Statistika a pravděpodobnost (MSP) – 2022/2023
Projekt

Domink Nejedlý (xnejed09)

Úkol 1: Český stát objednal průzkum, jak lidé vnímají střídání zimního a letního času. Průzkum zahrnoval větší města (Praha, Brno), menší města (Znojmo, Tišnov) a obce (Paseky, Horní Lomná, Dolní Věstonice). V průzkumu zjišťovali, co lidem lépe vyhovuje – zda střídání letního a zimního času, pouze zimní čas nebo pouze letní čas. Odpovědi respondentů vidíte v tabulce:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**Okolí studenta**
počet respondentů	**1327**	**915**	**681**	**587**	**284**	**176**	**215**	**71**
zimní čas	510	324	302	257	147	66	87	21
letní čas	352	284	185	178	87	58	65	25
střídání časů	257	178	124	78	44	33	31	8
nemá názor	208	129	70	74	6	19	32	17

Na hladině významnosti $\alpha = 0,05$ ($\alpha = 0,05$ je celková chyba 1. druhu pro a) až e)) prověřte hypotézy:

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

Řešení: K ověřování využijeme test dobré shody pro Multinomické rozdělení resp. chí-kvadrát test v kontingenčních tabulkách (přednáška 13 sekce Kategoriální analýza – Test dobré shody) (v tomto případě lze použít označení chí-kvadrát test homogenity, jelikož tento porovnává rozložení diskrétních veličin ve dvou či více populacích, přičemž testuje zda se napříč nimi tato rozložení neliší – pouze jiný způsob uvažování o chí-kvadrát testu nezávislosti, výpočet je však stejný). Mějme následující kontingenční tabulku, která vznikla z tabulky v zadání sloučením řádků *letní čas*, *střídání časů* a *nemá názor*:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**okolí studenta**	$\boldsymbol{n_{i,\bullet}}$
zimní čas	510	324	302	257	147	66	87	21	**1714**
ostatní	817	591	379	330	137	110	128	50	**2542**
$\boldsymbol{n_{\bullet,j}}$	**1327**	**915**	**681**	**587**	**284**	**176**	**215**	**71**	$\boldsymbol{n_{\bullet,\bullet} = 4256}$

Zjevně

$$n_{i,\bullet} = \sum_{j=1}^{c} n_{i,j},\ n_{\bullet,j} = \sum_{i=1}^{r} n_{i,j} \text{ a } n = n_{\bullet,\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{i,j},$$

kde r značí počet řádků a c počet sloupců tabulky (bez řádku $n_{\bullet,j}$ a sloupce $n_{i,\bullet}$).

Ověřujeme $H_0 : \forall i, j : p_{i,j} = p_{i,\bullet} \cdot p_{\bullet,j}$ proti $H_A : \exists i, j : p_{i,j} \neq p_{i,\bullet} \cdot p_{\bullet,j}$, kde $p_{i,\bullet} = \frac{n_{i,\bullet}}{n}$ a $p_{\bullet,j} = \frac{n_{\bullet,j}}{n}$

Alternativně lze v tomto případě uvažovat $H_0 : \forall i, j : p_{\textit{zimní čas},i} = p_{\textit{zimní čas},j}$ proti $H_A : \exists i, j : p_{\textit{zimní čas},i} \neq p_{\textit{zimní čas},j}$, kde i a j představují oblasti, kde bylo prováděno pozorování (tedy sloupce tabulky bez $n_{i,\bullet}$).

Spočtěme teoretické četnosti pomocí vzorce $n_{i,j} \overset{odhad}{=} \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}; \forall i, j$:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**okolí studenta**
zimní čas	534.4168	368.4939	274.2561	236.3999	114.3741	70.8797	86.5860	28.5935
ostatní	792.5832	546.5061	406.7439	350.6001	169.6259	105.1203	128.4140	42.4065

Vidíme, že získané hodnoty splňují podmínku $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i, j$. Není tedy nutné slučovat sloupce (řádky v tomhle případě slučovat nelze, jelikož jejich minimální vyžadovaný počet je 2, což platí i pro sloupce – pro čtyřpolní tabulku by se pak postupovalo dle speciálních metod (Yatesova korekce, Fischerův exaktní test)) a můžeme tedy pokračovat výpočtem testovacího kritéria:

$$t = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{i,j} - \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n})^2}{\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}} = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 38.0913$$

