Mjere zavisnosti-svojstva i zamke

Lamot, Lorena

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni preddiplomski studij matematike

Lorena Lamot

Mjere zavisnosti - svojstva i zamke

Završni rad

Osijek, 2021.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni preddiplomski studij matematike

Lorena Lamot

Mjere zavisnosti - svojstva i zamke

Završni rad

Mentor: izv. prof. dr. sc. Nenad Šuvak

Osijek, 2021.

Sažetak

U ovom ćemo se radu baviti mjerama povezanosti dviju slučajnih varijabli. Najprije ćemo uvesti definicije osnovnih numeričkih karakteristika slučajnog vektora koje su nam potrebne za definiranje koeficijenta korelacije. Zatim ćemo detaljno opisati Pearsonov korelacijski koeficijent koji nam daje informaciju o "jakosti" linearne veze između dvije varijable. Navest ćemo njegovu definiciju, svojstva te pretpostavke za njegovo korištenje. Nakon toga ćemo definirati dvije neparametarske mjere asocijacije - Spearmanov koeficijent korelacije ranga te Kendallov τ. Opisat ćemo njihova svojstva i vidjeti njihovu primjenu kroz nekoliko primjera.

Ključne riječi

Koeficijent korelacije, linearna veza, Pearsonov koeficijent korelacije, Spearmanov koeficijent korelacije ranga, Kendallov τ

Summary

In this bachelor's thesis we will be studying measures of correlation or association between two variables. First of all, we will introduce some definitions of basic numerical characteristics of random vector that we need for defying correlation coefficient. After that, we will describe Pearson's correlation coefficient which gives us information about the strength of linear relationship between two variables. We will define Pearson's correlation coefficient, say something about its properties and assumptions for its usage. After that, we will define two non-parametric measures of association - Spearman's rank-order correlation coefficient and Kendall's τ. We will describe their properties and see their usage on few examples.

Key words

Correlation coefficient, linear relationship, Pearson's correlation coefficient, Spearman's rank-order correlation coefficient, Kendall's τ

Sadržaj

1 Uvod

Korelacija predstavlja stupanj međusobne povezanosti dviju slučajnih varijabli. Povezanost ili asocijacija između varijabli znači da vrijednost jedne varijable ovisi o vrijednosti druge varijable: kako vrijednost jedne varijable raste, raste ili opada i vrijednost druge varijable. Koeficijent korelacije se u statistici koristi kao mjera povezanosti dviju varijabli. U prvom poglavlju definirat ćemo koefcijent korelacije te navesti njegova osnovna svojstva.

Za procjenu koeficijenta korelacije možemo koristiti nekoliko procjenitelja. U drugom poglavlju detaljnije ćemo obraditi jedan od tih procjenitelja, Pearsonov koeficijent korelacije. On se koristi kao mjera jakosti i daje informaciju o monotonosti linearne veze između dvije neprekidne varijable. Opisat ćemo njegova svojstva te pretpostavke koje moraju biti zadovoljene za njegovo korištenje. Navest ćemo i slučajeve u kojima nije prikladno koristiti Pearsonov koeficijent korelacije.

U trećem poglavlju bavit ćemo se neparametarskim mjerama asocijacije koje se koriste kada nije zadovoljena neka od pretpostavki za korištenje Pearsonovog koeficijenta korelacije te za opisivanje nelinearnih monotonih veza među varijablama. Spearmanov koeficijent korelacije ranga i Kendallov τ dvije su takve mjere asocijacije. Oni nam daju informaciju o postojanju monotone veze između dviju varijabli.

2 Koeficijent korelacije

2.1 Osnovni pojmovi

Najprije je potrebno definirati momente, osnovne numeričke karakteristike slučajnog vektora koji će nam poslužiti za definiranje koeficijenta korelacije.

