

**»Mladi za napredek Maribora 2019«
36. srečanje**

**Kriterij za določitev regresijske premice po metodi najmanjše vsote
ploščin pravokotnih trikotnikov**

Raziskovalno področje: Matematika

Raziskovalna naloga

PROSTOR ZA NALEPKO

Avtor: VAN DEČKO

Mentor: DAVID GAJSER
Šola: II. GIMNAZIJA MARIBOR
Število točk: 165
Mesto: 1
Priznanje: zlato

Maribor, februar 2019

**»Mladi za napredek Maribora 2019«
36. srečanje**

**Kriterij za določitev regresijske premice po metodi najmanjše vsote
ploščin pravokotnih trikotnikov**

Raziskovalno področje: Matematika

Raziskovalna naloga

PROSTOR ZA NALEPKO

Maribor, februar 2019

Kazalo vsebine

Povzetek	5
Zahvala	6
1. UVOD	7
2. METODA NAJMANJŠIH KVADRATOV	8
2.1. Izpeljava metode najmanjših kvadratov	9
2.2. Parcialni odvod	11
3. METODA NAJMANJŠE VSOTE PLOŠČIN PRAVOKOTNIH TRIKOTNIKOV	14
3.1. Izpeljava metode najmanjše vsote ploščin pravokotnih trikotnikov	16
4. ANALIZA OBEH REGRESIJSKIH PREMICE	20
5. PRAKTIČNI PRIMERI	21
5.1. Izbira podatkov (točk) za izračun regresijskih premic	22
5.3. Normalizacija podatkov	25
5.4. Regresijska premica po metodi najmanjših kvadratov glede na neodvisno spremenljivko	28
5.5. Simetrala kota	30
5.6. Uporaba inverzne funkcije	31
6. ZAKLJUČEK	32
7. DRUŽBENA ODGOVORNOST	33
8. VIRI IN LITERATURA	34

Kazalo slik

Slika 1: Metoda najmanjših kvadratov	8
Slika 2: Ploščina pravokotnega trikotnika.....	15
Slika 3: Pravokotni trikotniki in dolžine njihovih katet	16
Slika 4: Razsevni diagram v programu Logger Pro.....	22
Slika 5: Regresijski premici, ko je med spremenljivkama $r = 0,962$	23
Slika 6: Regresijski premici, ko je med spremenljivkama $r = 0,792$	24
Slika 7: Regresijski premici, ko je $r = 0,792$ in sem vsem podatkom dodal vrednost 20	25
Slika 8: Regresijskih premici za standardizirani spremenljivki.....	27
Slika 9: Regresijske premice na razsevnem diagramu	29

Povzetek

Ko sestavimo raziskavo in v njej merimo 2 statistični spremenljivki, med katerima iščemo linearno odvisnost, lahko meritve predstavimo z razsevnim diagramom in regresijsko premico. Standardna metoda za določitev regresijske premice je metoda najmanjših kvadratov. Iznášel sem novo metodo za določitev regresijske premice, metodo najmanjše vsote ploščin pravokotnih trikotnikov, ki imajo hipotenuzo na regresijski premici, oglišče nasproti hipotenuze, v točki razsevnega diagrama in kateti, vzporedni z osema koordinatnega sistema. Ta regresijska premica je drugačna od tiste, ki jo dobimo po standardni metodi t. j. metodi najmanjših kvadratov.

V tej raziskovalni nalogi bom izpeljal metodo najmanjših kvadratov in nato po analognem postopku tudi metodo najmanjše vsote ploščin pravokotnih trikotnikov. Nato bom primerjal obe enačbi regresijskih premic na podlagi praktičnih primerov in pojasnil prednosti in pomanjkljivosti obeh metod.

Število znakov (s presledki): 935

Ključne besede: regresijska premica, metoda najmanjših kvadratov, metoda najmanjše vsote ploščin pravokotnih trikotnikov

Zahvala

Rad bi se zahvalil svojemu mentorju, ki me je prepričal, da svojo idejo predstavim v obliki raziskovalne naloge in da me je pri tem podpiral.

1. UVOD

Med poukom matematike v programu mednarodne mature smo med obravnavanjem statistike kot ene izmed vej matematike spoznali tudi standardni kriterij za določitev regresijske premice, t. j. metoda najmanjših kvadratov. Ko nas je profesor med uro pozval, naj sami dodamo ideje, kako mislimo, da je regresijska premica definirana, me je ošinila preprosta, a vendar zanimiva ideja: lahko bi skušali minimalizirat seštevek ploščin pravokotnih trikotnikov na regresijsko premico, ki imajo hipotenuzo na regresijski premici, oglišče nasproti hipotenuze, v točki razsevnega diagrama in kateti, vzporedni z osema koordinatnega sistema. Profesorju in mentorju se je ideja zdela precej neznana, a vendar sem mu s pomočjo skice uspel razložiti način lastne metode. Za to idejo ni še nikoli slišal, a se je strinjal, da bi metoda lahko delovala, vendar da bi bili rezultati drugačni kot pri metodi najmanjših kvadratov.

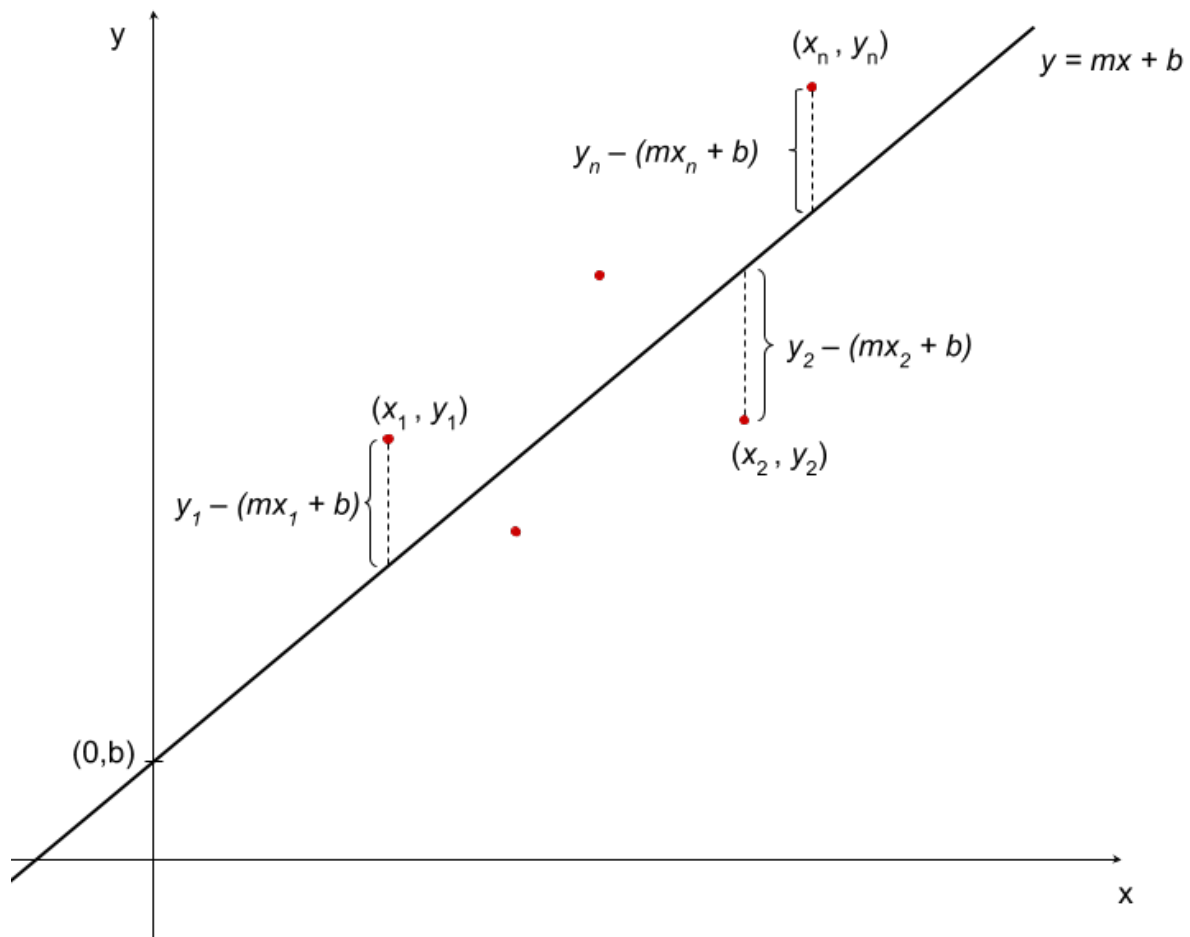
Izpeljava moje ideje, torej metode najmanjše vsote ploščin pravokotnih trikotnikov na regresijsko premico se mi je zdela odlična za raziskovalno nalogo, kjer bi lahko dobljene rezultate primerjal s standardno metodo najmanjših kvadratov.

