MSP projekt Timotej Ponek, xponek00

Úkol 1

Tabuľka zo zadania:

	Praha	Brno	Znojmo	Tišňov	Rokytnice nad Jizerou	Jablunkov	Dolní Věstonice	okolí studenta
Zimní čas	510	324	302	257	147	66	87	14
Letní čas	352	284	185	178	87	58	65	13
střídání časů	257	178	124	78	44	33	31	1
nemá názor	208	129	70	74	6	19	32	2

Nasledujúce hypotézy sú testované na hladine významnosti $\alpha = 0{,}05$.

a)

Pre riešenie si vytvoríme tabuľku odhadovaných četností (ktorú som vyrátal v python notebooku skrz funkciu `chi2_contingency`, ide o 4 parameter ktorý vracia táto funkcia)

	Praha	Brno	Znojmo	Tišňov	Rokytnice nad Jizerou	Jablunkov	Dolní Věstonice	okolí studenta
Zimní čas	537.411	370.559	275.793	237.725	115.015	71.277	87.071	12.149
Letní čas	384.720	265.274	197.433	170.181	82.336	51.025	62.332	8.698
střídání časů	234.862	161.943	120.528	103.891	50.264	31.150	38.052	5.310
nemá názor	170.007	117.224	87.246	75.203	36.384	22.548	27.544	3.843

Táto tabuľka bude používaná aj ďalej v a-c

Hypotéza: V městech, obcích a v okolí studenta (8. průzkumů) je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Použijem test dobrej zhody. Vypočítam si p-hodnotu (skrz funkciu `chisquare`), ktorá vyjadruje, akú veľkú oporu má naša hypotéza v pozorovaných dátach (čím nižšia hodnota, tým nižšia opora). Dá sa definovať aj ako najmenšia hladina významnosti testu, pri ktorej zamietnem nulovú hypotézu.
P-hodnozu porovnám s mojou hladinou významnosti α
p = 0.0067
p < α
Hypotézu zamietam.

b)

Hypotéza: V městech, obcích a v okolí studenta (8. průzkumů) je stejné procentuální zastoupení obyvatel, co preferují letní čas.

Použijem test dobrej zhody. Vypočítam si p-hodnotu skrz funkciu `chisquare` a tú porovnám s mojou hladinou významnosti α
p = 0.3258
p > α
Hypotézu nezamietam.

c)

Hypotéza: V městech, obcích a v okolí studenta (8. průzkumů) je stejné procentuální zastoupení obyvatel, co preferují střídání času.

Použijem test dobrej zhody. Vypočítam si p-hodnotu skrz funkciu `chisquare` a tú porovnám s mojou hladinou významnosti α
p = 2.4094*10^-5
p < α
Hypotézu zamietam.

d)

Pre riešenie si vytvoríme tabuľku odhadovaných četností (ktorú som vyrátal v python notebooku skrz funkciu `chi2_contingency`, ide o 4 parameter ktorý vracia táto funkcia)

	větší města	menší města	obce
Zimní čas	906.979	512.957	273.065
Letní čas	647.689	366.311	195.000
střídání časů	399.114	225.725	120.161
nemá názor	288.219	163.007	86.774

Táto tabuľka bude používaná aj ďalej v d-e

Hypotéza: U větších měst, menších měst a obcí (3. průzkumy) je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Použijem test dobrej zhody. Vypočítam si p-hodnotu skrz funkciu `chisquare` a tú porovnám s mojou hladinou významnosti α
p = 0.0018
p < α
Hypotézu zamietam.

d)

Hypotéza: U větších měst, menších měst a obcí (3. průzkumy) je stejné procentuální zastoupení nerozhodnutelných obyvatel.

Použijem test dobrej zhody. Vypočítam si p-hodnotu skrz funkciu `chisquare` a tú porovnám s mojou hladinou významnosti α
p = 3.5051*10^-5
p < α
Hypotézu zamietam.

f)

Použil som Pearsonov korelačný koeficient
hodnota pre koeficientu pre okolie študenta a veľké mestá: 0.9204
hodnota pre koeficientu pre okolie študenta a menšie mestá: 0.9085
hodnota pre koeficientu pre okolie študenta a dediny: 0.9261

Najväčšia korelácia je medzi okolím študenta a dedinou.