Definicija 1. *Neka je (X, Y) diskretan ili neprekidan dvodimenzionalan slučajni vektor. Očekivanje*

$$E(X^k Y^l), \quad k, l \in \mathbb{N}_0$$

slučajne varijable $X^k Y^l$ (ako postoji) nazivamo ***ishodišni moment*** *reda (k,l) slučajnog vektora (X, Y) i pišemo*

$$\mu_{kl} = E(X^k Y^l).$$

Definicija 2. *Očekivanje $E((X - EX)^k (Y - EY)^l)$ (ako postoji) nazivamo* ***centralni moment*** *reda (k,l) slučajnog vektora (X, Y) i pišemo*

$$m_{kl} = E((X - EX)^k (Y - EY)^l).$$

Centralni moment reda (1,1) nazivamo korelacijski moment ili kovarijanca dvodimenzionalnoga slučajnog vektora.

Definicija 3. *Kovarijanca dvodimenzionalnog slučajnog vektora (X, Y) definirana je izrazom*

$$Cov(X, Y) = E((X - EX)(Y - EY)).$$

Važnost kovarijance je u tome što ju možemo povezati s pojmom nezavisnosti slučajnih varijabli. Vrijedi sljedeći teorem:

Teorem 1. *Neka je (X, Y) neprekidan ili diskretan dvodimenzionalan slučajni vektor za koji postoje EX i EY. Ako su slučajne varijable X i Y nezavisne, onda je Cov(X, Y)=0.*

Dokaz. Neka su X i Y nezavisne slučajne varijable za koje postoje EX i EY. Zbog nezavisnosti je

$$E(X, Y) = EXEY.$$

Dakle, vrijedi:

$$Cov(X, Y) = E(XY) - EXEY = 0.$$

□

Kao posljedica tog teorema vrijedi sljedeća tvrdnja: ako za dvodimenzionalan slučajni vektor (X, Y) vrijedi $Cov(X, Y) \neq 0$, onda su varijable X i Y nužno zavisne. Obrat tog teorema općenito ne vrijedi, tj. ako je $Cov(X, Y) = 0$, to ne znači nužno da su X i Y nezavisne.

Definicija 4. *Neka je (X, Y) slučajni vektor za koji je Cov(X, Y)=0. Tada kažemo da su njegove komponente X i Y* ***nekorelirane****.*

Definicija 5. ***Koeficijent korelacije*** *dvodimenzionalnog slučajnog vektora* (X, Y) *definiramo kao broj*

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

gdje su

$$\sigma_X = \sqrt{VarX}, \quad \sigma_Y = \sqrt{VarY}.$$

2.2 Svojstva koeficijenta korelacije

Koeficijent korelacije je korisna numerička karakteristika za opisivanje veze među komponentama slučajnog vektora. To je vidljivo iz sljedećeg teorema koji pokazuje da se linearna veza među komponentama može uočiti na osnovu vrijednosti koeficijenta korelacije.

Teorem 2. *Neka je* (X, Y) *slučajni vektor za koji je* $0 < \sigma_X < \infty$ *i* $0 < \sigma_Y < \infty$. *Veza je među komponentama linearna, tj. postoje realni brojevi* a $(a \neq 0)$ *i* b *takvi da je*

$$Y = aX + b$$

onda i samo onda ako je $|\rho_{X,Y}| = 1$. *Pritom je koeficijent korelacije 1 ako je* $a > 0$, *odnosno -1 ako je* $a < 0$.

Dokaz. Neka je (X, Y) slučajni vektor takav da je $Y = aX + b$, $a \neq 0$. Po pretpostavkama teorema postoji kovarijanca pa vrijedi:

$$\begin{gathered} E(X - EX)(Y - EY) = \\ = E((X - EX)(aX + b) - E(aX + b)) = E((X - EX)a(X - EX)) = \\ = aE(X - EX)^2. \end{gathered}$$

Dakle,

$$Cov(X, Y) = a\sigma_X^2.$$

Kako je $VarY = a^2 VarX$, vrijedi sljedeće:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{a\sigma_X^2}{\sigma_X |a| \sigma_X} = \frac{a}{|a|}.$$

Ako je $a > 0$, onda je $\rho_{X,Y} = 1$.Ako je $a < 0$, onda je $\rho_{X,Y} = -1$.
Dokažimo obratnu tvrdnju. Pretpostavimo da je $\rho_{X,Y} = 1$. Definirajmo slučajnu varijablu

$$Z = \frac{1}{\sigma_X} X - \frac{1}{\sigma_Y} Y.$$

Tada je

$$VarZ = E\left(\frac{1}{\sigma_X} X - \frac{1}{\sigma_Y} Y\right)^2 - \left(E\left(\frac{1}{\sigma_X} X - \frac{1}{\sigma_Y} Y\right)\right)^2 = 1 - 2\rho_{X,Y} + 1 = 0.$$