Určíme doplněk kritického oboru ($k = (r-1) \cdot (c-1)$ je počet stupňů volnosti):

$$\overline{W}_\alpha = \langle 0, \chi^2_{1-\alpha}(k) \rangle = \langle 0, \chi^2_{1-\alpha}((2-1) \cdot (8-1)) \rangle = \langle 0, \chi^2_{1-\alpha}(7) \rangle = \langle 0, \chi^2_{1-\alpha}(7) \rangle = \langle 0, \chi^2_{1-\alpha}(7) \rangle$$
$$\overline{W}_{0.05} = \langle 0, \chi^2_{0.95}(7) \rangle = \langle 0, 14.067 \rangle$$

Jelikož $t \notin \overline{W}_{0.05}$, H_0 se **zamítá**. V městech, obcích a v okolí studenta tedy dle testu stejné procentuální zastoupení lidí, co preferují zimní čas, není.

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

Řešení: Postupujeme analogicky jako v bodě a), pouze vyměníme *zimní čas* za *letní čas*. Získáme tedy následující tabulku:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**okolí studenta**	$n_{i,\bullet}$
letní čas	352	284	185	178	87	58	65	25	**1234**
ostatní	975	631	496	409	197	118	150	46	**3022**
$n_{\bullet,j}$	**1327**	**915**	**681**	**587**	**284**	**176**	**215**	**71**	$n_{\bullet,\bullet} = 4256$

Ověřujeme $H_0 : \forall i, j : p_{i,j} = p_{i,\bullet} \cdot p_{\bullet,j}$ proti $H_A : \exists i, j : p_{i,j} \neq p_{i,\bullet} \cdot p_{\bullet,j}$, kde $p_{i,\bullet} = \frac{n_{i,\bullet}}{n}$ a $p_{\bullet,j} = \frac{n_{\bullet,j}}{n}$
Alternativně lze opět uvažovat $H_0 : \forall i, j : p_{letní\ čas,i} = p_{letní\ čas,j}$ proti $H_A : \exists i, j : p_{letní\ čas,i} \neq p_{letní\ čas,j}$, kde i a j představují oblasti, kde bylo prováděno pozorování (tedy sloupce tabulky bez $n_{i,\bullet}$).

Spočtěme teoretické četnosti dle vzorce $n_{i,j} \overset{odhad}{=} \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}; \forall i, j$:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**okolí studenta**
letní čas	384.7552	265.2984	197.4516	170.1969	82.3440	51.0301	62.3379	20.5860
ostatní	942.2448	649.7016	483.5484	416.8031	201.6560	124.9699	152.6621	50.4140

Opět všechny získané hodnoty splňují podmínku $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i, j$. Pokračujme tedy výpočtem testovacího kritéria:

$$t = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 10.5980$$

Doplněk kritického oboru je pak stejný jako v bodě a), tedy:

$$\overline{W}_{0.05} = \langle 0, \chi^2_{0.95}(7) \rangle = \langle 0, 14.067 \rangle$$

V tomto případě $t \in \overline{W}_{0.05}$, a proto se H_0 **nezamítá**. V městech, obcích a v okolí studenta je tedy dle testu stejné procentuální zastoupení obyvatel, co preferují letní čas.

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

Řešení: Opět postupujeme stejně jako v bodě a), tentokrát však vyměníme *zimní čas* za *střídání časů.* Upravená tabulka naměřených četností tedy vypadá následovně:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**okolí studenta**	$n_{i,\bullet}$
střídání č.	257	178	124	78	44	33	31	8	**753**
ostatní	1070	737	557	509	240	143	184	63	**3503**
$n_{\bullet,j}$	**1327**	**915**	**681**	**587**	**284**	**176**	**215**	**71**	$n_{\bullet,\bullet} = 4256$

H_0 a H_A zůstávají shodné jako v předcházejících bodech (viz bod a)), přičemž mohou být opět formulovány jako $H_0 : \forall i, j : p_{\textit{střídání časů},i} = p_{\textit{střídání časů},j}$ proti $H_A : \exists i, j : p_{\textit{střídání časů},i} \neq p_{\textit{střídání časů},j}$, kde i a j představují oblasti, kde bylo prováděno pozorování (tedy sloupce tabulky bez $n_{i,\bullet}$).

