Uogólnione modele liniowe

Laboratorium nr 3

3.1 Zbiór **finance** zawiera dane dotyczące kondycji finansowej 46 przedsiębiorstw na podstawie czterech wskaźników finansowych.

(a) Dopasować model logistyczny. Przetestować hipotezę, że zbiór zmiennych zawiera zmienne istotne i obliczyć procent dewiacji wyjaśnianej przez model.

(b) Za pomocą instrukcji step dokonać sekwencyjnego usunięcia z modelu nieistotnych zmiennych. Porównać mniejszy model z modelem wyjściowym. Obliczyć procent dewiacji wyjaśnianej.

(c) Rozpatrzyć rezydua oparte na dewiacjach. Podobnie jak w modelu liniowym, ich błędy standardowe są proporcjonalne do pierwiastka z $1-h_{i,i}$, gdzie $h_{i,i}$ jest odpowiednim elementem przekątnej macierzy daszkowej. Wyliczyć studentyzowane rezydua i narysować ich wykres kwantylowy.

(d) Wyrzucić obserwacje potencjalnie odstające, dopasować powtórnie model i obliczyć dla niego procent dewiacji wyjaśnionej.

3.2 SAheart4.data zawiera dane dotyczące zawałów serca wśród białych mężczyzn w wieku od 15 do 64 lat w Republice Południowej Afryki. Zmienne:

- chd (coronary heart disease) – objaśniana
- sbp – systolic blood pressure
- ldl – poziom cholesterolu ldl
- famhist – występowanie zawałów w rodzinie
- i inne.

(a) Dokonać konwersji zmiennej famhist na zmienną zero-jedynkową i nazwać ją family.

(b) Dokonać dopasowania modelu regresji logistycznej po pominięciu zmiennych famhist (zastąpiona przez family), row.names, adiposity (to jest przekształcona zmienna obesity). Zinterpretować wyniki.

(c) Dokonać sekwencyjnego wyrzucenia najmniej istotnych zmiennych z modelu za pomocą `drop1(nazwa.glm,test='Chi')`. Porównać z wynikiem działania `step(nazwa.glm, test='Chi')`. Porównać uzyskany model z modelem wyjściowym.

3.3 Jedną z metod stwierdzenia nieadekwatności modelu logistycznego w przypadku obserwacji indywidualnych jest dopasowanie modelu większego i porównanie adekwatności modelu mniejszego i większego. Rozpatrujemy zbiór kyphosis (jest on m.in. wbudowany w bibliotekę rpart). Pisząc ?kyphosis otrzymuje się opis zbioru.

(a) Dopasować dwa modele: `g: kyphosis~Age+Number+Start` oraz model `g` powiększony o kwadraty zmiennych objaśniających (nazwijmy go `g2`). Dla modelu `g2`:

```
g2=glm(Kyphosis~Age+I(Age^2)+Start+I(Start^2)+Number+I(Number^2),family=binomial,data=kyphosis).
```

(b) Stwierdzić, czy którakolwiek ze zmiennych w modelu `g` jest istotna.

(c) Za pomocą instrukcji step usunąć zmienne nieistotne z modelu `g2` (stworzony w ten sposób model nazwać `g1`). Porównać wartości AIC dla `g` i `g1` oraz przeprowadzić test $H_0 : g$ kontra $H_1 : g_1$.

3.4 Zbiór **discoveries** zawiera trajektorie szeregu czasowego z liczbą wielkich odkryć od 1860 do 1959 roku. Celem ćwiczenia jest stwierdzenie, czy średnia liczba odkryć w roku jest stała.

(a) Narysować wykres zależności liczby odkryć od czasu (discoveries są obiektem typu time series (ts), dlatego instrukcja plot(discoveries) daje na osi x zmienną o wartościach rzeczywistych).

(b) Zakładając, że liczba odkryć w roku ma rozkład Poissona i postulując model poissonowski:

i. przeprowadzić test hipotezy o stałości średniej liczby odkryć postulując najprostszy możliwy model, sprawdzając uprzednio jego dopasowanie

ii. metoda alternatywna: podobnie jak w postępowaniu z danymi ze zbioru kyphosis, dopasować do liczby odkryć trend kwadratowy względem czasu i stwierdzić, czy współczynniki odpowiadające członowi liniowemu i kwadratowemu są istotne.