Dakle, Z je konstanta. Označimo $Z = c$. Vrijedi:

$$c = \frac{1}{\sigma_X} X - \frac{1}{\sigma_Y} Y,$$

$$Y = \frac{\sigma_Y}{\sigma_X} X - \sigma_Y c.$$

Pretpostavimo sada da je $\rho_{X,Y} = -1$. Definirajmo slučajnu varijablu

$$Z_1 = \frac{1}{\sigma_X} X + \frac{1}{\sigma_Y} Y.$$

Tada je $VarZ_1 = 0$. Dakle, Z_1 je konstanta. Označimo $Z_1 = c_1$. Vrijedi:

$$c_1 = \frac{1}{\sigma_X} X + \frac{1}{\sigma_Y} Y,$$

$$Y = -\frac{\sigma_Y}{\sigma_X} X + \sigma_Y c_1.$$

□

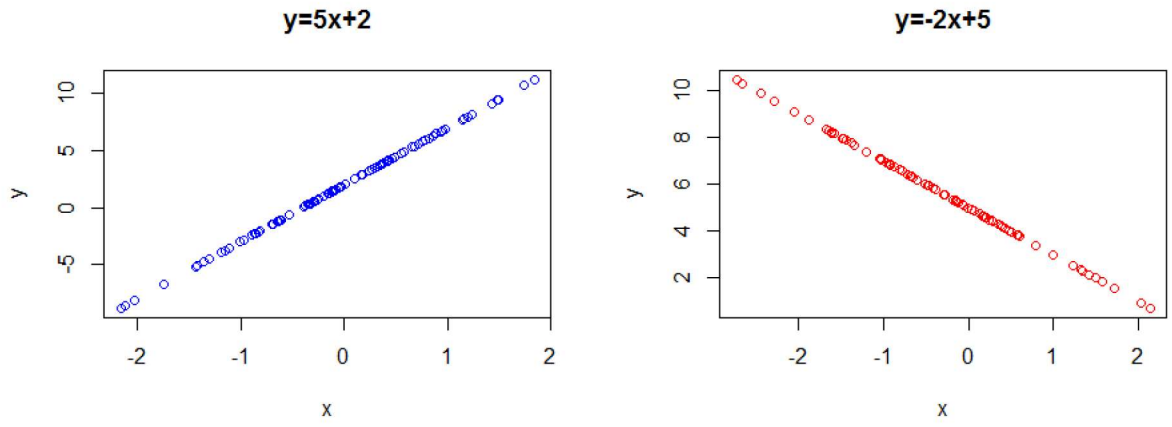
Koeficijent korelacije ima i sljedeća svojstva:

- $-1 \leq \rho_{XY} \leq 1$
- ako su X i Y nezavisne slučajne varijable, onda je $\rho_{XY} = 0$
- ako je $\rho_{XY} = 0$ kažemo da su slučajne varijable X i Y nekorelirane

Iz prethodnih svojstava primjećujemo kako je za potvrđivanje zavisnosti između slučajnih varijabli X i Y dovoljno pokazati da je njihov koeficijent korelacije različit od 0. Također, ako je koeficijent korelacije 1 ili -1, veza između X i Y je linearna.

