MSP projekt

Jakub Kryštůfek - xkryst02

Prosinec 2022

Úvod

Veškeré výpočty byly provedeny v jazyce python s využitím Jupyter Notebook. Zdrojové kódy pro první úkol jsou v souboru `task1.ipynb` a pro druhý úkol v souboru `task2.ipynb`. Data pro druhý úkol byly vyexportovány do csv souboru `data.csv`.

Úkol 1

Data pro první úkol								
	Praha	Brno	Znojmo	Tišnov	Rokytn.	Jabl.	Dolní Věstonice	Mé okolí
Zimní	510	324	302	257	147	66	87	3
Letní	352	284	185	178	87	58	65	24
Střídání	257	178	124	78	44	33	31	18
Nemá názor	208	129	70	74	6	19	32	15

Tabulka 1: Data pro první úkol

Pro podúkoly **a) - e)** bylo využito testu dobré shody, pouze se počítalo s jinými daty a pro **d) - e)** byly sjednoceny některé sloupce. Z toho důvodu proces výpočtu popíšu detailněji pouze pro **a)** a u ostatních pouze ukážu odlišnosti, mezivýsledky a závěry.

Úkol 1 a)

Budeme pracovat s řádkem `zimní čas` z tabulky 1. Nejprve bylo nutné bodově odhadnout pravděpodobnosti a následně teoretické četnosti (tabulka 2).

	Praha	Brno	Znojmo	Tišnov	Rokytn.	Jabl.	Dolní Věstonice	Mé okolí
Zimní	530.17	365.56	272.07	234.52	113.46	70.31	85.89	23.97

Tabulka 2: Teoretická četnost pro a)

Následně se vypočítalo testovací kritérium, které vyšlo **39.476**. Jelikož jsme museli bodově odhadnout pravděpodobnosti, pak je stupeň volnosti pro rozdělení chi kvadrát **8-1-1 = 6**. Doplněk kritického oboru nám tedy vyjde **(0; 12.591)**. Jelikož testovací kritérium nepatří do doplňku kritického oboru, pak hypotézu, že je ve městech, obcích a okolí studenta stejné procentuální zastoupení obyvatel, co preferují zimní čas **zamítáme**.

Úkol 1 b)

Budeme pracovat s řádkem `zimní čas` z tabulky 1. Odhadnuté teoretické četnosti lze vidět v tabulce 3.

	Praha	Brno	Znojmo	Tišnov	Rokytn.	Jabl.	Dolní Věstonice	Mé okolí
Letní	385.43	265.77	197.80	170.49	82.49	51.12	62.44	17.42

Tabulka 3: Teoretická četnost pro b)

Testovací kritérium vyšlo **9.065**. Stupeň volnosti opět $8-1-1=$ **6** a doplněk kritického oboru je tedy **(0; 12.591)**. Testovací kritérium tedy spadá do doplňku kritického oboru a tím pádem hypotézu, že je ve městech, obcích a okolí studenta je stejné procentuální zastoupení obyvatel, co preferují letní čas **nezamítáme**.

Úkol 1 c)

Budeme pracovat s řádkem `nemá názor` z tabulky 1. Odhadnuté teoretické četnosti lze vidět v tabulce 4.

	Praha	Brno	Znojmo	Tišnov	Rokytn.	Jabl.	Dolní Věstonice	Mé okolí
Nemá názor	238.51	164.46	122.40	105.50	51.04	31.63	38.64	10.78

Tabulka 4: Teoretická četnost pro c)

Testovací kritérium vyšlo **17.110**. Stupeň volnosti opět $8-1-1=$ **6** a doplněk kritického oboru je tedy **(0; 12.591)**. Testovací kritérium tedy nespadá

do doplňku kritického oboru a tím pádem hypotézu, že je ve městech, obcích a okolí studenta je stejné procentuální zastoupení obyvatel, co preferují letní čas **zamítáme**.

Úkol 1 d)

Pro tento podúkol a také pro podúkol e) jsem sjednotil sumou sloupce které patřily do stejných kategorií (Větší města, Menší města a Obce). Tato data jsou zobrazena v tabulce 5.