Ko sem doma brskal po spletu, da bi našel kakršnekoli podatke o moji metodi, nisem našel ničesar. To me je zelo presenetilo, saj je statistična obdelava podatkov in iskanje regresijske premice zelo pomembna pri večini naravoslovnih predmetov, kot so biologija, kemija in fizika, saj nam omogoča nadaljnjo računanje in interpretacijo podatkov. Začel sem se spraševati, zakaj sem našel tako malo metod za določitev regresijske premice? Kako vemo, katera metoda je najbolj učinkovita?

V tej raziskovalni nalogi bom raziskal, kako se izpelje standardna enačba regresijske premice, t. j. po metodi najmanjših kvadratov. Nato bom po analognem postopku izpeljal enačbo po lastni, t. j. po metodi najmanjše vsote ploščin pravokotnih trikotnikov. Na koncu bom primerjal med obema enačbama, tako teoretično kot na praktičnih primerih z uporabo računalniškega programa za obdelavo podatkov, imenovanega Logger Pro.

2. METODA NAJMANJŠIH KVADRATOV

Standardna metoda za določitev regresijske premice, imenovana metoda najmanjših kvadratov, se uporablja v večini programov informacijske tehnologije za obdelavo podatkov in velja kot standardna metoda za določitev regresijske premice med podatki. Ideja te metode je minimalizirati vertikalne razdalje od točke do regresijske premice. Vse razdalje so kvadrirane, saj predznačene razdalje med premico in točkami, ki se nahajajo pod premico, dajejo negativne vrednosti (glej Slika 1). S kvadriranjem vseh razdalj se znebimo negativnih vrednosti.



Slika 1: Metoda najmanjših kvadratov. Pomnimo, da je vrednost predznačenih razdalj med regresijsko premico in točko pod regresijsko premico negativna in če se točka nahaja nad regresijsko premico, potem je pozitivna. Vir: avtor

2.1. Izpeljava metode najmanjših kvadratov

Izpeljavo metode najmanjših kvadratov sem povzel iz spletne učilnice z razlagami o različnih matematičnih problemih in definicijah (Khan, 2019).

Prevedimo, kar smo opisali z besedami, v matematični jezik. Enačba regresijske premice je

$$y = mx + b .$$

Imamo točke $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ in želimo minimalizirati vsoto kvadratov navpičnih razdalj od vsake točke do regresijske premice. Vrednosti razdalj vidimo na *Sliki 1*. Če je na grafu n točk, je

$$SD_{line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots + (y_n - (mx_n + b))^2 \quad [1]$$

vsota kvadratov navpičnih razdalj od točk do regresijske premice. Ker poznamo koordinate vsake točke, želimo najti vrednosti m (koeficient regresijske premice) in b (začetna vrednost regresijske premice), ki minimalizirata vsoto kvadratov razdalj na regresijsko premico (SD_{line}). Da se dokopamo do rezultata (da bomo določili m in b), ki ga bomo obravnavali z odvodi, moramo algebraično manipulirati z izrazom. Najprej lahko odpravimo oklepaje

$$\begin{aligned} SD_{line} &= y_1^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 \\ &\quad + y_2^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 \\ &\quad \vdots \\ &\quad + y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2 \\ \\ &= y_1^2 - 2y_1mx_1 - 2y_1b + m^2x_1^2 + 2mx_1b + b^2 \\ &\quad + y_2^2 - 2y_2mx_2 - 2y_2b + m^2x_2^2 + 2mx_2b + b^2 \\ &\quad \vdots \\ &\quad + y_n^2 - 2y_nmx_n - 2y_nb + m^2x_n^2 + 2mx_nb + b^2 . \end{aligned}$$

Iz razširjenega izraza lahko seštejemo podobne dele enačbe in izpostavimo njihov skupen faktor

$$SD_{line} = (y_1^2 + y_2^2 + \dots + y_n^2) - 2m(x_1y_1 + x_2y_2 + \dots + x_ny_n) - 2b(y_1 + y_2 + \dots + y_n) + m^2(x_1^2 + x_2^2 + \dots + x_n^2) + 2mb(x_1 + x_2 + \dots + x_n) + nb^2.$$

V nadaljevanju lahko poenostavimo naslednje dele enačbe tako, da uvedemo oznako za povprečje (črta nad spremenljivko).

$$\frac{(y_1^2 + y_2^2 + \dots + y_n^2)}{n} = \overline{y^2} \Rightarrow y_1^2 + y_2^2 + \dots + y_n^2 = n\overline{y^2}$$

$$\frac{(x_1y_1 + x_2y_2 + \dots + x_ny_n)}{n} = \overline{xy} \Rightarrow x_1y_1 + x_2y_2 + \dots + x_ny_n = n\overline{xy}$$

$$\frac{(y_1 + y_2 + \dots + y_n)}{n} = \overline{y} \Rightarrow y_1 + y_2 + \dots + y_n = n\overline{y}$$

$$\frac{(x_1^2 + x_2^2 + \dots + x_n^2)}{n} = \overline{x^2} \Rightarrow x_1^2 + x_2^2 + \dots + x_n^2 = n\overline{x^2}$$

$$\frac{(x_1 + x_2 + \dots + x_n)}{n} = \overline{x} \Rightarrow x_1 + x_2 + \dots + x_n = n\overline{x}.$$

Nova enačba je sedaj

$$SD_{line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\overline{y} + m^2n\overline{x^2} + 2mbn\overline{x} + nb^2. \quad [2]$$

2.2. Parcialni odvod

Iz poenostavljene enačbe želimo minimalizirati vrednost vsote kvadriranih navpičnih razdalj med točkami in premico. Predpostavimo, da poznamo povprečni vrednosti od vseh x -ov in y -ov, vse kar moremo najti, je vrednosti m in b , ki bosta minimalizirali vsoto vseh kvadriranih razdalj. Da ju dobimo, moremo vpeljati uporabo parcialnih odvodov v prej izpeljano enačbo [2].

Parcialni odvodi delujejo na enak način kot običajni odvodi, ampak odvajajo funkcijo samo po določni spremenljivki, ostale spremenljivke pa delujejo kot konstante (Wikipedia, 2019). Najti je potrebno parcialni odvod SD_{line} v odvisnosti od m in parcialni odvod SD_{line} v odvisnosti od b . Oba odvoda moramo enačiti z nič (želimo najti stacionarno točko teh dveh odvodov oz. minimum po spremenljivki m in minimum po spremenljivki b).

$$\frac{\partial SD_{line}}{\partial m} = 0 \wedge \frac{\partial SD_{line}}{\partial b} = 0 .$$

Dobimo

$$\begin{aligned} \frac{\partial}{\partial m} \left(n\overline{y^2} - 2mn\overline{xy} - 2bn\overline{y} + m^2n\overline{x^2} + 2mbn\overline{x} + nb^2 \right) = \\ -2n\overline{xy} + 2mn\overline{x^2} + 2bn\overline{x} \end{aligned}$$

in

$$\begin{aligned} \frac{\partial}{\partial b} \left(n\overline{y^2} - 2mn\overline{xy} - 2bn\overline{y} + m^2n\overline{x^2} + 2mbn\overline{x} + nb^2 \right) = \\ -2n\overline{y} + 2mn\overline{x} + 2nb . \end{aligned}$$

Ker želimo za oba izraza najti stacionarno točko, ju enačimo z 0 in dobimo dve ločeni enačbi

$$-2n\overline{xy} + 2mn\overline{x^2} + 2bn\overline{x} = 0 ,$$

$$-2n\overline{y} + 2mn\overline{x} + 2nb = 0 .$$

Obe enačbi lahko delimo z $2n$

$$-\overline{xy} + m\overline{x^2} + b\overline{x} = 0 ,$$

$$-\overline{y} + m\overline{x} + b = 0 .$$

S preureditvijo členov dobimo

$$m\overline{x^2} + b\overline{x} = \overline{xy} ,$$

$$m\overline{x} + b = \overline{y} .$$

Kot zanimivost naj poudarim, da točko $A(\overline{x}, \overline{y})$ najdemo na dobljeni regresijski premici (glej 2. enačbo). Prav tako, če prvo enačbo delimo z $\overline{x}$, dobimo še eno točko $B\left(\frac{\overline{x^2}}{\overline{x}}, \frac{\overline{xy}}{\overline{x}}\right)$, ki jo prav tako lahko najdemo na dobljeni regresijski premici. Pridobljeni enačbi lahko rešimo za m in b z uporabo sistema enačb.

$$m\frac{\overline{x^2}}{\overline{x}} + b = \frac{\overline{xy}}{\overline{x}} ,$$

$$m\overline{x} + b = \overline{y} .$$

Enačbi odštejemo in dobimo

$$m \left(\frac{\overline{x^2}}{\bar{x}} - \bar{x} \right) = \frac{\overline{xy}}{\bar{x}} - \bar{y}$$

$$m = \frac{\frac{\overline{xy}}{\bar{x}} - \bar{y}}{\frac{\overline{x^2}}{\bar{x}} - \bar{x}},$$

torej

$$m = \frac{\overline{xy} - \bar{y} \bar{x}}{\overline{x^2} - \bar{x}^2}.$$

To je enačba za izračun koeficienta naše regresijske premice. Lahko ga izrazimo v obliki s standardno deviacijo σ in korelacijskim koeficientom r , če vemo da