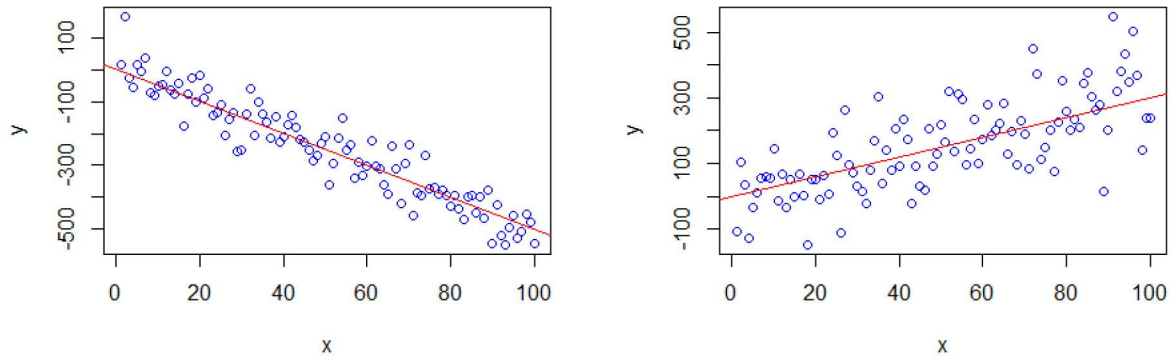
Za procjenu koeficijenta korelacije možemo koristiti nekoliko procjenitelja, ali detaljnije ćemo opisati Pearsonov koeficijent korelacije čija nam vrijednost daje informaciju o "jakosti" linearne veze između varijabli X i Y. U sljedećem primjeru ilustrirat ćemo pojam linearne veze između varijabli X i Y te objasniti što podrazumijevamo pod terminom "jakost" linearne veze.

Primjer 1. *Na sljedećim dijagramima vidljiva je linearna veza između varijabli X i Y. Na prvom dijagramu veza između varijabli je monotono rastuća jer se rastom vrijednosti od X povećavaju i vrijednosti od Y. Na drugom dijagramu veza između varijabli je monotono padajuća jer se rastom vrijednosti od X smanjuju vrijednosti od Y.*



Slika 1: Linearna veza izmedu varijabli X i Y

U statističkim analizama ne možemo očekivati linearnu vezu bez određenih odstupanja. U sljedećim dijagramima vidimo kako se parovi vrijednosti varijabli X i Y grupiraju oko pravca, ali ne leže svi na njemu.



Slika 2: Sugerirana linearna veza između varijabli X i Y

To sugerira da bi vezu između ovih varijabli mogli modelirati linearnom funkcijom, ali do na neku grešku. Što su ta odstupanja od pravca manja, "jakost" linearne veze je veća.

3 Pearsonov koeficijent korelacije

Definicija 6. *Neka su* $(x_1, y_1), ..., (x_n, y_n)$ *nezavisne realizacije neprekidnog slučajnog vektora* (X, Y). ***Pearsonov koeficijent korelacije*** *definiramo izrazom*

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_n)^2}},$$

gdje su

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Da bismo na osnovu uzorka potvrdili zavisnost korištenjem koeficijenta korelacije, treba odbaciti hipotezu:

$$H_0 : \rho_{XY} = 0.$$

Ovdje ćemo navesti jedan od testova koji se može koristiti u tu svrhu. On je kreiran pod pretpostavkom normalnosti distribucije slučajnog vektora (X, Y) korištenjem Pearsonova korelacijskog koeficijenta. Za testiranje navedene nul-hipoteze računamo vrijednost test statistike po formuli:

$$\hat{t} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Ako je nul-hipoteza istinita, statistika kojoj smo tako izračunali distribuciju ima Studentovu distribuciju s $(n-2)$ stupnja slobode:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(n-2).$$

Označimo li s T slučajnu varijablu koja ima studentovu distribuciju s $(n-2)$ stupnja slobode, pripadnu p-vrijednost određujemo na sljedeći način:

$p = P\{T \geq t\}$ ako je alternativna hipoteza oblika $H_1 : \rho_{XY} > 0$

$p = P\{T \leq t\}$ ako je alternativna hipoteza oblika $H_1 : \rho_{XY} < 0$.

Tako dobivenu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku:

ako je $p < \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da su slučajne varijable X i Y zavisne

ako je $p > \alpha$, nemamo dovoljno argumenata kako bi odbacili nul-hipotezu, tj. kažemo da nemamo dovoljno argumenata tvrditi da su slučajne varijable X i Y zavisne.

3.1 Svojstva i interpretacija

Vrijednost Pearsonovog koeficijenta korelacije daje nam informaciju o "jakosti" linearne veze između varijabli X i Y.

