SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK
AK. GOD. 2020. / 2021.

SEMINARSKI RAD IZ STATISTIČKOG PRAKTIKUMA 1

ZADATAK 53.

AUTOR: Josipa Radnić

ZAGREB, 2020. GOD

Proučavamo vezu između težine tijela *x* i određenog svojstva metabolizma *y*. Dakle, statistička obilježja koja promatramo su

X = težina tijela Y = određeno svojstvo metabolizma

Datoteka **zad53r.dat** sadrži informacije o realizaciji slučajnih uzoraka X_i i Y_i, $i = 1, \ldots, n$, gdje će n u nastavku teksta biti jednak 14.

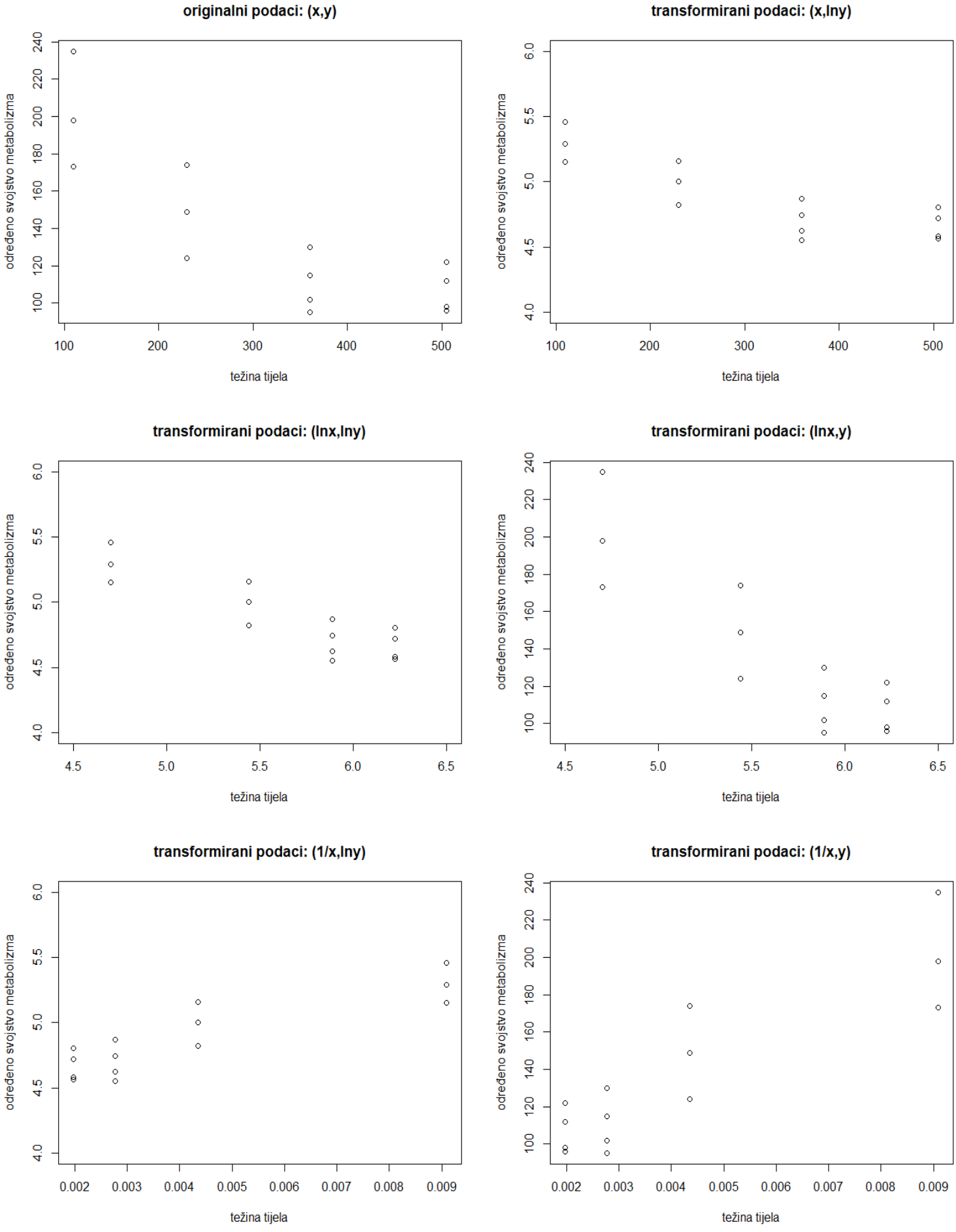
Podaci su zadani tablicom:

X	Y
110	235
110	198
110	173
230	174
230	149
230	124
360	115
360	130
360	102
360	95
505	122
505	112
505	98
505	96

Tablica 1: podaci

Prikazat ćemo podatke *(x, y)* u Kartezijevom koordinatnom sustavu, te ćemo isto napraviti za transformirane originalne podatke:

(1) $(x', y') = (x, \ln y)$
(2) $(x', y') = (\ln x, \ln y)$
(3) $(x', y') = (\ln x, y)$
(4) $(x', y') = (\frac{1}{x}, \ln y)$
(5) $(x', y') = (\frac{1}{x}, y)$



Slika 1: prikaz originalnih i transformiranih podataka

Također ćemo izračunati Pearsonov koeficijent korelacije pomoću formule $r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}$ kako bi mogli pretpostaviti postoji li linearna zavisnost između podataka.

