VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

MSP – projekt

9. prosince 2022

Jozef Makiš (xmakis00)

1 Úloha 1

Vypracovanie tejto úlohy sme začali zberom dát z okolia študenta. Táto časť prieskumu bola vykonaná na 32 respondentoch. Výsledok uceleného prieskumu, s ktorým sme pracovali v tejto úlohe môžeme vidieť na obrázku 1.

	Praha	Brno	Znojmo	Tišnov	Paseky	Horní Lomná	Dolní Věstvonice	okolie studenta
0	1327	915	681	587	284	176	215	32
1	510	324	302	257	147	66	87	11
2	352	284	185	178	87	58	65	15
3	257	178	124	78	44	33	31	4
4	208	129	70	74	6	19	32	2

Obrázek 1: Kompletný prieskum pre 1. úlohu.

Prvý riadok (0) v tabuľke reprezentuje celkový počet respondentov pre jednotlivé mesta ako i okolie študenta. Druhý riadok (1) je počet odpovedí v prospech zimného času, tretí (2) je počet respondentov v prospech letného času. Štvrtý riadok (3) obsahuje respodentov, ktorý su spokojný so zmenami času a piaty (4) je pre respondentov bez názoru.

Následne sme overovali hypotézy na hladine významnosti 0.05. Podrobný postup rišenia úloh je v jupyter notebooku.

1.1 A

Testovali sme, že či je rovnaké zastúpenie respondentov preferujúcich zimný čas v mestách, obciach a okolí študenta. Museli sme spočítať zastúpenia respodentov, ktorý by chceli zimný čas, letný čas, striedanie času a tých, čo sú bez názoru. Tie sú na obrázku 2.

```
Zimný čas: 0.40407872895423286
Letny čas: 0.29025373488261796
Zmena času: 0.17761441783258242
Bez nazoru: 0.12805311833056676
Súčet: 1.0
```

Obrázek 2: Zastúpenie respondentov v celom prieskume pre preferované nastavenie času.

Aby sme mohli teda overiť, že zastúpenie respondentov v prospech zimného času vo všetkých mestách, obciach a okolia študenta je rovnake, budeme ich testovať s očakávaným percentualnym zastupenim v kompletnom prieskume a to je podľa obrázku 2 40,40%. To znamená, že ak v celom prieskume majú respondenti preferujúci zimný čas 40,40% zastúpenie, potom by mali mať aj naprieč mestami tiež 40,40%.

$$p_{zimny\,cas} = \frac{pocet\,respondentov\,pre\,zimny\,cas}{kompletny\,pocet\,respondentov} = \frac{1704}{4217} = 0.4040$$

Očakávané početnosti pre mesta získame súčinom bodového odhadu pravdepodobnosti zastúpenia v celom prieskume a počtom respondentov v danom meste. Pre prahu a pre respondentov, ktorý preferujú zimu to bude:

$$teor - pocetnost_{praha} = p_{zimny\,cas} * respondenti\,praha = 0.4040 * 1327 = 536,2124$$

	Mesta	Početnosť	Zastupenie teor - Zima	Početnosť teor	rozdiel^2/teor_poc
0	Praha	510	0.404079	536.212473	1.281383
1	Brno	324	0.404079	369.732037	5.656581
2	Znojmo	302	0.404079	275.177614	2.614458
3	Tišnov	257	0.404079	237.194214	1.653789
4	Paseky	147	0.404079	114.758359	9.058368
5	Horní Lomná	66	0.404079	71.117856	0.368296
6	Dolní Věstonice	87	0.404079	86.876927	0.000174
7	Okolie študenta	11	0.404079	12.930519	0.288225

Obrázek 3: Tabuľka pre rovnake zastúpenie zimného času.

Aby sme teda overili túto hypotézu použili sme test dobrej zhody. Testovacie kritérium nám vyšlo

$$\chi^2 = 20.9212$$

a doplnok kritického oboru je

$$\overline{W_\alpha} = \langle 0, 12.5915 \rangle.$$

Hypotézu, že v mestách, obciach a v okolí študenta je rovnaké percentuálne zastúpenie obyvateľov čo preferujú zimný čas zamietame.