Sljedeća svojstva važna su za interpretaciju dobivene vrijednosti:

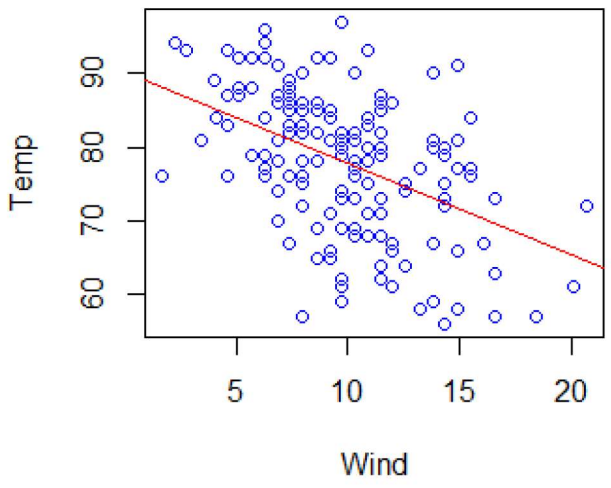
- $r \in [-1, 1]$
- vrijednost $r \approx 0$ sugerira da između X i Y nema linearne veze, tj. da su varijable X i Y nekorelirane
- vrijednost $r \approx 1$ ili $r \approx -1$ sugerira da vezu između X i Y ima smisla modelirati linearnom funkcijom, tj. da su varijable X i Y korelirane
- ako je $r < 0$, analiziramo "jakost" padajuće linearne veze između X i Y
- ako je $r > 0$, analiziramo "jakost" rastuće linearne veze između X i Y.

"Jakost" linearne veze predstavlja informaciju o tome koliko je linearna funkcija prikladna za opisivanje veze između varijabli X i Y. Interpretiramo ju na sljedeći način:

- ako je između varijabli X i Y sugerirana rastuća linearna veza, tj. dobivena vrijednost $r > 0$, tada kažemo da je rast x vrijednosti povezan s rastom y vrijednosti
- ako je između varijabli X i Y sugerirana padajuća linearna veza, tj. dobivena vrijednost $r < 0$, tada kažemo da je rast x vrijednosti povezan s padom y vrijednosti.

Često se "jakost" linearne veze pogrešno interpretira tako što se u slučaju sugerirane rastuće linearne veze kaže kako rast x vrijednosti uzrokuje rast y vrijednosti. Također, u slučaju sugerirane padajuće linearne veze pogrešno je reći kako rast x vrijednosti uzrokuje pad y vrijednosti zbog toga što koeficijent korelacije ne sadrži informaciju o uzročnosti.

Primjer 2. *U bazi podataka* airquality *iz R paketa* datasets *varijabla* Wind *sadrži podatke o izmjerenim brzinama vjetra u miljama na sat, a varijabla* Temp *podatke o temperaturi zraka u Fahrenheitima. Želimo analizirati vezu između te dvije varijable. Pogledajmo najprije dijagram raspršenosti vrijednosti tih varijabli.*



Slika 3: Dijagram raspršenosti vrijednosti varijabli Wind i Temp

Iz dijagrama vidimo kako se parovi vrijednosti varijabli Wind *i* Temp *grupiraju oko pravca koji ima negativan koeficijent smjera. Sada želimo izračunati vrijednost Pearsonovog koeficijenta korelacije tih dviju varijabli. Dobivena vrijednost je približno* -0.458 *što nam ukazuje na postojanje negativne korelacije između varijabli* Wind *i* Temp. *Želimo to potvrditi provođenjem korelacijskog testa. Označimo li s* X *varijablu* Wind, *a s* Y *varijablu* Temp, *nul-hipoteza je oblika:*

$$H_0 : \rho_{XY} = 0,$$

dok je alternativna hipoteza oblika:

$$H_1 : \rho_{XY} < 0.$$

Dobivena p-vrijednost je 1.321×10^{-9} *pa na razini značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. možemo tvrditi da što je brzina vjetra veća, temperatura zraka je manja.*

3.2 Pretpostavke

Za uzorak $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ pretpostavljamo da dolazi iz dvodimenzionalne normalne distribucije $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.
Pod tom je pretpostavkom poznata egzaktna distribucija Pearsonovog koeficijenta korelacije $\rho_{X,Y}$.
U tom slučaju test-statistika kojom testiramo hipotezu o nekoreliranosti:

$$H_0 : \rho_{XY} = 0$$

u uvjetima kada je H_0 istinita, ima Studentovu distribuciju s $(n-2)$ stupnja slobode:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(n-2).$$