Računajući dobili smo ove rezultate:

MODEL	PEARSONOV KOEFICIJENT KORELACIJE
(x, y)	-0.8346941
(1)	-0.8456404
(2)	-0.8861268
(3)	-0.8866195
(4)	0.879965
(5)	0.8922631

Tablica 2: Pearsonov koeficijent korelacije

Iz prikaza podataka na Slici 1 vidimo da je moguće pretpostaviti linearnu zavisnost između podataka u svim slučajevima, a pomoću Tablice 2 vidimo da je apsolutna vrijednost Pearsonovog koeficijenta korelacije najveća u modelima (2), (3) i (5), te možemo pretpostaviti da je najbolja linearna zavisnost u tim modelima.

Uzeti ćemo model (2) i (5) te ćemo za oba modela napraviti prilagodbu linearnog modela $y' = \theta_0 + \theta_1 x'$ transformiranim podacima. Kako bi mogli u ostatku teksta dobiti dojam o kakvim se podacima radi ispisati ćemo transformirane podatke za oba modela.

$\frac{1}{X}$	Y
0.009090909	235
0.009090909	198
0.009090909	173
0.004347826	174
0.004347826	149
0.004347826	124
0.002777778	115
0.002777778	130
0.002777778	102
0.002777778	95
0.001980198	122
0.001980198	112
0.001980198	98
0.001980198	96

Tablica 3: podaci za model (5)

lnX	lnY
4.700480	5.459586
4.700480	5.288267
4.700480	5.153292
5.438079	5.159055
5.438079	5.003946
5.438079	4.820282
5.886104	4.744932
5.886104	4.867534
5.886104	4.624973
5.886104	4.553877
6.224558	4.804021
6.224558	4.718499
6.224558	4.584967
6.224558	4.564348

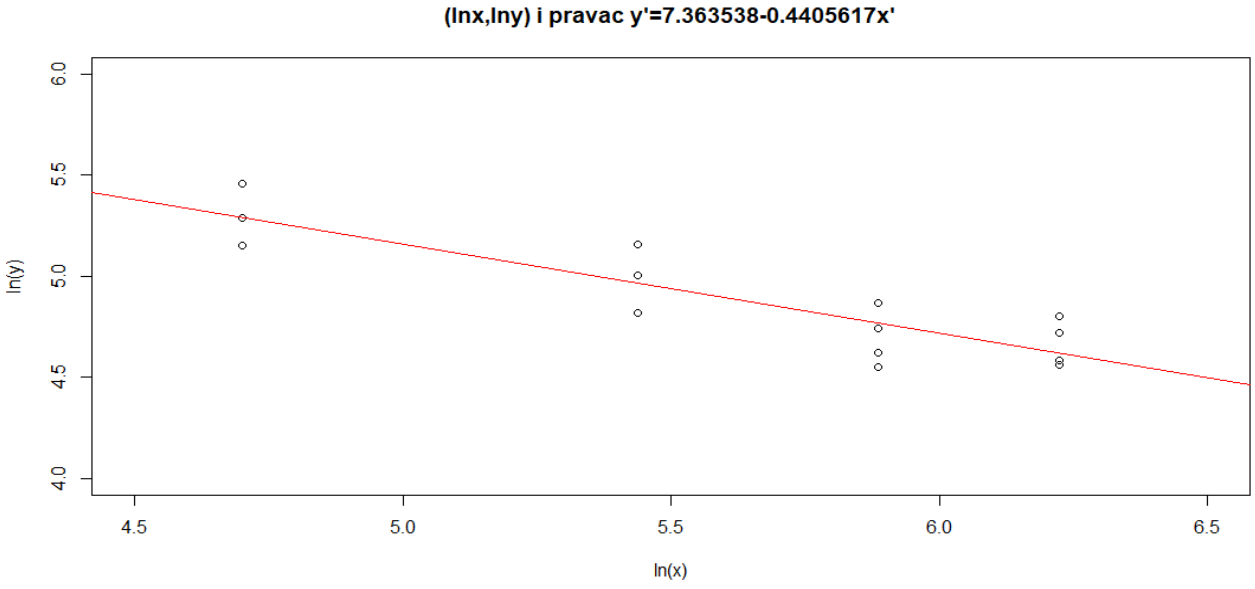
Tablica 4: podaci za model (2)