1.2 B

Pri overení rovnakého percentuálneho zastúpenia čo preferujú letný čas sme postupovali podobne ako u zimného času. Tu majú respondenti v prospech letného času zastúpenie podľa obrázku 2 iba 29,02%. Aby sme mohli potvrdiť hypotézu rovnakého percentuálneho zastúpenia v jednotlivých mestách budeme ju porovnávať s percentualnym zastupenim 29,02%. Tu je bodový odhad teoretickej pravdepodobnosti 0.2902 pre všetky mesta.

	Mesta	Početnosť	Zastupenie teor - Leto	Početnosť teor	rozdiel^2/teor_poc
0	Praha	352	0.290254	385.166706	2.855985
1	Brno	284	0.290254	265.582167	1.277257
2	Znojmo	185	0.290254	197.662793	0.811212
3	Tišnov	178	0.290254	170.378942	0.340890
4	Paseky	87	0.290254	82.432061	0.253131
5	Horní Lomná	58	0.290254	51.084657	0.936132
6	Dolní Věstonice	65	0.290254	62.404553	0.107946
7	Okolie študenta	15	0.290254	9.288120	3.512614

Obrázek 4: Tabuľka pre rovnake zastúpenie letného času.

Teoretickú početnosť sme získali tak, že sme zobrali bodový odhad pravdepodobnostného zastúpenia pre jednotlivé mesta 0,2902 a prenásobili sme ju s celkovým počtom respondentov v jednotlivých mestách. V prípade Prahy to je napríklad

$$0.290254 * 1327 = 385, 166706.$$

Podobne ako u predchádzajúcej hypotézy získame tak teoreticky počet respondentov, ktorým vyhovuje letný čas.

Testovacie kritérium nám vyšlo

$$\chi^2 = 10.0951$$

a doplnok kritického oboru je

$$\overline{W_\alpha} = \langle 0, 12.5915 \rangle.$$

Hypotézu, že v mestách, obciach a v okolí študenta je rovnaké percentuálne zastúpenie obyvateľov čo preferujú letný čas nezamietame.

1.3 C

Rovnaké percentualne zastúpenie respondentov, ktorý preferuju zmenu času sme overovali podobne ako predchádzajúce hypotézy.

	Mesta	Početnosť	Zastupenie teor - Zmena	Početnosť teor	rozdiel^2/teor_poc
0	Praha	257	0.177614	235.694332	1.925933
1	Brno	178	0.177614	162.517192	1.475028
2	Znojmo	124	0.177614	120.955419	0.076635
3	Tišnov	78	0.177614	104.259663	6.613966
4	Paseky	44	0.177614	50.442495	0.822833
5	Horní Lomná	33	0.177614	31.260138	0.096836
6	Dolní Věstonice	31	0.177614	38.187100	1.352666
7	Okolie študenta	4	0.177614	5.683661	0.498748

Obrázek 5: Tabuľka rovnakého zastúpenia pre zmenu času.

Testovacie kritérium nám vyšlo

$$\chi^2 = 12.8626$$

a doplnok kritického oboru je

$$\overline{W_\alpha} = \langle 0, 12.5915 \rangle.$$

Hypotézu, že v mestách, obciach a v okolí študenta je rovnaké percentuálne zastúpenie obyvateľov čo preferujú zmenu času zamietame.

1.4 D

Aby sme porovnali 3 prieskumy medzi veľkými mestami, menšími a obcami, tak sme združili hodnoty miest, ktoré patria do týchto skupín. Teoretická pravdepodobnosť zastúpenia sa nemení, akurát početnosť sme upravili súčtom respondentov v mestách. Pre väčšie mesta napríklad:

$$teor - pocetnost_{velke\,mesto} = 0.4040 * (1327 + 195) = 906,9787$$

	Mesta	Početnosť	Zastupenie teor	Početnosť teor	rozdiel^2/teor_poc
0	Väčšie mesta	834	0.40454	906.978734	5.872128
1	Menšie mesta	559	0.40454	512.956750	4.132865
2	Obce	300	0.40454	273.064516	2.656956

Obrázek 6: Tabuľka združených hodnôt pre zimu.