Postavlja se pitanje može li se pretpostavka o normalnosti uzorka relaksirati. C.J. Kowalski, nakon diskusije o relaksaciji pretpostavke o normalnosti u članku [4], dolazi do zaključka kako distribucija Pearsonovog koeficijenta korelacije ovisi o tome dolazi li uzorak iz normalne distribucije te kako je zbog toga za analizu povezanosti varijabli bitno da uzorak bude barem približno normalno distribuiran. Nadalje, distribucija Pearsonovog koeficijenta korelacije se za velik broj mjerenja može dobro aproksimirati normalnom distribucijom (pokazano u knjizi [5]). Stoga zaključujemo kako se pretpostavka o normalnosti može relaksirati za velike uzorke.

Ipak, pretpostavka normalnosti često se naglašava i kod velikih uzoraka jer se za testiranje hipoteze o nekoreliranosti varijabli:

$$H_0 : \rho_{XY} = 0$$

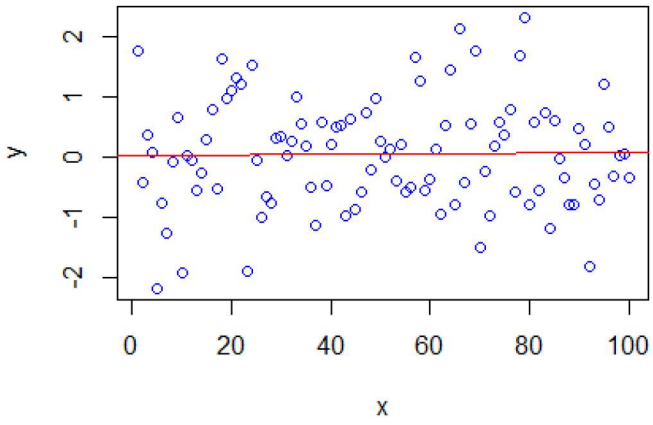
korištenjem gotovih funkcija u softverima kao što su R ili Statistica provodi T-test, tj. pretpostavlja se da uzorak $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ dolazi iz dvodimenzionalne normalne distribucije.

3.3 Neprikladnost

Navest ćemo nekoliko slučajeva u kojima nije prikladno koristiti Pearsonov koeficijent korelacije.

Prvi takav slučaj je kada nije sugerirano postojanje veze između varijabli X i Y. To možemo uočiti iz dijagrama raspršenosti, što je vidljivo u sljedećem primjeru.

Primjer 3. *Dijagram raspršenosti vrijednosti varijabli X i Y izgleda ovako:*

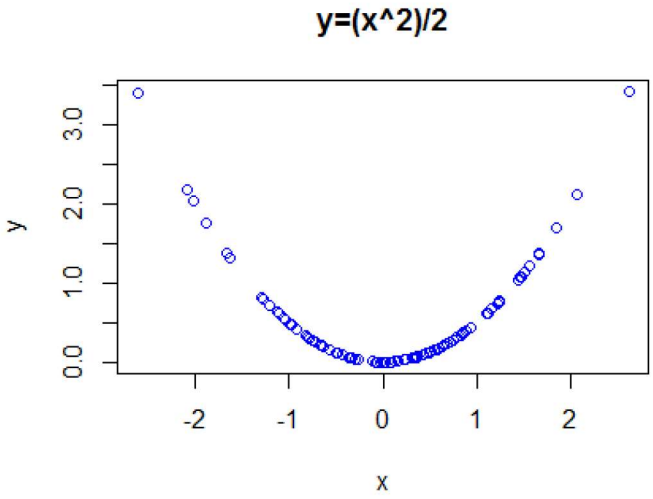


Slika 4: Dijagram raspršenosti varijabli X i Y

Iz dijagrama raspršenosti ne možemo uočiti vezu između tih dviju varijabli. Vrijednost Pearsonovog koeficijenta korelacije je približno 0.015, što nam sugerira kako ne postoji veza između varijabli X i Y.