	Větší města	Menší města	Obce
Zimní	834	559	300
Letní	636	363	210
Střídání	435	202	108
Nemá názor	337	144	57

Tabulka 5: Data pro první úkol

Budeme pracovat s řádkem `Zimní` z tabulky 5. Zbytek postupu je stejný jako v předcházejících příkladech. Odhadnuté teoretické četnosti lze vidět v tabulce 6.

	Větší města	Menší města	Obce
Zimní	906.97	512.95	273.06

Tabulka 6: Teoretická četnost pro d)

Testovací kritérium vyšlo **12.661**. Stupeň volnosti je $3-1-1=\mathbf{1}$ a doplněk kritického oboru je tedy **(0; 3.841)**. Testovací kritérium tedy nespadá do doplňku kritického oboru a tím pádem hypotézu, že je ve větších městech, menších městech a obcích je stejné procentuální zastoupení obyvatel, co preferují zimní čas **zamítáme**.

Úkol 1 e)

Budeme pracovat s řádkem `Nemá názor` z tabulky 5. Zbytek postupu je stejný jako v předcházejících příkladech. Odhadnuté teoretické četnosti lze vidět v tabulce 7.

	Větší města	Menší města	Obce
Nemá názor	288.21	163.00	86.77

Tabulka 7: Teoretická četnost pro e)

Testovací kritérium vyšlo **20.688**. Stupeň volnosti je opět $3-1-1=\mathbf{1}$ a doplněk kritického oboru je tedy **(0; 3.841)**. Testovací kritérium tedy nespadá do doplňku kritického oboru a tím pádem hypotézu, že je ve větších městech,

menších městech a obcích je stejné procentuální zastoupení obyvatel, co nemají názor **zamítáme**.

Úkol 1 d)

Pro zjištění, zdali byla moje data sbírána z většího města, menšího města nebo obce jsem se rozhodl použít test dobré shody z kategoriální analýzy nad všemi dvojicemi z mých dat a referenčních dat v tabulce 5 a následně vybrat to s nejvyšším testovacím kritériem.

Postup bude u všech následující. Nejprve si vypočítáme tabulku četností pomocí vzorečku:

$$\frac{n_{i,\circ} \cdot n_{\circ,j}}{n}$$

V této tabulce zkontrolujeme, zdali jsou všechno hodnoty > 5 a následně můžeme vytvořit novou tabulku pomocí vzorečku:

$$\frac{(n_{i,j} - \frac{n_{i,\circ} \cdot n_{\circ,j}}{n})^2}{\frac{n_{i,\circ} \cdot n_{\circ,j}}{n}}$$

Sumou přes všechny prvky této tabulky dostaneme testovací kritérium.

V následujících 3 podsekcích nebudu tento postup již vypisovat, pouze uvedu vypočítané hodnoty.

Větší města x Moje okolí

	Větší města	Moje okolí
Zimní čas	815.184188	21.815812
Letní čas	642.797567	17.202433
Střídání časů	441.192876	11.807124
Nemá názor	342.825369	9.174631

Tabulka 8: Tabulka četností pro Větší města

Všechny hodnoty v tabulce četností jsou > 5.
Testovací kritérium pro Větší města = **26.553**.

Menší města x Moje okolí

	Menší města	Moje okolí
Zimní čas	536.608434	25.391566
Letní čas	369.515060	17.484940
Střídání časů	210.060241	9.939759
Nemá názor	151.816265	7.183735

Tabulka 9: Tabulka četností pro Větší města

Všechny hodnoty v tabulce četností jsou > 5.
Testovací kritérium pro Větší města = **38.975**.

Obce x Moje okolí

	Obce	Moje okolí
Zimní čas	278.265306	24.734694
Letní čas	214.897959	19.102041
Střídání časů	115.714286	10.285714
Nemá názor	66.122449	5.877551

Tabulka 10: Tabulka četností pro Větší města

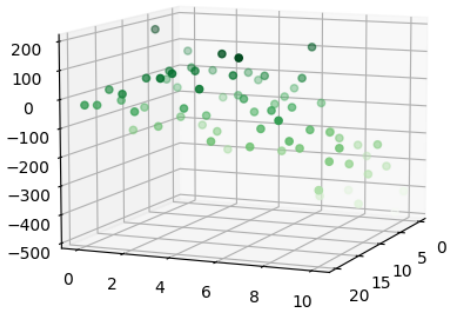
Všechny hodnoty v tabulce četností jsou > 5.
Testovací kritérium pro Větší města = **43.881**.

