

# Dzień 3 - Regresja, modele liniowe

Patryk Czortek, Marcin K. Dyderski

24 kwietnia 2024

## Zadania do wykonania

1. Wczytaj plik 'prunus.csv' dostępny na githubie, link: [https://github.com/mkdyderski/BSS/blob/BSS2024/datasets/prunus.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
prunus<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasets/prunus.csv',  
                sep=';')
```

Zawiera on dane wykorzystane w pracy Dyderski i Jagodzinski 2015 [https://www.forestry.actapol.net/pub/2\_2\_2015.pdf]. Opis zmiennych: typ - typ roślinności (Car-Aln to ols, Fra-Aln to łęg olszowo-jesionowy, transit to zbiorowisko przejściowe - między olsem a łęgiem, LZZ - skrajnie zdegenerowany, przesuszony i brzydki łęg), a - zwarcie warstwy drzew (%) b - zwarcie warstwy krzewów (%) c - pokrycie runa zielnego (%) d - pokrycie warstwy mszystej (%) prunusc - liczba osobników czeremchy w warstwie zielnej, prunusb - liczba osobników czeremchy w warstwie krzewów, richness - bogactwo gatunkowe runa, shannon - wskaźnik różnorodności Shannona dla runa, L - wskaźnik świetlny Ellenberga (1-9, 1-cień, 9-pełne słońce), M - wskaźnik wilgotności Ellenberga (1-12, 1- pustynia, 12 - rośliny zanurzone), SR - wskaźnik odczynu gleby (1-9, 1-kwaśne, 9-lekko zasadowe, 7- obojętne), N - wskaźnik żyzności (1-9, 1-ubogie, 9-bardzo żyzne)

2. Przygotuj model liniowy prunusc w oparciu o kilka zmiennych - wybierz najlepszy model w oparciu o AIC. Możesz zrobić to ręcznie przy użyciu `AIC()` lub półautomatycznie używając `step()` lub `MuMIn::dredge()`, jednak wtedy zastanów się które parametry jest sens potraktować jako potencjalne predyktory.
3. Przygotuj wizualizację modelu z punktu 2. przy użyciu pakietu `ggpredict`.
4. Wczytaj zbiór danych `hotspots` link: [https://github.com/mkdyderski/BSS/blob/BSS2024/datasets/hotspots.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
hotspots<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasets/hotspots.csv',  
                  sep=';')
```

5. Stwórz model mieszany (funkcja `lmer` z pakietu `lmerTest`) bogactwa gatunkowego roślin (kolumna `plants`) z efektami losowymi (`continent`) oraz stałymi (wybierz interesujące Cię kolumny;), sprawdź VIFy za pomocą funkcji `car::vif()`, przeprowadź selekcję zmiennych za pomocą `MuMIn::dredge()` i wybierz model finalny.
6. Sprawdź `summary()` modelu i za pomocą funkcji `r.squaredGLMM()` z pakietu `MuMIn` sprawdź R<sup>2</sup>c i R<sup>2</sup>m.
7. Za pomocą `ggeffects::ggpredict()` przygotuj wizualizację danych.

##Propozycje do pracy z własnym zbiorem danych 9. Wczytaj *własny zbiór danych* i sprawdź korelacje pomiędzy zmiennymi liczbowymi - przygotuj ładną wizualizację macierzy korelacji, którą będzie można pokazać promotorowi;) 10. Wykonaj model liniowy przedstawiający relacje pomiędzy cechami dla których zakładasz występowanie pewnych zależności. Najlepiej spróbuj przetestować zależności które udało Ci się wczoraj zwizualizować. Zacznij od modeli z jedną zmienną objaśniającą. Sprawdź potencjalne problemy z modelami przy użyciu wykresów diagnostycznych. 11. Zastanów się, czy modele które przygotowałeś mogą mieć problem związany z obserwacjami odstającymi. Jeśli tak, przetestuj wariant z ich wyłączeniem. Jeśli nie, zastanów się czy problemem słabego dopasowania modeli jest rozkład danych. 12. Sprawdź czy dołożenie do modeli kolejnych zmiennych spowoduje wzrost mocy predykcyjnej. Przetestuj modele w oparciu o AIC oraz R<sup>2</sup>. W przypadku problemów z naturą danych (rozkłady itp.) poproś o pomoc prowadzących aby przejść od razu do modeli uogólnionych. 13. Przygotuj wykres i tabelę z wybranym modelem liniowym. Wzoruj się na publikacjach ze swojej działości lub zapytaj co musi się tam znaleźć.