w metrach do kwadratu)),

`diabetes` – wskaźnik wystąpienia cukrzycy określający ryzyko genetyczne zachorowania na cukrzycę poprzez ocenę historii rodzinnej,

`age` – wiek w latach,

`test` – wynik testu sprawdzającego, czy u pacjentki występują objawy cukrzycy (0 w przypadku wyniku negatywnego, 1 w przypadku wyniku pozytywnego).

(A) Na podstawie zbioru `pima` zbuduj model liniowy za pomocą którego możemy wyjaśnić wpływ stężenia glukozy we krwi (`glucose`) oraz wieku (`age`) na grubość fałdu skórniego tricepsa (`triceps`). Sprawdź, czy istnieje interakcja między zmiennymi objaśniającymi w zbudowanym modelu liniowym. (6 pkt.)

(B) Zmodyfikuj zbudowany model liniowy w punkcie (A) logarytmując zmienną objaśnianą `age`. Podaj interpretacje współczynników w otrzymanym modelu. (8 pkt.)

(C) Zbadaj, czy w zbudowanym modelu liniowym w punkcie (A) zostało naruszone założenie braku autokorelacji błędów. Użyj wszystkich poznanych testów na wykładzie. (5 pkt.)

(D) Jaką grubość fałdu tricepsa możemy przewidzieć na podstawie modelu liniowego zbudowanego w punkcie (A) w przypadku 50-letniej Indianki, u której stężenie glukozy we krwi wynosi 100 mg/dL. (5 pkt.)

(E) Na podstawie zbioru `pima` zbuduj model, za pomocą którego możemy wyjaśnić szanse na to, że u losowo wybranej Indianki Pima występują objawy cukrzycy. Do modelu włącz wszystkie zmienne. Na podstawie zbudowanego modelu określ jaki jest wpływ wskaźnika wystąpienia cukrzycy (`diabetes`) oraz ciśnienia rozkurczowego (`diastolic`) na to czy dana Indianka Pima jest chora na cukrzycę. (8 pkt.)

(F) Przyjmując próg klasyfikacyjny 0.7 stwórz zmienną zero-jedynkową `pred`, za pomocą której można określić, czy u badanej Indianki Pima występują objawy cukrzycy, a następnie oblicz miary oceny jakości prognoz i na ich podstawie dokonaj oceny zbudowanego modelu. (10 pkt.)

ZADANIE 6. W zbiorze `faraway::coagulation` zostały zapisane czasy krzepnięcia krwi pobranej od 24 zwierząt oraz sposób ich żywienia.

(A) Za pomocą **odpowiedniego** testu zweryfikuj hipotezę, że średni czas krzepnięcia krwi różni się w zależności sposobu żywienia. (8 pkt.)

(B) Przeprowadź analizę kontrastów Helmerta aby sprawdzić, czy średni czas krzepnięcia krwi różni się w zależności od sposobu żywienia. (10 pkt.)

ZADANIE 7. W zbiorze `agricolae::sweetpotato` znajdują się dane pochodzące z eksperymentu przeprowadzonego na słodkich ziemniakach (Costanero) w departamencie Tacna, południowe Peru. Badano wpływ dwóch wirusów `Spcsv` i `Spfnv` na plon roślin. Na każdej spośród 12 działek zasiano 50 roślin, przy czym każdy wariant powtarzano 3 razy. Po upływie pewnego czasu zmierzono masę plonów w kilogramach (`yield`). Zmienna `virus` przyjmuje wartości:

CC – zakażenie wirusem `Spcsv` powodującym żółknięcie i niższy wzrost,

FF – zakażenie wirusem `Spfnv` powodującym plamy na liściach,

FC – zakażenie obu wirusami,

00 – brak zakażenia.

Na podstawie zbioru `sweetpotato` chcemy zweryfikować, czy średni plon słodkich ziemniaków różni się w zależności od sposobu zakażenia wirusami `Spcsv` i `Spfnv`. W tym celu wykonaj poniższe polecenia. Zweryfikuj postawioną hipotezę za pomocą

(A) odpowiedniego testu. (8 pkt.)

