

# Dzień 4 - Uogólnione modele liniowe, modele addytywne (GLM, GLMM, GAM) - zadania

Marcin K. Dyderski, Patryk Czortek

20 kwietnia 2023

## 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/datasets/afis.csv',  
              sep=';')
```

W zbiorze danych mamy informacje o udziale procentowym terenów otwartych (agricultural, semi-natural & wetlands, 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** przygotuj wykres na którym pokażesz zależność pomiędzy **AFIS** a **Forests** z linią regresji zakładającą rozkład Poissona w oparciu o `geom_smooth(method='glm', method.args=list(family='poisson'))` dla celów poglądowych.
  - b. Wykonaj model dla liczby gatunków wskaźnikowych starych lasów (kolumna **AFIS**) w oparciu o trzy predyktory: **Water**, **Urban.dense** oraz **OLDFR**. Z uwagi na charakter danych skorzystaj z rozkładu Poissona używając funkcji `glm(..., family=poisson)`.
  - c. Za pomocą `testDispersion()` sprawdź czy jest problem z dyspersją, jeśli jest to zastosuj odpowiedni rozkład używając `glmmTMB`
  - d. Sprawdź VIFy, dokonaj selekcji zmiennych za pomocą `dredge()`
  - e. Sprawdź model finalny - obejrzyj `summary()`, współczynniki  $r^2$ , przygotuj wizualizację za pomocą `ggpredict()`
2. Wykonaj analogiczny model używając zamiast **OLDFR** kolumny **Forests**. Potraktuj te dane odpowiednim rozkładem (Beta - proporcja udziału lasów), pamiętaj o zamianie na wartości z zakresu 0-1. Przeprowadź cały cykl budowy modelu (VIFy, dredge), oceń model finalny i przygotuj wizualizację.
  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 zerojedynekowo - użyj `glm(..., family = binomial(link='logit'))`. Przeprowadź cały cykl budowy modelu (VIFy, dredge), oceń model finalny i przygotuj wizualizację.
  4. Wczytaj zbiór danych **survi** link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasets/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 zbiorze tym sprawdź wpływ pH, light, cover i stand\_type na przeżywalność siewek (kolumna surv). Stwórz GLMM (funkcja `glmer` z pakietu `lmerTest`) z rozkładem dwumianowym używając `family=binomial(link='logit')` - jako efekt losowy sprawdź rok oraz blok - pomiń efekty związane z plotem. Przeprowadź cały cykl budowy modelu (VIFy, dredge), oceń model finalny i przygotuj wizualizację.

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

```
survi<-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 m<sup>2</sup> 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 legi. 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. 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`
- c. przygotuj model dla jesionu i porównaj go z pozostałymi
- d. dla wybranego przez siebie modelu przygotuj wizualizację obejmującą punkty (dane rzeczywiste) i krzywe z modelu

##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`