Uogólnione modele liniowe

Laboratorium nr 6

6.1 Zbiór **lungcanc.dat** zawiera dane dotyczące badania kohortowego miliona osób w latach 1982-1988. Każdego roku sprawdzano status (dead, alive) każdej z osób oraz rejestrowano wartości zmiennych objaśniających: cigcat (0,1,2 – w zależności od tego, czy osoba w ogóle nie pali, pali do 20 papierosów dziennie, pali powyżej 20 papierosów dziennie), age (wiek), follow (kolejny rok badania, tzn. liczba lat od startu kohorty). Kolejne wiersze zbioru podają liczności (freq) osób w kohocie przy wszystkich możliwych wartościach zmiennych. Przykładowo, pierwszy wiersz orzeka, że w kohorcie była 1 osoba w wieku 35 lat, niepaląca, która zmarła w pierwszym roku ewolucji kohorty. Celem zadania jest zbadanie zależności liczby zgonów od zmiennych występujących w zbiorze, dla pierwszego roku kohorty.

(a) Przekodować osoby, które były obserwowane dłużej niż rok jako „osoby, które nie zmarły w pierwszym roku".

(b) Wyrysować wykres zależności frakcji zgonów od wieku, w rozbiciu na kategorie wynikające z różnych wartości zmiennej *cigcat*.

(c) Wyrysować wykres zależności frakcji zgonów od wieku, w rozbiciu na kategorie osób palących i niepalących.

(d) Ponieważ liczba zgonów zależy w sposób oczywisty od liczby osób w danej kategorii wiekowej, będziemy modelować $\log(l.zgonow/l.osob)$ (a nie $\log(l.zgonow)$).

(e) Ocenić intensywność zgonu w kategorii smoker przy ustalonym wieku w porównaniu z intensywnością zgonu w kategorii nonsmoker.

(f) Porównać wartości dopasowane z frakcjami empirycznymi przez wyrysowanie krzywej wartości prognozowanych: ponieważ jest 46 kategorii wiekowych (wiek od 35 do 80) i dwie kategorie smoker, tworzymy 92 grupy po 20 osób:

```
new<-data.frame(age=rep(35:80,2),smoker<-rep(0:1,each=46),freq=rep(20,92))
```

prognozujemy oczekiwaną liczbę zgonów w każdej grupie, a następnie dzielimy ją przez 20:

```
pred<-predict(lung.glm,newdata=new,type="response")/20
```

(g) Ponieważ krzywe dla większych wartości wieku nie wzrastają dostatecznie szybko, dodać do modelu człon kwadratowy wieku. Ocenić jego istotność i wyrysować nowe krzywe (jak w poprzednim punkcie).

6.2 (Model gamma)

(a) Przypomnienie: narysować wykresy gęstości rozkładu gamma dla przykładowych wartości parametru kształtu: 0.5,1.5,3 i parametru skali 1.

(b) Dane ze zbioru **clot.data** opisują czasy krzepnięcia krwi w zależności od koncentracji plazmy (9 poziomów) oraz poziomu tromboblastyny (2 poziomy). Przekształcić dane do postaci: jedna obserwacja dla każdego poziomu obydwu czynników.

(c) Narysować wykres zależności czasu krzepnięcia od koncentracji plazmy w rozbiciu na lot1 i lot2: znaleźć przekształcenie x i y, które doprowadzi obie krzywe do przybliżonej liniowości.

(d) Jedną z możliwości jest zależność: $(czaskrzepniecia)^{-1} \sim \log(conc)$, co sugeruje zastosowanie modelu gamma. Dopasować model gamma (za pomocą polecenia glm z wyszczególnieniem family=Gamma - zwrócić uwagę na wielką literę).

(e) Obejrzeć wykres rezyduów opartych na dewiancji od $\log(conc)$. Zidentyfikować problem, rozwiązać go i dokonać powtórne dopasowanie modelu. Czy jakość dopasowania poprawiła się?

(f) Do wykresu odwrotności czasu krzepnięcia względem $\log(conc)$ w rozbiciu na lot1 i lot2, dorysować wyestymowane proste.

(g) Załóżmy, że powodem wystąpienia obserwacji odstających było błędne zapisanie najniższej koncentracji (zamiast faktycznych 6 zapisano 5). Zmienić odpowiednią wartość koncentracji i ocenić wpływ zmiany na jakość dopasowania.