Závěr

Nejvyšší testovací kritérium měla dvojice Obce a Má data, což znamená, že má data nejvíce odpovídají datům z obcí. Toto i odpovídá realitě, jelikož jsem data sbíral v okolí menší obce.

Úkol 2

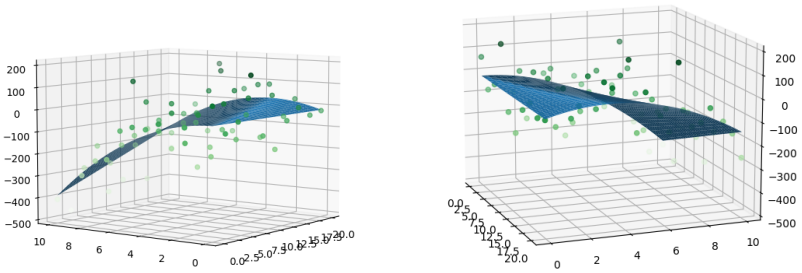
Má sada dat odpovídá číslu **17**. Z důvodu velkého počtu hodnot nebudu uvádět žádné tabulkové výpisy dlouhých mezivýpočtů. Na obrázku 1 jsou referenční data. Jak je vidět, tak data mají na první pohled velice zašuměné hodnoty, což se projeví na výsledcích mých výpočtů. Z experimentálních důvodů jsem svůj pokus vyzkoušel i na datech jiného studenta, který mě již od pohledu data přesně kopírující regresní funkci a pro tento případ mi vycházely mnohem přesnější a hezčí výsledky. O těch se však dále už nebudu zmiňovat.



Obrázek 1: Vstupní data

Úkol 2 a)

Pomocí testu nulovosti regresních parametrů jsem postupným zjednodušováním modelu došel k závěru, že výchozí regresní funkce $Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 XY$ lze zjednodušit na $Z = \beta_1 + \beta_5 Y^2 + \beta_6 XY$ bez závažnějšího dopadu na vhodnost. Koeficient determinace původního modelu vychází $R^2 = 0.556$ a zjednodušený model $R^2 = 0.541$. Následně jsem iterativním zjednodušováním modelu došel až na $Z = \beta_1$, kde ale s každým zjednodušením koeficient determinace znatelně klesal. Z toho důvodu jsem za nejvhodnější model vybral $Z = \beta_1 + \beta_5 Y^2 + \beta_6 XY$.



Obrázek 2: Výsledný model v porovnání s výchozími daty

Úkol 2 b)

Bodové odhady metodou nejmenších čtverců a 95% intervaly spolehlivosti:

	Bodový odhad	95% interval spolehlivosti
β_1	-20.08	(-54.53 ; 14.35)
β_5	-3.71	(-4.56 ; -2.86)
β_6	1.21	(0.63 ; 1.79)

Tabulka 11: Bodové a intervalové odhady vybraného modelu.

Úkol 2 c)

Rozptyl nezávislé proměnné vyšel 9215.015.

Úkol 2 d)

Testoval jsem hypotézu, že β_5 a β_6 jsou nulové.

Pro otestování této hypotézy jsem použil F-test. P hodnota testu vyšla $4.807 * 10^{-12}$ a jelikož je tato hodnota $< \alpha$, tak hypotézu, že jsou tyto dva parametry nulové zamítáme.

Úkol 2 e)

Testoval jsem hypotézu, že β_5 a β_6 jsou stejné.

Pro otestování této hypotézy jsem použil T-test. P hodnota testu vyšla $1.308 * 10^{-10}$ a jelikož je tato hodnota $< \alpha$, tak hypotézu, že jsou tyto dva parametry stejné zamítáme.