TECHNICKÁ UNIVERZITA V KOŠICIACH

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

KATEDRA MATEMATIKY A TEORETICKEJ INFORMATIKY

Výpočet korelačnej a regresnej analýzy pre reálny príklad v jazyku R

Semestrálny projekt

Aplikovaná štatistika

Zimný semester 2022/2023

Meno: Bc. Erik Žiak

Študijný program: Hospodárska informatika
II. stupeň, 1. ročník

Obsah

1. Úvodné informácie o projekte ... 3

2. Tabuľky hodnôt štatistických súborov ... 4

2.1. Náhodná premenná $\boldsymbol{X}$ ... 4

2.2. Náhodná premenná $\boldsymbol{Y}$ ... 5

3. Korelačná analýza ... 6

3.1. Overenie lineárnej závislosti náhodných premenných ... 6

3.2. Test nekorelovanosti ... 8

4. Regresná analýza ... 12

4.1. Lineárny regresný model ... 12

4.1.1. Test vhodnosti regresného modelu pomocou koeficientu determinácie ... 14

4.1.2. Test významnosti regresného modelu na hladine významnosti ... 15

4.1.3. Test významnosti regresných parametrov na hladine významnosti ... 16

4.1.4. Určenie intervalu spoľahlivosti pre regresné parametre ... 18

4.1.5. Grafická analýza rezíduí modelu ... 19

4.1.6. Overenie normálneho rozdelenia pravdepodobnosti náhodných chýb ... 20

4.1.7. Overenie nulovej strednej hodnoty náhodných chýb ... 22

4.1.8. Overenie konštantného rozptylu náhodných chýb ... 23

4.1.9. Overenie miery závislosti (korelovanosti) rezíduí ... 24

4.2. Kvadratický regresný model ... 25

4.2.1. Test vhodnosti regresného modelu pomocou koeficientu determinácie ... 27

4.2.2. Test významnosti regresného modelu na hladine významnosti ... 28

4.2.3. Test významnosti regresných parametrov na hladine významnosti ... 29

4.2.4. Určenie intervalu spoľahlivosti pre regresné parametre ... 32

4.2.5. Grafická analýza rezíduí modelu ... 33

4.2.6. Overenie normálneho rozdelenia pravdepodobnosti náhodných chýb ... 34

4.2.7. Overenie nulovej strednej hodnoty náhodných chýb ... 36

4.2.8. Overenie konštantného rozptylu náhodných chýb ... 37

4.2.9. Overenie miery závislosti (korelovanosti) rezíduí ... 38

Záver ... 39

1. Úvodné informácie o projekte

Projekt z predmetu Aplikovaná štatistika v zimnom semestri 2022/2023 pozostáva z výpočtu korelačnej a regresnej analýzy pre reálny príklad v jazyku R. Jeho cieľom je spracovať vybrané časti korelačnej a regresnej analýzy vrátane uvedenia všeobecných vzorcov, výpočtov pre reálny príklad, výsledných hodnôt a slovného vyhodnotenia získaných výsledkov s príslušnou dokumentáciou. Na výpočty a vizualizáciu grafov bol vytvorený skript v jazyku R s využitím softvéru RStudio.

Hodnoty štatistických súborov, ktoré som si vybral na účely tohto projektu, sú prevzaté z otvorených dát pre verejnosť – dostupné na webovej stránke Finančnej správy Slovenskej republiky: https://opendata.financnasprava.sk/opendata/show/humanitarna-pomoc-pre-ukrajinu-vysne-nemecke

Tieto dva štatistické súbory obsahujú dáta o **množstve vozidiel s humanitárnou pomocou, ktoré prešli cez slovensko-ukrajinský hraničný priechod Vyšné Nemecké a o váhe prepraveného tovaru pre Ukrajinu**.

Hodnoty prvého štatistického súboru predstavujú počet dopravných prostriedkov, ktoré prešli cez hraničný priechod Vyšné Nemecké. V druhom štatistickom súbore sa nachádzajú údaje o množstve prepraveného tovaru humanitárnej pomoci v kilogramoch.

Údaje pochádzajú z obdobia od 1. marca 2022 do 30. júna 2022 s dennou aktualizáciou. Oba štatistické súbory obsahujú po 122 údajov, a teda spolu ich máme 244. Tieto hodnoty sa testovali na hladine významnosti $\alpha = 0{,}1$, teda $10\ \%$.

2. Tabuľky hodnôt štatistických súborov

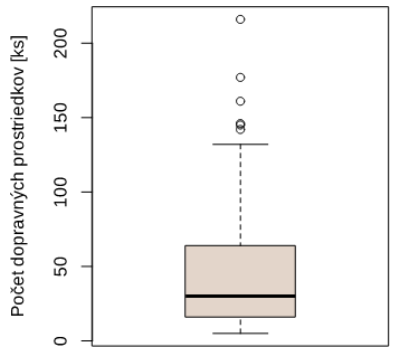
2.1. Náhodná premenná X

Tabuľka 1 uvádza hodnoty náhodnej premennej X. Tieto údaje predstavujú počet dopravných prostriedkov, ktoré prešli cez hraničný priechod Vyšné Nemecké.

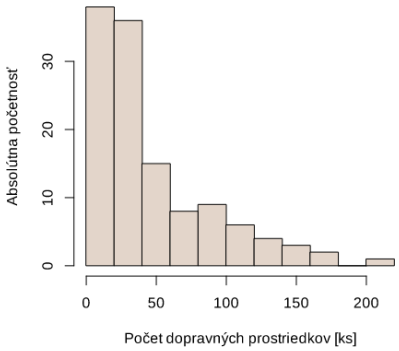
108	132	129	177	216	146	104	101	142	131	145	161
109	90	86	98	112	90	120	66	44	64	22	98
84	123	42	46	51	81	73	98	86	58	40	64
43	64	63	61	44	31	48	51	42	58	46	52
13	30	56	42	36	35	25	26	30	30	33	38
26	30	13	36	33	27	24	69	23	14	19	27
30	23	22	14	15	18	21	24	22	20	8	14
25	27	19	22	22	10	19	23	26	33	9	15
8	19	9	28	13	16	12	10	12	20	15	22
15	23	11	13	16	14	5	8	11	10	14	16
10	7										

Tabuľka 1. *Hodnoty náhodnej premennej X*

Krabicový diagram graficky znázorňuje hodnoty kvartilov aj s odľahlými hodnotami. Histogramom zobrazíme rozdelenie absolútnej početnosti hodnôt premennej X.



Obrázok 1. *Krabicový diagram premennej X*



Obrázok 2. *Histogram premennej X*

Minimálna hodnota	5
Priemerná hodnota	46,83
Maximálna hodnota	216
Prvý kvartil	16,5
Medián – druhý kvartil	30
Tretí kvartil	63,75
Koeficient šikmosti (asymetrie)	1,531
Koeficient špicatosti (excesu)	1,957

Tabuľka 2*. Číselné charakteristiky premennej X*

Pre koeficient šikmosti platí $\gamma_3 > 0$, a teda rozdelenie početností je zošikmené kladne. Pre koeficient špicatosti platí $\gamma_4 > 0$, teda krivka rozdelenia početností je špicatejšia ako Gaussova krivka normálneho rozdelenia.

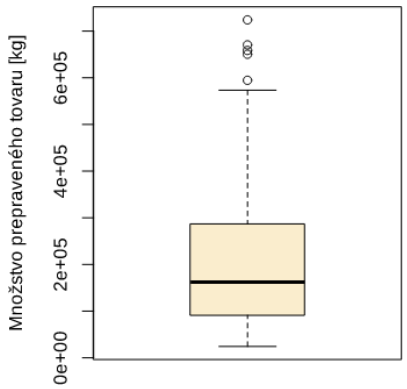
2.2. Náhodná premenná Y

Tabuľka 3 obsahuje hodnoty náhodnej premennej Y, ktoré predstavujú množstvo prepraveného tovaru humanitárnej pomoci v kilogramoch.

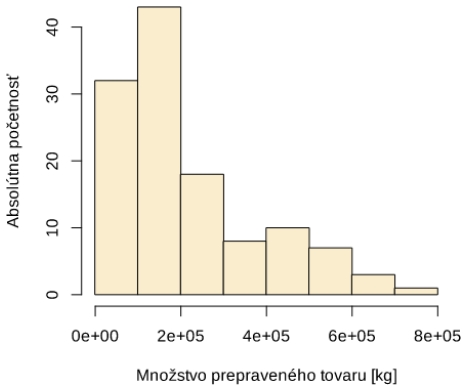
286901	272539	377609	650367	493115	594579	571583	519730
515740	556509	670470	493819	384928	315463	458803	428551
658584	414545	455357	262520	191543	304085	217358	573364
411250	541133	94583	186756	245735	419997	236559	362913
267133	331218	119472	272456	348386	467590	342409	259173
195986	247509	232540	214262	275415	497260	289322	186155
31099	63343	189116	149578	164263	241959	136602	149643
155584	154265	173183	168147	168924	281726	38246	247019
136846	149024	191267	724016	154623	71280	91021	178826
222286	139425	63943	60364	110613	64843	85769	177796
148899	71717	53295	85875	114305	175033	57191	146088
115905	87383	123327	158870	150954	128018	79203	160483
59517	121772	62365	167904	88751	29207	53051	24308
79268	138883	130564	119321	106850	68823	37144	62401
147757	119925	43455	59054	30623	61107	127471	68164
76437	122485						

Tabuľka 3. *Hodnoty náhodnej premennej Y*

Krabicový diagram graficky znázorňuje hodnoty kvartilov aj s odľahlými hodnotami. Histogramom zobrazíme rozdelenie absolútnej početnosti hodnôt premennej Y.



Obrázok 3. *Krabicový diagram premennej Y*



Obrázok 4. *Histogram premennej Y*

Minimálna hodnota	24 308
Priemerná hodnota	220 582
Maximálna hodnota	724 016
Prvý kvartil	91 912
Medián – druhý kvartil	162 373
Tretí kvartil	285 607
Koeficient šikmosti (asymetrie)	1,135
Koeficient špicatosti (excesu)	0,41

Tabuľka 4*. Číselné charakteristiky premennej Y*

Podľa koeficientov je rozdelenie početností taktiež zošikmené kladne (platí $\gamma_3 > 0$) a krivka je špicatejšia ako Gaussova krivka normálneho rozdelenia (platí $\gamma_4 > 0$).