Testovacie kritérium nám vyšlo

$$\chi^2 = 12.6619$$

a doplnok kritického oboru je

$$\overline{W_\alpha} = \langle 0, 3.8414 \rangle.$$

Zamietame teda, že by medzi veľkými, menšími mestami a obcami bolo rovnake zastúpenie obyvateľov preferujúcich zimný čas.

1.5 E

Overovali sme medzi tromi prieskumami rovnake zastúpenie nerozhodných respondentov. Postupovali sme podobne ako u predchádzajúcej úlohy. Sčítali sme hodnoty miest podľa ich kategorie.

	Mesta	Početnosť	Zastupenie teor	Početnosť teor	rozdiel^2/teor_poc
0	Väčšie mesta	337	0.128554	288.218877	8.256218
1	Menšie mesta	144	0.128554	163.006930	2.216245
2	Obce	57	0.128554	86.774194	10.216201

Obrázek 7: Tabuľka združených hodnôt pre leto.

Testovacie kritérium nám vyšlo

$$\chi^2 = 20.6886$$

a doplnok kritického oboru je

$$\overline{W_\alpha} = \langle 0, 3.8414 \rangle.$$

Zamietame teda, že by medzi veľkými, menšími mestami a obcami bolo rovnake zastúpenie nerozhodných obyvateľov.

1.6 F

Snažili sme sa odhadnúť z dat okolia študenta, že v ktorom väčšiom meste prevádzal prieskum. Aby sme porovnali meranie študenta s jednotlivými kategóriami miest rozhodli sme sa použiť T-test. Ten ale počíta s tým, že merania maju identický rozptyl, to ale vzhľadom na počet nameraných dát študenta neplatí. Použili sme teda jeho variantu Welchov t-test. Tá pracuje s rozdielnymi rozptylmi pri našom rozdielnom počte meraní. Testovali sme prieskum študenta s jednotlivými združenými hodnotami miest. Výsledky testov môžeme vidieť na obrázku 8.

```
Vele mesta: Ttest_indResult(statistic=4.905307766028999, pvalue=0.016164043363197746)
Mensie mesta: Ttest_indResult(statistic=3.3200864309574745, pvalue=0.04491165367338969)
Obce: Ttest_indResult(statistic=2.9673376530248934, pvalue=0.05873355141467863)
```

Obrázek 8: P hodnoty pre porovnanie okolie študenta.

Keďže sme prevádzali testy na hladine významnosti 0.05, tak nezamietame že by prieskum študenta bol z obce. To robíme na základe p hodnoty, ktorá ma hodnotu 0.058 a je väčšia ako naša hladina významnosti. Zamietame ale, že by bol jeho prieskum z veľkého alebo menšieho mesta.

Výsledok, ktorý sme získali neodzrkadluje realitu. Merania z okolia študenta prebehli na kamarátoch a rodine, ktorá pochádza z menšieho mesta. Kamaráti pochádzajú z väčšinou z menších miest a obci. Väčšina výsledkov prieskumu je ale z menšieho mesta. Najmenšia p-hodnota pre veľké mestá sedí, pretože žiadne z meraní sa nekonalo vo veľkom meste. To že sme zamietli aj menšie mesto môže byť spôsobene nepresnosťou Welchovho T-testu v porovnaní s klasickým T-testom. Ak by sme dosiahli porovnatelny počet meraní a rovnaký rozptyl mohli by sme menšie mesta nezamietnuť. Prieskum, ktorý sa nekonal v jednom meste určite tiež nepomohlo výsledku.