Teoretické četnosti jsou pak následující:

	Praha	**Brno**	**Znojmo**	**Tišnov**	**Paseky**	**Horní Lomná**	**Dolní Věstonice**	**okolí studenta**
střídání časů	234.7817	161.8879	120.4871	103.8560	50.2472	31.1391	38.0392	12.5618
ostatní	1092.2183	753.1121	560.5129	483.1440	233.7528	144.8609	176.9608	58.4382

Všechny získané hodnoty splňují podmínku $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i, j$. Přejděme tedy k výpočtu testovacího kritéria:

$$t = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 17.1222$$

Doplněk kritického oboru zůstává stejný jako v předcházejících bodech (viz bod a)), tedy:

$$\overline{W}_{0.05} = \langle 0, \chi^2_{0.95}(7) \rangle = \langle 0, 14.067 \rangle$$

Vidíme, že $t \notin \overline{W}_{0.05}$, H_0 se proto **zamítá**. V městech, obcích a v okolí studenta tedy dle testu stejné procentuální zastoupení lidí, co preferují střídání časů, není.

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

Řešení: Zde postupujme opět obdobně jako v bodě a), ovšem kromě sloučení řádků aplikujme rovněž sloučení sloupců do skupin podle velikosti pozorovaných oblastí (viz zadání), přičemž sloupec *okolí studenta* úplně vyloučíme. Tímto způsobem získáme následující tabulku:

	větší města	**menší města**	**obce**	$n_{i,\bullet}$
zimní čas	834	559	300	**1693**
ostatní	1408	709	375	**2492**
$n_{\bullet,j}$	**2242**	**1268**	**675**	$n_{\bullet,\bullet} = 4185$

V obecné rovině ověřujeme opět $H_0 : \forall i, j : p_{i,j} = p_{i,\bullet} \cdot p_{\bullet,j}$ proti $H_A : \exists i, j : p_{i,j} \neq p_{i,\bullet} \cdot p_{\bullet,j}$, kde $p_{i,\bullet} = \frac{n_{i,\bullet}}{n}$ a $p_{\bullet,j} = \frac{n_{\bullet,j}}{n}$. Alternativně lze pak H_0 a H_A opět chápat jako:

$$H_0 : p_{\textit{zimní čas,větší města}} = p_{\textit{zimní čas,menší města}} = p_{\textit{zimní čas,obce}}$$
$$H_A : \exists i, j : p_{\textit{zimní čas},i} \neq p_{\textit{zimní čas},j}, \text{kde } i, j \in \{\textit{větší města, menší města, obce}\}$$

Nyní opět standardním způsobem (viz bod a)) získejme teoretické četnosti:

	větší města	**menší města**	**obce**
zimní čas	906.9787	512.9568	273.0645
ostatní	1335.0213	755.0432	401.9355

Vzhledem k tomu, že všechny získané hodnoty splňují podmínku $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i,j$, spočtěme dále testovací kritérium:

$$t = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 21.2641$$

Určeme dále doplněk kritického oboru ($k = (r-1)\cdot(c-1)$ je počet stupňů volnosti):

$$\overline{W}_\alpha = \langle 0, \chi^2_{1-\alpha}(k)\rangle = \langle 0, \chi^2_{1-\alpha}((2-1)\cdot(3-1))\rangle = \langle 0, \chi^2_{1-\alpha}(2)\rangle$$
$$\overline{W}_{0.05} = \langle 0, \chi^2_{0.95}(2)\rangle = \langle 0, 5.991\rangle$$

Jelikož $t \notin \overline{W}_{0.05}$, tak se H_0 **zamítá**. U větších měst, menších měst a obcí dle testu stejné procentuální zastoupení obyvatel, co preferují zimní čas, není.

e) U větších měst, menších měst a obcí (3 průzkumy) je stejné procentuální zastoupení nerozhodnutých obyvatel.