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\overline{x^2} - \bar{x}^2},$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}} = \sqrt{\overline{y^2} - \bar{y}^2},$$

in

$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right) = \frac{\overline{xy} - \bar{x} \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \sqrt{\overline{y^2} - \bar{y}^2}}.$$

(Wazir, 2012, str. 293, 313)

Dobimo torej

$$m = r \frac{\sigma_y}{\sigma_x}.$$

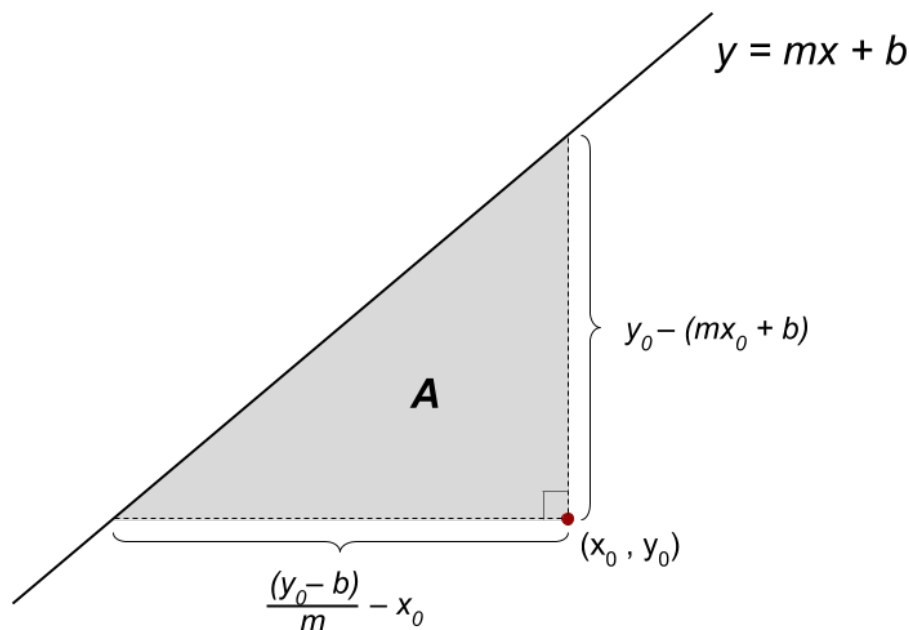
V tej obliki smo formulo za koeficient m izrazili s korelacijskim koeficientom med obema spremenljivkama, s standardno deviacijo po spremenljivki x in standardno deviacijo po spremenljivki y . Prav tako lahko vpeljemo m v drugo enačbo iz sistema enačb in z uporabo korelacijskega koeficienta in standardnega odklona dobimo enačbo za b :

$$b = \bar{y} - \bar{x}r \frac{\sigma_y}{\sigma_x}.$$

3. METODA NAJMANJŠE VSOTE PLOŠČIN PRAVOKOTNIH TRIKOTNIKOV

Moj kriterij za določitev regresijske premice temelji na najmanjši vsoti ploščin pravokotnih trikotnikov. Enačbo za pridobitev regresijske premice po moji metodi sem izpeljal po analognem načinu kot pri metodi najmanjših kvadratov.

Iz vsake točke v koordinatnem sistemu lahko narišemo dve daljici, ki se stikata v tej točki, drugo oglišče pa imata na regresijski premici. Ena od teh daljic je horizontalna, vzporedna z x -osjo, druga pa vertikalna, vzporedna z y -osjo. Tako dobimo pravokotni trikotnik, v katerem sta obe kateti daljici, ki povezujeta točko z regresijsko premico, hipotenuza pa je del regresijske premice. Nato lahko iz obeh katet izračunamo ploščino A pravokotnega trikotnika, kot je prikazano na *Sliki 2*.



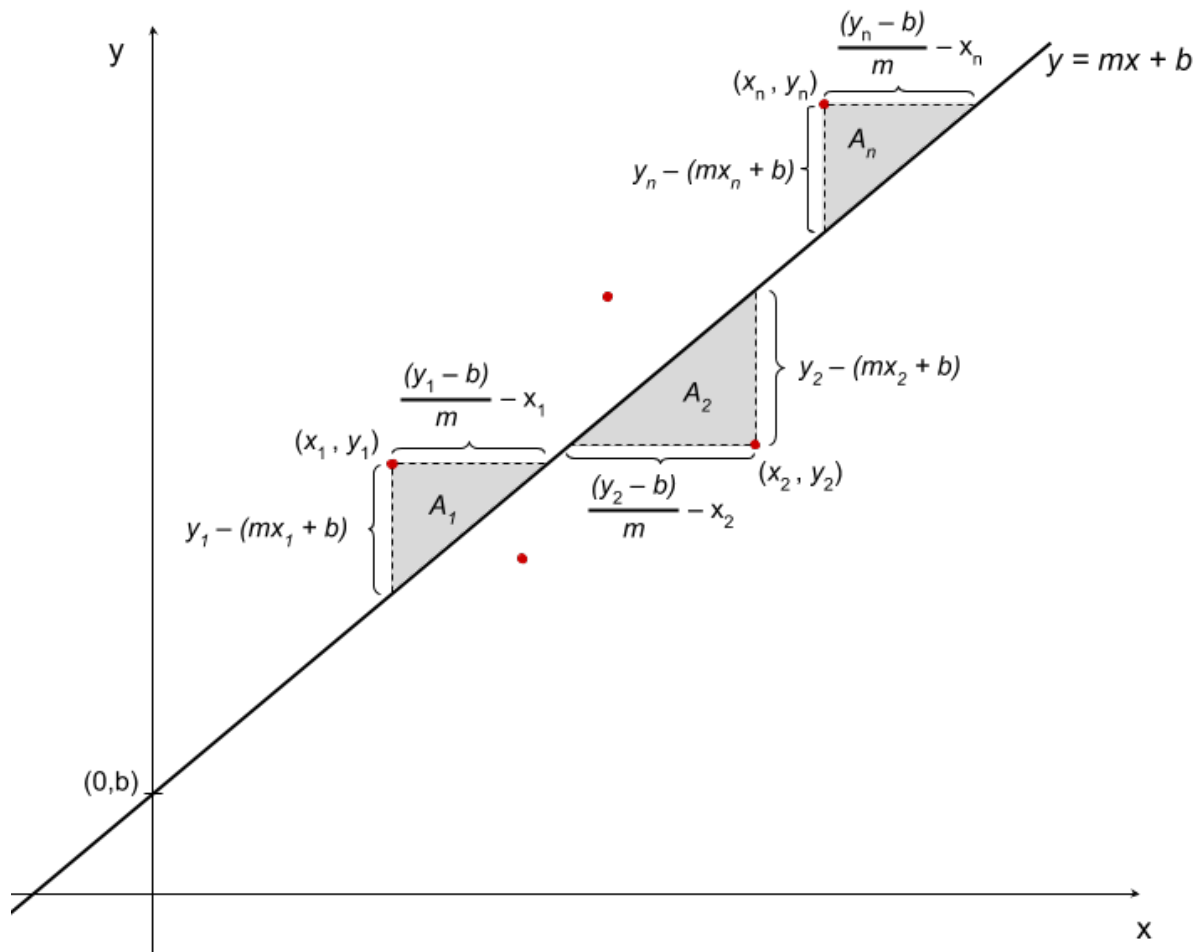
Slika 2: Ploščino pravokotnega trikotnika (A) lahko dobimo, če pomnožimo dolžini daljic, ki sta med seboj pravokotni, ena je vzporedna z x -osjo, druga pa z y -osjo koordinatnega sistema. Obe daljici imata skupno točko (x_0, y_0) in se dotikata regresijske premice. Ploščina dobljenega pravokotnega trikotnika je seveda vedno pozitivna, ne glede na to, ali se točka nahaja pod ali nad regresijsko premico. Vir: avtor

Ploščino pravokotnega trikotnika lahko izrazimo kot

$$A = \frac{(y_0 - (mx_0 + b)) \times \left(\frac{y_0 - b}{m} - x_0 \right)}{2}.$$

Ta izraz je zmeraj pozitiven, ne glede na to, ali se točka nahaja pod ali nad regresijsko premico. Če se točka nahaja pod regresijsko premico, sta oba faktorja negativna (glej Sliko 2), če se točka nahaja nad premico, pa sta oba faktorja pozitivna. Po enačbi za računanje ploščine pravokotnega trikotnika se obe kateti zmnožita in delita z 2 in če sta oba faktorja negativna, je rezultat pozitiven. Pri trikotniku, kateremu točka se nahaja nad regresijsko premico, bo rezultat za ploščino prav tako pozitiven saj bosta oba faktorja pozitivna.

Ideja moje metode je, da minimaliziramo vsoto ploščin vseh pravokotnih trikotnikov, ki jih pridobimo iz točk na razsevnem grafikonu. Če se pravokotni trikotniki prekrivajo, izračunamo ploščino vsakega trikotnika posebej (trikotniki so neodvisni drug od drugega).