2 Úloha 2

2.1 A

Poprvé sme natrénovali základný model

$$Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 XY.$$

Súhrn štatistik tohto modelu môžeme nájsť na obrázku 9. Môžeme vidieť, že nám vyšiel slušný koeficient determinácie $R^2 = 0.942$. Hodnota F-štatistiky, konkrétne jej p hodnota nám vyšla hlboko pod 0.05 a teda zamietame, že by náš model mal všetky koeficienty nulove.

```
                           OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.942
Model:                            OLS   Adj. R-squared:                  0.937
Method:                 Least Squares   F-statistic:                     206.8
Date:                Sat, 03 Dec 2022   Prob (F-statistic):           4.17e-38
Time:                        11:13:20   Log-Likelihood:                -413.00
No. Observations:                  70   AIC:                             838.0
Df Residuals:                      64   BIC:                             851.5
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         62.0036     44.220      1.402      0.166     -26.335     150.342
x1            -1.2625      6.898     -0.183      0.855     -15.044      12.519
x2            -6.9407     13.025     -0.533      0.596     -32.962      19.080
x3            -1.9199      0.308     -6.240      0.000      -2.535      -1.305
x4            -3.1013      1.148     -2.702      0.009      -5.394      -0.808
x5            10.9502      0.519     21.100      0.000       9.913      11.987
==============================================================================
Omnibus:                        0.880   Durbin-Watson:                   1.855
Prob(Omnibus):                  0.644   Jarque-Bera (JB):                0.970
Skew:                          -0.191   Prob(JB):                        0.616
Kurtosis:                       2.569   Cond. No.                         839.
...
==============================================================================
```

Obrázek 9: Výsledok natrénovaneho modelu.

Chceme ale zlepšiť koeficient determinace našeho modelu a začneme spätnou metodou. Ak sa pozrieme na p hodnoty koeficientov $\beta_2 = 0.855$, $\beta_3 = 0.596$. Tak nezamietame pre ne, že by boli rovné nule. Preto začneme s odstránením koeficientu β_2. Po postupnom odstránení koeficientu β_2 a β_3 sme nedosiahli zlepšenie koeficientu determinácie.

```
                           OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.941
Model:                            OLS   Adj. R-squared:                  0.939
Method:                 Least Squares   F-statistic:                     353.7
Date:                Sat, 03 Dec 2022   Prob (F-statistic):           1.36e-40
Time:                        11:13:21   Log-Likelihood:                -413.17
No. Observations:                  70   AIC:                             834.3
Df Residuals:                      66   BIC:                             843.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         44.0109     20.749      2.121      0.038       2.585      85.437
x1            -1.9489      0.131    -14.878      0.000      -2.210      -1.687
x2            -3.6243      0.514     -7.055      0.000      -4.650      -2.599
x3            10.8227      0.440     24.581      0.000       9.944      11.702
==============================================================================
Omnibus:                        0.620   Durbin-Watson:                   1.834
Prob(Omnibus):                  0.734   Jarque-Bera (JB):                0.750
Skew:                          -0.143   Prob(JB):                        0.687
Kurtosis:                       2.581   Cond. No.                         389.
==============================================================================
```

Obrázek 10: Výsledok natrénovaneho modelu bez β_2, β_3.

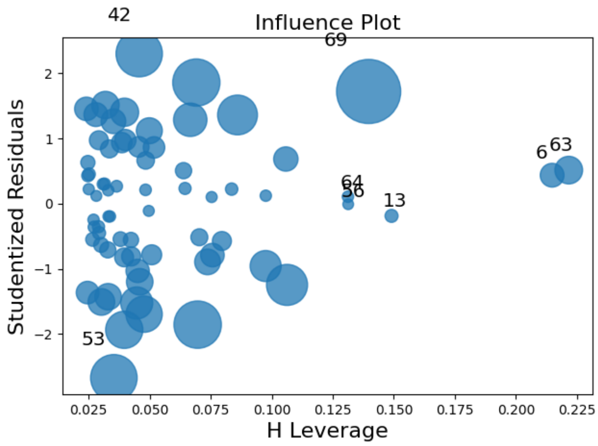
Dosiahli sme ale zjednodušenie modelu. Pre zvyšné Bety podla ich p hodnoty zamietame, že by sa rovnali nule.

