Dzień 3 - Testy statystyczne i regresja - zadania

Patryk Czortek, Marcin K. Dyderski

3 kwietnia 2019

Zadania do wykonania

1. Dane zawarte w pliku `lichenes1.csv` reprezentują bogactwo (kolumna `Rich`) i różnorodność gatunkową (kolumna `Shan`) oraz proporcję gatunków porostów epifitycznych o różnych wymaganiach względem zasobności podłoża w azot (kolumna `EIV_N`) w Puszczy Białowieskiej na 144 powierzchniach historycznych z 1992 roku (kolumna `time=='h'`) oraz na 144 powierzchniach powtórnie przebadanych w roku 2014 (kolumna `time=='n'`) wraz z danymi odnośnie typu zbiorowiska leśnego dla każdej powierzchni (kolumna `habitat`). link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/lichenes1.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
lichenes<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/lichenes1.csv',
                   sep=';')
```

   a) Korzystając z funkcji `hist()` lub `ggplot2::geom_histogram()` ocenić, czy bogactwo i różnorodność gatunkowa prób historycznych i powtórnie przebadanych reprezentują rozkład normalny
   b) Zakładając, że dane reprezentują rozkład normalny, zaproponować rodzaj testu statystycznego, odpowiedniego do zbadania różnic w różnorodności gatunkowej pomiędzy dwoma typami zbiorowisk leśnych. W którym zbiorowisku różnorodność gatunkowa była większa? Czy różnice były istotne statystycznie? A biologicznie?
   c) Zakładając, że dane reprezentują rozkład normalny, zaproponować rodzaj testu statystycznego, odpowiedniego do zbadania różnic w proporcji gatunków porostów epifitycznych o różnych wymaganiach względem zasobności podłoża w azot pomiędzy danymi historycznymi i powtórnie przebadanymi. Kiedy średni udział porostów o wyższych wymaganiach względem azotu był większy – w 1992 roku, czy w roku 2014? Czy różnice były istotne statystycznie? Ocenić, czy różnice w czasie były duże, czy niewielkie.

2. Po ponad 90 latach od pierwszych obserwacji florystycznych badano zmiany w bogactwie gatunkowym wyleżysk (plik `wylezyska.csv`; kolumna `rich`). Zakładając, że zarówno dane historyczne (kolumna `time=='k'`), jak i powtórnie przebadane (kolumna `time=='n'`) nie reprezentują rozkładu normalnego, oraz że dane w 2015 roku były pobrane dokładnie z tych samych lokalizacji, co w 1927 roku, zaproponować rodzaj testu statystycznego, odpowiedniego do zbadania różnic w bogactwie gatunkowym pomiędzy dwoma okresami badawczymi. Czy różnice w bogactwie gatunkowym były istotne statystycznie? link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/wylezyska.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

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

3. W pliku `freq.epiphytes.csv` zawarto zmiany we frekwencji 10 gatunków porostów epifitycznych po 30 latach od pierwszych badań. Ile gatunków istotnie zwiększyło/zmniejszyło częstość występowania w porównaniu do stanu sprzed 30 lat? link: [https://github.com/mkdyderski/BSS/blob/BSS2019/datasety/freq.epiphytes.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
freq.epiphytes<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/freq.epiphy



4. Wczytaj plik 'prunus.csv' dostepny na githubie, link: [https://github.com/mkdyderski/BSS/blob/BSS201
```

```
prunus<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/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 - miedzy olsem a łęgiem, LZZ - skrajnie zdegenerowany, przesuszony i brzydki łeg), 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 żyznosci (1-9, 1-ubogie, 9-bardzo żyzne)

5. Sprawdź korelację L z a, richness z shannon oraz prunusc z M
6. Wykonaj macierz korelacji dla wszystkich zmiennych liczbowych w tym datasecie i zwizualizują ją za pomocą pakietu `corrplot`.
7. Przygotuj model liniowy prunusb jako funkcji N i wykonaj wykresy diagnostyczne.Jaki jest współczynnik determinacji (R^2)? czy model jest istotny statystycznie? Sprawdź efekt usunięcia potencjalnie odstających obserwacji.
8. Przygotuj model liniowy prunusc w oparciu o kilka zmiennych - wybierz najlepszą w oparciu o AIC. Możesz zrobić to ręcznie przy użyciu `AIC()` lub półautomatycznie używając `step()` lib `MuMIn::dredge()`, jednak wtedy zastanów się które parametry jest sens potraktować jako potencjalne predyktory. W razie wątpliwości skonsultuj się z prowadzącymi.
9. Za pomocą jednoczynnikowej analizy wariancji sprawdź czy są różnice w L pomiędzy typami roślinności. Jeśli są, za pomocą testu Tukeya sprawdź pomiędzy którymi.

Propozycje do pracy z własnym zbiorem danych

10. Wczytaj *własny zbiór danych* i sprawdź korelacje pomiędzy zmiennymi liczbowymi - przygotuj ładną wizualizację macierzy korelacji, którą będzie można pokazać promotorowi;)
11. 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.
12. 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.
13. Sprawdź czy dołożenie do modeli kolejnych zmiennych spowoduje wzrost mocy predykcyjnej. Przetestuj modele w oparciu o AIC oraz R2. W przypadku problemów z naturą danych (rozkłady itp.) poproś o pomoc prowadzących aby przejść od razu do modeli uogólnionych.
14. Sprawdź czy badane cechy różnią sie pomiędzy grupami za pomocą testów t-Studenta/chi-kwadrat lub analizy wariancji. Jeśli wykonujesz analizę wariancji, pamiętaj o testach post-hoc (Tukeya).
15. Przygotuj wykres i tabelę z wybranym modelem liniowym lub analizą wariancji. Wzoruj się na publikacjach ze swojej działki lub zapytaj co musi się tam znaleźć.