Slika 3: Pravokotni trikotniki in dolžine njihovih katet. Vir: avtor

3.1. Izpeljava metode najmanjše vsote ploščin pravokotnih trikotnikov

Želomo minimalizirati vsoto ploščin vseh pravokotnih trikotnikov na regresijsko premico

$$AT = \frac{(y_1 - (mx_1 + b)) \times \left(\frac{y_1 - b}{m} - x_1\right)}{2} + \frac{(y_2 - (mx_2 + b)) \times \left(\frac{y_2 - b}{m} - x_2\right)}{2} + \dots$$

$$+ \frac{(y_n - (mx_n + b)) \times \left(\frac{y_n - b}{m} - x_n\right)}{2}.$$

Če izraz razpišemo, dobimo

$$\begin{aligned}
AT &= \left(\frac{\frac{y_1^2 - y_1 b}{m} - x_1 y_1 - (x_1 y_1 - x_1 b) + x_1^2 m + \frac{-y_1 b + b^2}{m} + x_1 b}{2} \right) \\
&+ \left(\frac{\frac{y_2^2 - y_2 b}{m} - x_2 y_2 - (x_2 y_2 - x_2 b) + x_2^2 m + \frac{-y_2 b + b^2}{m} + x_2 b}{2} \right) \\
&\quad \vdots \\
&+ \left(\frac{\frac{y_n^2 - y_n b}{m} - x_n y_n - (x_n y_n - x_n b) + x_n^2 m + \frac{-y_n b + b^2}{m} + x_n b}{2} \right) \\
&= \left(\frac{y_1^2}{2m} - \frac{y_1 b}{m} + \frac{b^2}{2m} - x_1 y_1 + \frac{x_1^2 m}{2} + x_1 b \right) \\
&+ \left(\frac{y_2^2}{2m} - \frac{y_2 b}{m} + \frac{b^2}{2m} - x_2 y_2 + \frac{x_2^2 m}{2} + x_2 b \right) \\
&\quad \vdots \\
&+ \left(\frac{y_n^2}{2m} - \frac{y_n b}{m} + \frac{b^2}{2m} - x_n y_n + \frac{x_n^2 m}{2} + x_n b \right).
\end{aligned}$$

Iz razširjenega izraza lahko sorodne člene seštejemo med sabo, tako kot smo to naredili pri izpeljavi metod najmanjših kvadratov. Tako dobimo skrajšano obliko enačbe s povprečnimi vrednostmi

$$AT = \frac{\overline{ny^2}}{2m} - \frac{n\bar{y}b}{m} + \frac{nb^2}{2m} - n\overline{xy} + \frac{\overline{nx^2}m}{2} + n\bar{x}b.$$

Zdaj lahko izraz parcialno odvajamo in odvod enačimo z nič, da izračunamo vrednosti m in b , ki minimalizirata AT

$$\frac{\partial AT}{\partial m} = 0 \wedge \frac{\partial AT}{\partial b} = 0 .$$

Dobimo

$$\begin{aligned} \frac{\partial}{\partial m} \left(\frac{\overline{ny^2}}{2m} - \frac{n\overline{y}b}{m} + \frac{nb^2}{2m} - n\overline{x}y + \frac{n\overline{x^2}m}{2} + n\overline{x}b \right) = \\ -\frac{\overline{ny^2}}{2m^2} + \frac{n\overline{y}b}{m^2} - \frac{nb^2}{2m^2} + \frac{n\overline{x^2}}{2} \end{aligned}$$

in

$$\begin{aligned} \frac{\partial}{\partial b} \left(\frac{\overline{ny^2}}{2m} - \frac{n\overline{y}b}{m} + \frac{nb^2}{2m} - n\overline{x}y + \frac{n\overline{x^2}m}{2} + n\overline{x}b \right) = \\ -\frac{n\overline{y}}{m} + \frac{nb}{m} + n\overline{x} . \end{aligned}$$

Obe enačbi enačimo z nič, da minimalno vrednosti AT . Z vsakim izrazom lahko manipuliramo individualno.

1. Enačbo

$$-\frac{\overline{ny^2}}{2m^2} + \frac{n\overline{y}b}{m^2} - \frac{nb^2}{2m^2} + \frac{n\overline{x^2}}{2} = 0$$

pomnožimo z $2m^2$, da se znebimo ulomkov. Prav tako jo lahko delimo z n :

$$-\overline{y^2} + 2\overline{y}b - b^2 + \overline{x^2}m^2 = 0 .$$

2. Enačbo

$$-\frac{n\bar{y}}{m} + \frac{nb}{m} + n\bar{x} = 0$$

pomnožimo z m in delimo z n :

$$-\bar{y} + b + \bar{x}m = 0.$$

Tudi na tej regresijski premici se pojavi točka $A(\bar{x}, \bar{y})$. Zdaj, ko imamo dve enačbi za dve neznanki, ju lahko rešimo z uporabo reševanja sistema enačb. Če iz 2. enačbe izrazimo b in ga vstavimo v 1. enačbo, dobimo

$$\begin{aligned} -\bar{y}^2 + 2\bar{y}b - b^2 + \bar{x}^2m^2 &= 0 \\ -\bar{y} + b + \bar{x}m &= 0 \rightarrow b = \bar{y} - \bar{x}m \end{aligned}$$

$$\begin{aligned} -\bar{y}^2 + 2\bar{y}(\bar{y} - \bar{x}m) - (\bar{y} - \bar{x}m)^2 + \bar{x}^2m^2 &= 0 \\ -\bar{y}^2 + 2\bar{y}^2 - 2\bar{y}\bar{x}m - (\bar{y}^2 - 2\bar{y}\bar{x}m + \bar{x}^2m^2) + \bar{x}^2m^2 &= 0 \\ -\bar{y}^2 + 2\bar{y}^2 - 2\bar{y}\bar{x}m - \bar{y}^2 + 2\bar{y}\bar{x}m - \bar{x}^2m^2 + \bar{x}^2m^2 &= 0 \\ -\bar{y}^2 + \bar{y}^2 - \bar{x}^2m^2 + \bar{x}^2m^2 &= 0 \\ \bar{x}^2m^2 - \bar{x}^2m^2 &= \bar{y}^2 - \bar{y}^2 \\ m^2(\bar{x}^2 - \bar{x}^2) &= \bar{y}^2 - \bar{y}^2 \end{aligned}$$

$$m^2 = \frac{\bar{y}^2 - \bar{y}^2}{\bar{x}^2 - \bar{x}^2}$$

$$m = \pm \sqrt{\frac{\bar{y}^2 - \bar{y}^2}{\bar{x}^2 - \bar{x}^2}}.$$

Kot smo videli na str. 11, lahko m izrazimo s standardnima deviacijama

$$m = \pm \frac{\sigma_y}{\sigma_x}.$$

Če dobljen m vstavimo v 2. enačbo, lahko rešimo za b

$$b = \bar{y} - \bar{x} \left(\pm \frac{\sigma_y}{\sigma_x} \right).$$

4. ANALIZA OBEH REGRESIJSKIH PREMICE

Zdaj, ko smo izpeljali obe enačbi, se lahko vprašamo, kateri od kriterijev za določitev regresijske premice je boljši, bolj natančen. Metoda, ki bi nam zagotavljala in nam dokazovala, katera enačba je boljša, ne obstaja, saj sta obe enačbi pravilni po njunem kriteriju. Kljub temu lahko obe metodi primerjamo teoretično in praktično.

V obeh primerih pri računanju smernega koeficienta m dobimo ulomek standardnih deviacij y in x :

$$m_{NK} = r \frac{\sigma_y}{\sigma_x}$$

in

$$m_{NPT} = \pm \frac{\sigma_y}{\sigma_x}.$$

To nam pove, da je razpršenost med vsemi y -ni in x -i zelo pomembna za določitev obeh regresijskih premic. Kljub temu opazimo razliko v obeh enačbah. Pri izračunu koeficienta regresijske premice m po metodi najmanjših kvadratov lahko zasledimo korelacijski koeficient r v enačbi. Če se korelacijski koeficient približa vrednosti 1 ali -1 (to pomeni da je korelacija točk na premici zelo močna-točke od regresijske premice odstopajo zelo malo), bi moral

koeficient regresijske premice biti podoben tistemu, ki bi ga zračunali po moji metodi (ki ne vsebuje korelacijskega koeficienta r med točkami). Prednost uporabe korelacijskega koeficienta r v enačbi je prav tako določitev predznaka koeficienta regresijske premice. Če bo r negativen, bo naklon premice negativen. Če bo pozitiven, bo naklon premice pozitiven. Če se bo korelacijski koeficient r bližal vrednosti nič, bo absolutna vrednost koeficienta regresijske premice m vedno manjša in to bo vodilo v vedno bolj »položno« regresijsko premico na grafu.