V rámci regresnej analýzi sme náš model podrobili následným testom :

- Heteroskedasticitu sme overili testom Breush-Pagan. P hodnota nám vyšla 0.3107 >0.05. Heteroskedasticitu teda nezamietame.

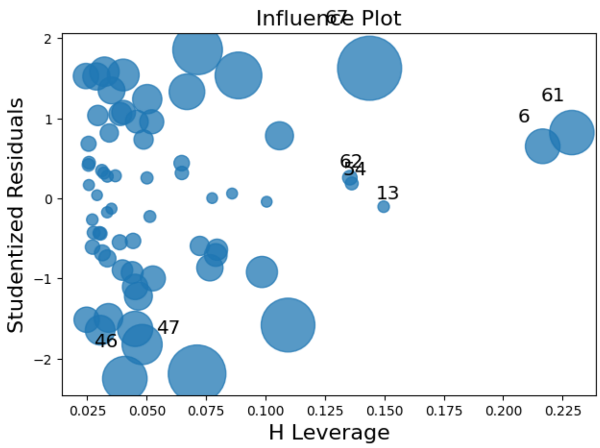
- Autokorelaciu sme overovali pomocou testu Durbin–Watson. Testovacia hodnota je 1.834. Keďže hodnota tohto testu je menšia ako 2, tak je tu známka o menšej seriovej korelacii. Napriek tomu nezávislosť nezamietame.
- Normalitu sme overili pomocou testu Jarque-Bera, konkrétne jeho p hodnotou. P hodnota má hodnotu 0.687, platí že 0.687 >0.05. Normalitu modelu nezamietame.

Po overení metody a závislosti modelu sme prešli ku kritike dát. Použili sme cookovu vzdialenosť.

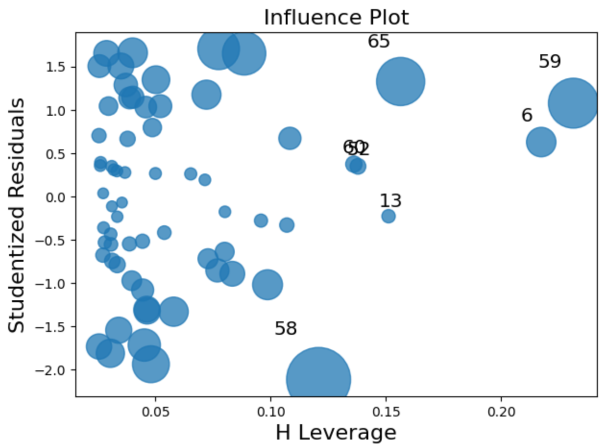


Obrázek 11: Cookova metoda.

Z obrázku 11 môžeme vidieť, najviac vyplvne body. To sú 42, 53, 63 a 6. Rozhodli sme sa odstrániť body 42 a 53, pretože sa nám pomocou nich podarilo zlepšiť koeficient determinizacie. Snažili sme sa odstrániť podozrivé body, ktoré su nad hranicou 2 a pod hranicou -2. Alebo tie, ktoré vybočuju veľmi vpravo. Keďže body 6 a 61 nejak neovplyvnili chovanie modelu ponechali sme ich.



Obrázek 12: Cookova metoda bez bodov 42 a 53.