Promatramo linearni model $y' = \theta_0 + \theta_1 x' + \varepsilon$, gdje je ε slučajna pogreška, a θ_0 i θ_1 parametri modela. Pretpostavljamo da vrijede Gauss – Markovljevi uvjeti:
(i) $\mathbb{E}[\varepsilon_i] = 0 \ \forall i = 1, 2, \ldots, n$
(ii) $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0 \ \forall i, j = 1, 2, \ldots, n$ takve da je $i \neq j$
(iii) $\mathrm{Var}[\varepsilon_i] = \sigma^2 > 0 \ \forall i = 1, 2, \ldots, n$
Pravac prilagođavamo metodom najmanjih kvadrata, tj. želimo minimizirati funkciju $L(\theta_0, \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$. Stavimo:

$$\mathrm{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

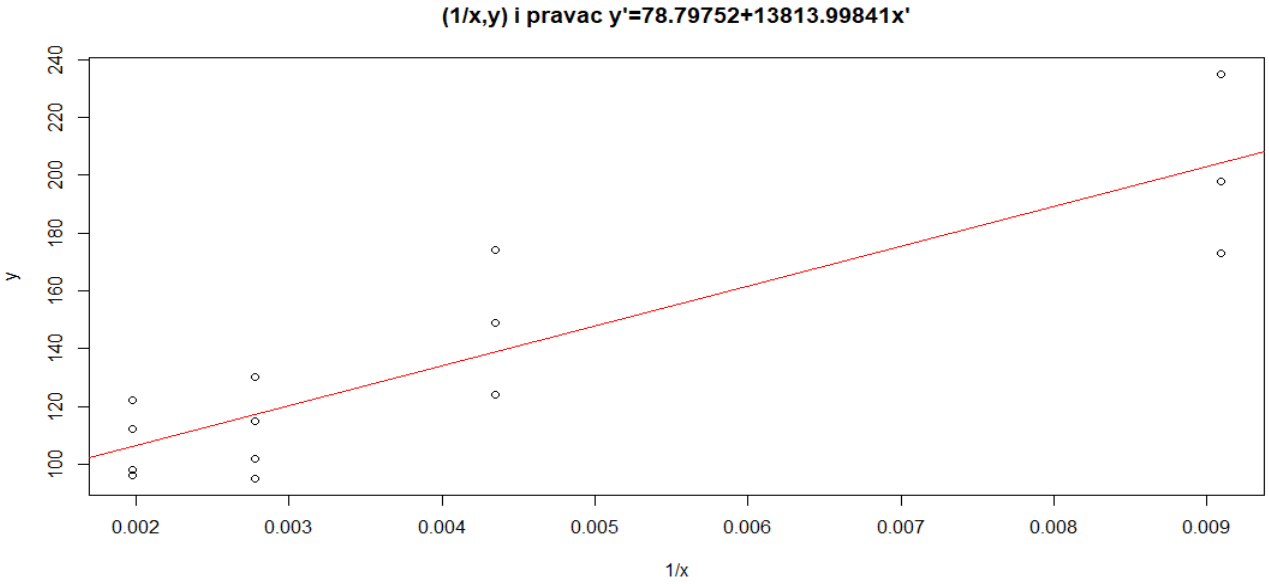
Nepristrani procjenitelj za θ metodom najmanjih kvadrata je $\hat{\theta} = (X^T X)^{-1} X^T Y$ uz procjenu $(X^T X)^{-1} X^T y$, a procjenitelji za Yi su $\hat{Y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$, $i = 1, \ldots, n$.

Tim postupkom za model (2) dobijemo procjenitelj za θ da je
$\hat{\theta} = \begin{bmatrix} 7.3635382 \\ -0.4405617 \end{bmatrix}$, pa je regresijski pravac za model (2): $y' = 7.3635382 - 0.4405617x'$, a na Slici 2 možemo vidjeti prikaz dobivenog pravca na istom grafu s transformiranim podacima iz modela (2).



Slika 2: prikaz podataka *(lnx, lny)* i pripadni regresijski pravac $y' = 7.3635382 - 0.4405617x'$

Na isti način za model (5) dobijemo procjenitelj za θ da je
$\hat{\theta} = \begin{bmatrix} 78.79752 \\ 13813.99841 \end{bmatrix}$, pa je regresijski pravac za model (5): $y' = 78.79752 + 13813.99841x'$, a na Slici 3 možemo vidjeti prikaz dobivenog pravca na istom grafu s transformiranim podacima iz modela (5).



Slika 3: prikaz podataka $(\frac{1}{x}, y)$ i pripadni regresijski pravac $y' = 78.79752 + 13813.99841x'$

Kako bi izračunali statistiku R^2 za oba modela, trebati će nam njihovi pripadni reziduali. Reziduali su slučajne varijable $E_i = Y_i - \hat{Y}_i$ a njihove realizacije označavamo e_i , $i = 1,\ldots, n$. Stavimo:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad \text{suma kvadrata reziduala}$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Koeficijent determinacije je $R^2 \cdot 100\%$, gdje je $R^2 = 1 - \frac{SSE}{S_{YY}} \in [0,1]$. Što je R^2 bliži 1, to je prilagodba modelu bolja.