Řešení: Postupujeme analogicky jako v předcházejícím bodě d), pouze vyměníme *zimní čas* za *nemá názor*. Upravená výchozí tabulka k tomuto bodu tedy vypadá následovně:

	větší města	**menší města**	**obce**	$\boldsymbol{n_{i,\bullet}}$
nemá názor	337	144	57	**538**
ostatní	1905	1124	618	**3647**
$\boldsymbol{n_{\bullet,j}}$	**2242**	**1268**	**675**	$\boldsymbol{n_{\bullet,\bullet} = 4185}$

H_0 a H_A jsou opět stejné jako v předcházejícím bodě d, jen v jejich alternativním znění zaměníme *zimní čas* za *nemá názor*, tedy:

$$H_0 : p_{\textit{nemá názor,větší města}} = p_{\textit{nemá názor,menší města}} = p_{\textit{nemá názor,obce}}$$
$$H_A : \exists i,j : p_{\textit{nemá názor},i} \neq p_{\textit{nemá názor},j}, \text{kde } i,j \in \{\textit{větší města, menší města, obce}\}$$

Pokračujeme opět výpočtem teoretických četností:

	větší města	**menší města**	**obce**
nemá názor	288.2189	163.0069	86.7742
ostatní	1953.7811	1104.9931	588.2258

Všechny teoretické četnosti splňují podmínku $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i,j$, přejděme k výpočtu testovacího kritéria:

$$t = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 23.7406$$

Doplněk kritického oboru je pak stejný jako v bodě d), tedy:

$$\overline{W}_{0.05} = \langle 0, \chi^2_{0.95}(2)\rangle = \langle 0, 5.991\rangle$$

Vidíme, že $t \notin \overline{W}_{0.05}$, a proto se H_0 **zamítá**. U větších měst, menších měst a obcí tedy dle testu stejné procentuální zastoupení nerozhodnutých obyvatel není.

f) Na základě odpovědí z okolí studenta zkuste určit z dat, zda student prováděl výzkum ve větším městě, menším městě nebo v obci. Porovnejte výsledek se skutečností a okomentujte.

Řešení: V tomto bodě využijeme opět chí-kvadrát test homogenity (klasický chí-kvadrát test v kontingenčních tabulkách použitý ve všech předcházejících bodech). Jelikož tento test porovnává rozložení diskrétních veličin ve dvou či více populacích a testuje, zdali se napříč nimi tato rozložení neliší, můžeme proti sobě testovat po dvojicích vždy *okolí studenta* a jednu z oblastí – *větší města*, *menší města*, *obce* – a jako nejpodobnější pak určíme dvojici, jež dosáhne nejnižšího testovacího kritéria (to lze totiž chápat jako míru, jak moc se od sebe rozložení diskrétních veličin v populacích liší – čím je jeho hodnota vyšší, tím více se liší i jednotlivá rozložení) tzn. vybereme dvojici, jež má nejpodobnější rozložení diskrétních veličin (naměřených hodnot/získaných odpovědí). Tímto způsobem odhadneme, kde byl s největší pravděpodobností prováděn výzkum (zdali rozložení odpovědí z *okolí studenta* odpovídá nejvíce rozložení odpovědí ve *větších městech*, *menších městech* nebo *obcích*), tedy jestli byl prováděn ve *větším městě*, *menším městě*, nebo *obci*.

Začněme porovnáním *okolí studenta* s *většími městy*. Výchozí tabulka vypadá následovně:

	okolí studenta	**větší města**	$\boldsymbol{n_{i,\bullet}}$
zimní čas	21	834	**855**
letní čas	25	636	**661**
střídání časů	8	435	**443**
nemá názor	17	337	**354**
$\boldsymbol{n_{\bullet,j}}$	**71**	**2242**	$\boldsymbol{n_{\bullet,\bullet} = 2313}$