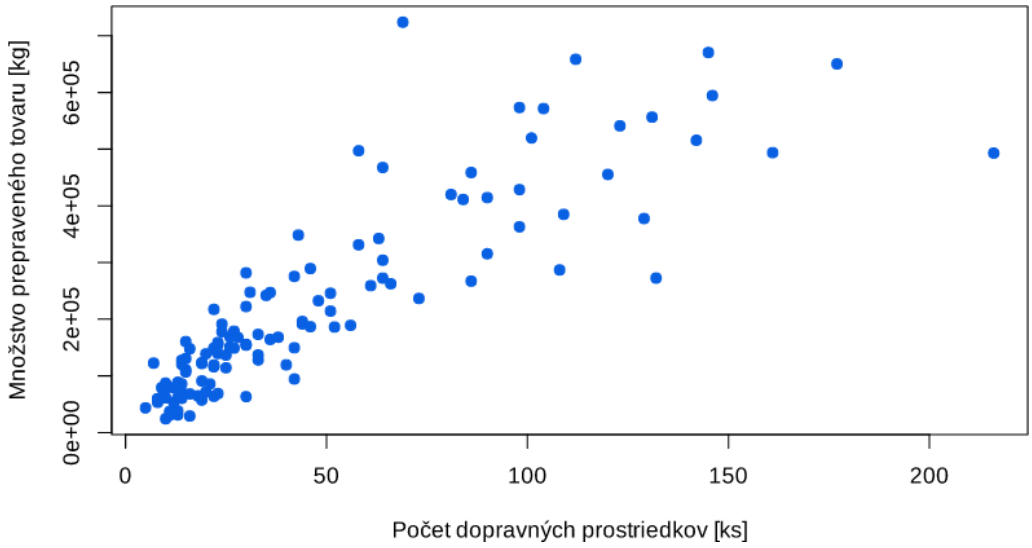
3. Korelačná analýza

Úlohou korelačnej analýzy je zaoberať sa vzájomnými závislosťami medzi náhodnými premennými, pričom sa overuje predovšetkým kvalita (sila a tesnosť) ich vzájomného vzťahu. V našom prípade sa vyskytuje základný prípad štatistickej závislosti, ktorým je jednoduchá závislosť, teda závislosť len medzi dvoma náhodnými premennými X a Y.

Máme dve náhodné premenné X a Y, ktoré predstavujú vybrané štatistické súbory:

- X – počet dopravných prostriedkov, ktoré prešli cez hraničný priechod [ks]
- Y – množstvo prepraveného tovaru humanitárnej pomoci pre Ukrajinu [kg]

Obrázok 5 zobrazuje bodový graf týchto dvoch náhodných premenných.



Obrázok 5. *Bodový graf náhodných premenných X a Y*

3.1. Overenie lineárnej závislosti náhodných premenných

Pre naše dve náhodné premenné X a Y označíme strednú hodnotu ako $E(X)$, resp. $E(Y)$ a disperziu ako $D(X)$, resp. $D(Y)$.

Na začiatku vypočítame hodnotu kovariancie $\mathrm{cov}(X,Y)$ náhodných premenných X a Y, ktorá má tvar reálneho čísla. Na jej výpočet použijeme vzťah:

$$\mathrm{cov}(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y).$$

Následne určíme hodnotu korelačného koeficientu $\rho(X,Y)$, ktorý charakterizuje lineárny vzťah dvoch náhodných veličín a je definovaný vzťahom:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{D(X)} \cdot \sqrt{D(Y)}}.$$

Korelačný koeficient ρ nadobúda hodnoty z intervalu $\langle -1; 1 \rangle$, pričom ak platí:

- $\rho(X,Y) < 0$, tak medzi náhodnými premennými X a Y je nepriama lineárna závislosť, a teda sú korelované negatívne,
- $\rho(X,Y) = 0$, tak náhodné premenné X a Y nie sú lineárne závislé, a teda sú nekorelované,
- $\rho(X,Y) > 0$, tak medzi náhodnými premennými X a Y je priama lineárna závislosť, a teda sú korelované pozitívne.

Hodnota kovariancie $\mathbf{cov}(X,Y)$ a koeficient korelácie $\boldsymbol{\rho}(X,Y)$

Hodnota kovariancie $\mathrm{cov}(X,Y)$:

$$\mathrm{cov}(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y) = 6033029$$

Koeficient korelácie $\rho(X,Y)$:

$$\boldsymbol{\rho(X,Y)} = \frac{\mathbf{cov}(X,Y)}{\sqrt{\boldsymbol{D(X)}} \cdot \sqrt{\boldsymbol{D(Y)}}} = \mathbf{0,8551756}$$

Pre hodnotu korelačného koeficientu $\boldsymbol{\rho(X,Y)} = \mathbf{0,8551756}$ platí $\rho(X,Y) > 0$, a preto medzi náhodnými premennými X a Y **je priama lineárna závislosť a sú korelované pozitívne**.

Keďže v praxi sa intervalový odhad pre korelačný koeficient ρ nepoužíva často, zaoberať sa budeme len jeho bodovým odhadom. Pearsonov výberový korelačný koeficient r_{xy} vypočítame pomocou vzťahu:

$$r_{xy} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}}.$$

Hodnoty $\bar{x}$, $\bar{y}$ vyjadrujú výberové priemery náhodných premenných X a Y, pričom:

$$\overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} {x_i}^2, \qquad \overline{y^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} {y_i}^2, \qquad \overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i.$$

Pearsonov výberový korelačný koeficient r_{xy} nadobúda taktiež hodnoty z intervalu $\langle -1; 1 \rangle$. Čím je hodnota $|r_{xy}|$ bližšia k 1, tým je lineárna korelačná závislosť silnejšia a naopak, ak sa číselná hodnota $|r_{xy}|$ blíži k 0, tým je lineárna korelačná závislosť slabšia. Hodnota $r_{xy} = 0$, indikuje, že lineárna závislosť neexistuje.

Nakoľko je to len výberová charakteristika, nevieme s istotou prehlásiť, že korelačný koeficient ρ základného súboru je nulový, a preto to otestujeme pomocou testu nekorelovanosti.

Vyhodnotiť kvalitu (silu, tesnosť) lineárnej závislosti medzi náhodnými premennými X a Y môžeme aj takto:

- ak $|r_{xy}| \leq 0{,}3$, tak lineárna závislosť nie je preukázaná,
- ak $0{,}3 < |r_{xy}| \leq 0{,}5$, tak lineárna závislosť je mierna,
- ak $0{,}5 < |r_{xy}| \leq 0{,}7$, tak lineárna závislosť je významná,
- ak $0{,}7 < |r_{xy}| \leq 0{,}9$, tak lineárna závislosť je vysoká,
- ak $0{,}9 < |r_{xy}|$, tak lineárna závislosť je veľmi vysoká.

Pearsonov výberový korelačný koeficient r_{xy}

Výberové aritmetické priemery $\bar{x}$ a $\bar{y}$:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = 46{,}827 \qquad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = 220582{,}508$$

Hodnoty $\overline{x^2}$, $\overline{y^2}$ a $\overline{x \cdot y}$:

$$\overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} {x_i}^2 = 3960{,}139 \qquad \overline{y^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} {y_i}^2 = 76358154612{,}147$$

$$\overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i = 16362437{,}606$$

Pearsonov výberový korelačný koeficient r_{xy}:

$$\boldsymbol{r_{xy} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}} = 0{,}8622431}$$

Výpočtom Pearsonovho výberového korelačného koeficientu $\boldsymbol{r_{xy} = 0{,}8622431}$ sme zistili, že medzi náhodnými premennými X a Y **existuje vysoká lineárna závislosť**, pretože platí $0{,}7 < |r_{xy}| \leq 0{,}9$, a teda hodnota sa blíži k 1.

3.2. Test nekorelovanosti

Test nekorelovanosti slúži na overenie štatistickej významnosti korelačného koeficientu ρ na určitej hladine významnosti α. Nulová hypotéza H_0 vyjadruje, že korelačný koeficient ρ nie je štatisticky významný. Naopak, alternatívna hypotéza H_1 vyjadruje, že korelačný koeficient ρ je štatisticky významný:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Pre výpočet hodnoty testovacieho kritéria t platí vzťah:

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - {r_{xy}}^2}}.$$

Kritickú oblasť K_α potom bude tvoriť interval:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

ktorého súčasťou sú aj hodnoty $-t_{1-\frac{\alpha}{2}}(n-2)$ a $t_{1-\frac{\alpha}{2}}(n-2)$, ktoré predstavujú kvantily Studentovho t-rozdelenia. Tieto hodnoty sú tabelované alebo ich získame pomocou funkcie `qt()` v jazyku R. V závere urobíme vyhodnotenie:

- ak hodnota testovacieho kritéria t patrí do kritickej oblasti K_α, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak hodnota testovacieho kritéria t nepatrí do kritickej oblasti K_α, tak nulovú hypotézu H_0 nezamietame.

Test nekorelovanosti – štatistická významnosť

Hodnota testovacieho kritéria t:

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - {r_{xy}}^2}} = 18{,}648$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0{,}95}(120) = 1{,}657651$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right)$$

$$\boldsymbol{K_{0,1} = (-\infty; -1{,}657651) \cup (1{,}657651; \infty) \Rightarrow t \in K_{0,1}}$$

Hodnota testovacieho kritéria $t = 18{,}648$ **patrí do kritickej oblasti $K_{0,1}$**, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti $\alpha = 0{,}1$ je korelačný koeficient ρ štatisticky významný.**

V ďalšom kroku môžeme overiť, či náhodné premenné X a Y sú korelované pozitívne alebo negatívne.

Pre pozitívnu koreláciu zavedieme nulovú hypotézu H_0, pre ktorú platí, že náhodné premenné sú nekorelované. Alternatívna hypotézu H_1 tvrdí, že náhodné premenné sú korelované pozitívne:

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

Hodnota testovacieho kritéria t sa vypočíta rovnako, ako v predošlom prípade. Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = (t_{1-\alpha}(n-2); \infty)$$

Hodnota $t_{1-\alpha}(n-2)$ z kritickej oblasti K_α je kvantilom Studentovho t-rozdelenia. Táto hodnota je tabelovaná alebo ju získame pomocou funkcie `qt()` v jazyku R. Pomocou hodnoty testovacieho kritéria t a kritickej oblasti K_α urobíme záver:

- ak hodnota testovacieho kritéria t patrí do kritickej oblasti K_α, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak hodnota testovacieho kritéria t nepatrí do kritickej oblasti K_α, tak nulovú hypotézu H_0 nezamietame.

Test nekorelovanosti – pozitívna korelácia

Hodnota testovacieho kritéria t:

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - {r_{xy}}^2}} = 18{,}648$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\alpha}(n-2) = t_{0{,}9}(120) = 1{,}288646$$

Kritická oblasť K_α:

$$K_\alpha = (t_{1-\alpha}(n-2); \infty)$$
$$\boldsymbol{K_{0{,}1} = (1{,}288646; \infty) \Rightarrow t \in K_{0{,}1}}$$

Hodnota testovacieho kritéria $t = 18{,}648$ **patrí do kritickej oblasti $K_{0{,}1}$**, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti $\alpha = 0{,}1$ sú náhodné premenné X a Y korelované pozitívne.**

Pre negatívnu koreláciu zavedieme nulovú hypotézu H_0, pre ktorú platí, že náhodné premenné nie sú korelované. Alternatívna hypotézu H_1 tvrdí, že náhodné premenné sú negatívne korelované:

$$H_0: \rho = 0$$

$$H_1: \rho < 0$$

Hodnota testovacieho kritéria t sa vypočíta rovnako, ako v predošlom prípade. Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; -t_{1-\alpha}(n-2)\right).$$

Hodnota $-t_{1-\alpha}(n-2)$ z kritickej oblasti K_α je kvantilom Studentovho t-rozdelenia. Táto hodnota je tabelovaná alebo ju získame pomocou funkcie `qt()` v jazyku R. Pomocou hodnoty testovacieho kritéria t a kritickej oblasti K_α urobíme záver:

- ak hodnota testovacieho kritéria t patrí do kritickej oblasti K_α, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak hodnota testovacieho kritéria t nepatrí do kritickej oblasti K_α, tak nulovú hypotézu H_0 nezamietame.