Računajući gore navedenim postupkom dobijemo ove rezultate:

	SSE	S_{YY}
(2)	0.2324698	1.082366
(5)	4867.359	23875.21

Tablica 5: rezultati *SSE* i S_{YY} za modele

Tada je za model (2) statistika $R^2 = 0.7852206$, tj. koeficijent determinacije je 78.52206%. Dakle, 78.52206% ukupnog rasipanja dolazi od danog modela, a ostalo potječe od slučajnih pogrešaka pa možemo reći da je linearni model dobro prilagođen podacima, no moglo bi biti i bolje. Također se iz Slike 2 vidi da se dosta točaka nalaze dalje od pravca.
Za model (5) statistika $R^2 = 0.7961334$, tj. koeficijent determinacije je 79.61334%. Analogno kao za model (2), 79.61334% ukupnog rasipanja dolazi od danog modela, a ostalo potječe od slučajnih pogrešaka, pa možemo reći da je linearni model dobro prilagođen podacima, no moglo bi biti i bolje. Također, slično kao kod modela (2), iz Slike 3 vidimo da se dosta točaka nalazi dalje od pravca, no ipak je bolji od modela (2) jer je vrijednost statistike R^2 bliža 1, nego kod modela (2).

Sada ćemo provesti test adekvatnosti za oba modela.
Testiramo:

H_0: linearan model je adekvatan

H_1: ne H_0

Među vrijednostima $x_1, \ldots, x_n$ se nalazi točno l različitih vrijednosti. Označit ćemo ih sa $z_1, \ldots, z_l$. Za svaki z_i se Y mjeri n_i puta. Neka su y_{ij} dobivene realizacije, a Y_{ij} odgovarajuće slučajne varijable, za $j = 1, \ldots, n_i$, $i = 1, \ldots, l$.
Definiramo $SSE_{pe} = \sum_{i=1}^{l} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{n_i})^2$ (*pe=pure error*), a testna statistika je:

$$F = \frac{(SSE - SSE_{pe})/(l-k-1)}{SSE_{pe}/(n-l)} \sim F(l-k-1, n-l)$$

pri čemu je $l = 4$, $k = 1$ i $n = 14$ za oba modela.

Za model (2) dobijemo $SSE_{pe} = 0.2003867$, a realizacija statistike F na podacima iz Tablice 4 je $f = 0.8005306$ pa je stoga p-vrijednost

$$p\text{-vrijednost} = \mathbb{P}(F > f \mid H_0) = 0.4758953$$

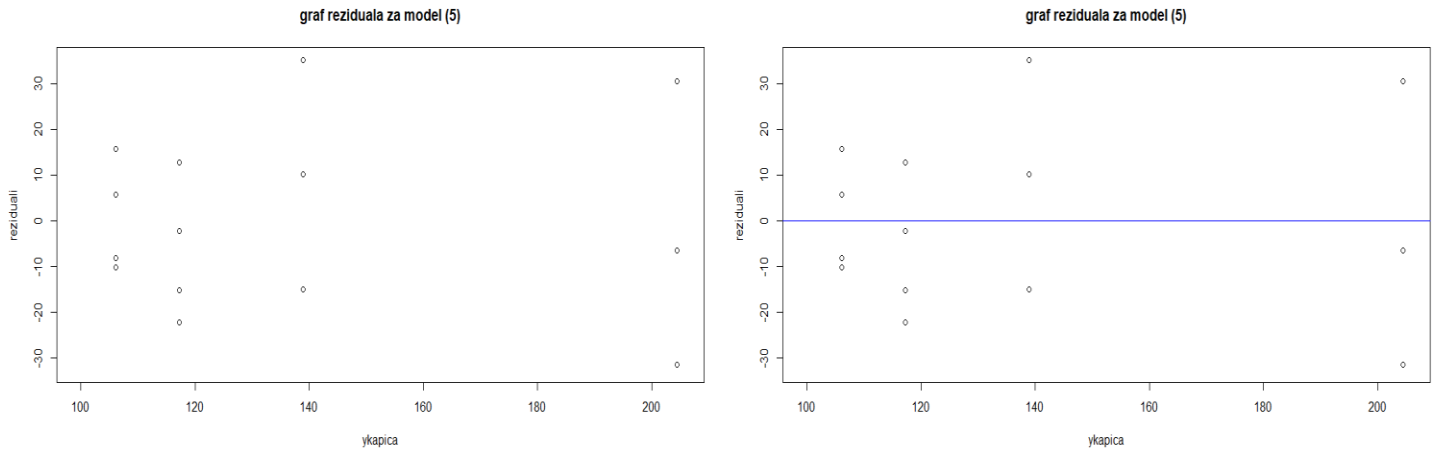
stoga nećemo odbaciti hipotezu H_0 niti za jednu $standardnu$ razinu značajnosti $\alpha = 0.01, 0.05, 0.1$.

Za model (5) dobijemo $SSE_{pe} = 4361$, a realizacija statistike F na podacima iz Tablice 3 je $f = 0.5805537$ pa je stoga p-vrijednost

$$p\text{-vrijednost} = \mathbb{P}(F > f \mid H_0) = 0.5773824$$

stoga nećemo odbaciti hipotezu H_0 niti za jednu $standardnu$ razinu značajnosti $\alpha = 0.01, 0.05, 0.1$.

