# Uogólnione modele liniowe

**Laboratorium nr 4**

4.1 W przypadku danych (zgrupowanych) ze zbioru **bliss** porównać wartości dopasowane uzyskane z zastosowanie regresji logistycznej, probitowej i funkcji łączącej ,,complementary log-log", zarówno w skali $\pi$, jak i predyktora liniowego $\eta$. Dla skali $\pi$ narysować wykresy odpowiednich krzywych w zakresie od $-2$ do 8 (zakres danych - zmiennej *conc* - jest od 0 do 4). Następnie dla wszystkich (trzech) par funkcji łączących narysować na jednym rysunku wykresy ilorazów $\pi_a/\pi_b$ i $(1-\pi_a)/(1-\pi_b)$, gdzie $\pi_a$, $\pi_b$ - dopasowane wartości $\pi$ wynikające z użycia odpowiednich funkcji łączących.

4.2 Zbiór **gala** zawiera informacje o liczbie gatunków żółwi znalezionych na każdej z 30 wysp należących do archipelagu Galapagos oraz o liczbie gatunków stale występujących na danej wyspie (endemicznych). Dodatkowo zbiór zawiera pięć zmiennych geograficznych, które opisują każdą z wysp.

(a) Dopasować model liniowy `Species~.` (oprócz zmiennej Endemics). Sporządzić wykres rezyduów (jako funkcji od wartości dopasowanych) i zauważyć wyraźną heteroskedastyczność (niestałość wariancji).

(b) Znaleźć (metodą Boxa–Coxa, funkcja boxcox w bibliotece MASS z opcją plotit=T i wybranym odpowiednio zakresem parametru $\lambda$) przekształcenie zmiennej Species poprawiające problem z poprzedniego punktu. Na podstawie analizy Boxa-Coxa wybrać najbardziej naturalną wartość $\lambda$. Dopasować nowy model i sporządzić wykres jego rezyduów.

(c) Dopasować model poissonowski. Stwierdzić, czy jest dopasowany.

(d) Obliczyć procent dewiacji objaśnianej przez model poissonowski i porównać go z wartością $R^2$ w modelu liniowym.

(e) Porównać (na wykresie) wartości dopasowane w obu modelach.

(f) Sprawdzić zbiory zmiennych istotnych w obu modelach.

4.3 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 (osoby, które były obserwowane dłużej niż rok, przekodowane zostaną jako nie zmarłe w pierwszym roku).

(a) Wykonać i przeanalizować poniższy kod:

```
lung=read.table("c:/.../lungcanc.dat",header=T)
lung1=lung
death1=lung$death
death1[lung$death==1 & lung$follow>1]=0
lung1$death=death1
attach(lung1)
den<-tapply(freq,list(age=lung1$age,smoker=lung1$smoker),FUN=sum)
num<-tapply(lung1$death,list(age=lung1$age,smoker=lung1$smoker),FUN=sum)
inc<-num/den
```

Obejrzeć wynikowy zbiór inc.

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

```
age1<-tapply(lung1$age,list(age=lung1$age,smoker=lung1$smoker),FUN=mean)
#obejrzec zbior age1
smoke1<-tapply(lung1$smoker,list(age=lung1$age,smoker=lung1$smoker),FUN=mean)
plot(age1,inc,type="n",ylab="Incidence rate",xlab="Age")
points(age1[smoke1==0],inc[smoke1==0],pch=1)
points(age1[smoke1==1],inc[smoke1==1],pch=2)
legend(40,.03,legend=c("non-smoker","smoker"),pch=c(1,2))
```

(c) Ponieważ liczba zgonów zależy w sposób oczywisty od liczby osób w danej kategorii wiekowej, będziemy modelować $\log(l.zgonow/l.osob)$ (a nie $\log(l.zgonow)$). Osiąga się to przez wymuszenie w procedurze glm współczynnika 1 przy zmiennej $\log(freq)$ (model dla intensywności, ang. rate model):

```
lung.glm=glm(death~smoker+age+offset(log(freq)),family=poisson,data=lung1)
```

(d) Ocenić intensywność zgonu w kategorii smoker przy ustalonym wieku w porównaniu z szansami zgonu w kategorii nonsmoker.

(e) 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
lines(35:80,pred[1:46])
lines(35:80,pred[47:92])
```

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

(g) Powtórzyć powyższe analizy dla wyższych lat kohorty (od drugiego do szóstego roku obserwacji kohorty).

4.4 (Nadwyżka rozproszenia dla modelu logitowego) Zbiór beetle.txt zawiera dane dotyczące działania środka owadobójczego na wołki zbożowe. Grupy wołków zbożowych poddawane były różnym stężeniom środka.

(a) Dopasować model logitowy z $\log(conc)$ jako zmienną objaśniającą.

(b) Narysować wykres affected/exposed względem $\log(conc)$ (pozwali to ocenić, czy logitowa funkcją łącząca jest w tym przypadku sensownym wyborem). Dorysować linię wartości prognozowanych do wykresu i ocenić jej dopasowanie.

(c) Narysować wykres logitów empirycznych:

$$\log\left((\text{affected} + 0.5)/(\text{exposed} - \text{affected} + 0.5)\right)$$

względem $\log(conc)$ i dorysować do niego prostą regresyjną.

(d) Sprawdzić, w oparciu o wykres normalny rezyduów, czy są w zbiorze wartości odstające.

(e) Jednym z możliwych powodów niedopasowania, po wyeliminowaniu ewentualnych obserwacji odstających i stwierdzeniu poprawności zastosowanej funkcji łączącej, jest występowanie większego rozproszenia w danych niż te przewidywane przez model logitowy (powodami mogą być np. niejednakowe zastosowanie trucizny na różnych poziomach, niejednakowe warunki traktowania wołków po zastosowaniu środka owadobójczego – nie ma on działania natychmiastowego, niejednakowa żywotność wołków przy rozpoczęciu eksperymentu).

W modelu z nadwyżką rozproszenia zakłada się, że wariancja odpowiedzi wynosi $\phi*$ (wartość wariancji dla rozkładu dwumianowego). Estymuje się ją jako $(N-p)^{-1}\sum {r_i^P}^2$ (gdzie $r_i^P$ to rezydua pearsonowskie.

i. Ocenić wielkość parametru rozproszenia $\phi$. Stwierdzić, jak zmieniły się błędy standardowe po uwzględnieniu nadwyżki rozproszenia.

ii. Ile wynosi $p$-wartość testu $F$ istotności dla zmiennej $\log(conc)$ (test $F$ stosowany jest w przypadku możliwej nadwyżki rozproszenia)?