Wrocław, 5.11.2022 r.

**PODSTAWY ANALIZY DANYCH – PRAKTYCZNE WARSZTATY**

**ZADANIA DO SAMODZIELNEGO PRZEĆWICZENIA, LISTA NR 3**

1. Proszę obejrzeć stronę internetową p. Przemysława Biecka poświęconą wykonywaniu grafiki przy pomocy pakietu *ggplot2* i pobawić się😊:

<https://pbiecek.gitbooks.io/przewodnik/content/Wizualizacja/jak_tworzyc_wykresy_ggplot2.html>

1. Proszę rzucić okiem na stronę:

<http://pbiecek.github.io/Przewodnik/Programowanie/generatory_2.html>

z internetowego przewodnika Przemysława Biecka. Znajdą na niej Państwo opis, jak pracować z popularnymi rozkładami zmiennych losowych (ciągłych i dyskretnych). Proszę pobawić się generowaniem krzywych pokazujących funkcje gęstości i dystrybuanty rozkładów. W szczególności, proszę obejrzeć rozkłady: normalny, Studenta, chi-kwadrat i F. Aby uzyskać rozkłady o różnych parametrach wpisujemy m.in.:

- dla rozkładu normalnego: dnorm(x, wartość\_oczekiwana, odchylenie\_standardowe),

- dla rozkładu Studenta: dt(x, liczba\_stopni\_swobody),

- dla rozkładu chi-kwadrat: dchisq(x, liczba\_stopni\_swobody),

- dla rozkładu F: df(x, liczba\_stopni\_swobody\_1, liczba\_stopni\_swobody\_2)

Proszę sprawdzić, jak zmieniają się funkcje gęstości dla różnej liczby stopni swobody.

1. W naszym zespole na platformie MS Teams znajdą Państwo trzy pliki Excela z danymi: „Proszki\_do\_prania”, „Produkcja\_koszt” i „Przestoje”. Dane pochodzą z podręcznika S. Ostasiewicz, Z. Rusnak i U. Siedleckiej „Statystyka. Elementy teorii i zadania”, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 1989, str. 330, odpowiednio zad. nr 12.34, 12.35 i 12.37 (opis danych został dostosowany do niniejszych zajęć).

„Proszki\_do\_prania” zawierają dane na temat zużycia proszków do prania w pewnym kraju, w różnych momentach czasowych. Zatem, dane przedstawiają szereg czasowy.

Dane z pliku „Produkcja\_koszt” pokazują, jak zmieniała się produkcja i koszty jednostkowe produkcji w pewnym zakładzie. Dane z pliku „Przestoje” pokazują wyniki działań podejmowanych przez firmę w celu zmniejszenia czasu trwania przestojów w pracy pewnej firmy.

Proszę obejrzeć dane na wykresach rozrzutu (scatterplots) i zasugerować modele, które mogą dobrze opisać omawiane zależności. Scatterplots można sporządzić korzystając z funkcji plot(x,y) z pakietu base lub wykorzystując funkcję scatterplot(x,y) z pakietu „car”. Korzystając z pakietu car warto wyrzucić prostą fegresji, która się automatycznie pokaże (parametr regLine=F) i efekt działania parametru smooth (wpisujemy smooth=F). Następnie proszę pobawić się modelami liniowymi (regresją liniową prostą) i spróbować ocenić istotność parametrów modelu, przyjrzeć się wartościom współczynnika R^2 oraz wstępnie ocenić, czy Państwa modele spełniają założenia metody najmniejszych kwadratów. Zauważmy, że liczba danych jest niewielka, co może wpłynąć na jakość oceny.

1. Proszę dodać do modeli z zadania nr 1 składnik I(x^2) i zobaczyć, czy i jak modele się popsuły/poprawiły. Jak teraz wyglądają wykresy diagnostyczne?
2. Proszę wyciągnąć reszty z modeli z zadania nr 1. Czy wartość bezwzględna którejkolwiek z nich jest większa od 2? Proszę wyciągnąć tzw. reszty standaryzowane i studentyzowane (w niedalekiej przyszłości powiemy sobie o nich więcej) z modeli używając odpowiednio komend: rstandard(obiekt\_w\_którym\_jest\_model) i rstudent(obiekt\_w\_którym\_jest\_model). Przykład: model2=lm(y~x), rstandard(model2), rstudent(model2). Jak teraz wyglądają reszty?
3. Budując model regresji liniowej można wymusić włączenie do niego początku układu współrzędnych. Robimy to przez komendę lm(y~x-1), (-1) wymusza wartość interceptu równą zero. Proszę sprawdzić, czy komenda działa na dowolnych przykładach. Proszę zauważyć, że po wywołaniu podsumowania modelu (summary(model)) nie widzimy oszacowanej wartości interceptu. Jest to logiczne, bo jego zerowa wartość została wymuszona. Jak można przedstawić model z zerowym interceptem na wykresie?