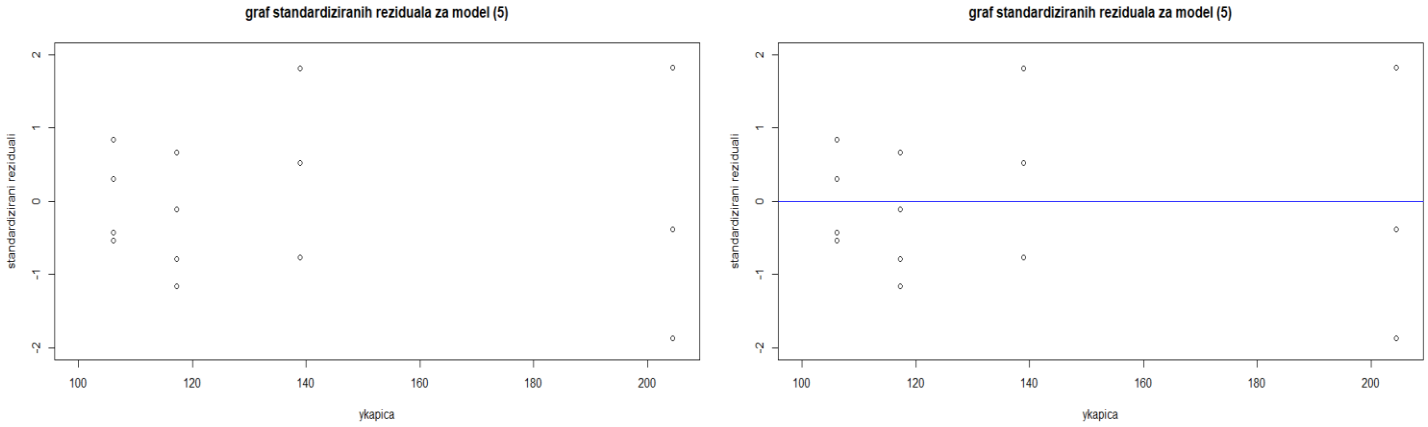
Možemo pretpostaviti na osnovi koeficijenta determinacije i testa adekvatnosti da je model (5) bolji od modela (2), te ćemo za njega prikazati graf reziduala.



Slika 4: graf reziduala za model (5)

Također ćemo prikazati graf standardiziranih reziduala za model (5), no kako bi ih mogli izračunati moramo dodatno pretpostaviti za slučajne pogreške
(iv) $\varepsilon \sim N(0, \sigma^2 I)$.

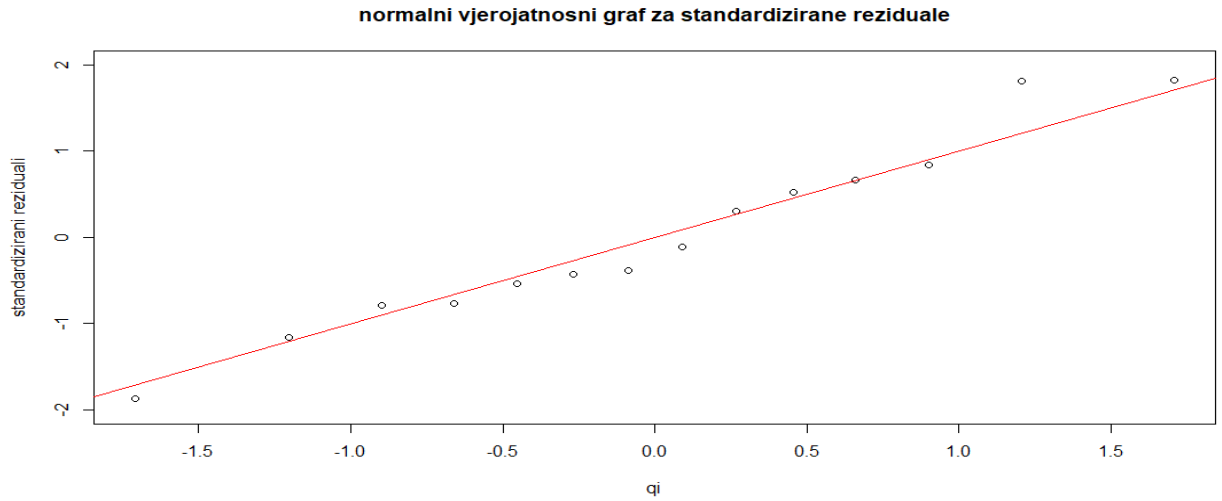
Standardizirani reziduali su slučajne varijable $E_i^S = \frac{E_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ gdje je $H = X(X^TX)^{-1}X^T = [h_{ij}]$, a njihove realizacije označavamo e_i^S, za $i = 1,\ldots,n$. Nepristrani procjenitelj za σ^2 je $\hat{\sigma}^2 = \frac{SSE}{n-k-1}$, gdje je $k = 1$ i $n = 14$.



Slika 5: graf standardiziranih reziduala za model (5)

S obzirom na graf reziduala sa Slike 4 i graf standardiziranih reziduala sa Slike 5 možemo zaključiti da je prilagodba linearnom modelu zadovoljavajuća jer su točke slučajno grupirane oko x-osi.