Prednost moje regresijske premice je, da korelacijski koeficient r ni potreben. Prav tako ni nobenega drugega člena, ki bi vplival na vrednost koeficienta m regresijske premice (m je določen samo kot koeficient standardne deviacije y in standardne deviacije x). Torej, če bodo podatki v šibki korelaciji na grafu, se koeficient m regresijske premice po moji metodi ne bo zmanjšal. Slabost moje enačbe je, da ni razviden predznak koeficienta m regresijske premice. Tukaj ne ostane drugega, kot kvalitativno opazovanje grafa ali poračunanje korelacijskega koeficienta r , iz katerega bi razbrali predznak smernega koeficienta. Za kvalitativno opazovanje bi lahko pomagala skica regresijske premice, po kateri bi ocenili pozitiven ali negativni naklon (predznak). Problem se pojavi pri zelo šibkih korelacijah oz. ko je r enak nič, saj takrat težko ocenimo lego regresijske premice med točkami. Prav tako je pomembno omeniti točko $A(\bar{x}, \bar{y})$, ki se nahaja na obeh regresijskih premicah.

Pri nadaljnji analizi sem se odločil obe regresijski premici testirati na praktičnih primerih in jih primerjati med seboj. Za praktično analizo sem uporabil računalniški program Logger Pro.

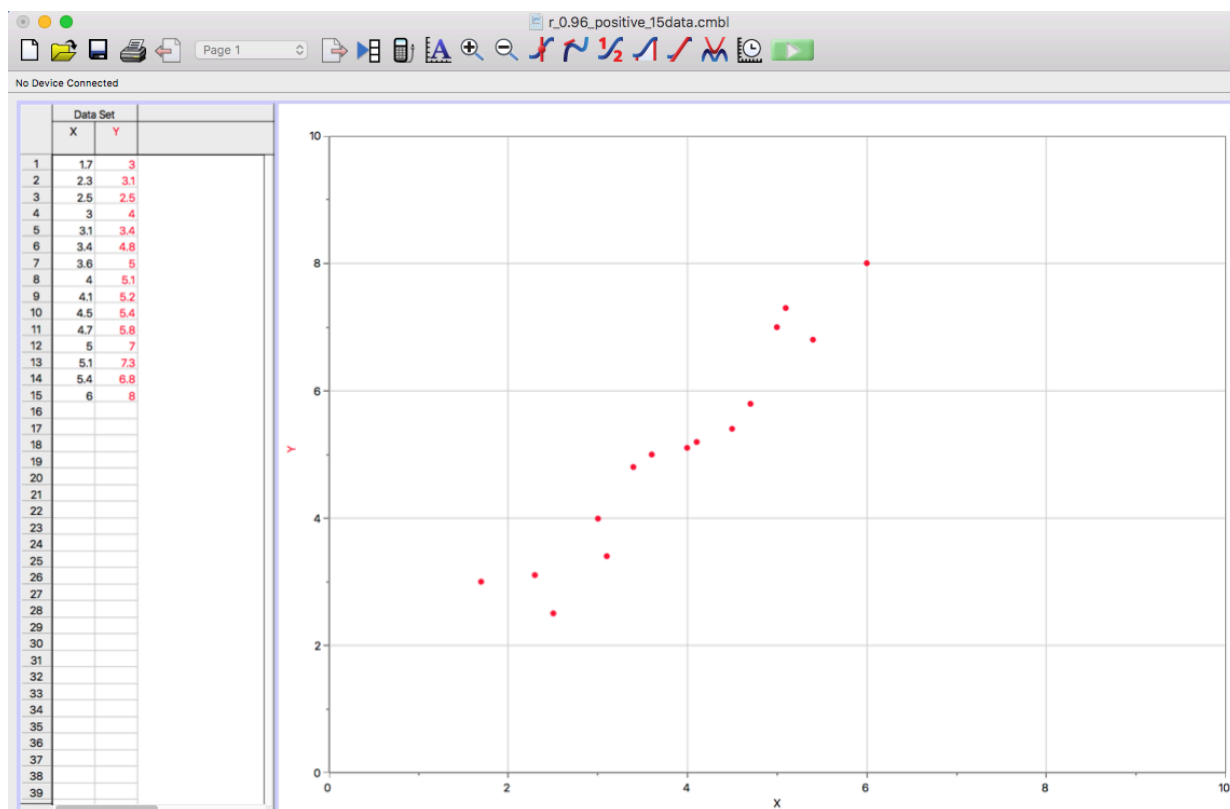
5. PRAKTIČNI PRIMERI

Program Logger Pro je namenjen statističnemu analiziranju podatkov, ki jih dobimo pri merjenju različnih stvari v vsakdanjem življenju (tlak, temperatura, absorpcija itd.), s pomočjo merilnikov, katere priključimo na računalnik. Program se pogosto uporablja pri naravoslovnih predmetih, kot so biologija, fizika in kemija. Za svojo raziskovalno nalogo se mi je program zdel odlična zamisel za analiziranje podatkov z obema regresijskima premicama. V nadaljevanju bom predstavil izbiro podatkov, izračun enačbe obeh regresijskih premic in njuno obnašanje

v odvisnosti od različnih podatkov ter vpeljal tudi normalizacijo svojih podatkov ter analiziral vpliv normalizacije podatkov na mojo regresijsko premico.

5.1. Izbira podatkov (točk) za izračun regresijskih premic

Na začetku sem naključno izbral 15 točk ter jih manipuliral, da sem dobil željen korelacijski koeficient. Po prvem izračunu se točke med seboj niso močno razlikovale, interval neodvisne spremenljivke se je gibal med 1,7 in 6, interval odvisne spremenljivke pa med 3 in 8. Izračunal sem korelacijski koeficient $r = 0,962$. Točke sem zapisal v tabelo in narisal na razsevni diagram. Za poimenovanje osi na grafu sem si preprosto izbral x za os z neodvisnimi spremenljivkami in y os za odvisne spremenljivke.

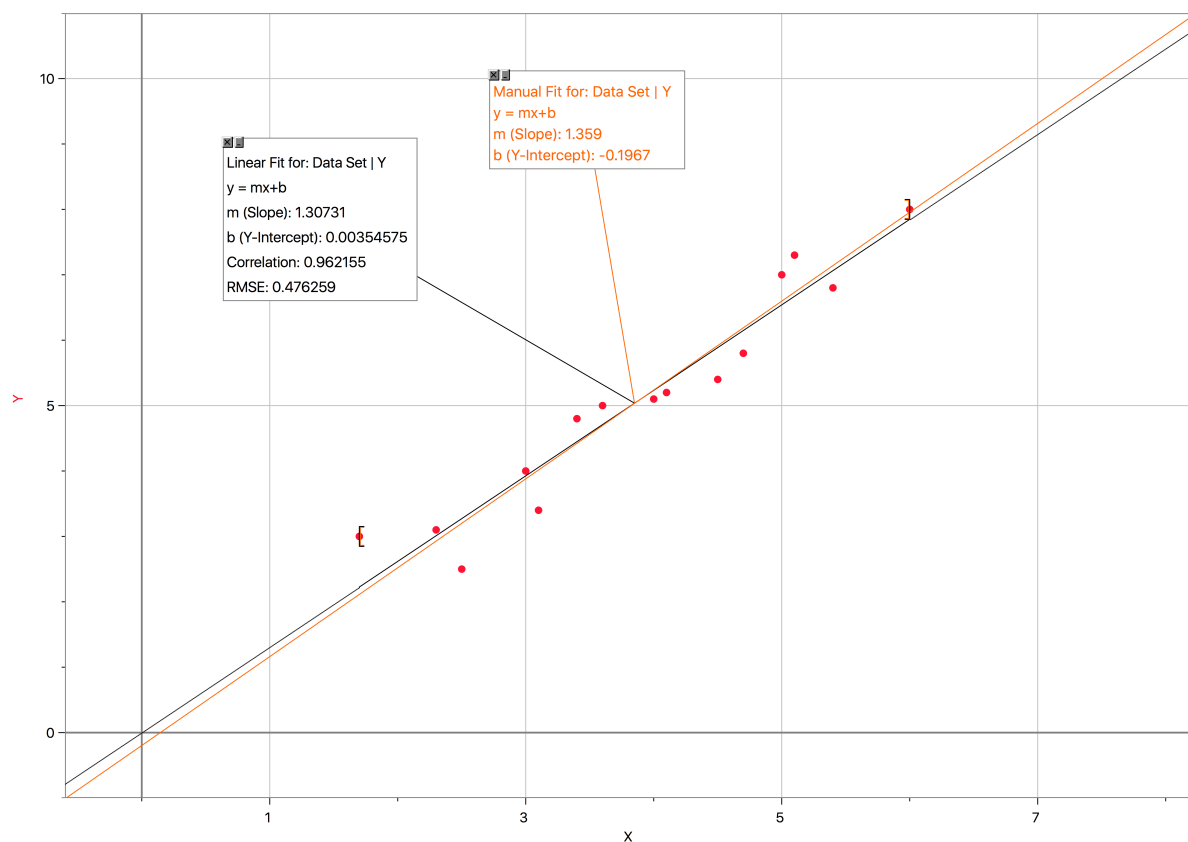


Slika 4: Razsevni diagram v programu Logger Pro. Vir: avtor

Ko sem želel spreminjati naklon regresijskih premic, sem manipuliral s osnovnimi (prvotnimi) podatki. Prav tako sem spreminjal podatke, da sem dobil različne korelacijske koeficiente med njimi.