Vypočítané teoretické četnosti (opět dle vzorce $n_{i,j} \overset{odhad}{=} \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}; \forall i, j$) jsou následující:

	okolí studenta	**větší města**
zimní čas	26.2451	828.7549
letní čas	20.2901	640.7099
střídání časů	13.5984	429.4016
nemá názor	10.8664	343.1336

Všechny teoretické četnosti splňují podmínku $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i, j$, vypočtěme tedy testovací kritérium:

$$t_{okolí\ studenta, větší\ města} = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 8.1589$$

Pokračujme porovnáním *okolí studenta* s *menšími městy*. Získáváme následující kontingenční tabulku:

	okolí studenta	**menší města**	$\boldsymbol{n_{i,\bullet}}$
zimní čas	21	559	**580**
letní čas	25	363	**388**
střídání časů	8	202	**210**
nemá názor	17	144	**161**
$\boldsymbol{n_{\bullet,j}}$	**71**	**1268**	$\boldsymbol{n_{\bullet,\bullet} = 1339}$

Dále spočtěme teoretické četnosti:

	okolí studenta	**menší města**
zimní čas	30.7543	549.2457
letní čas	20.5736	367.4264
střídání časů	11.1352	198.8648
nemá názor	8.5370	152.4630

Podmínka $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i,j$ je splněna. Přejděme tedy k výpočtu testovacího kritéria:

$$t_{okolí\ studenta, menší\ města} = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 14.0643$$

Nakonec porovnejme *okolí studenta* s *obcemi*. Pracujme tedy s následující tabulkou:

	okolí studenta	**obce**	$n_{i,\bullet}$
zimní čas	21	300	**321**
letní čas	25	210	**235**
střídání časů	8	108	**116**
nemá názor	17	57	**74**
$n_{\bullet,j}$	**71**	**675**	$n_{\bullet,\bullet} = 746$

Opět spočtěme teoretické četnosti:

	okolí studenta	**obce**
zimní čas	30.5509	290.4491
letní čas	22.3660	212.6340
střídání časů	11.0402	104.9598
nemá názor	7.0429	66.9571

Podmínka $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} > 5; \forall i,j$ opět platí. Vypočítejme tedy testovací kritérium:

$$t_{okolí\ studenta, obce} = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{i,j}^2}{n_{i,\bullet} \cdot n_{\bullet,j}} - n = 20.1259$$

Vidíme, že nejnižším testovacím kritériem je $t_{okolí\ studenta, větší\ města}$ a platí, že

$$t_{okolí\ studenta, větší\ města} < t_{okolí\ studenta, menší\ města} < t_{okolí\ studenta, obce}.$$

Naměřené hodnoty z *okolí studenta* jsou tedy nejpodobnější hodnotám naměřeným ve *větších městech* a nejméně podobné datům získaným v *obcích.*

Na základě odpovědí z *okolí studenta* (mého okolí) bychom odhadovali, že průzkum byl prováděn v nějakém větším městě. Mnou dodané odpovědi však pochází z vesnice a jejího bližšího okolí (vedlejší vši/městysy). Tento nesoulad může být způsoben menším vstupním vzorkem a případně i složením respondentů, jimiž byli převážně lidé mladšího věku a studenti, kteří se převážnou část dne právě v městech pohybují (pracují/studují). Vliv na výsledný odhad mohlo mít také větší zastoupení lidí bez názoru na předmět výzkumu v získaném vzorku. Dalším možným důvodem odchylky výsledného odhadu pak může být též nepoměr počtu respondentů z větších měst, menších měst a obcí (respondentů z měst je zásadně více než respondentů z menších měst a i těch je stále skoro dvojnásobek oproti počtu respondentů z obcí). Celkově by tedy mohl větší a vyrovnanější vzorek respondentů z jednotlivých oblastí vést k lepšímu a ucelenějšímu odrazu reality.

Úkol 2: Data sestávají ze 70 realizací 3 náhodných veličin. První dva sloupce v tabulce (Excel – Úkol 2 – Data) obsahují vysvětlující proměnné X a Y (regresory – pro všechny zadání stejné), třetí sloupec – viz. číslo zadání – udává hodnoty závislé (vysvětlované) proměnné Z. Testy provádějte na hladině významnosti 0,05 %, intervalové odhady vypočítejte se spolehlivostí 95 %. Pro zpřehlednění textu označte jednotlivé kroky.