Želimo provjeriti dolaze li standardizirani reziduali iz jedinične normalne distribucije *N(0,1)* što ćemo testirati pomoću grafa normalnih vjerojatnosti. Neka su $y_{(1)}, \ldots, y_{(n)}$ sortirane realizacije standardiziranih reziduala modela (5). Pomoću funkcije distribucije jedinične normalne razdiobe ϕ definiramo $q_i = \phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$, $i = 1, \ldots, n$. Točke $(q_i, y_{(i)})$, $i = 1, \ldots, n$ u normalnom vjerojatnosnom grafu moraju biti aproksimativno na pravu $y = x$, što se i vidi iz Slike 6.



Slika 6: normalni vjerojatnosni graf

Odredimo sada 95% pouzdani interval za parametre θ_0 i θ_1 za model (5). Pouzdane intervale za θ_j konstruiramo koristeći

$$\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k-1), \text{ gdje je } C = (X^T X)^{-1} = [c_{ij}], \text{ za } j = 1, \dots, k$$

gdje je $k = 1$ i $n = 14$

Dakle, traženi interval pouzdanosti dan je s

$$-t_{\frac{\alpha}{2}} \leq \frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{c_{jj}}} \leq t_{\frac{\alpha}{2}}$$

gdje je $\alpha = 0.05$, a $t_{\frac{\alpha}{2}}$ 97.5% kvantil t distribucije s ($n - k - 1$) stupnjem slobode

($t_{\frac{\alpha}{2}} = 2.178813$). Rješavanjem ovih nejednadžbi po θ_j dobivamo da je traženi 95% interval pouzdanosti za $j = 0, 1$

$$\theta_j \in \left[\hat{\theta}_j - t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{c_{jj}},\ \hat{\theta}_j + t_{\frac{\alpha}{2}}\hat{\sigma}\sqrt{c_{jj}}\right].$$

Uvrštavajući već prije dobivene rezultate dobili smo 95% pouzdani intervale za parametre θ_0 i θ_1 za model (5) su:

$$\theta_0 \in [56.77647,\ 100.8186] \qquad \theta_1 \in [9417.27818,\ 18210.7186]$$

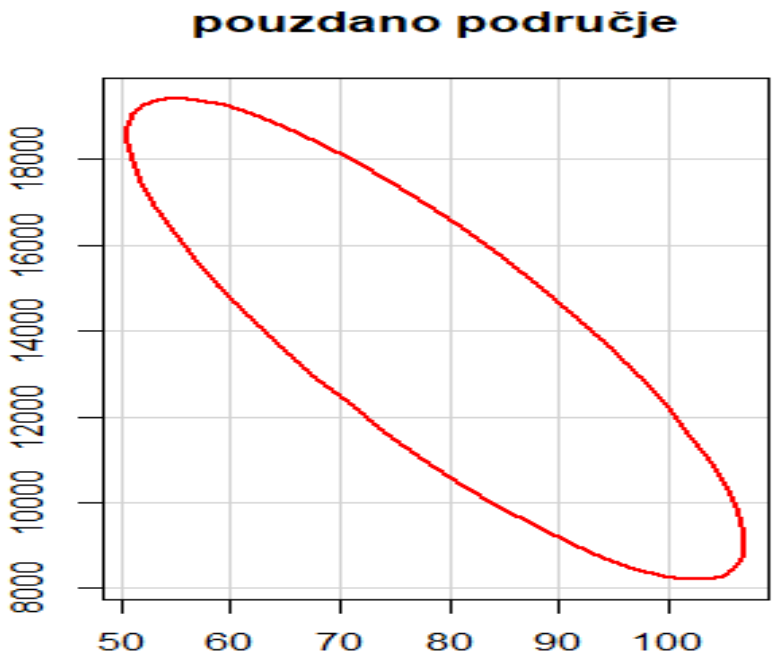
Želimo naći 95% pouzdano područje za $\theta = (\theta_0, \theta_1)$ za model (5). Pouzdano područje za $\theta = (\theta_0, \theta_1)$ konstruiramo koristeći

$$F = \frac{1}{(k+1)\cdot\hat{\sigma}^2}\left((X^T X)\left(\theta - \hat{\theta}\right), \left(\theta - \hat{\theta}\right)\right) \sim F(k+1, n-k-1)$$

gdje je $k = 1$ i $n = 14$.

Zbog $\mathbb{P}(F \leq f_\alpha) = 1 - \alpha$ se iz $F - f_\alpha \leq 0$ dobije $(1 - \alpha) \cdot 100\%$ pouzdano područje za θ, a kako je $k = 1$ to će biti unutrašnjost elipse. Rješavajući nejednadžbu za $\alpha = 0.05$, za koji $je\ f_\alpha = 3.885294$, dobijemo

$$14\theta_0^2 + 0.1186962\theta_0\theta_1 + 0.00035119936\theta_1^2 - 3846\theta_0 - 19.05574\theta_1 + 279993.7 \leq 0.$$