Test nekorelovanosti – negatívna korelácia

Hodnota testovacieho kritéria t:

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - {r_{xy}}^2}} = 18{,}648$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\alpha}(n-2) = t_{0{,}9}(120) = 1{,}288646$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\alpha}(n-2)\right)$$
$$\boldsymbol{K_{0{,}1} = (-\infty; -1{,}288646) \Rightarrow t \notin K_{0{,}1}}$$

Hodnota testovacieho kritéria $t = 18{,}648$ **nepatrí do kritickej oblasti $\boldsymbol{K_{0{,}1}}$**, preto nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti $\boldsymbol{\alpha = 0{,}1}$ nie sú náhodné premenné $\boldsymbol{X}$ a $\boldsymbol{Y}$ korelované negatívne.**

4. Regresná analýza

Regresná analýza sa zaoberá skúmaním tvaru závislosti medzi náhodnými premennými, pokiaľ korelačná analýza rieši existenciu samotnej závislosti – korelácie.

Takto hľadáme matematickú funkciu, tzv. regresnú funkciu, ktorá čo najlepšie popisuje tvar – priebeh závislosti kvantitatívnych náhodných premenných. Kvalitu regresného modelu vieme posúdiť pomocou korelačnej analýzy alebo iných štatistických metód.

Regresný model môžeme zapísať vo všeobecnom tvare ako:

$$Y = f(X_1, X_2, \dots, X_n, \beta_0, \beta_1, \dots, \beta_n) + \varepsilon,$$

kde f je regresná funkcia s n premennými $X_1, X_2, \dots, X_n$ a $\beta_0, \beta_1, \dots, \beta_n$ sú parametre modelu a ε je náhodná zložka, ktorá predstavuje pôsobenie náhodných vplyvov a iných faktorov, ktoré neboli zaradené do modelu.

Regresný model	Počet parametrov	Koeficient R^2
Lineárny	2	$R^2 = 0{,}7434$
Kvadratický	3	$R^2 = 0{,}7947$
Hyperbolický	2	$R^2 = 0{,}7589$
Logaritmický	2	$R^2 = 0{,}7741$

Tabuľka 5. *Porovnanie viacerých regresných modelov*

V práci porovnáme výsledky lineárneho a kvadratického regresného modelu, nakoľko z linearizovateľných modelov nadobúda kvadratický najvyššiu hodnotu koeficientu determinácie R^2, ktorý popisuje časť, ktorú je možné regresným modelom popísať. Jednotlivé modely si zadefinujeme a vykonáme testy ich vhodnosti pomocou koeficientu determinácie R^2, testy ich významnosti na hladine významnosti α, testy významnosti regresných parametrov $\beta_0, \beta_1, \dots, \beta_n$ na hladine významnosti α, určíme $100(1-\alpha)\%$-ný interval spoľahlivosti pre regresné parametre a štatistickú analýzu rezíduí. Použitím Shapiro-Wilkovho testu overíme, či rozdelenie pravdepodobnosti náhodných chýb je z normálneho rozdelenia a v závere overíme nulové stredné hodnoty náhodných chýb.

4.1. Lineárny regresný model

Ak predpokladáme, že medzi sledovanými premennými je lineárny vzťah, tak potom lineárny regresný model je definovaný v tvare:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

a bodovým odhadom lineárneho regresného modelu je rovnica:

$$\hat{Y} = b_0 + b_1 \cdot X,$$

kde β_0, β_1 sú parametre modelu, ε je náhodná chyba, b_0, b_1 sú bodové odhady a X, Y sú náhodné premenné.

Parametre b_0, b_1 odhadneme použitím metódy najmenších štvorcov a potrebujeme tak minimalizovať štatistiku reziduálneho súčtu štvorcov (*Sum of Square Errors*):

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Vyriešením nasledujúcej sústavy rovníc získame bodové odhady b_0 a b_1 parametrov β_0 a β_1 lineárneho regresného modelu:

$$n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i \cdot y_i.$$

V jazyku R sa hodnoty bodových odhadov b_0, b_1 nachádzajú v stĺpci `Estimate` tabuľky `Coefficients` – v prehľade o modeli, ktorý zobrazí funkcia `summary()`.

Po vytvorení lineárneho regresného modelu pomocou funkcie `lm()` v jazyku R si teda zobrazíme súhrn informácií o vytvorenom modeli funkciou `summary()`:

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-304973  -44696   -9628   33064  427744