Při řešení vycházíme převážně ze vzorců a principů představených v přednášce MSP – 07 – Regresní analýza, Vzorce pro výpočty jednotlivých veličin a jejich hodnot jsou tedy dostupné v tomto zdroji a dále zde tedy z důvodu jejich většího počtu a komplexnosti uváděny nebudou. Výpočty jsou pak prováděny v přiloženém jupyter notebooku převážně pomocí dostupných knihovních funkcí.

a) Určete vhodný model pomocí zpětné metody a regresní diagnostiky. V úvahu vezměte model polynomiální – kvadratický (v obou proměnných). Vycházejte tedy z regresní funkce:

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

až po $Z = \beta_1$. Vhodnost nalezených modelů porovnejte pomocí koeficientu determinace R^2. Možnost zjednodušení jednoho modelu na jeho submodel (model získaný vynecháním některého sloupce matice plánu) ověřte pomocí vhodného testu nulovosti regresních parametrů.

Řešení: Nejprve spočítáme koeficient determinace (R_1^2) přímo pro výchozí podobu regresní funkce. Získáváme hodnotu

$$R_1^2 = 0.995.$$

Následně spočítáme bodové odhady regresních koeficientů a u každého vypočítáme *p-hodnotu* testu na nulovost, která nám pomůže určit, které koeficienty můžeme případně vynechat (nemají na hodnotu koeficientu determinace téměř žádný vliv). U testů jednotlivých koeficientů β_j, kde $j \in \{1, 2, 3, 4, 5, 6\}$, na nulovost předpokládáme následující hypotézy:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

Uveďme si nyní odhady jednotlivých regresních koeficientů včetně jejich *p-hodnot* v následující tabulce:

koeficient	**bodový odhad**	**p-hodnota**
$\boldsymbol{\beta_1}$	68.9179	0.005
$\boldsymbol{\beta_2}$	-4.4562	0.238
$\boldsymbol{\beta_3}$	-14.6426	0.042
$\boldsymbol{\beta_4}$	-2.9886	0.000
$\boldsymbol{\beta_5}$	-3.6310	0.000
$\boldsymbol{\beta_6}$	-4.5350	0.000

Zde vidíme, že nejvyšší *p-hodnoty* nabývá koeficient β_2 (má tedy nejvyšší pravděpodobnost, že nemá vliv na závislou proměnnou Z – konkrétně 23.8 %) a zásadně převyšuje i zadanou hladinu významnosti. Při jeho odstranění pak dostáváme koeficient determinace R_2^2, jež je shodný s R_1^2, tedy $R_2^2 = R_1^2 = 0.995$. Výsledná regresní funkce pak vypadá následovně:

$$Z = \beta_1 + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 X \cdot Y$$

Opět i pro tento model spočteme bodové odhady hodnot koeficientů a *p-hodnoty*:

koeficient	**bodový odhad**	**p-hodnota**
$\boldsymbol{\beta_1}$	49.3070	0.006
$\boldsymbol{\beta_3}$	-13.3819	0.060
$\boldsymbol{\beta_4}$	-3.1659	0.000
$\boldsymbol{\beta_5}$	-3.6310	0.000
$\boldsymbol{\beta_6}$	-4.6610	0.000

Zde vidíme, že dalším odebraným koeficientem by mohl být koeficient β_3 (jeho *p-hodnota* je vyšší než α – hladina významnosti), při jeho odebrání se však koeficient determinace lehce sníží ($R_3^2 = 0.994 < R_2^2 = 0.995$). To je však zanedbatelná cena za zjednodušení modelu, které může zvýšit jeho odolnost proti přetrénování, a proto dále pokračujme s funkcí:

$$Z = \beta_1 + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 X \cdot Y$$

Nyní opět spočtěme bodové odhady hodnot koeficientů a *p-hodnot*:

koeficient	bodový odhad	p-hodnota
$\boldsymbol{\beta_1}$	24.0281	0.043
$\boldsymbol{\beta_4}$	-3.1227	0.000
$\boldsymbol{\beta_5}$	-4.6939	0.000
$\boldsymbol{\beta_6}$	-4.8477	0.000