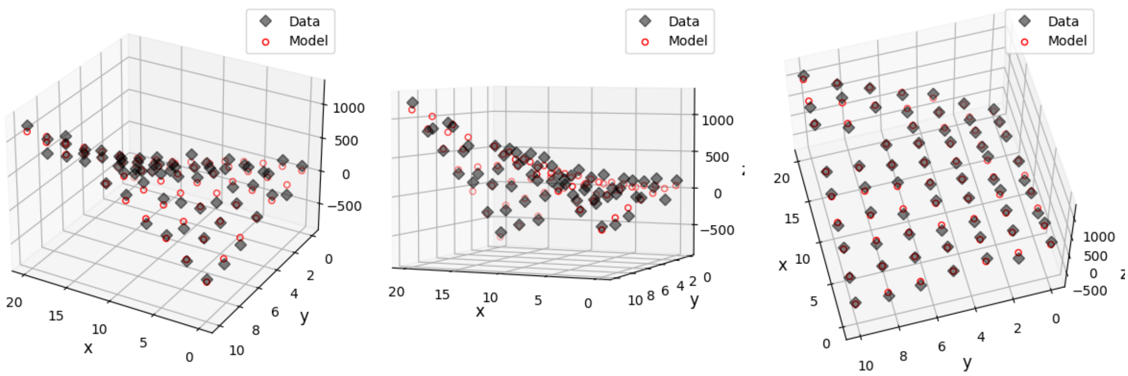


Obrázek 13: Cookova metoda bez bodov 46 a 47.

Výsledny model sme zobrazili na obrázku 14. Podarilo sa nám zlepšiť koeficient determinizacie z 0.942 na 0.957.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.957
Model:                            OLS   Adj. R-squared:                  0.955
Method:                 Least Squares   F-statistic:                     463.4
Date:                Sat, 03 Dec 2022   Prob (F-statistic):           2.16e-42
Time:                        11:13:22   Log-Likelihood:                -379.70
No. Observations:                  66   AIC:                             767.4
Df Residuals:                      62   BIC:                             776.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         40.9320     17.986      2.276      0.026       4.978      76.886
x1            -2.0226      0.114    -17.709      0.000      -2.251      -1.794
x2            -3.6777      0.446     -8.252      0.000      -4.568      -2.787
x3            11.2531      0.389     28.924      0.000      10.475      12.031
==============================================================================
Omnibus:                        5.454   Durbin-Watson:                   2.182
Prob(Omnibus):                  0.065   Jarque-Bera (JB):                2.374
Skew:                          -0.103   Prob(JB):                        0.305
Kurtosis:                       2.094   Cond. No.                         378.
==============================================================================
```

Obrázek 14: Výsledny model.



Obrázek 15: Graficke zobrazenie vysledneho modelu.

2.2 B

Výsledne odhady regresných parametrov a ich intervaly spolahlivosti s 95% spolahlivosťou sú na obrázku 15.

	Bety	Intervalovy odhad
0	40.9320	[4.977879884146553, 76.88604731499493]
1	-2.0226	[-2.2509496711618673, -1.794335027527551]
2	-3.6777	[-4.568497472359534, -2.7868267038373005]
3	11.2531	[10.475411386734429, 12.030834095390489]

Obrázek 16: Regresne parametry a ich odhady.

2.3 C

Nestranný odhad rozptylu závislej premennej sme získali podielom sumy štvorcov chýb podelených počtom hodnôt od ktorých sme odčítali počet regresných parametrov.

$$s^2 = MSE = \frac{SSE}{n-4} = 6189.3175$$

2.4 D

Regresne parametry, ktoré sme sa rozhodli testovať na rovnosť s nulou sú β_2 a β_3. K tomu sme použili Waldov test. Vyšla nám p hodnota 2.4542315831430707e-25. Teda zamietame, že by tieto dva regresne parametry boli rovné nule spoločne.

2.5 E

Ako u predchádzajúcej podúlohy budeme testovať na rovnosť β_2 a β_3. Ak by platilo, že tieto dva regresne parametry sú si rovne, potom vytvoríme následony model:

$$\beta_2 - \beta_3 = 0$$

$$\beta_2 = \beta_3$$

$$Z = \beta_1 + \beta_2 X^2 + \beta_2 Y^2 + \beta_4 XY.$$

$$Z = \beta_1 + \beta_2 (X^2 + Y^2) + \beta_3 XY.$$

Novovzniknutý model sme natrenovali s rovnakými dátami ako náš finálny. Test sme previedli ANOVou pre lineárne modely. Výsledkom tohto testu je p hodnota 0.000067. Teda zamietame, že by tieto dva linearne modely boli rovnake. Z toho vyplýva, že β_2 a β_3 nie sú si rovné.