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60725.4    11519.6   5.271 6.06e-07 ***
x             3413.7      183.1  18.649  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85000 on 120 degrees of freedom
Multiple R-squared:  0.7435,    Adjusted R-squared:  0.7413
F-statistic: 347.8 on 1 and 120 DF,  p-value: < 2.2e-16
```

Obrázok 6. *Súhrn informácií o lineárnom regresnom modeli*

Lineárny regresný model – bodové odhady b_0, b_1

Vyriešením sústavy rovníc získame hodnoty parametrov b_0 a b_1, ktoré doplníme do bodového odhadu lineárneho regresného modelu $\hat{Y}$:

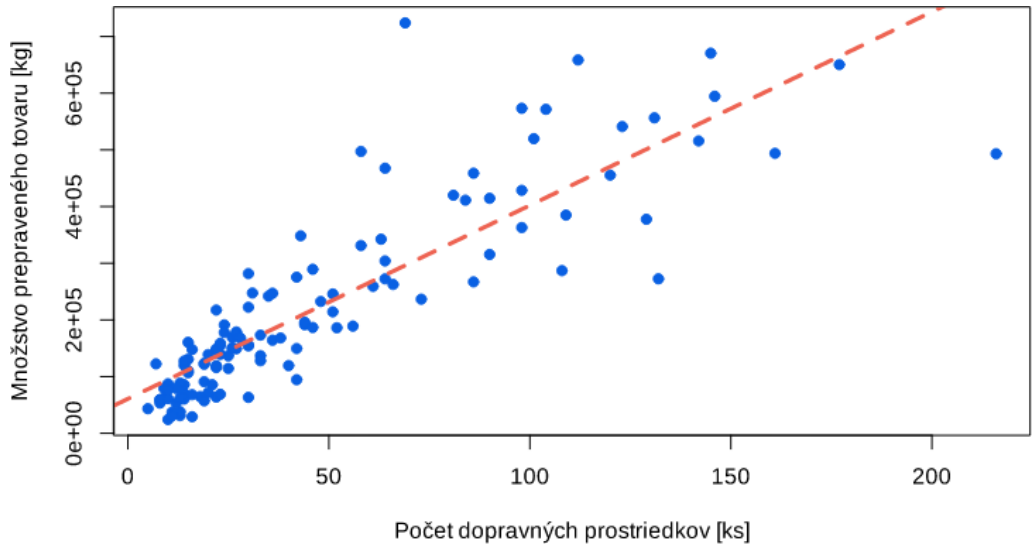
$$\hat{Y} = b_0 + b_1 \cdot X$$

$$\boldsymbol{\hat{Y} = 60725{,}4 + 3413{,}7 \cdot X}$$

Regresná priamka vytvoreného lineárneho regresného modelu nadobúda tvar $\hat{Y} = 60725,4 + 3413,7 \cdot X$.

Vo výstupe funkcie `summary()` pre vytvorený lineárny regresný model v jazyku R sú hodnoty parametrov b_0 a b_1 uvedené v tabuľke **`Coefficients`**, v stĺpci **`Estimate: 60725.4, 3413.7`**.

Obrázok 7 vykresľuje lineárnu regresnú priamku v bodovom grafe náhodných premenných X, Y a je zobrazená červenou čiarou:



Obrázok 7. *Bodový graf náhodných premenných s lineárnou regresnou priamkou*

4.1.1. Test vhodnosti regresného modelu pomocou koeficientu determinácie R^2

Posúdiť vhodnosť zvoleného regresného modelu odhadnutého pomocou metódy najmenších štvorcov nám umožňuje výberový koeficient determinácie R^2, ktorý je popisnou mierou vhodnosti použitia regresnej funkcie na predikciu. Koeficient je definovaný ako:

$$R^2 = 1 - \frac{\text{SSE}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

kde $\bar{y}$ je výberový aritmetický priemer náhodnej premennej Y a n špecifikuje rozsah náhodného výberu. Hodnota výberového koeficientu determinácie R^2 nadobúda hodnoty z intervalu $\langle 0; 1 \rangle$ a určuje časť, ktorú je možné daným regresným modelom popísať. Čím sa hodnota R^2 viac približuje k 1, tým je model vhodnejší na opísanie tvaru závislosti.

V jazyku R môžeme tento koeficient vyčítať z prehľadu, ktorý vytvorí funkcia `summary()` a nachádza sa tam hodnota `Multiple R-squared`.

Test vhodnosti regresného modelu – koeficient determinácie R^2

Výberový aritmetický priemer $\bar{y}$:

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = 220582{,}508$$

Reziduálny súčet štvorcov SSE:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 866987621804{,}453$$

Hodnota v menovateli:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = 3379584426154{,}49$$

Výberový koeficient determinácie R^2:

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 0{,}7434632}$$

Z hodnoty výberového koeficientu determinácie $\boldsymbol{R^2 = 0{,}7434632}$ vyplýva, že $\boldsymbol{74{,}34\ \%}$ **hodnôt je možné popísať vytvoreným lineárnym regresným modelom**.

Vo výstupe funkcie `summary()` pre vytvorený regresný model v jazyku R je uvedená hodnota koeficientu ako **`Multiple R-squared: 0.7435`**.

4.1.2. Test významnosti regresného modelu na hladine významnosti α

Test významnosti regresného modelu testuje významnosť výberového koeficienta determinácie a všetkých parametrov modelu. Budeme testovať nulovú hypotézu H_0, ktorá tvrdí, že regresný model nie je štatisticky významný a alternatívnu hypotézu H_1, ktorá tvrdí, že regresný model je štatisticky významný:

$$H_0: \beta_0 = \beta_1 = 0$$

$$H_1: \beta_0 \neq \beta_1 \neq 0$$

Hodnota testovacieho kritéria F je definovaná ako:

$$F = \frac{(n-m) \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{(m-1) \cdot \text{SSE}} = \frac{(n-m) \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{(m-1) \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2},$$

kde $\bar{y}$ je výberový aritmetický priemer náhodnej premennej Y, m je počet odhadovaných parametrov regresnej funkcie a n určuje rozsah náhodného výberu.

Kritickú oblasť K_α vytvára interval:

$$K_\alpha = (F_{1-\alpha}(m-1, n-m); \infty),$$

ktorého súčasťou je kvantil Fisherovho F-rozdelenia $F_{1-\alpha}(m-1, n-m)$. Hodnota je tabelovaná alebo ju získame pomocou funkcie `qf()` v jazyku R.

Test významnosti regresného modelu

Hodnota testovacieho kritéria F:

$$F = \frac{(n-m) \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{(m-1) \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = 347{,}7692$$

Kvantil Fisherovho F-rozdelenia:

$$F_{1-\alpha}(m-1, n-m) = F_{0,9}(1; 120) = 2{,}747807$$

Kritická oblasť K_α:

$$K_\alpha = (F_{1-\alpha}(m-1, n-m); \infty)$$
$$\boldsymbol{K_{0,1} = (2{,}747807; \infty) \Rightarrow F \in K_{0,1}}$$

Hodnota testovacieho kritéria $F = 347{,}7692$ **patrí do kritickej oblasti** $\boldsymbol{K_{0,1}}$, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti** $\boldsymbol{\alpha = 0{,}1}$ **je lineárny regresný model štatisticky významný.**

4.1.3. Test významnosti regresných parametrov β_0, β_1 na hladine významnosti α

V prípade testu významnosti regresných parametrov β_0, β_1 budeme testovať nulovú hypotézu H_0, ktorá tvrdí, že daný regresný parameter nie je štatisticky významné a alternatívnu hypotézu H_1, ktorá tvrdí, že daný regresný parameter je štatisticky významný:

$$H_0: \beta_0 = 0 \qquad H_0: \beta_1 = 0$$
$$H_1: \beta_0 \neq 0 \qquad H_1: \beta_1 \neq 0$$

Pre parametre β_0, β_1 sú hodnoty testovacieho kritéria t definované:

$$t = \frac{b_0 - b}{s(b_0)} \qquad t = \frac{b_1 - b}{s(b_1)},$$

kde $s(b_0), s(b_1)$ sú výberové smerodajné odchýlky príslušných odhadnutých parametrov b_0, b_1, pre ktoré platia nasledovné vzťahy:

$$s(b_0) = \sqrt{s^2(b_0)} = \hat{\sigma}^2 \cdot \frac{\sum_{i=1}^{n} x_i{}^2}{n \cdot \sum_{i=1}^{n}(x_i - \overline{x_i})^2}$$

$$s(b_1) = \sqrt{s^2(b_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

v ktorých $s^2(b_0)$, $s^2(b_1)$ sú výberové rozptyly príslušných odhadnutých parametrov b_0, b_1 a $\hat{\sigma}^2$ predstavuje výberový reziduálny rozptyl MSE.

Stanovením bodových odhadov b_0, b_1 parametrov β_0, β_1 lineárneho regresného modelu vieme pomocou reziduálneho súčtu štvorcov SSE vypočítať výberový reziduálny rozptyl – priemernú kvadratickú chybu (*Mean Square of Error*) v tvare:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-m} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m}.$$

Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

ktorého súčasťou sú aj hodnoty $-t_{1-\frac{\alpha}{2}}(n-2)$ a $t_{1-\frac{\alpha}{2}}(n-2)$, ktoré predstavujú kvantily Studentovho t-rozdelenia. Tieto hodnoty sú tabelované alebo ich získame pomocou funkcie `qt()` v jazyku R.

Test významnosti regresného parametra $\boldsymbol{\beta_0}$

Hodnota testovacieho kritéria t:

$$t = \frac{b_0 - b}{s(b_0)} = 5{,}27148$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0{,}95}(120) = 1{,}657651$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

$$\boldsymbol{K_{0,1} = (-\infty; -1{,}657651) \cup (1{,}657651; \infty) \Rightarrow t \in K_{0,1}}$$

Hodnota testovacieho kritéria $t = 5{,}27148$ **patrí do kritickej oblasti $\boldsymbol{K_{0,1}}$**, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti $\boldsymbol{\alpha = 0{,}1}$ je regresný parameter $\boldsymbol{\beta_0}$ štatisticky významný.**

Test významnosti regresného parametra $\boldsymbol{\beta_1}$

Hodnota testovacieho kritéria t:

$$t = \frac{b_1 - b}{s(b_1)} = 18{,}64857$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0{,}95}(120) = 1{,}657651$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

$$\boldsymbol{K_{0,1} = (-\infty; -1{,}657651) \cup (1{,}657651; \infty) \Rightarrow t \in K_{0,1}}$$

Hodnota testovacieho kritéria $t = 18{,}64857$ **patrí do kritickej oblasti** $\boldsymbol{K_{0,1}}$, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti$\boldsymbol{\alpha = 0{,}1}$ je regresný parameter $\boldsymbol{\beta_1}$ štatisticky významný.**

4.1.4. Určenie $\boldsymbol{100(1-\alpha)\%}$-ného intervalu spoľahlivosti pre regresné parametre $\boldsymbol{\beta_0, \beta_1}$

Obojstranný $100(1-\alpha)\%$-ný interval spoľahlivosti pre regresné parametre β_i, $i = 0; 1$ má tvar:

$$\langle b_i - t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i); b_i + t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i) \rangle$$

Určenie obojstranného 90%-ného intervalu spoľahlivosti pre regresné parametre $\boldsymbol{\beta_0, \beta_1}$

Všeobecný tvar intervalu pre regresný parameter β_i, $i = 0; 1$:

$$\langle b_i - t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i); b_i + t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i) \rangle$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0{,}95}(120) = 1{,}657651$$

Výberová smerodajná odchýlka odhadnutého parametra b_0:

$$s(b_0) = \sqrt{s^2(b_0)} = 132701080$$

Výberová smerodajná odchýlka odhadnutého parametra b_1:

$$s(b_1) = \sqrt{s^2(b_1)} = 33509{,}19$$

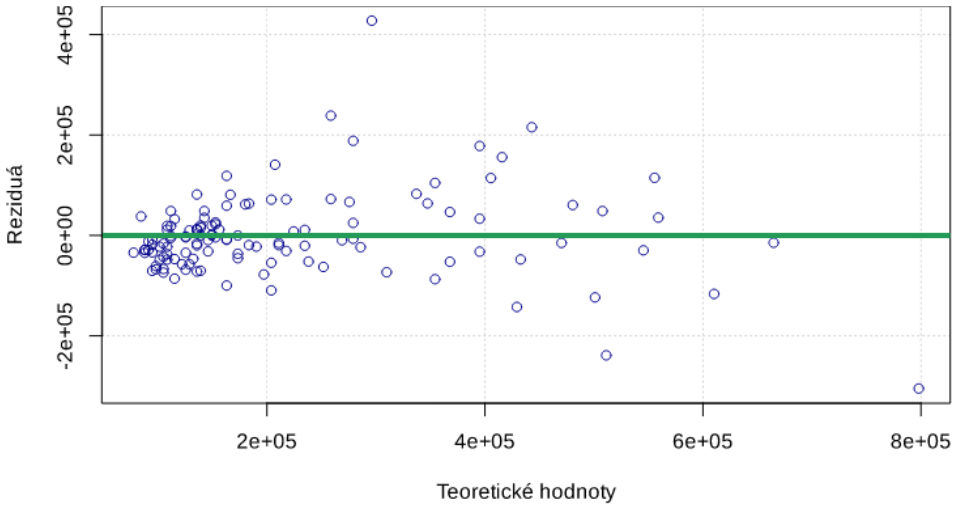
Obojstranný 90%-ný interval spoľahlivosti pre regresný parameter $\boldsymbol{\beta_0}$:

$$\boldsymbol{\langle 41629{,}93; 79820{,}86 \rangle}$$

Obojstranný 90%-ný interval spoľahlivosti pre regresný parameter β_1: $\langle 3110,276; 3717,159 \rangle$

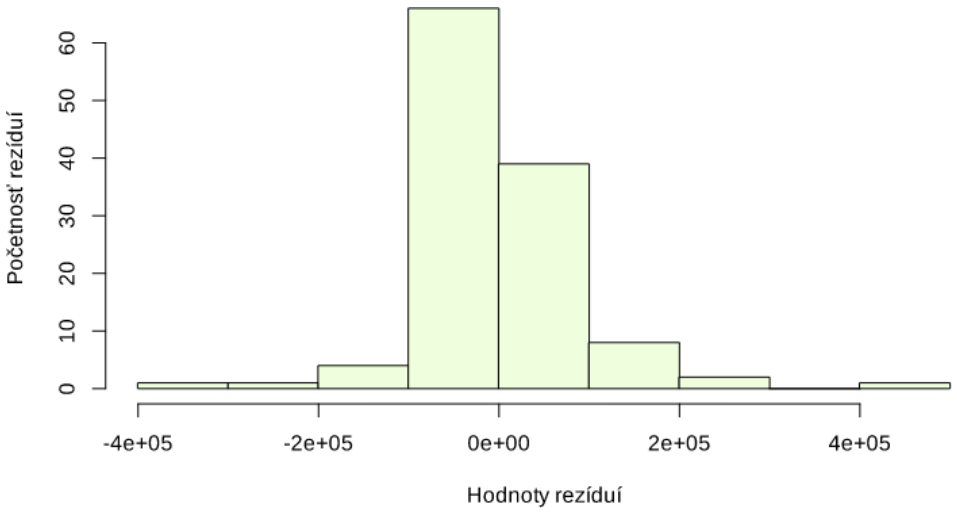
4.1.5. Grafická analýza rezíduí modelu

Rezíduá predstavujú bodové odhady náhodných chýb ε_i. Grafickou analýzou znázornime bodový graf rezíduí proti teoretickým hodnotám. Platí, že rezíduá sú náhodne rozptýlené okolo nuly, môžeme ich ohraničiť dvomi priamkami rovnobežnými s osou x a graf nenaznačuje potenciálny trend:



Obrázok 8. *Grafická analýza rezíduí lineárneho regresného modelu*

Obrázok 9 zobrazuje histogram početnosti rezíduí pre vytvorený lineárny regresný model:



Obrázok 9. *Histogram početnosti rezíduí lineárneho regresného modelu*

4.1.6. Overenie normálneho rozdelenia pravdepodobnosti náhodných chýb

Použitím Shapiro-Wilkovho testu normality vieme overiť, či rozdelenie pravdepodobnosti náhodných chýb je normálnym rozdelením. Budeme testovať hypotézy:

$$H_0: F(x) = G(x), kde\ G(x) \sim N(\mu, \sigma^2)$$

$$H_1: F(x) \neq G(x)$$

Hodnotu testovacieho kritéria vypočítame ako:

$$W = \frac{\left(\sum_{i=1}^{m} a_{i,n}\left(x_{(n-i+1)} - x_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2},$$

kde $a_{i,n}$ sú tabuľkové váhy, $\bar{x}$ je výberový aritmetický priemer a $m = \left\lfloor \frac{n}{2} \right\rfloor$ pre párny rozsah náhodného výberu n.

Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; W_\alpha(n)\right).$$

Hodnota $W_\alpha(n)$ je tabelovaná a pre hodnotu testovacieho kritéria W platí, že čím viac sa blíži k 1, tým je zhoda medzi teoretickým a empirickým rozdelením lepšia.

Tento test vieme vyhodnotiť aj pomocou hodnoty p:

- ak platí $p \leq \alpha$, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame.

Overenie normálneho rozdelenia pravdepodobnosti náhodných chýb

Hodnota testovacieho kritéria F:

$$W = \frac{\left(\sum_{i=1}^{m} a_{i,n}\left(x_{(n-i+1)} - x_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2} = 0{,}88963$$

Overenie Shapiro-Wilkovým testom na základe hodnoty p:

$$p = 4{,}921337 \cdot 10^{-8}$$

$$\boldsymbol{p \leq \alpha}$$
$$\boldsymbol{4{,}921337 \cdot 10^{-8} \leq 0{,}1}$$

Keďže platí $p \leq \alpha$, nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti $\alpha = 0{,}1$ rozdelenie pravdepodobnosti náhodných chýb nie je normálnym rozdelením.**

V jazyku R sme funkciou `shapiro.test()` z balíka `stats` získali hodnoty:

```
        Shapiro-Wilk normality test

data:  resid(model)
W = 0.88963, p-value = 4.921e-08
```

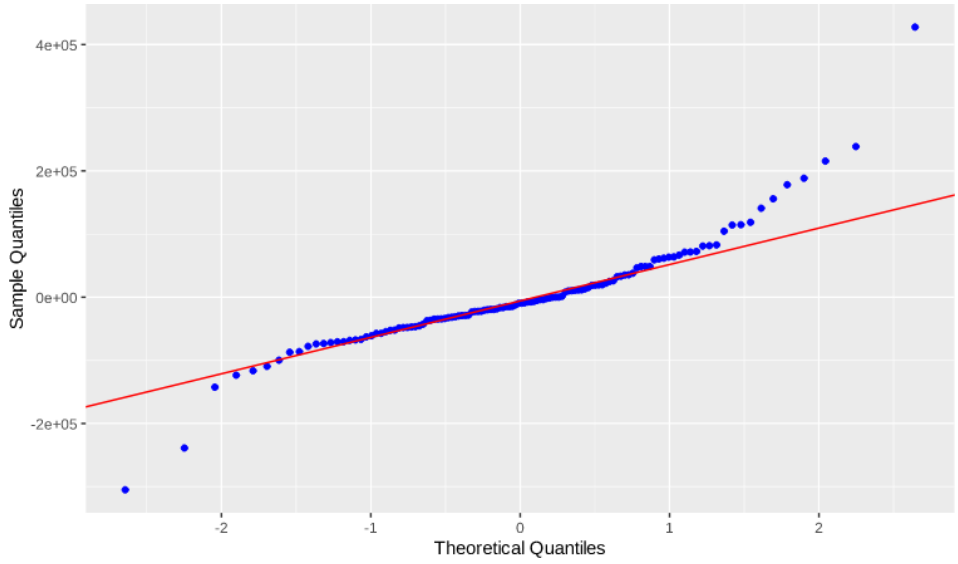
Obrázok 10. *Shapiro-Wilkov test v prostredí jazyka R*

Obrázok 11 pozostáva z tabuľky porovnania testovacích kritérií a hodnôt p pre viacero testov dobrej zhody – Shapiro-Wilkov, Kolmogorov-Smirnovov Cramér-von Misésov a Anderson-Darlingov, ktorú sme zostrojili v jazyku R použitím funkcie `ols_test_normality()` z balíka `olsrr`:

Test	Statistic	pvalue
Shapiro-Wilk	0.8896	0.0000
Kolmogorov-Smirnov	0.1214	0.0548
Cramer-von Mises	11.1585	0.0000
Anderson-Darling	3.3165	0.0000

Obrázok 11. *Porovnanie hodnôt testovacích kritérií pre testy dobrej zhody*

Nakoľko Shapiro-Wilkov test normality je založený na zisťovaní skutočnosti, či sa body zostrojeného kvantil-kvantilového grafu významne líšia od regresnej priamky preloženej týmito bodmi, použitím funkcie `ols_test_resid_qq()` z balíka `olsrr` sme zostrojili Q-Q graf:



Obrázok 12. *Kvantil-kvantilový graf pre rezíduá lineárneho regresného modelu*

4.1.7. Overenie nulovej strednej hodnoty náhodných chýb

Na overenie nulovej strednej hodnoty náhodných chýb použijeme jednovýberový test strednej hodnoty, pričom nepoznáme rozptyl (jednovýberový t-test). Budeme testovať:

$$H_0: \bar{e} = 0$$

$$H_1: \bar{e} \neq 0$$

Hodnotu testovacieho kritéria vypočítame:

$$t = \frac{\bar{e} - u_0}{s} \cdot \sqrt{n},$$

kde s je výberová smerodajná odchýlka a n je rozsah náhodného výberu.

Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

ktorého súčasťou sú aj hodnoty $-t_{1-\frac{\alpha}{2}}(n-2)$ a $t_{1-\frac{\alpha}{2}}(n-2)$, ktoré predstavujú kvantily Studentovho t-rozdelenia. Tieto hodnoty sú tabelované alebo ich získame pomocou funkcie `qt()` v jazyku R.

Tento test vieme vyhodnotiť aj pomocou hodnoty p:

- ak platí $p \leq \alpha$, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame.

Overenie nulovej strednej hodnoty náhodných chýb

Hodnota testovacieho kritéria t:

$$t = \frac{\bar{e} - u_0}{s} \cdot \sqrt{n} = 1{,}4437 \cdot 10^{-15}$$

Vyhodnotenie na základe hodnoty p:

$$p = 1$$

$$\boldsymbol{p > \alpha}$$
$$\boldsymbol{1 > 0{,}1}$$

Keďže platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti $\boldsymbol{\alpha = 0{,}1}$ sa stredná hodnota náhodných chýb nelíši významne od nuly.**

4.1.8. Overenie konštantného rozptylu náhodných chýb

Na overenie konštantného rozptylu náhodných chýb použijeme Goldfeld-Quandtov test. Hodnoty náhodnej premennej X, ktoré tvoria neklesajúcu postupnosť, je potrebné rozdeliť do dvoch skupín.

Nulová hypotéza H_0 tvrdí, že je splnený predpoklad rovnosti rozptylov oboch skupín a alternatívna hypotéza H_1 tvrdí, že nie je splnený predpoklad rovnosti rozptylov oboch skupín:

$$H_0: {\sigma_d}^2 = {\sigma_h}^2$$

$$H_1: {\sigma_d}^2 \neq {\sigma_h}^2$$

Testovacie kritérium vypočítame ako:

$$F = \frac{\mathrm{SSE}_d}{\mathrm{SSE}_h} \cdot \frac{n_h - m}{n_d - m} = \frac{\mathrm{MSE}_d}{\mathrm{MSE}_h},$$

kde m predstavuje počet odhadovaných parametrov regresnej funkcie.

Kritickú oblasť K_α reprezentuje interval:

$$K_\alpha = (F_{1-\alpha}(n_d - m, n_h - m); \infty).$$

ktorého súčasťou je kvantil Fisherovho F-rozdelenia $F_{1-\alpha}(n_d - m, n_h - m)$. Hodnota je tabelovaná alebo ju získame pomocou funkcie `qf()` v jazyku R.

Tento test vieme vyhodnotiť aj pomocou hodnoty p:

- ak platí $p \leq \alpha$, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame.

Overenie konštantného rozptylu náhodných chýb

Overenie Goldfeld-Quandtovým testom na základe hodnoty p:

$$p = 1$$

$$\boldsymbol{p > \alpha}$$
$$\boldsymbol{1 > 0,1}$$

Keďže platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti $\alpha = 0,1$ je splnený predpoklad rovnosti rozptylov náhodných chýb pre obidve skupiny.**

V jazyku R sme funkciou `gqtest()` z balíka `lmtest` získali hodnoty:

```
        Goldfeld-Quandt test

data:  model
GQ = 0.25609, df1 = 59, df2 = 59, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
```

Obrázok 13. *Goldfeld-Quandtov test v prostredí jazyka R*

4.1.9. Overenie miery závislosti (korelovanosti) rezíduí

Mieru závislosti (korelovanosti) rezíduí charakterizuje Durbin-Watsonova štatistika, ktorá má tvar:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2},$$

kde e_i predstavujú hodnoty rezíduí. Durbin-Watsonova štatistika nadobúda hodnoty z intervalu $\langle 0; 4\rangle$. Na interpretáciu jej výsledkov slúži stupnica:

- ak $DW < 1{,}4$, tak rezíduá e_i sú kladne korelované a model je nevyhovujúci,
- ak $DW \in \langle 1{,}4; 2{,}6\rangle$, tak rezíduá e_i nevykazujú autokoreláciu (majú náhodný charakter, sú nezávislé) a model je dobrý,
- ak $DW > 2{,}6$, tak rezíduá e_i sú záporne korelované a model je nevyhovujúci.

Overenie miery závislosti (korelovanosti) rezíduí

Overenie na základe Durbin-Watsonovej štatistiky DW:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2} = 1{,}433623$$

Keďže platí $\boldsymbol{DW \in \langle 1,4; 2,6\rangle}$**, tak rezíduá** $\boldsymbol{e_i}$ **nevykazujú autokoreláciu (majú náhodný charakter, sú nezávislé) a model je dobrý.**

4.2. Kvadratický regresný model

Ak predpokladáme, že vzťah medzi sledovanými premennými vyjadruje kvadratická funkcia, tak potom kvadratický regresný model je definovaný v tvare:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \varepsilon$$

a bodovým odhadom kvadratického regresného modelu je rovnica:

$$\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot X^2,$$

kde $\beta_0, \beta_1, \beta_2$ sú parametre modelu, ε je náhodná chyba, b_0, b_1, b_2 sú bodové odhady a X, Y sú náhodné premenné.

Parametre b_0, b_1, b_2 odhadneme metódou najmenších štvorcov a potrebujeme tak minimalizovať štatistiku reziduálneho súčtu štvorcov (*Sum of Square Errors*):

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2.$$

Vyriešením nasledujúcej sústavy rovníc získame bodové odhady b_0, b_1 a b_2 parametrov β_0, β_1 a β_2 kvadratického regresného modelu:

$$n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i + b_2 \cdot \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} y_i$$

$$b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} {x_i}^2 + b_2 \cdot \sum_{i=1}^{n} {x_i}^3 = \sum_{i=1}^{n} x_i \cdot y_i$$

$$b_0 \cdot \sum_{i=1}^{n} {x_i}^2 + b_1 \cdot \sum_{i=1}^{n} {x_i}^3 + b_2 \cdot \sum_{i=1}^{n} {x_i}^4 = \sum_{i=1}^{n} {x_i}^2 \cdot y_i.$$

V jazyku R sa hodnoty bodových odhadov b_0, b_1, b_2 nachádzajú v stĺpci `Estimate` tabuľky `Coefficients` – v prehľade o modeli, ktorý zobrazí funkcia `summary()`.

Po vytvorení kvadratického regresného modelu pomocou funkcie `lm()` v jazyku R si teda zobrazíme súhrn informácií o vytvorenom modeli funkciou `summary()`:

```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-239851  -41186     930   34923  382399

