Dzień 3 - Testy statystyczne i regresja - zadania

Marcin K. Dyderski, Patryk Czortek

3 kwietnia 2019

Zadania do wykonania

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

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

Używając funkcji `nls()` stwórz model nieliniowy biomasy części nadziemnej `AB` jako funkcji `Hg` używając formuł przedstawionych na wykładzie. Wykonaj wykres używając pakietu `ggplot` i dodając linię modelu za pomocą `geom_smooth(method='nls'....)`, pamiętaj o `SE=FALSE`. 2. Wczytaj zbiór danych dotyczący występowania gatunków wskaźnikowych starych lasów w Poznaniu. Pochodzi on z publikacji Dyderski et al. 2017. Urban For Urban Greening [https://www.sciencedirect.com/science/article/pii/S161886671730078X?via%3Dihub]. link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/afis.csv]. 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 (`OLDFOR`,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 (`AFIS`) jako funkcji `Water`, `Urban.dense` oraz `OLDFOR`. Z uwagi na charakter danych skorzystaj z rozkładu Poisson używając funkcji `glm(...., family=poisson)` b. Wykonaj analogiczny model używając zamiast `OLDFOR` kolumny forest. Sprawdź który z modeli jest lepszy używając funkcji `AIC()` c. 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'))`. *d. Sprawdź jak zmieni się jakość dopasowania modelu (wykresy diagnostyczne) po zastosowaniu modelu typu zero-inflated (zadanie dla chętnych)

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

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

Stwórz model liniowy bogactwa gatunkowego ptaków z efektami losowymi (`continent`) oraz stałymi (wybierz interesujące Cię;) i za pomocą funkcji `r.squaredGLMM()` z pakietu `MuMIn` sprawdź R2c i R2m. 5. 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 zbiorze tym sprawdź wpływ pH na przeżywalność siewek. Stwórz GLMM z rozkładem dwumianowym

używając `family=binomial(link='logit')` - jako efekt losowy sprawdź rok oraz blok - pomiń efekty związane z plotem.

6. Badano skład gatunkowy gatunków roślin runa na powierzchniach z: (i) usuniętym martwym świerkiem (litera „c" przy id powierzchni), (ii) nieusuniętym martwym świerkiem zabitym przez kornika drukarza (litera „d" przy id powierzchni) oraz (iii) drzewostanem nietkniętym przez gradację kornika (litera „f" przy id powierzchni). Dane zawarto w pliku `dend.csv`. link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/dend.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

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

a) Zaimportować dane do R
b) Za pomocą funkcji `vegdist()` w bibliotece `vegan` stworzyć macierz niepodobieństwa Bray-Curtisa pomiędzy próbami. Przed stworzeniem macierzy należy przeprowadzić transpozycję kolumn z wierszami.
c) Wyniki macierzy zobrazować za pomocą dendrogramów używając metody najbliższego i najdalszego sąsiada oraz metodą centroidów. Czy są obecne różnice? Która metoda okazuje się najlepsza do analizy powyższych danych?
d) Dendrogram najlepiej obrazujący wyniki

- korzystając z biblioteki `ape` (jednocześnie pamiętając o konflikcie `ape` z `vegan`) zobrazować za pomocą różnych metod graficznych (`triangle`, `unrooted`, `fan` i `radial`)
- zmienić kolor, typ linii oraz kolor tekstu
- podzielić na trzy klasy

e) Korzystając z funkcji `beta.pair()` (biblioteka `betapart`) stworzyć macierz niepodobieństwa Sorensena pomiędzy próbami. Przed stworzeniem macierzy należy dokonać transformacji danych do postaci binarnej (przy użyciu funkcji `vegan::decostand()`).
f) Dane z macierzy zobrazować w postaci dendrogramu wybierając najlepszą strategię
g) Dla każdego typu powierzchni obliczyć średni wskaźnik różnorodności Shannona-Wienera i wskazać siedlisko o największej różnorodności gatunkowej. Dane id siedliska:

```
id<-c(rep("clearcut", 30), rep("dead", 19), rep("forest", 25), rep("dead", 2),
      "forest", rep("dead", 3), rep("forest", 4), rep("dead", 4), "forest")`
```

należy skleić kolumnami z obiektem zawierającym wskaźniki Shannona-Wienera policzone dla każdej powierzchni w następujący sposób:

```
szanon<-cbind(as.data.frame(Shannon.index), id)
```

Wtedy średni wskaźnik Shannona-Wienera dla każdego typu powierzchni można policzyć używając następującego kodu:

```
mean(szanon$Shannon.index[szanon$id=="typ_siedliska"])
```

albo korzystając z pakietu `dplyr`:

```
library(dplyr)
szanon%>%group_by(id)%>%summarise(m=mean(Shannon.index))
```

h) Używając tej samej procedury jak w podpunkcie (g), dla każdego typu powierzchni obliczyć średni wskaźnik równocenności Pielou. O czym mówi wskaźnik i czy są różnice pomiędzy trzema siedliskami?

Propozycje do pracy z własnym zbiorem danych

10. 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.

11. 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)?
12. Jeśli korzystasz z analizy wariancji zastanów się czy nie włączyć do niej efektów losowych - spróbuj wrzucić w `anova()` obiekt typu `lmer` zamiast `lm`
13. Pracując na danych różnorodnościowych oblicz dla swoich danych wskaźniki różnorodności i porównaj ze swoimi wcześniejszymi obliczeniami (jeśli masz). Czy są jakieś różnice? Z czego mogą wynikać?
14. Czy można pogrupować obserwacje wg cech? Używając dendrogramów można np. sprawdzić czy obserwacje przyporządkowane do pewnych grup (jednostek fitosocjologicznych, lat, grup poletek, gatunków) grupują się wg tego klucza czy inaczej. Może się okazać, że np. fitosocjologia nie odzwierciedla rzeczywistości;)