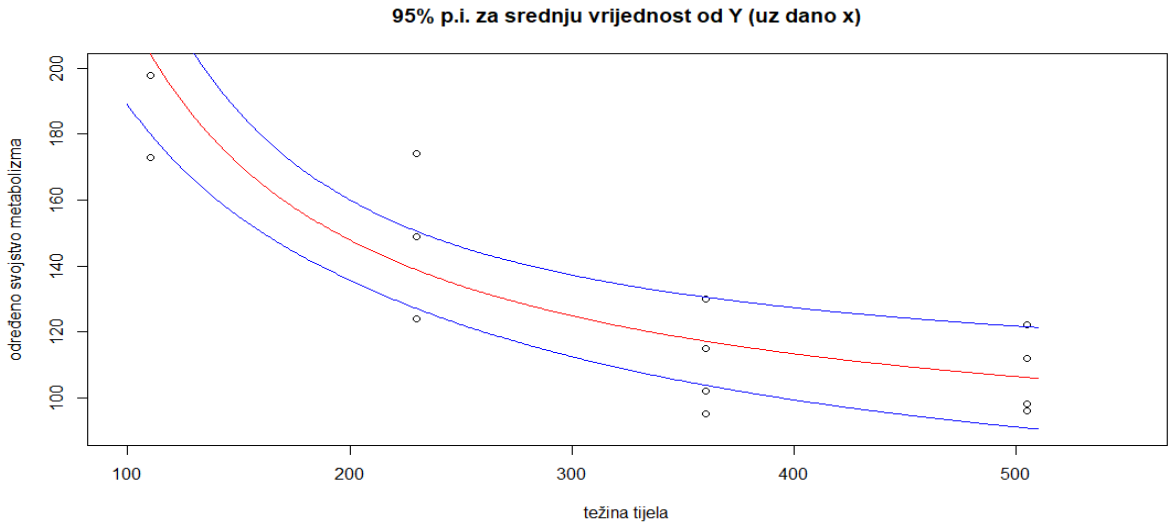
Slika 7: pouzdano područje za $\theta = (\theta_0, \theta_1)$ modela (5)

Model (regresijska funkcija) za originalne podatke *(x, y)* je $y = 78.79752 + 13813.99841\frac{1}{x}$. Prikazat ćemo originalne podatke zajedno s regresijskom funkcijom i krivuljama koje definiraju gornje i donje 95% pouzdane intervale za srednju vrijednost Y (uz dano x), no prvo trebamo naći te pouzdane intervale. Za zadani $x_0 = (x_0^{(1)}, x_0^{(2)}, \dots, x_0^{(k)})$, gdje je *k* = 1, stavimo $z = [1 \quad x_0^{(1)} \quad x_0^{(2)} \dots \quad x_0^{(k)}]$, tj. u našem slučaju $z = [1 \quad \frac{1}{x_0}]$. Za $\alpha = 0.05$, (1 – α) · 100% pouzdani interval za $\mathbb{E}$ [Y | $\frac{1}{X}$ = x_0] uz dano $\frac{1}{x}$ = x_0 je

$$\hat{\theta}_0 + \hat{\theta}_1 \frac{1}{x_0} \pm t_{\frac{\alpha}{2}}(n-k-1)\hat{\sigma}\sqrt{z(X^TX)^{-1}z^T}$$

gdje je $\hat{\theta} = \begin{bmatrix} 78.79752 \\ 13813.99841 \end{bmatrix}$, $t_{\frac{\alpha}{2}}(n-k-1) = 2.178813$, X = $\begin{bmatrix} 1 & \frac{1}{x_1} \\ 1 & \frac{1}{x_2} \\ \vdots & \vdots \\ 1 & \frac{1}{x_n} \end{bmatrix}$ i $\hat{\sigma}^2 = \frac{SSE}{n-k-1}$, gdje je *k* = 1,

n = 14 i *SSE* iz Tablice 5 za model (5).



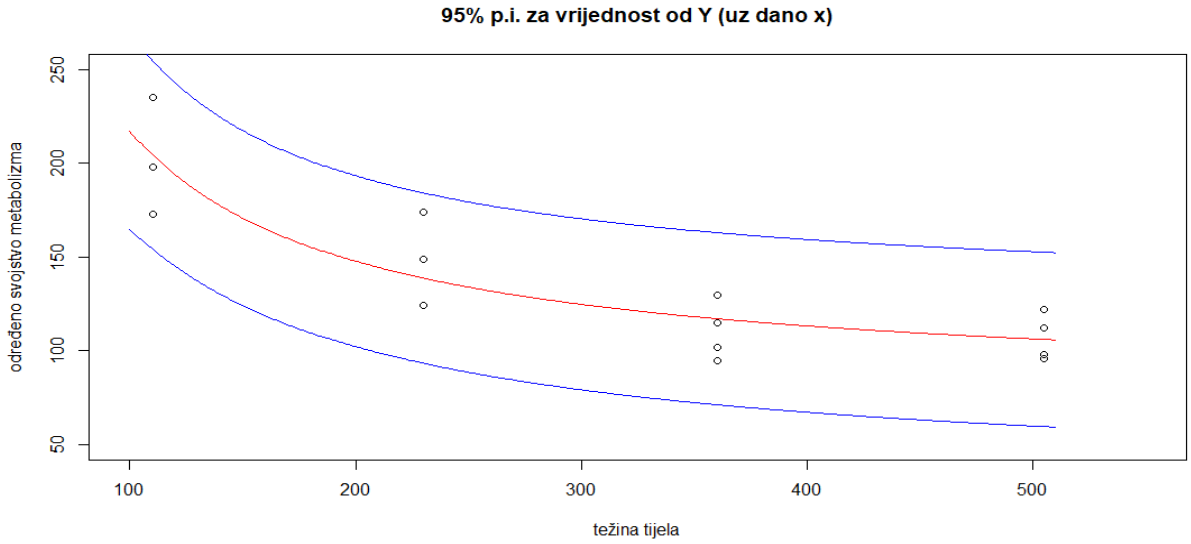
Slika 8: 95% pouzdani interval za srednju vrijednost od Y (uz dano x) uz regresijsku funkciju za *(x, y)* i podatke iz Tablice 1