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2770.713  14834.025   0.187    0.852
x            6060.878    512.592  11.824  < 2e-16 ***
I(x^2)        -16.668      3.057  -5.452 2.74e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76350 on 119 degrees of freedom
Multiple R-squared:  0.7947,    Adjusted R-squared:  0.7913
F-statistic: 230.4 on 2 and 119 DF,  p-value: < 2.2e-16
```

Obrázok 14. *Súhrn informácií o kvadratickom regresnom modeli*

Kvadratický regresný model – bodové odhady b_0, b_1, b_2

Vyriešením sústavy rovníc získame hodnoty parametrov b_0, b_1, b_2, ktoré doplníme do bodového odhadu kvadratického regresného modelu $\hat{Y}$:

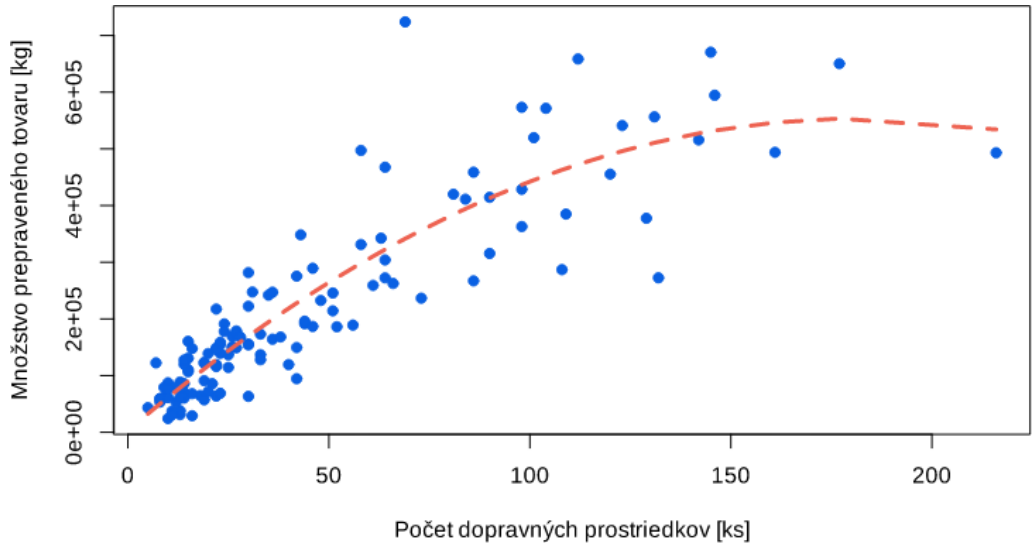
$$\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot X^2$$

$$\boldsymbol{\hat{Y} = 2770,713 + 6060,878 \cdot X - 16,668 \cdot X^2}$$

Regresná krivka vytvoreného kvadratického regresného modelu má tvar $\boldsymbol{\hat{Y} = 2770,713 + 6060,878 \cdot X - 16,668 \cdot X^2}$.

Vo výstupe funkcie `summary()` pre vytvorený kvadratický regresný model v jazyku R sú hodnoty parametrov b_0, b_1 a b_2 uvedené v tabuľke **`Coefficients`**, v stĺpci **`Estimate: 2770.713, 6060.878, -16.668`**.

Obrázok 15 vykresľuje kvadratickú regresnú krivku v bodovom grafe náhodných premenných X, Y a je zobrazená červenou čiarou:



Obrázok 15. *Bodový graf náhodných premenných s kvadratickou regresnou krivkou*

4.2.1. Test vhodnosti regresného modelu pomocou koeficientu determinácie R^2

Posúdiť vhodnosť zvoleného regresného modelu odhadnutého pomocou metódy najmenších štvorcov nám umožňuje výberový koeficient determinácie R^2, ktorý je popisnou mierou vhodnosti použitia regresnej funkcie na predikciu. Koeficient je definovaný ako:

$$R^2 = 1 - \frac{\text{SSE}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

kde $\bar{y}$ je výberový aritmetický priemer náhodnej premennej Y a n špecifikuje rozsah náhodného výberu. Hodnota výberového koeficientu determinácie R^2 nadobúda hodnoty z intervalu $\langle 0; 1\rangle$ a určuje časť, ktorú je možné daným regresným modelom popísať. Čím sa hodnota R^2 viac približuje k 1, tým je model vhodnejší na opísanie tvaru závislosti.

V jazyku R môžeme tento koeficient vyčítať z prehľadu, ktorý vytvorí funkcia `summary()` a nachádza sa tam hodnota `Multiple R-squared`.

Test vhodnosti regresného modelu – koeficient determinácie R^2

Výberový aritmetický priemer $\bar{y}$:

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = 220582{,}508$$

Reziduálny súčet štvorcov SSE:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 693689528976{,}205$$

Hodnota v menovateli:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = 3379584426154{,}49$$

Výberový koeficient determinácie R^2:

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 0{,}7947412}$$

Z hodnoty výberového koeficientu determinácie $\boldsymbol{R^2 = 0{,}7947412}$ vyplýva, že $\boldsymbol{79{,}47\ \%}$ **hodnôt je možné popísať vytvoreným kvadratickým regresným modelom**.

Vo výstupe funkcie `summary()` pre vytvorený regresný model v jazyku R je uvedená hodnota koeficientu ako **`Multiple R-squared: 0.7947`**.

4.2.2. Test významnosti regresného modelu na hladine významnosti α

Test významnosti regresného modelu testuje významnosť výberového koeficienta determinácie a všetkých parametrov modelu. Budeme testovať nulovú hypotézu H_0, ktorá tvrdí, že regresný model nie je štatisticky významný a alternatívnu hypotézu H_1, ktorá tvrdí, že regresný model je štatisticky významný:

$$H_0: \beta_0 = \beta_1 = \beta_2 = 0$$

$$H_1: \beta_0 \neq \beta_1 \neq \beta_2 \neq 0$$

Hodnota testovacieho kritéria F je definovaná ako:

$$F = \frac{(n-m) \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{(m-1) \cdot \text{SSE}} = \frac{(n-m) \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{(m-1) \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2},$$

kde $\bar{y}$ je výberový aritmetický priemer náhodnej premennej Y, m je počet odhadovaných parametrov regresnej funkcie a n určuje rozsah náhodného výberu.

Kritickú oblasť K_α vytvára interval:

$$K_\alpha = (F_{1-\alpha}(m-1, n-m); \infty),$$

ktorého súčasťou je kvantil Fisherovho F-rozdelenia $F_{1-\alpha}(m-1, n-m)$. Hodnota je tabelovaná alebo ju získame pomocou funkcie `qf()` v jazyku R.

Test významnosti regresného modelu

Hodnota testovacieho kritéria F:

$$F = \frac{(n-m)\cdot\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{(m-1)\cdot\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} = 230{,}3779$$

Kvantil Fisherovho F-rozdelenia:

$$F_{1-\alpha}(m-1, n-m) = F_{0,9}(2; 119) = 2{,}347719$$

Kritická oblasť K_α:

$$K_\alpha = (F_{1-\alpha}(m-1, n-m); \infty)$$
$$\boldsymbol{K_{0,1} = (2{,}347719; \infty) \Rightarrow F \in K_{0,1}}$$

Hodnota testovacieho kritéria $F = 230{,}3779$ **patrí do kritickej oblasti** $\boldsymbol{K_{0,1}}$, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti** $\boldsymbol{\alpha = 0{,}1}$ **je kvadratický regresný model štatisticky významný.**

4.2.3. Test významnosti regresných parametrov $\boldsymbol{\beta_0, \beta_1, \beta_2}$ na hladine významnosti $\boldsymbol{\alpha}$

V prípade testu významnosti regresných parametrov β_0, β_1 budeme testovať nulovú hypotézu H_0, ktorá tvrdí, že daný regresný parameter nie je štatisticky významné a alternatívnu hypotézu H_1, ktorá tvrdí, že daný regresný parameter je štatisticky významný:

$$H_0: \beta_0 = 0 \qquad H_0: \beta_1 = 0 \qquad H_0: \beta_2 = 0$$
$$H_1: \beta_0 \neq 0 \qquad H_1: \beta_1 \neq 0 \qquad H_1: \beta_2 \neq 0$$

Pre parametre $\beta_0, \beta_1, \beta_2$ sú hodnoty testovacieho kritéria t definované:

$$t = \frac{b_0 - b}{s(b_0)} \qquad t = \frac{b_1 - b}{s(b_1)} \qquad t = \frac{b_2 - b}{s(b_2)},$$

kde $s(b_0)$, $s(b_1)$, $s(b_2)$ sú výberové smerodajné odchýlky príslušných odhadnutých parametrov b_0, b_1, b_2, pre ktoré platia vzťahy:

$$s(b_0) = \sqrt{s^2(b_0)} = \hat{\sigma}^2 \cdot \frac{\sum_{i=1}^{n} {x_i}^2}{n\cdot\sum_{i=1}^{n}(x_i-\bar{x_i})^2}$$

$$s(b_1) = \sqrt{s^2(b_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$s(b_2) = \sqrt{s^2(b_2)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

v ktorých $s^2(b_0)$, $s^2(b_1)$, $s^2(b_2)$ sú výberové rozptyly príslušných odhadnutých parametrov b_0, b_1, b_2 a $\hat{\sigma}^2$ predstavuje výberový reziduálny rozptyl MSE.