Vidíme, že koeficient β_1 nemá nulovou *p-hodnotu* (zde záleží, jak si vyložíme hladinu významnosti ze zadání, pokud jako $\alpha = 0.05$, tak je *p-hodnota* menší – ignorujeme znak % – jinak můžeme chápat α jako 0.0005, a potom je *p-hodnota* větší), a když jej odděláme, tak se i zvýší hodnota koeficientu determinace na 0.998. Tohle však vynutí průchod modelu počátkem soustavy souřadnic (bod $[0, 0, 0]$), což může vést k jeho celkovému zkreslení a ke snížení úspěšnosti jeho odhadů pro neznámé hodnoty. Výjimkou může být situace, kdy je konstanta již dle odhadu parametrů téměř rovna nule (a pokud to s velkou jistotou prokáže test na nulovost). To však není náš případ, a tedy koeficient (konstantu) β_1 ponecháme. Odebrání ostatních koeficientů pak již vždy vyústí ve výraznější pokles koeficientu determinace a i jejich *p-hodnoty* jsou téměř nulové. Proto jsou tyto parametry ponechány.

b) Pro takto získaný model (dostatečný submodel) uveďte v jedné tabulce odhady regresních parametrů metodou nejmenších čtverců a jejich 95 % intervaly spolehlivosti.

Řešení: Následující tabulka shrnuje bodové odhady regresních parametrů a jejich 95 % intervaly spolehlivosti:

koeficient	bodový odhad	95 % interval spolehlivosti
$\boldsymbol{\beta_1}$	24.0281	$\langle 0.741, 47.315 \rangle$
$\boldsymbol{\beta_4}$	-3.1227	$\langle -3.270, -2.976 \rangle$
$\boldsymbol{\beta_5}$	-4.6939	$\langle -5.270, -4.117 \rangle$
$\boldsymbol{\beta_6}$	-4.8477	$\langle -5.342, -4.354 \rangle$

Vzorce pro výpočet intervalů viz přednáška MSP – 07 – Regresní analýza – sekce Intervalové odhady a testování hypotéz.

c) Nestranně odhadněte rozptyl závisle proměnné.

Řešení: K výpočtu můžeme opět využít vzorec ze sedmé přednášky – sekce Bodové odhady. Výsledný rozptyl závislé proměnné je

$$D(Z_i) = \sigma^2 = 2627.1949.$$

d) Vhodným testem zjistěte, že vámi zvolené dva regresní parametry jsou současně nulové.

Řešení: Pracujeme s naším submodelem a zvolíme například koeficienty β_1 a β_4. Uvažujeme následující sdružené hypotézy:

$$H_0 : (\beta_1, \beta_4) = (0, 0)$$
$$H_A : (\beta_1, \beta_4) \neq (0, 0)$$

Jelikož testujeme sdružené hypotézy, tak použijeme `f-test`. Testovací kritérium zde tedy vychází z Fisher-Snedecorova rozdělení. Získaná *p-hodnota* je pak téměř nulová (2.38e-54), je tedy dozajista menší než naše uvažovaná hladina významnosti, a proto H_0 **zamítáme**. Dle testu tedy platí $H_A : (\beta_1, \beta_4) \neq (0, 0)$.

e) Vhodným testem zjistěte, že vámi zvolené dva regresní parametry jsou stejné.

Řešení: Opět pracujme s naším submodelem, tentokrát však zvolme koeficienty β_5 a β_6. Uvažovanými hypotézami v tomto případě jsou:

$$H_0 : \beta_5 = \beta_6$$
$$H_A : \beta_5 \neq \beta_6$$

V tomto případě můžeme k výpočtu použít `t-test` (nepracujeme se sdruženou hypotézou), kde testovací kritérium vychází ze Studentova rozdělení. Výsledná *p-hodnota* je 0.763 (pozn. stejnou dostaneme i za použití `f-testu`). Jelikož je pak získaná *p-hodnota* větší než hladina významnosti α pro tento úkol, tak H_0 **nezamítáme**.