Analogno ćemo napraviti za vrijednost od Y (uz dano x). Za $\alpha = 0.05$, $(1-\alpha)\cdot 100\%$ pouzdani interval za Y uz dano $\frac{1}{x} = x_0$ je

$$\hat{\theta}_0 + \hat{\theta}_1 \frac{1}{x_0} \pm t_{\frac{\alpha}{2}}(n-k-1)\hat{\sigma}\sqrt{1+z(X^TX)^{-1}z^T}$$

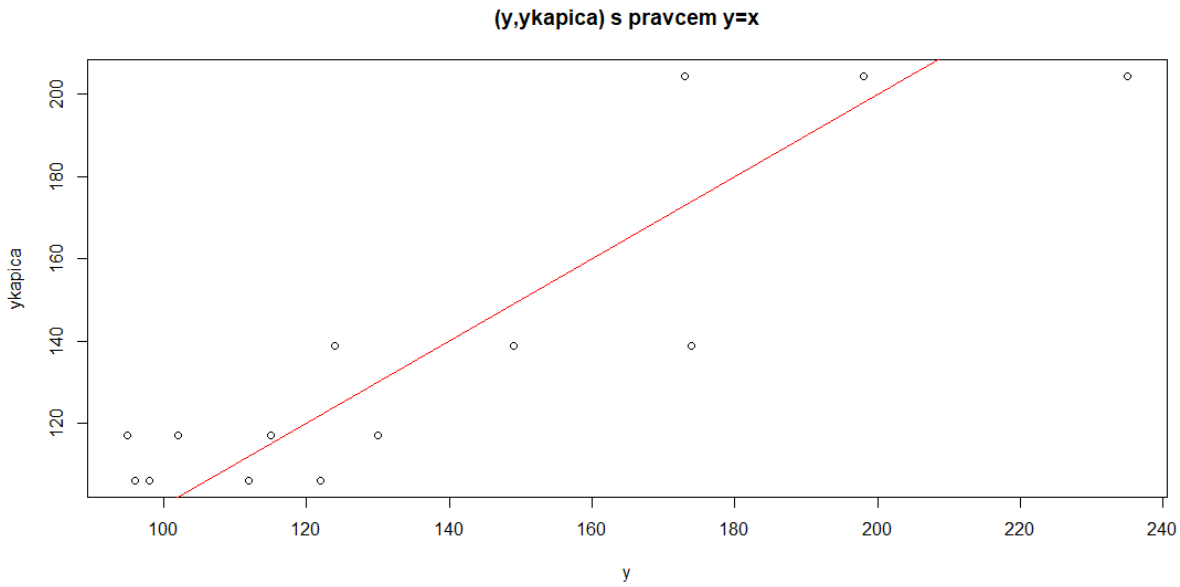
gdje je $\hat{\theta} = \begin{bmatrix} 78.79752 \\ 13813.99841 \end{bmatrix}$, $t_{\frac{\alpha}{2}}(n-k-1) = 2.178813$, $X = \begin{bmatrix} 1 & \frac{1}{x_1} \\ 1 & \frac{1}{x_2} \\ \vdots & \vdots \\ 1 & \frac{1}{x_n} \end{bmatrix}$ i $\hat{\sigma}^2 = \frac{SSE}{n-k-1}$, gdje je $k = 1$,

$n = 14$ i *SSE* iz Tablice 5 za model (5).



Slika 9: 95% pouzdani interval za vrijednost od Y (uz dano x) uz regresijsku funkciju za *(x, y)* i podatke iz Tablice 1

Još ćemo prikazati točke $(y, \hat{y})$ zajedno s pravcem $y = x$, gdje je $\hat{y}$ procjena od y na osnovu modela za originalne podatke.



Slika 10: prikaz točaka $(y, \hat{y})$ zajedno s pravcem $y = x$

Možemo vidjeti iz Slike 10 da je model dobar jer su podaci otprilike simetrično oko dijagonale, te se vidi da problem stvaraju ponovljena mjerenja za isti x. No, testiranjem smo pokazali da se transformacijom podataka kao u modelu (5) dobije dobar model, te je on ipak bolji.