Dzień 4 - Modele: GLM, GAM - zadania

Marcin K. Dyderski, Patryk Czortek

13 stycznia 2022

Zadania do wykonania

1. Wczytaj zbiór danych dotyczący występowania gatunków wskaźnikowych starych lasów w Poznaniu. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
afis<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/afis.csv',
               sep=';')
```

W zbiorze danych mamy informacje o udziale procentowym terenów otwartych (agricultural, semi-natural & wetlans, kolumna `ASW`), lasów (`Forests`), terenów przemysłowych (`Industrial`), wód ('`Water`), zabudowy gęstej (`Urban.dense`) i rzadkiej (`Urban.sparse`), typ lasów w kwadracie (`OLDFR`,stare, nowe i brak lasów), liczbę gatunków wskaźnikowych starych lasów (`AFIS`) oraz obecność (0/1) pięciu wybranych gatunków.

   a. Używając zbioru danych `afis` wykonaj model dla liczby gatunków wskaźnikowych starych lasów (kolumna `AFIS`) w opraciu o trzy predyktory: `Water`, `Urban.dense` oraz `OLDFR`. Z uwagi na charakter danych skorzystaj z rozkładu Poissona używając funkcji `glm(...., family=poisson)`

2. Wykonaj model proporcji lasów w kwadracie (kolumna `Forests`) w zależności od wybranych predyktorów. Potraktuj te dane odpowiednim rozkładem (zero-inflated Beta - proporcja udziału lasów, wykorzystaj pakiet `glmmTMB`), pamiętaj o zamianie na wartości z zakresu 0-1 (podziel przez 100). Sprawdź który z modeli jest lepszy używając funkcji `AIC()`

3. Korzystając ze zbioru danych `afis` przygotuj model występowania wybranego gatunku (np. `Ficavern`) używając jako predyktorów wybranych cech. Pamiętaj że występowanie gatunków w tym zbiorze danych jest wyrażone zerojedynkowo - użyj `glm(...., family = binomial(link='logit'))`

4. Wczytaj zbiór danych `survi` link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/survi.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
survi<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/survi.csv',
                sep=';')
```

W tym zbiorze sprawdź wpływ pH na przeżywalność siewek (kolumna `surv`). Stwórz GLMM (funkcja `glmer` z pakietu `lmerTest` lub `glmmTMB` z pakietu `glmmTMB`) z rozkładem dwumianowym używając `family=binomial(link='logit')` - jako efekt losowy sprawdź rok oraz blok - pomiń efekty związane z plotem.

5. Wczytaj zbiór danych `regen.plots` link: [https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/regen.plots.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
regen<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/regen.plots.csv',
                sep=';')
```

zawiera on liczbę jaworów (Acer), buków (Fagus) i jesionów (Fraxinus) na 100 m2 na 32 poletkach oraz trzy zmienne: DIFN (miarę dostępności światła), ph gleby i cyfrowy zapis kompozycji gatunkowej d-stanu: wartości ujemne - świerkowe, blisko zera - bukowe, dodatnie - jaworzyny i łęgi. riv to kod doliny rzecznej - potraktuj go jako random intercept

a. przygotuj model liczebności jawora używając funkcji `glmmTMB()` z pakietu `glmmTMB`. Sprawdź za pomocą funkcji `testDispersion(simulateResiduals(model))` i `testZeroInflation(simulateResiduals(model))` dyspersję i zera (pakiet `DHARMa`. Po zbudowaniu modelu globalnego zredukuj go używając funkcji `dredge`. Jaki model jest ostateczny?

b. przygotuj model dla buka - czy można zastosować rozkład Poissona, czy ujemny dwumianowy? sprawdź wielkość efektów używając funkcji `ggpredict` z pakietu `ggeffects`

Propozycje do pracy z własnym zbiorem danych

5. Przetestuj hipotezy o wpływie czynników na zmienną zależną używając odpowiednich modeli. Weź pod uwagę rozkłady i logikę badanych zmiennych - np. tempo wzrostu korzeni nie może być ujemne, a temperatura ciała poniżej pewnej wartości oznacza śmierć.
6. Sprawdź czy do modelu należy włączyć efekty losowe - czasem może to przewrócić wnioskowanie do góry nogami, ale lepiej zinterpretować to teraz niż po uwagach recenzenta;) Zastanów się co może być modyfikowane przez czynniki losowe - nachylenie krzywej (tempo odpowiedzi) czy też tylko jej położenie (intercept)?
7. Jeśli korzystasz z analizy wariancji zastanów się czy nie włączyć do niej efektów losowych - spróbuj wrzucić w `anova()` obiekt typu `lmer` zamiast `lm`. Sprawdź odpowiedzi brzegowe używając funkcji `cld` i `emmeans`