7. Dane: *ptakiZag.csv*

Dane dotyczą występowania gatunków ptaków zagrożonych w pasmach śródpolnych. Celem analizy jest wybór modelu zależności prawdopodobieństwa występowania gatunków ptaków zagrożonych w skali logitowej (`ptz.logit`) od warunków, w których żyją ptaki. Zmienne `PHRAGMITES,URTICA,PRUNUS` opisują gęstość występowania trzcin, pokrzyw i drzew śliwowych w paśmie, zmienne `DrzKrzOb,rów,droga,rowPrzekroj,odl_lasu, Hnorm,szer,lupr200,lkep,odl_wieś` opisują objętość drzew i krzewów, występowanie rowu (1 – tak, 0 – nie), występowanie drogi (1 – tak, 0 – nie), przekrój rowu (o ile występuje), odległość od lasu, unormowaną entropię opisującą różnorodność upraw w okolicy, szerokość pasma, liczba upraw w odległości do 200 m od pasma, liczbę kęp roślinności i odległość od wsi. Zmienne te zostały przekształcone tak, aby relacje między zmienną wynikową a zmiennymi objaśniającymi były liniowe. Zmienne po przekształceniu mają przedrostek „`t`". W modelu występują zmienne w postaci przekształconej (z wyjątkiem zmiennej `rów`) :

```
ptz.logit=logit(Pt_lgat.Z/Pt_lgat),
tDrzKrzOb=ifelse(DrzKrzOb>0,log(DrzKrzOb),0),
trowPrzekroj=sqrt(rowPrzekroj),
todlas=abs(odl_lasu-600)^1.25,
tphr=ifelse(PHRAGMITES>0,PHRAGMITES^-1,0),
turt=log(URTICA),
thnorm=Hnorm^2,
tprun=ifelse (PRUNUS>0,sqrt(PRUNUS),0),
tszer=log(szer),
tlupr200=abs(lupr200-20)^3,
tlkep=log(lkep+.5),
todwies=sqrt(odl_wieś))
```

a) Oblicz macierz korelacji zmiennych objaśniających. Zwróć uwagę na pary zmiennych o współczynniku korelacji > 0.5 (co do modułu)

b) Zbuduj addytywny model wiążący prawdopodobieństwo występowania gatunków zagrożonych ze zmiennymi opisującymi środowisko, w których występują (*ptaki.reg*)

c) Oblicz wskaźnik VIF. Porównaj jego wartość ze współczynnikami korelacji dla zmiennych objaśniających

d) Oznacz model regresji ridge symbolem *ptaki.ridge* i oblicz jego współczynniki.

e) Wybierz optymalny współczynnik lambda (np. estymator Hoerla i Kennarda)

f) Narysuj wykres wartości wszystkich współczynników regresji dla lambda od 0 do 50 skokiem co 5. Zauważ, że na prawo od współczynnika lambda Hoerla i Kennarda estymatory się stabilizują

g) Dla współczynnika lambda Hoerla i Kennarda wyznacz model regresji ridge *ptaki.ridge0*. Oblicz jego współczynniki.

h) Oblicz ilorazy współczynników regresji modeli *ptaki.ridge0* i *ptaki.reg*. Skomentuj te wyniki i porównaj z tabelą korelacji z zad a)

i) Dla każdej z metod Cp, adjr2 wybierz najlepszy zestaw zmiennych (regsubsets{leaps}). Wybierz opcję `nbest=3` . Zauważ, jakie zmienne najczęściej a jakie najrzadziej występują na liście proponowanych modeli. Co oznacza Cp<0?

j) Dla zmiennych wybranych metodą BIC porównaj model z interakcją i bez interakcji. Porównaj skorygowany współczynnik determinacji, błąd resztowy i p-wartość testu F w obu modelach. Który z nich wybierzesz jako najlepszy?

k) Korzystając z przybliżonego wzoru dla przyrostu prawdopodobieństwa $\pi(x)$ przy przyroście zmiennej x o współczynniku β w regresji logitowej o wartość δ (pozostałe zmienne są na stałym poziomie)

$$\pi(x+\delta)-\pi(x)\approx\delta\beta\pi(x)(1-\pi(x)),$$

oszacuj w tym modelu o ile punktów procentowych wzrośnie prawdopodobieństwo występowania liczby gatunków zagrożonych, gdy % gatunków zagrożonych wynosi 13%, 18%, 23%, 40%[1] gdy liczba kęp wzroście dwukrotnie oraz gdy pierwiastek z przekroju rowu wzrośnie o 0.1.

l) Z listy najlepszych modeli wg kryterium adjr2 wybierz najlepszy. Oszacuj parametry modelu pełnego, modelu po jego redukcji z kryterium BIC i modelu addytywnego (bez interakcji). Porównaj te modele i model addytywny z zadania j). Jakie wnioski wyciągniesz, gdy kryterium porównania jest poziom istotności 0.05? Jaką decyzje podejmiesz dla najlepszego modelu uzyskanego metodą Cp?

[1] Są to wartości 1 – 3 kwartyla (w przybliżeniu) oraz wartość pomiędzy 3. kwartylem a maksimum