(B) analizy kontrastów sum. (10 pkt.)

ZADANIE 8. Zbiór `datarium::titanic.raw` zawiera informacje o losach pasażerów podczas tragicznego pierwszego rejsu oceanicznego liniowca Titanic. W zbiorze znajdują się zmienne

`Class` – klasa podróży, `Sex` – płeć, `Age` – wiek, `Survived` – przeżycie pasażera.

(A) Na podstawie zbioru `titanic.raw` zbuduj model określający szanse przeżycia pasażera Titanica i podaj interpretacje współczynników zbudowanego modelu. (8 pkt.)

(B) Przyjmując próg klasyfikacyjny 0.4 stwórz zmienną zero-jedynkową `pred`, za pomocą której można określić, czy na podstawie zbudowanego modelu pasażer Titanica przeżył katastrofę, a następnie oblicz miary oceny jakości prognoz i na ich podstawie dokonaj oceny zbudowanego modelu. (10 pkt.)

ZADANIE 9. Zbiór `caret::Sacramento` zawiera informacje o domach i cenach sprzedaży 932 nieruchomości w Sacramento, Kalifornia. W zbiorze znajdują się zmienne

`city` – miasto, w którym znajduje się nieruchomość,
`zip` – kod pocztowy nieruchomości,
`beds` – liczba sypialni w domu,
`baths` – liczba łazienek w domu,
`sqft` – powierzchnia domu w stopach kwadratowych,
`type` – typ nieruchomości,
`price` – cena sprzedaży nieruchomości w dolarach,
`latitude` – szerokość geograficzna nieruchomości w stopniach,
`longitude` – długość geograficzna nieruchomości w stopniach.

(A) Na podstawie danych zawartych w zbiorze **Sacramento** zbuduj model, za pomocą którego można przewidzieć cenę domu na podstawie liczby sypialni (`beds`) i powierzchni (`sqft`). Zbadaj, czy w zbudowanym modelu istnieje interakcja między zmiennymi objaśniającymi. (6 pkt.)

(C) Zbadaj, czy w zbudowanym modelu (A) zostało naruszone założenie braku autokorelacji błędów. W przypadku odpowiedzi pozytywnej zlogarytmuj zmienną `sqft` i sprawdź, czy taka transformacja wymusiła brak autokorelacji błędów. (10 pkt.)

(D) Podaj interpretacje współczynników modelu ze zlogarytmowaną zmienną `sqft`. (8 pkt.)

ZADANIE 10. Rozważamy hipotetyczne dane dotyczące przyjęć na studia drugiego stopnia pewnego amerykańskiego uniwersytetu. Dane trzeba pobrać ze strony internetowej za pomocą kodu

```
dane <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
```

W zbiorze znajdują się zmienne

`admit` – informuje, czy kandydat został przyjęty na studia (1 = tak, 0 = nie),

`gre` – wynik testu GRE (Graduate Record Examination), standaryzowanego egzaminu używanego w USA przy rekrutacji na studia magisterskie,

`gpa` – średnia ocen z ukończonych studiów pierwszego stopnia,

`rank` – prestiż uniwersytetu, na który kandydat aplikuje (1–4), gdzie 1 = najbardziej prestiżowy, 4 = najmniej.

Zmienne `gre` i `gpa` są ilościowe. Zmienną `rank` należy traktować jako jakościową, co można uczynić za pomocą polecenia

```
dane$rank <- factor(dane$rank)
```

Na podstawie tego zbioru chcemy ocenić szanse kandydata na dostanie się na studia drugiego stopnia.

(A) Zbuduj model, w którym zmienną objaśnianą jest zmienna `admit`, a zmiennymi objaśniającymi pozostałe zmienne. (3 pkt)

(B) Podaj interpretacje współczynników otrzymanego modelu. (10 pkt)

(C) Przyjmując próg klasyfikacyjny 0.5 stwórz zmienną zero-jedynkową `pred_admit`, za pomocą której można określić, czy na podstawie zbudowanego modelu kandydat dostanie się na studia, a następnie oblicz miary oceny jakości prognoz i na ich podstawie dokonaj oceny zbudowanego modelu. (10 pkt)