Drugi slučaj je kada je sugerirana veza između varijabli X i Y, ali ona nije linearna. Pearsonov koeficijent korelacije daje nam informaciju o tome koliko je linearna veza prikladna za modeliranje veze među varijablama. Neprikladno ga je koristiti kada je sugerirana neka druga vrsta veze, npr. kvadratna.

Primjer 4. *Na dijagramu raspršenosti vrijednosti varijabli X i Y uočavamo kako postoji kvadratna veza.*

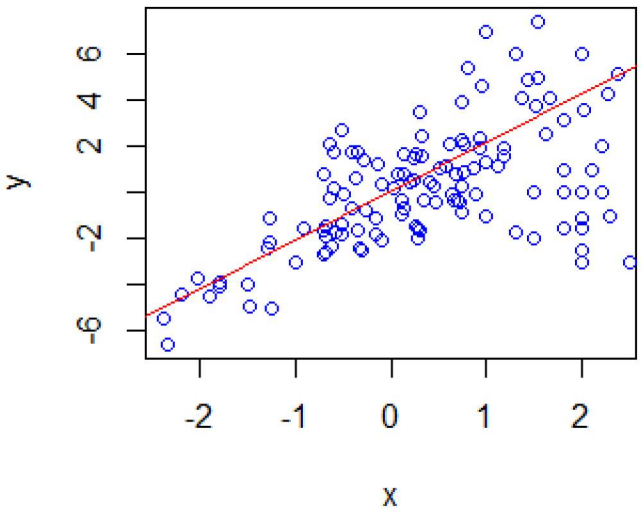


Slika 5: Kvadratna veza varijabli X i Y

Iako su varijable X i Y povezane determinističkom vezom, Pearsonov koeficijent korelacije iznosi približno 0.036 zbog toga što on ne prepoznaje veze koje nisu linearne.

Treći slučaj neprikladnosti korištenja Pearsonovog koeficijenta korelacije je narušena homoskedastičnost u uzorku. Homoskedastičnost je narušena kada linearna veza nije jednako prikladna za modeliranje veze između promatranih obilježja na cijelom skupu njihovih vrijednosti.

Primjer 5. *Iz dijagrama raspršenosti uočavamo narušenu homoskedastičnost u uzorku koja nam sugerira neprikladnost korištenja linearne veze za modeliranje veze između ovih varijabli.*



Slika 6: Heteroskedastičnost u uzorku

4 Neparametarske mjere asocijacije

Neparametarske mjere asocijacije prikladne su za opisivanje veza među varijablama X i Y kada nije zadovoljena neka od pretpostavki na uzorak kod Pearsonovog koeficijenta korelacije te za opisivanje nelinearne monotone veze među obilježjima s proizvoljnim skupom vrijednosti.
Kod takvih mjera nema pretpostavke na distribuciju uzorka $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ te su neosjetljive na prisutnost stršećih vrijednosti.

Spearmanov koeficijent korelacije i Kendallov τ su dvije takve mjere asocijacije koje se ne baziraju na sam uzorak nego na rangove podataka iz uzorka.

Za određivanje ranga nekog podatka skup podataka $x_1, x_2, ..., x_n$ najprije sortiramo po veličini u rastućem ili padajućem poretku. Ako su svi podaci međusobno različiti, rang podatka x_i je njegov redni broj u tako sortiranom nizu podataka. Ako postoje jednaki podaci, rang podatka x_i je aritmetička sredina rednih brojeva svih podataka koji su jednaki x_i.

4.1 Spearmanov koeficijent korelacije ranga

Spearmanov koeficijent korelacije ranga ρ_S daje nam informaciju u kojoj se mjeri veza između slučajnih varijabli X i Y može opisati monotonom funkcijom.

Neka je $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ realizacija jednostavnog slučajnog uzorka iz slučajnog vektora (X, Y). Označimo s r_{x_i} rang podatka x_i u sortiranom nizu podataka $(x_1, x_2, ..., x_n)$, a s r_{y_i} rang podatka y_i u sortiranom nizu podataka $(y_1, y_2, ..., y_n)$.