5.2. Vizualni prikaz razlik med regresijskima premicama

V vsakem grafu sem regresijsko premico po metodi najmanjših kvadratov pobarval črno. Prav tako jo je računalnik izračunal sam. Podatke o koeficientu (m) in začetni vrednosti (b) so vidni v črnem kvadratu s podatki. Regresijsko premico po metodi najmanjše vsote ploščin pravokotnih trikotnikov sem vedno pobarval z oranžno barvo. Podatki o njenem koeficientu (m) in začetni vrednosti (b) so vidni v oranžnem kvadratu. Izračunati sem jih moral sam. Rdeče pike na grafu predstavljajo točke.

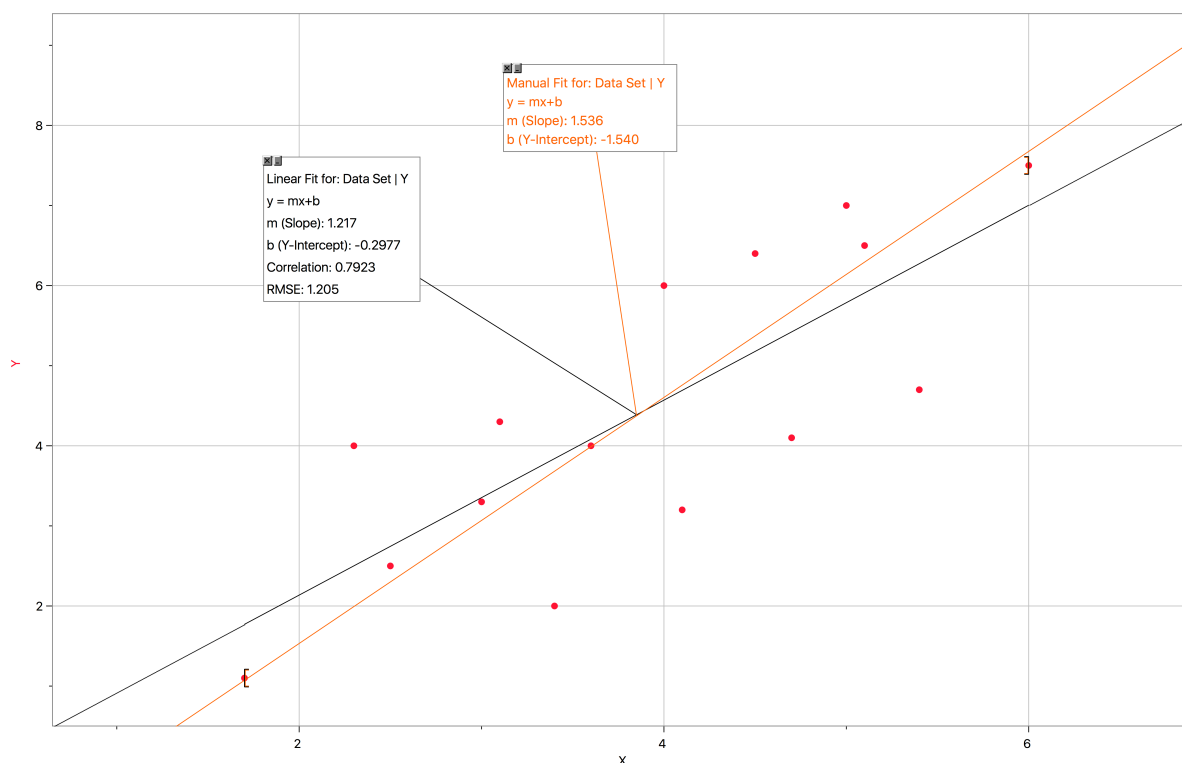


Slika 5: Regresijski premici, ko je med spremenljivkama $r = 0,962$. Vir: avtor

Iz Slike 5 je korelacijski koeficient med podatki relativno visok ($r = 0,962$). Ker se pri regresijski premici po metodi najmanjših kvadratov koeficient premice izračuna po enačbi $m = r \frac{\sigma_y}{\sigma_x}$, je premica le malenkost bolj položna oz. je njen koeficient manjši kot pri moji regresijski premici, ki je definirana po metodi najmanjših vsot ploščin pravokotnih trikotnikov, kjer koeficient izračunamo po enačbi $m = \pm \frac{\sigma_y}{\sigma_x}$. Ravno zaradi visokega korelacijskega

koeficienta, sta koeficienta obeh premic zelo podobna. Zelo pomembna na razsevnem diagramu je točka $A(\bar{x}, \bar{y})$, v kateri se regresijski premici sekata.

Ker sem želel dobiti večja odstopanja med obema premicama, sem manipuliral s podatki tako, da sem dobil absolutno manjši korelacijski koeficient.

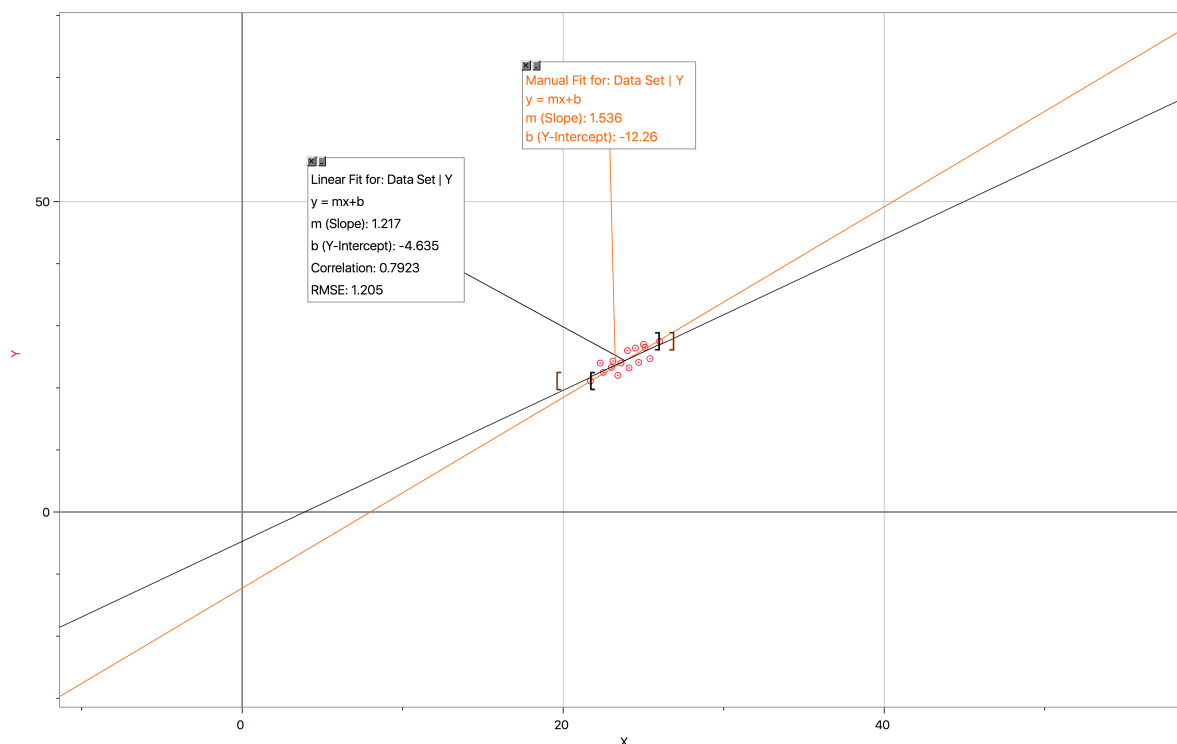


Slika 6: Regresijski premici, ko je med spremenljivkama $r = 0,792$. Vir: avtor

Kot je vidno na Sliki 6, so odstopanja med premicama opazno večje zaradi manjšega korelacijskega koeficienta ($r = 0,792$). Koeficienta se med seboj razlikujeta za $\Delta m = 0.319$. Razvidno je, da z manjšo korelacijo med podatki pride do večji razlik med regresijskima premicama, regresijska premica po metodi najmanjših kvadratov postaja vse bolj položna, glede na regresijsko premico po metodi najmanjše vsote ploščin pravokotnih trikotnikov. Na njeno položnost vpliva korelacijski koeficient. Pri prvem primeru standarda odklona predstavljata vrednosti $\sigma_x = 1,240$ in $\sigma_y = 1,684$ (smerni koeficient premice pa vrednost $m = 1,348$) medtem ko pri korelacijskem koeficientu $r = 0,792$ standardna odklona

predstavljata vrednosti: $\sigma_x = 1,197$ in $\sigma_y = 1,839$ (smerni koeficient premice pa vrednost $m = 1,536$). Pri negativnih koeficientih situacija ostaja enaka.

Če so podatki večji in se nahajajo stran od središča razsevnega diagrama (vendar na kratkem intervalu) bodo absolutne razlike med začetnimi vrednostmi narasle (glej *Slika 7*).