Stanovením bodových odhadov b_0, b_1, b_2 parametrov $\beta_0, \beta_1, \beta_2$ kvadratického regresného modelu vieme pomocou reziduálneho súčtu štvorcov SSE vypočítať výberový reziduálny rozptyl – priemernú kvadratickú chybu (*Mean Square of Error*) v tvare:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-m} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m}.$$

Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

ktorého súčasťou sú aj hodnoty $-t_{1-\frac{\alpha}{2}}(n-2)$ a $t_{1-\frac{\alpha}{2}}(n-2)$, ktoré predstavujú kvantily Studentovho t-rozdelenia. Tieto hodnoty sú tabelované alebo ich získame pomocou funkcie `qt()` v jazyku R.

Test významnosti regresného parametra $\boldsymbol{\beta_0}$

Hodnota testovacieho kritéria t:

$$t = \frac{b_0 - b}{s(b_0)} = 0{,}2677694$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0{,}95}(120) = 1{,}657651$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$
$$\boldsymbol{K_{0,1} = (-\infty; -1{,}657651) \cup (1{,}657651; \infty) \Rightarrow t \notin K_{0,1}}$$

Hodnota testovacieho kritéria $t = 0{,}2677694$ **nepatrí do kritickej oblasti** $\boldsymbol{K_{0,1}}$, preto nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti $\boldsymbol{\alpha = 0{,}1}$ regresný parameter $\boldsymbol{\beta_0}$ nie je štatisticky významný.**

Test významnosti regresného parametra $\boldsymbol{\beta_1}$

Hodnota testovacieho kritéria t:

$$t = \frac{b_1 - b}{s(b_1)} = 36{,}86042$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0,95}(120) = 1{,}657651$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

$$\boldsymbol{K_{0,1} = (-\infty; -1{,}657651) \cup (1{,}657651; \infty) \Rightarrow t \in K_{0,1}}$$

Hodnota testovacieho kritéria $t = 36{,}86042$ **patrí do kritickej oblasti** $\boldsymbol{K_{0,1}}$, preto nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti** $\boldsymbol{\alpha = 0,1}$ **je regresný parameter** $\boldsymbol{\beta_1}$ **štatisticky významný.**

Test významnosti regresného parametra β_2

Hodnota testovacieho kritéria t:

$$t = \frac{b_2 - b}{s(b_2)} = -0{,}1013675$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0,95}(120) = 1{,}657651$$

Kritická oblasť K_α:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

$$\boldsymbol{K_{0,1} = (-\infty; -1{,}657651) \cup (1{,}657651; \infty) \Rightarrow t \notin K_{0,1}}$$

Hodnota testovacieho kritéria $t = -0{,}1013675$ **nepatrí do kritickej oblasti** $\boldsymbol{K_{0,1}}$, preto nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti** $\boldsymbol{\alpha = 0,1}$ **regresný parameter** $\boldsymbol{\beta_2}$ **nie je štatisticky významný.**

4.2.4. Určenie $100(1-\alpha)\%$-ného intervalu spoľahlivosti pre regresné parametre $\beta_0, \beta_1, \beta_2$

Obojstranný $100(1-\alpha)\%$-ný interval spoľahlivosti pre regresné parametre β_i, $i = 0; 1$ má tvar:

$$\langle b_i - t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i); b_i + t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i) \rangle$$

Určenie obojstranného 90%-ného intervalu spoľahlivosti pre regresné parametre $\beta_0, \beta_1, \beta_2$

Všeobecný tvar intervalu pre regresný parameter β_i, $i = 0; 1$:

$$\langle b_i - t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i); b_i + t_{1-\frac{\alpha}{2}}(n-2) \cdot s(b_i) \rangle$$

Kvantil Studentovho t-rozdelenia:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0,95}(120) = 1{,}657651$$

Výberová smerodajná odchýlka odhadnutého parametra b_0:

$$s(b_0) = \sqrt{s^2(b_0)} = 132701080$$

Výberová smerodajná odchýlka odhadnutého parametra b_1:

$$s(b_1) = \sqrt{s^2(b_1)} = 33509{,}19$$

Výberová smerodajná odchýlka odhadnutého parametra b_2:

$$s(b_2) = \sqrt{s^2(b_2)} = 33509{,}19$$

Obojstranný 90%-ný interval spoľahlivosti pre regresný parameter β_0:

$$\langle \mathbf{-14381,63; 19923,06} \rangle$$

Obojstranný 90%-ný interval spoľahlivosti pre regresný parameter β_1:

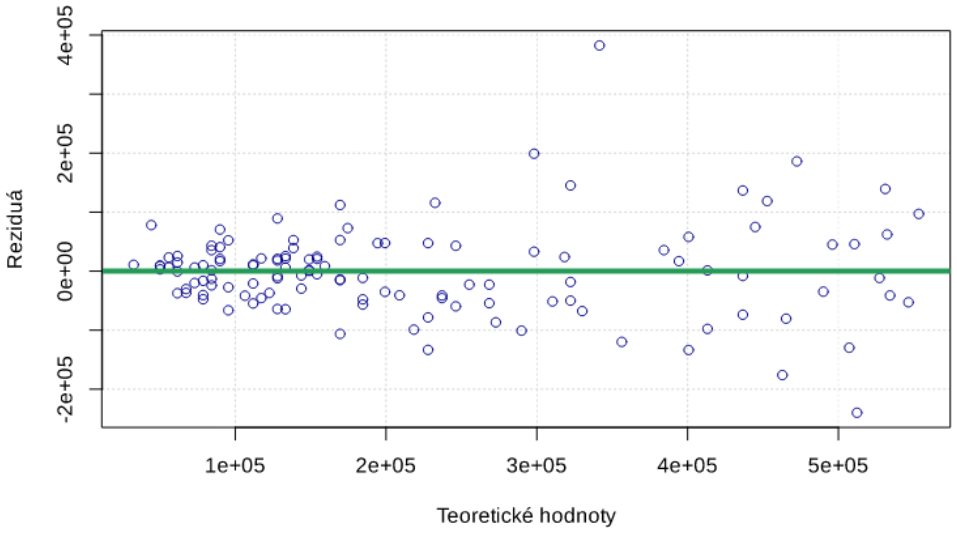
$$\langle \mathbf{5788,314; 6333,442} \rangle$$

Obojstranný 90%-ný interval spoľahlivosti pre regresný parameter β_2:

$$\langle \mathbf{-289,2315; 255,8963} \rangle$$

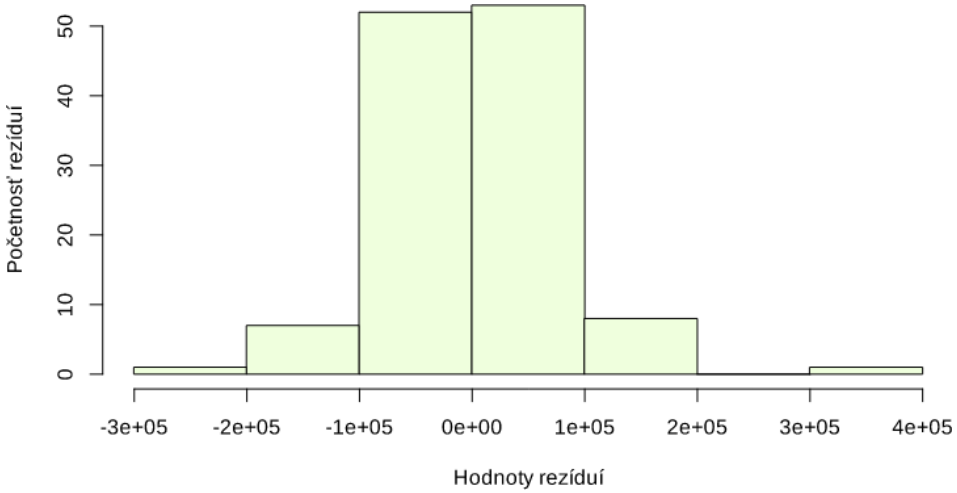
4.2.5. Grafická analýza rezíduí modelu

Rezíduá predstavujú bodové odhady náhodných chýb ε_i. Grafickou analýzou znázornime bodový graf rezíduí proti teoretickým hodnotám. Platí, že rezíduá sú náhodne rozptýlené okolo nuly, môžeme ich ohraničiť dvomi priamkami rovnobežnými s osou x a graf nenaznačuje potenciálny trend:



Obrázok 16. *Grafická analýza rezíduí kvadratického regresného modelu*

Obrázok 17 zobrazuje histogram početnosti rezíduí pre náš vytvorený kvadratický regresný model:



Obrázok 17. *Histogram početnosti rezíduí kvadratického regresného modelu*

4.2.6. Overenie normálneho rozdelenia pravdepodobnosti náhodných chýb

Použitím Shapiro-Wilkovho testu normality vieme overiť, či rozdelenie pravdepodobnosti náhodných chýb je normálnym rozdelením. Budeme testovať hypotézy:

$$H_0: F(x) = G(x), kde\ G(x) \sim N(\mu, \sigma^2)$$

$$H_1: F(x) \neq G(x)$$

Hodnotu testovacieho kritéria vypočítame ako:

$$W = \frac{\left(\sum_{i=1}^{m} a_{i,n}\left(x_{(n-i+1)} - x_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2},$$

kde $a_{i,n}$ sú tabuľkové váhy, $\bar{x}$ je výberový aritmetický priemer a $m = \left\lfloor \frac{n}{2} \right\rfloor$ pre párny rozsah náhodného výberu n.

Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; W_\alpha(n)\right).$$

Hodnota $W_\alpha(n)$ je tabelovaná a pre hodnotu testovacieho kritéria W platí, že čím viac sa blíži k 1, tým je zhoda medzi teoretickým a empirickým rozdelením lepšia.

Tento test vieme vyhodnotiť aj pomocou hodnoty p:

- ak platí $p \leq \alpha$, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame.

Overenie normálneho rozdelenia pravdepodobnosti náhodných chýb

Hodnota testovacieho kritéria F:

$$W = \frac{\left(\sum_{i=1}^{m} a_{i,n}\left(x_{(n-i+1)} - x_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2} = 0{,}92683$$

Overenie Shapiro-Wilkovým testom na základe hodnoty p:

$$p = 5{,}214831 \cdot 10^{-6}$$

$$\boldsymbol{p \leq \alpha}$$
$$\boldsymbol{5{,}214831 \cdot 10^{-6} \leq 0{,}1}$$

Keďže platí $p \leq \alpha$, nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1. **Môžeme predpokladať, že na hladine významnosti $\alpha = 0{,}1$ rozdelenie pravdepodobnosti náhodných chýb nie je normálnym rozdelením.**

V jazyku R sme funkciou `shapiro.test()` z balíka `stats` získali hodnoty:

