Zadanie kolokwium 2

Wojciech Ganobis 310519, Bartosz Troszka 309912

05/05/20

Celem zadania jest obliczenie wspolczynnika korelacji miedzy zachorowaniami, na podstawie dostarczonych danych. Najpierw omowmy metode zrobienia tego zadnaia.
Do wykonania zadnia uzyjemy wspolczynnika korelacji Pearsona. Wyraza sie on wzorem:

$$r_{xy} = \frac{\operatorname{cov}(r, y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} =$$

$$= \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i)^2] \cdot [n \cdot \sum y_i^2 - (\sum y_i)^2]}},$$

gdzie $r_{XY} \in [-1, 1]$. Im wieksza wartosc bezwzgledna, tym silniejsza zaleznosc miedzy cechami. Jesli $r_{XY} == 1$ to zaleznosc jest dokladnie liniowa. Natomiast jesli $r_{XY} == 0$ to nie ma liniowej zaleznosci pomiedzy cechami. Zaleznoscli mozna opisac za pomoca tabelki:

Korelacje	Ujemne	Dodatnie
Slabe	-0.5 do 0.0	0.0 do 0.5
Silne	-1.0 do -0.5	0.5 do 1.0

Kolejna rzecza wazna w zadaniu jest wybranie punktu poczatkowego. "Pierwsze" dane w kazdym panstwie sa niedokladne oraz niezetelne. Zaczniemy wiec obliczanie korelacji gdy liczba zakazonych bedzie zwiekszac sie kazdego dnia.
Przejdzmy teraz do obliczenia pierwszej korelacji.

- Pierwsze panstwa miedzy ktorymi bedziemy szukac korelacji to Wlochy i Hiszpania. Jako pinkt poczatkowy dla Wloch wybierzemy 22 luty, poniewaz po tym dniu kazdego nastepnego dnia liczba zarazonych przybywa. Dla Hiszpani natomiast wybierzemy 27 luty, z tego samego powodu. Z racji ze dane mialy byc do 30 kwietnia punkt koncowy dla Hiszpani to 30 kwietnia, a dla Wloszech 25 kwiatnia, alby liczba dni byla taka sama. Do obliczenia wykorzystamy wzor $\frac{n\cdot\sum x_i y_i - \sum x_i \cdot \sum y_i}{\sqrt{[n\cdot\sum x_i^2 - (\sum x_i)^2]\cdot[n\cdot\sum y_i^2 - (\sum y_i)^2]}}$.

  Dla ulatwienia obliczen zrobimy tableke:

x	y	$x_i y_i$	x_i^2	y_i^2
6	10	60	36	100
67	13	871	4489	169
48	7	336	2304	49
105	13	1365	11025	169
...	...	...	...	...

$\sum$ kolumna1	$\sum$ kolumna2	$\sum$ kolumna3	$\sum$ kolumna4	$\sum$ kolumna5
192991	225045	988355044	810008099	1286773571

  Teraz mozna latwo obliczyc wspolczunnik wstawiajac do wzoru:

$$\frac{64 \cdot 988355044 - 192991 \cdot 225045}{\sqrt{[64 \cdot 810008099 - (192991)^2] \cdot [64 \cdot 1286773571 - (225045)^2]}} = 0.8827529322$$

  Wynik ten oznacza ze korelacja jest silna dodatnia.

- Obliczny teraz korealacje dla innych panstw. Wezmy np Polske oraz Francje. Za punkt poczatkowy Polski wezmy 7 marzec, a koncowy 3 maja. Dla francji bedzie to 27 luty i 24 kwietnia. Metoda ta sama wiec nie bede zapisywal obliczen tylko sam wynik ktory wynosi:

$$0,7512537726$$

  Oznacza to rowniez ze korelacja jest silna, jednak slabsza niz miedzy Hiszpania a Wlochami.

- Obliczmy teraz korelacje miedzy naszymi sasiadami mianowicie miedzy Czechami a Slowacja. Za punkt poczatkowy Slowacji wezmy 5 marca, a koncowy 2 maja. W Czechach namiosast dzien pozniej czyli 6 marca do 3 maja. Korelacja wynosi:

$$0,60372796$$

  Korelacja jest juz na skraju korelacji silnej i korelacji slabej.

- Porownjamy teraz USA oraz Polske, przypuszczam ze korelacja bedzie slaba, poniewaz po polsce wirus nie roprzestrzenial sie tak szybko z powodu wprowadzenia ograniczen oraz z powodu mniejszej ilosci zaludnienia. Sprawdzmy wiec matematycznie czy nasz przypuszczenia sa poprawne. W USA za pierwszy dzien wezmieemy 22 luty, a ostatni. W Polsce natomiast pierwszy bedzie 7 marca, a ostatni 3 maja. Nasz wynik to:

  $$0,6353313598$$

  Widzimy wiec ze mimo roznej gestosci zaludnienia korelacja zachodzi calkiem niezle. Jest nawet wieksza niz miedzy Czechami a Slowacja, co jest zaskakujace.

- Sprawdzmy teraz korelacje miedzy panstwami beienych panstw innych kontynentow. Wezmy za przyklad Brazylie oraz Azerbejdzan. W Azerbejdzanie wyniki bedziemy brac z dni od 13 marca do 3 maja, natomiast z Brazyli od 5 maja do 25 kwietnia. Nasz korelacja to:

  $$0,1799759062$$

  Jak widac korelacja jest fatalna, oznacza to ze dane nie sa zetelne, wynika to prawdopodobnie z miedzy innymi ilosci robionych testow.

- Sprawdzmy jeszcze korelacje miedzy USA a Azerbejdzanem. Azerbejdzan bedzie zawieral dni od 13 marca do 3 maja, a USA od 22 luty do 13 kwietnia. Wynik to:

  $$-0,08316804048$$

  Oznacza to ze korelacja praktycznie nie istnieje, jest bardzo bliska 0.

Z powyzszych obliczen mozna wywnioskowac ze zachodzi korelacja miedzy zachorowaniami miedzy wiekszoscia panstw. Jednak model nie jest uniwersalny co widac miedzy korelacja Brazylii a Azerbejdzamu oraz Azerbejdzanu oraz USA.
Zadanie zostalo wykonane na wersje punktowa 10+4 poniewaz obliczono korelacje dla wiekszej ilosci panstw.