Slika 7: Regresijski premici, ko je $r = 0,792$ in sem vsem podatkom dodal vrednost 20. Vir: avtor

Absolutne razlike med začetnima vrednostnima, ko sem pri $r = 0,792$ vsako točko povečal za vrednost 20, so močno opazne. Prvotno je bila absolutna razlika v začetni vrednosti enaka 1,2423 (*Slika 6*), pri dodani vrednosti točkam pa je absolutna razlika 7,625 (*Slika 7*).

5.3. Normalizacija podatkov

Preden analiziramo vpliv normalizacije na obe regresijski premici, jo moramo definirati. Novim spremenljivkam se reče standardizirane spremenljivke. Da iz navadne spremenljivke dobimo standardizirane spremenljivke, je najprej potreben vzporeden premik. To storimo tako, da od vsake vrednosti spremenljivke odštejemo povprečno vrednost. Da uvedemo novo enoto,

vsako spremenljivko delimo s standardno deviacijo (tako bo nova enota standardna deviacija).
Formula za izračun standardnih spremenljivke je

$$x_s = \frac{x - \bar{x}}{\sigma_x}$$

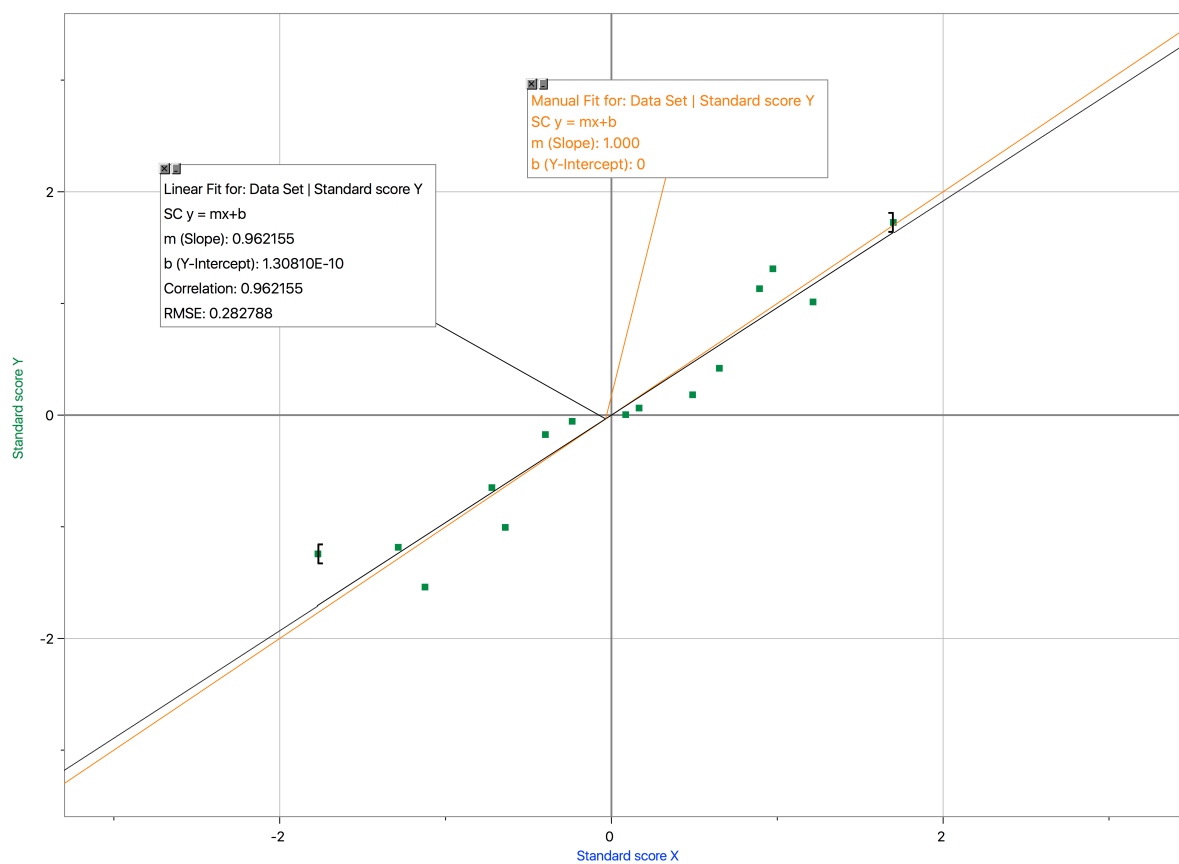
oz.

$$y_s = \frac{y - \bar{y}}{\sigma_y}.$$

Z normalizacijo dobimo nov nabor odvisnih in neodvisnih spremenljivk. Če izračunamo povprečno vrednost teh spremenljivk, bo nova aritmetična sredina enaka 0 ($\bar{x} = \bar{y} = 0$).

Prav tako bo standardni odklon v obeh primerih, pri neodvisni in odvisni spremenljivki enak 1 ($\sigma_x = \sigma_y = 1$).

Odločil sem se normalizirati podatke, za katere sem izračunal korelacijski koeficient $r = 0,962$. Korelacijski koeficient se pri standardizaciji ne spremeni. Nato sem iz standardiziranih podatkov izračunal obe regresijski premici.



Slika 8: Regresijskih premici za standardizirani spremenljivki. Vir: avtor

Na Sliki 8 lahko vidimo, kako novi vrednosti standardnih odklonov vplivata na novi enačbi obeh regresijskih premic. Regresijska premica po metodi najmanjše vsote ploščin pravokotnih trikotnikov je dobila vrednost koeficienta $m = 1$ in s tem postala simetrala lihih kvadrantov (če bi bil naklon premice negativen, potem bi dobili simetralo sodih kvadrantov). Po enačbi za izračun začetne vrednosti je razvidno, da je začetna vrednost enaka 0 ($b = 0$).

Regresijska premica po metodi najmanjših kvadratov ima koeficient definiran samo po vrednosti korelacijskega koeficienta. Povedano drugače, korelacijski koeficient je hkrati koeficient premice ($r = m = 0,962$). Prav tako je začetna vrednost te premice enaka 0 (na Sliki 8 je rezultat začetne vrednosti zelo majhen, kar je posledica računalnikovega zaokroževanja).

Ker se povedano sklada s teorijo, lahko enačbi premic zapišemo v obliki

$$y = rx,$$

za enačbo premice po metodi najmanjših kvadratov, in

$$y = \pm x$$

za enačbo premice po metodi najmanjše vsote ploščin pravokotnih trikotnikov.

5.4. Regresijska premica po metodi najmanjših kvadratov glede na neodvisno spremenljivko

Enačbo za regresijsko premico po metodi najmanjših kvadratov minimalizirano po y sem prilagodil tako, da sem dobil enačbo za regresijsko premico po metodi najmanjših kvadratov minimalizirano po x .

Enačba regresijske premice po metodi najmanjših kvadratov minimalizirano po y je

$$y = r \frac{\sigma_y}{\sigma_x} x + b .$$

Enačba regresijske premice po metodi najmanjših kvadratov minimalizirano po x je torej