Môjho prieskumu sa zúčastnili moji spolužiaci zo strednej školy, ktorí sú poväčšinou z dedín a jedného menšieho mesta, takže výsledok by sedel.

Úkol 2

a)

Zostrojím prvý model pre pôvodnú regresnú funkciu $Z = \beta_1 + \beta_2X + \beta_3Y + \beta_4X^2 + \beta_5Y^2 + \beta_6X*Y$.
Použijem na to funkciu `OLS`, ktorá mi vytvorí regresný model s nasledujúcimi parametrami:

					Interval spoľahlivosti	
	coef	std err	t	P>|t|	[0.025	0.975]
----------	----------	----------	----------	----------	----------	----------

β_1	-124.7148	24.310	-5.130	0.000	-173.280	-76.149
β_2	2.8798	3.793	0.759	0.450	-4.697	10.456
β_3	2.3539	7.161	0.329	0.743	-11.952	16.659
β_4	9.9022	0.169	58.542	0.000	9.564	10.240
β_5	-1.9379	0.631	-3.071	0.003	-3.198	-0.677
β_6	3.1224	0.285	10.944	0.000	2.552	3.692

hodnota $R^2 = 0.99$ -> model je validný

Na základe intervalov spoľahlivosti z regresného modelu odstránime parametre, ktorých interval spoľahlivosti zahŕňa hodnotu 0. Odstránime teda parametre β_2 a β_3.

Vytvorím nový model pre upravenú regresnú funkciu $Z = \beta_1 + \beta_4X^2 + \beta_5Y^2 + \beta_6X*Y$.
Funkcia `OLS`, mi vytvorí regresný model s nasledujúcimi parametrami:

					Interval spoľahlivosti	
	coef	std err	t	P>|t|	[0.025	0.975]
β_1	-109.1347	11.436	-9.543	0.000	-131.967	-86.302
β_4	10.0118	0.072	138.671	0.000	9.868	10.156
β_5	-1.8156	0.283	-6.413	0.000	-2.381	-1.250
β_6	3.2253	0.243	13.291	0.000	2.741	3.710

hodnota $R^2 = 0.99$ -> model je validný

Odstraňovanie ďalších regresných parametrov β by neviedlo k nájdeniu nového lepšieho modelu, keďže interval spoľahlivosti žiadneho z regresných parametrov neobsahuje hodnotu 0. Preto ukončujem hľadanie nového modelu, a prehlasujem model $Z = \beta_1 + \beta_4X^2 + \beta_5Y^2 + \beta_6X*Y$ za najlepší

b)

Funkcia `OLS` vytvára regresný model a vypočítava jeho parametre metódou najmenších štvorcov, takže pre tento pod stačí do tabuľky skopírovať hodnoty ktoré sú výstupom tejto funkcie, a sú zapísané aj vyššie.

		Interval spoľahlivosti 95%	
Parameter	Odhad hodnoty parametru	0.025	0.975
β_1	-109.1347	-131.967	-86.302
β_4	10.0118	9.868	10.156
β_5	-1.8156	-2.381	-1.250
β_6	3.2253	2.741	3.710

c)

nestranný odhad rozptylu získam z druhého modelu skrz premennú `mse_resid`
nestranný odhad rozptylu = 2525.6887

d)

zvolím si 2 ľubovoľné regresné parametre z najlepšieho modelu, napr. β_4 a β_5 a tie budem ďalej uvažovať v c-d.
Hypotéza: Zvolené dva regresní parametry jsou současně nulové.
Použijem f-test, $\alpha = 0.05$
p-hodnota = 9.5423*10^-90
$p < \alpha$
Hypotézu zamietam.

e)

Hypotéza: Zvolené dva regresní parametry jsou stejné.
Použijem t-test, $\alpha = 0.05$
p-hodnota = 3.5725*10^-53
$p < \alpha$
Hypotézu zamietam.