```
        Shapiro-Wilk normality test

data:  resid(model)
W = 0.92683, p-value = 5.215e-06
```

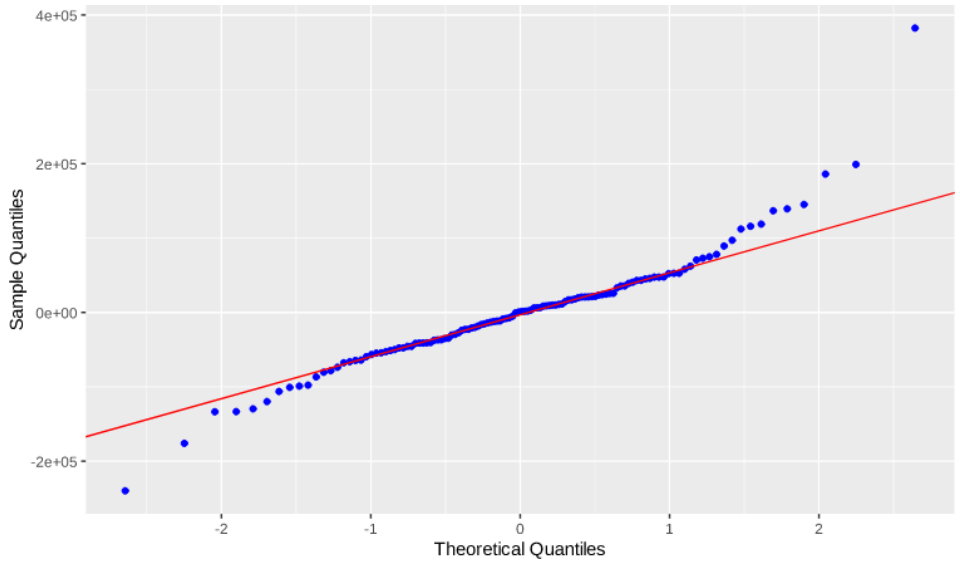
Obrázok 18. *Shapiro-Wilkov test v prostredí jazyka R*

Obrázok 19 pozostáva z tabuľky porovnania testovacích kritérií a hodnôt p pre viacero testov dobrej zhody – Shapiro-Wilkov, Kolmogorov-Smirnovov Cramér-von Misésov a Anderson-Darlingov, ktorú sme zostrojili v jazyku R použitím funkcie `ols_test_normality()` z balíka `olsrr`:

Test	Statistic	pvalue
Shapiro-Wilk	0.9268	0.0000
Kolmogorov-Smirnov	0.105	0.1357
Cramer-von Mises	10.1749	0.0000
Anderson-Darling	1.7455	2e-04

Obrázok 19. *Porovnanie hodnôt testovacích kritérií pre testy dobrej zhody*

Nakoľko Shapiro-Wilkov test normality je založený na zisťovaní skutočnosti, či sa body zostrojeného kvantil-kvantilového grafu významne líšia od regresnej priamky preloženej týmito bodmi, použitím funkcie `ols_test_resid_qq()` z balíka `olsrr` sme zostrojili Q-Q graf:



Obrázok 20. *Kvantil-kvantilový graf pre rezíduá kvadratického regresného modelu*

4.2.7. Overenie nulovej strednej hodnoty náhodných chýb

Na overenie nulovej strednej hodnoty náhodných chýb použijeme jednovýberový test strednej hodnoty, pričom nepoznáme rozptyl (jednovýberový t-test). Budeme testovať:

$$H_0: \bar{e} = 0$$

$$H_1: \bar{e} \neq 0$$

Hodnotu testovacieho kritéria vypočítame:

$$t = \frac{\bar{e} - u_0}{s} \cdot \sqrt{n},$$

kde s je výberová smerodajná odchýlka a n je rozsah náhodného výberu.

Kritickú oblasť K_α bude tvoriť interval:

$$K_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); \infty\right),$$

ktorého súčasťou sú aj hodnoty $-t_{1-\frac{\alpha}{2}}(n-2)$ a $t_{1-\frac{\alpha}{2}}(n-2)$, ktoré predstavujú kvantily Studentovho t-rozdelenia. Tieto hodnoty sú tabelované alebo ich získame pomocou funkcie `qt()` v jazyku R.

Tento test vieme vyhodnotiť aj pomocou hodnoty p:

- ak platí $p \leq \alpha$, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame.

Overenie nulovej strednej hodnoty náhodných chýb

Hodnota testovacieho kritéria t:

$$t = \frac{\bar{e} - u_0}{s} \cdot \sqrt{n} = -5{,}0713 \cdot 10^{-16}$$

Vyhodnotenie na základe hodnoty p:

$$p = 1$$

$$\boldsymbol{p > \alpha}$$
$$\boldsymbol{1 > 0{,}1}$$

Keďže platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti $\boldsymbol{\alpha = 0{,}1}$ sa stredná hodnota náhodných chýb nelíši významne od nuly.**

4.2.8. Overenie konštantného rozptylu náhodných chýb

Na overenie konštantného rozptylu náhodných chýb použijeme Goldfeld-Quandtov test. Hodnoty náhodnej premennej X, ktoré tvoria neklesajúcu postupnosť, je potrebné rozdeliť do dvoch skupín.

Nulová hypotéza H_0 tvrdí, že je splnený predpoklad rovnosti rozptylov oboch skupín a alternatívna hypotéza H_1 tvrdí, že nie je splnený predpoklad rovnosti rozptylov oboch skupín:

$$H_0: {\sigma_d}^2 = {\sigma_h}^2$$

$$H_1: {\sigma_d}^2 \neq {\sigma_h}^2$$

Testovacie kritérium vypočítame ako:

$$F = \frac{\mathrm{SSE}_d}{\mathrm{SSE}_h} \cdot \frac{n_h - m}{n_d - m} = \frac{\mathrm{MSE}_d}{\mathrm{MSE}_h},$$

kde m predstavuje počet odhadovaných parametrov regresnej funkcie.

Kritickú oblasť K_α reprezentuje interval:

$$K_\alpha = (F_{1-\alpha}(n_d - m, n_h - m); \infty).$$

ktorého súčasťou je kvantil Fisherovho F-rozdelenia $F_{1-\alpha}(n_d - m, n_h - m)$. Hodnota je tabelovaná alebo ju získame pomocou funkcie `qf()` v jazyku R.

Tento test vieme vyhodnotiť aj pomocou hodnoty p:

- ak platí $p \leq \alpha$, tak nulovú hypotézu H_0 zamietame a prijímame alternatívnu hypotézu H_1,
- ak platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame.

Overenie konštantného rozptylu náhodných chýb

Overenie Goldfeld-Quandtovým testom na základe hodnoty p:

$$p = 1$$

$$\boldsymbol{p > \alpha}$$
$$\boldsymbol{1 > 0,1}$$

Keďže platí $p > \alpha$, tak nulovú hypotézu H_0 nezamietame. **Môžeme predpokladať, že na hladine významnosti $\alpha = 0,1$ je splnený predpoklad rovnosti rozptylov náhodných chýb pre obidve skupiny.**

V jazyku R sme funkciou `gqtest()` z balíka `lmtest` získali hodnoty:

```
        Goldfeld-Quandt test

data:  model
GQ = 0.17323, df1 = 58, df2 = 58, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
```

Obrázok 21. *Goldfeld-Quandtov test v prostredí jazyka R*

4.2.9. Overenie miery závislosti (korelovanosti) rezíduí

Mieru závislosti (korelovanosti) rezíduí charakterizuje Durbin-Watsonova štatistika, ktorá má tvar:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2},$$

kde e_i predstavujú hodnoty rezíduí. Durbin-Watsonova štatistika nadobúda hodnoty z intervalu $\langle 0; 4\rangle$. Na interpretáciu jej výsledkov slúži stupnica:

- ak $DW < 1{,}4$, tak rezíduá e_i sú kladne korelované a model je nevyhovujúci,
- ak $DW \in \langle 1{,}4; 2{,}6\rangle$, tak rezíduá e_i nevykazujú autokoreláciu (majú náhodný charakter, sú nezávislé) a model je dobrý,
- ak $DW > 2{,}6$, tak rezíduá e_i sú záporne korelované a model je nevyhovujúci.

Overenie miery závislosti (korelovanosti) rezíduí

Overenie na základe Durbin-Watsonovej štatistiky DW:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2} = 1{,}416667$$

Keďže platí $\boldsymbol{DW \in \langle 1,4; 2,6\rangle}$**, tak rezíduá** $\boldsymbol{e_i}$ **nevykazujú autokoreláciu (majú náhodný charakter, sú nezávislé) a model je dobrý.**

Záver

V práci sú spracované určité časti korelačnej a regresnej analýzy. Zadefinovali sme všetky potrebné vzťahy k jednotlivým častiam, zaviedli potrebné hypotézy a testovacie kritériá.

Sivý rámec na konci každej podkapitoly obsahuje konkrétne výpočty s vyhodnotenými výsledkami na reálnom príklade, ktorý sme zvolili na začiatku – dáta o množstve vozidiel s humanitárnou pomocou, ktoré prešli cez slovensko-ukrajinský hraničný priechod Vyšné Nemecké a o váhe prepraveného tovaru pre Ukrajinu (náhodné premenné X, Y). Súčasťou práce je skript v jazyku R, ktorý poslúžil na všetky potrebné výpočty a vykreslenie znázornených grafov.

Výpočtom Pearsonovho výberového korelačného koeficientu $\boldsymbol{r_{xy} = 0,8622431}$ sme zistili, že medzi náhodnými premennými X a Y **existuje vysoká lineárna závislosť**.

Z hodnoty výberového koeficientu determinácie $\boldsymbol{R^2 = 0,7434632}$ vyplýva, že $\boldsymbol{74,34}$ % **hodnôt je možné popísať vytvoreným lineárnym regresným modelom**.

Z hodnoty výberového koeficientu determinácie $\boldsymbol{R^2 = 0,7947412}$ vyplýva, že $\boldsymbol{79,47}$ % **hodnôt je možné popísať vytvoreným kvadratickým regresným modelom**.

Niektoré použité funkcie v jazyku R

- `boxplot()` na vykreslenie krabicového diagramu
- `dwtest()` na výpočet Durbin-Watsonovej štatistiky
- `gqtest()` na vykonanie Goldfeld-Quandtovho testu
- `hist()` na vykreslenie histogramu početností
- `lm()` na vytvorenie regresného modelu
- `ols_test_normality()` na porovnanie výsledkov testov dobrej zhody
- `ols_test_resid_qq()` na vykreslenie kvantil-kvantilového grafu
- `plot()` na vykreslenie grafov s regresnými krivkami
- `qf()` na výpočet kvantilu Fisherovho F-rozdelenia
- `qt()` na výpočet kvantilu Studentovho t-rozdelenia
- `shapiro.test()` na vykonanie Shapiro-Wilkovho testu
- `summary()` na zobrazenie informácií o regresnom modeli
- `t.test()` na vykonanie jednovýberového t-testu

Použitá literatúra

[1] Andrejiová, M. ***Štatistické metódy v praxi.*** Technická univerzita v Košiciach, Košice (2016).

[2] Grinčová, A., Petrillová, J. ***Aplikovaná štatistika.*** Technická univerzita v Košiciach, Košice (2019).

[3] Grinčová, A., Petrillová, J. ***Aplikovaná štatistika – zbierka riešených úloh v jazyku R.*** Technická univerzita v Košiciach, Košice (2019).

[4] Török, Cs. ***Úvod do teórie pravdepodobnosti a matematickej štatistiky.*** Technická univerzita v Košiciach, Košice (1992).

Zoznam obrázkov

Obrázok 1. Krabicový diagram premennej X 4
Obrázok 2. Histogram premennej X 4
Obrázok 3. Krabicový diagram premennej Y 5
Obrázok 4. Histogram premennej Y 5
Obrázok 5. Bodový graf náhodných premenných X a Y 6
Obrázok 6. Súhrn informácií o lineárnom regresnom modeli 13
Obrázok 7. Bodový graf náhodných premenných s lineárnou regresnou priamkou 14
Obrázok 8. Grafická analýza rezíduí lineárneho regresného modelu 19
Obrázok 9. Histogram početnosti rezíduí lineárneho regresného modelu 19
Obrázok 10. Shapiro-Wilkov test v prostredí jazyka R 21
Obrázok 11. Porovnanie hodnôt testovacích kritérií pre testy dobrej zhody 21
Obrázok 12. Kvantil-kvantilový graf pre rezíduá lineárneho regresného modelu 21
Obrázok 13. Goldfeld-Quandtov test v prostredí jazyka R 24
Obrázok 14. Súhrn informácií o kvadratickom regresnom modeli 26
Obrázok 15. Bodový graf náhodných premenných s kvadratickou regresnou krivkou 27
Obrázok 16. Grafická analýza rezíduí kvadratického regresného modelu 33
Obrázok 17. Histogram početnosti rezíduí kvadratického regresného modelu 33
Obrázok 18. Shapiro-Wilkov test v prostredí jazyka R 35
Obrázok 19. Porovnanie hodnôt testovacích kritérií pre testy dobrej zhody 35
Obrázok 20. Kvantil-kvantilový graf pre rezíduá kvadratického regresného modelu 35
Obrázok 21. Goldfeld-Quandtov test v prostredí jazyka R 38