$$x = r \frac{\sigma_x}{\sigma_y} y + d ,$$

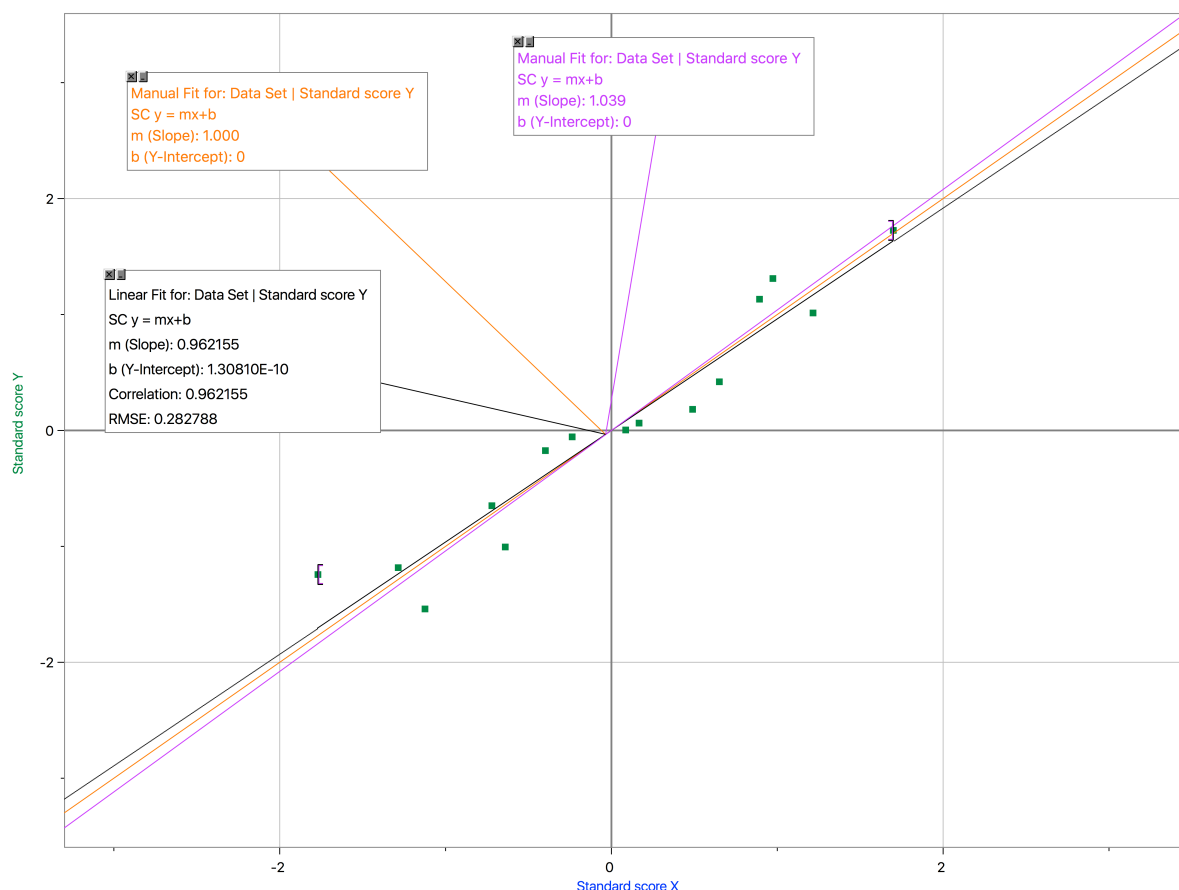
oz. če so spremenljivke standardizirane

$$y = \frac{1}{r} x .$$

Vidimo lahko, da sta si enačbi regresijskih premic po metodi najmanjših kvadratov za standardizirane spremenljivke zelo podobni, razlika je samo v koeficientu premice. Torej, če regresijsko premico minimaliziramo po x , je enačba za njen koeficient enaka

$$m_x = \frac{1}{r}.$$

Novo regresijsko premico sem vrisal v graf k ostalima regresijskima premicama in dobil novo sliko. Novo premico sem pobarval z vijolično.



Slika 9: Regresijske premice po metodi najmanjših kvadratov minimalizirano po x (vijolična), po metodi najmanjših kvadratov minimalizirano po y (črna) in po metodi najmanjše vsote ploščin pravokotnih trikotnikov (oranžna). Vir: avtor

Nova regresijska premica po metodi najmanjših kvadratov minimalizirano po x ima koeficient $r = 1,039$. Na *Sliki 9* je na prvi pogled razvidno, da regresijska premica po metodi najmanjše vsote ploščin pravokotnih trikotnikov predstavlja simetralo kota, ki ga oklepata regresijski premici po metodi namanajših kvadratov minimalizirani po x in y . Vrh kota se nahaja v točki $(0,0)$.

5.5. Simetrala kota

Da bi se prepričal, ali regresijska premica po metodi najmanjše vsote plosčin pravokotnih trikotnikov predstavlja simetralo kota, kateri vrh se nahaja v točki (0,0), njegova kraka pa sta regresijski premici, po metodi najmanjših kvadratov minimalizirani po y in x , sem se odločil to preizkusiti z izračuni kotov, ki jih oklepa vsaka premica z x osjo. Enačba za izračun kota med premico in x osjo je

$$\tan \alpha = m ,$$

torej

$$\tan^{-1} m = \alpha .$$

Izračunal sem kote za vsako regresijsko premico. Za boljšo natančnost sem računal s koeficienti s šestimi decimalnimi vrednostmi. Tako sem zaokroževal tudi kote. Za bolj nazoren prikaz kotov sem za oznake kotov uporabil barve kot za premice na *Sliki 9*. Naklonski koti premic pri korelacijskem koeficientu $r = 0,962$

$$\alpha_1 = 0,766113324 ,$$

$$\alpha_2 = \frac{\pi}{4} ,$$

$$\alpha_3 = 0,804703727 .$$

Simetralo kota lahko dokažemo po enačbi

$$\alpha_3 - \alpha_2 = \alpha_2 - \alpha_1 .$$

Torej

$$0,804703727 - \frac{\pi}{4} = \frac{\pi}{4} - 0,766113324$$

$$0,019305564 \approx 0,019284839 ,$$

če zaokrožimo na štiri decimalna mesta

$$0,0193 = 0,0193 .$$

Vidimo torej, da regresijska premica po metodi najmanjše vsote ploščin pravokotnih trikotnikov predstavlja simetralo kota med regresijskima premicama po metodi najmanjših kvadratov minimaliziranimi po y in x . Da je enačba regresijske premice po metodi najmanjše vsote ploščin pravokotnih trikotnikov simetrala kota, velja samo, ker imamo v razsevnem diagramu standardizirane spremenljivke.

5.6. Uporaba inverzne funkcije

Regresijska premica po metodi najmanjše vsote ploščin pravokotnih trikotnikov ima torej funkcijo simetrale kota med obema regresijskima premicama po metodi najmanjših kvadratov. Prav tako sem omenili, da predstavlja simetralo lihih kvadrantov. Simetrala lihih kvadrantov ima pri operacijah s funkcijami posebno vlogo. Z zrcaljenjem grafa funkcije preko nje dobimo graf inverzne funkcije te funkcije. Z inverzno funkcijo lahko teoretično utemeljimo, da regresijska premica najmanjše vsote ploščin pravokotnih trikotnikov predstavlja simetralo kota regresijskih premic po metodi najmanjših kvadratov minimaliziranih po x in y . Z inverzno funkcijo je potrebno utemeljiti enačbo

$$m_x = \frac{1}{r} .$$

Enačbo za regresijsko premico, minimalizirano po y lahko zapišemo na naslednji način

$$f(x) = rx$$

$$y = rx.$$

Funkcijo želimo pretvoriti v inverzno ($f^{-1}(x)$)

$$x = ry$$

$$y = \frac{1}{r}x.$$

Vidimo torej, da se koeficient pri inverzni funkciji spremeni iz r v r^{-1} . S tem smo dokazali, da je moja regresijska premica res simetrala kota med obema premicama po metodi najmanjših kvadratov.

6. ZAKLJUČEK

Po izpeljavi regresijskih premic in teoretičnih in praktičnih analizah sem spoznal, kako se regresijske premice razlikujejo med seboj. Kot glavne ugotovitve bi rad izpostavil, da regresijska premica po metodi najmanjše vsote ploščin pravokotnih trikotnikov v enačbi za izračun koeficienta m regresijske premice ne vsebuje korelacijskega koeficienta r , kar pomeni, da je njen naklon v primerjavi z regresijsko premico, definirano po metodi najmanjših kvadratov, bolj strm. Pri regresijski premici najmanjše vsote ploščine pravokotnih trikotnikov je potrebno npr. vizualno razbrati, ali je koeficient pozitiven ali negativen. Če sta spremenljivki normalizirani, sem dokazal, da regresijska premica, definirana po metodi najmanjše vsote ploščin pravokotnih trikotnikov, predstavlja simetralo sodih ali lihih kvadrantov (odvisno od korelacijskega koeficienta) ter hkrati simetralo kota, ki ga oklepata regresijski premici po metodi najmanjših kvadratov minimalizirani po y oz. x . Prav tako ima velik pomen točka $A(\bar{x}, \bar{y})$, ki leži na vseh treh regresijskih premicah in hkrati predstavlja sekališče vseh treh premic v eni sami točki.

Vprašanje, ki je ostalo neodgovorjeno, je, zakaj, kljub dolgemu iskanju, nikjer na spletu nisem našel moje metode določanja regresijske premice oz. nisem našel več kot dva različna načina določanja regresijske premice, ki bi pokazala relacijo med podatki na drugačen način, kot to opravi regresijska premica po metodi najmanjših kvadratov. Domnevam, da z izbrano univerzalno metodo (metoda najmanjših kvadratov) drugi načini določevanja regresijskih premic niso več tako pomembni, saj se je vse dobljene podatke iz raziskav primerjalo na podlagi univerzalne metode, t. j. metode najmanjših kvadratov.

7. DRUŽBENA ODGOVORNOST

Menim, da je moja raziskovalna naloga ravnala v načelih družbene odgovornosti in ni prišlo do kršenja nobenih izmed sedmih načel družbene odgovornosti po ISO 26000. Nobena meritev ni škodovala družbi in naravi v okviru dela moje raziskovalne naloge, prav tako sem spoštoval že dosežene rezultate, definicije in izpeljave tako, da sem navajal avtorja le-teh in se s tem izognil plagiatu. Z izvedbo raziskovalne naloge nisem ogrožal nobene že obstoječe raziskave ali ji škodoval. Nisem zlorabljal dela drugih in kršil pravila o kraji podatkov, tako uradnih kot tudi neuradnih.

8. VIRI IN LITERATURA

- Khan, S., (2018), Proof of minimizing squared error to regression line, Khan Academy. [online] <https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/more-on-regression/v/squared-error-of-regression-line> (dostopno 9. 2. 2019)
- Wikipedia, (2019), Partial derivate. [online] https://en.wikipedia.org/wiki/Partial_derivative (dostopno 9. 2. 2019)
- Wazir, I., Garry, T., Ashbourne, P., Barclay, P., Flynn, P., Frederick, K., Wakeford, M., (2012), Standard Level Mathematics (2012 edition). Edinburgh Gate, Pearson Education Limited