Tada ρ_S možemo procijeniti s

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(r_{x_i} - r_{y_i})^2}{n(n^2-1)}.$$

Pretpostavimo da su svi $x_1, ...x_n$ i $y_1, ..., y_n$ različiti zbog jednostavnosti računa.
Vrijedi sljedeće:

$$\overline{r}_x = \frac{1}{n}\sum_{i=1}^{n} r_{x_i} = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$\begin{aligned}\sum_{i=1}^{n}(r_{x_i} - \overline{r}_x)^2 &= \sum_{i=1}^{n} r_{x_i}^2 - 2\overline{r}_x\sum_{i=1}^{n} r_{x_i} + n\overline{r}_x^2 = \sum_{i=1}^{n} r_{x_i}^2 - n\overline{r}_x \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}.\end{aligned}$$

Na analogan način dobivamo:

$$\overline{r}_y = \frac{n+1}{2}$$

te

$$\sum_{i=1}^{n}(r_{y_i} - \overline{r}_y)^2 = \frac{n(n^2-1)}{12}.$$

Sada r_S možemo zapisati kao:

$$\begin{aligned}
r_S &= \frac{12}{n(n^2-1)}\left(\frac{n(n^2-1)}{12} - \frac{1}{2}\left(\sum_{i=1}^{n} r_{x_i}^2 - 2\sum_{i=1}^{n} r_{x_i} r_{y_i} + \sum_{i=1}^{n} r_{y_i}^2\right)\right) \\
&= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} r_{x_i} r_{y_i} - \frac{n(n+1)(2n+1)}{6} + \frac{n(n-1)(n+1)}{12}\right) \\
&= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} r_{x_i} r_{y_i} - \frac{n(n+1)^2}{4}\right) \\
&= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} r_{x_i} r_{y_i} - \frac{(n+1)}{2}\frac{n(n+1)}{2} - \frac{(n+1)}{2}\frac{n(n+1)}{2} + n\left(\frac{n+1}{2}\right)^2\right) \\
&= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} r_{x_i} r_{y_i} - \overline{r}_y \sum_{i=1}^{n} r_{x_i} - \overline{r}_x \sum_{i=1}^{n} r_{y_i} + \sum_{i=1}^{n} \overline{r}_x \overline{r}_y\right) \\
&= \frac{\sum_{i=1}^{n}(r_{x_i} - \overline{r}_x)(r_{y_i} - \overline{r}_y)}{\sqrt{\sum_{i=1}^{n}(r_{x_i} - \overline{r}_x)^2}\sqrt{\sum_{i=1}^{n}(r_{y_i} - \overline{r}_y)^2}}.
\end{aligned}$$

Time smo pokazali kako je Spearmanov koeficijent zapravo Pearsonov koeficijent korelacije, ali ne podataka, nego rangova podataka.

Dakle, Spearmanov koeficijent korelacije ranga računamo kao

$$r_s = \frac{\sum_{i=1}^{n}(r_{x_i} - \overline{r}_x)(r_{y_i} - \overline{r}_y)}{\sqrt{\sum_{i=1}^{n}(r_{x_i} - \overline{r}_x)^2}\sqrt{\sum_{i=1}^{n}(r_{y_i} - \overline{r}_y)^2}},$$

gdje su

$$\overline{r}_x = \frac{1}{n}\sum_{i=1}^{n} r_{x_i}, \quad \overline{r}_y = \frac{1}{n}\sum_{i=1}^{n} r_{y_i}.$$

Za testiranje hipoteze o nepostojanju monotone veze između varijabli X i Y:

$$H_0 : \rho_S = 0$$

može se koristiti pristup temeljen na egzaktnoj distribuciji od ρ_S u H_0 ako uzorak nije prevelik (do 1000 u R-u) te ako među podacima $(x_1, x_2, ..., x_n)$ te $(y_1, y_2, ..., y_n)$ nema jednakih. U suprotnom, koristi se asimptotski test za koji je realizacija test statistike

$$\hat{t} = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$$

što odgovara strukturi test statistike kod Pearsonovog koeficijenta korelacije.
Distribucija test statistike, u uvjetima kada je H_0 istinita, može se aproksimirati Studentovom distribucijom s $(n-2)$ stupnja slobode ako je uzorak velik (dovoljno $n > 10$).

Alternativna hipoteza može biti jednostrana:

$$H_1 : \rho_S \neq 0$$

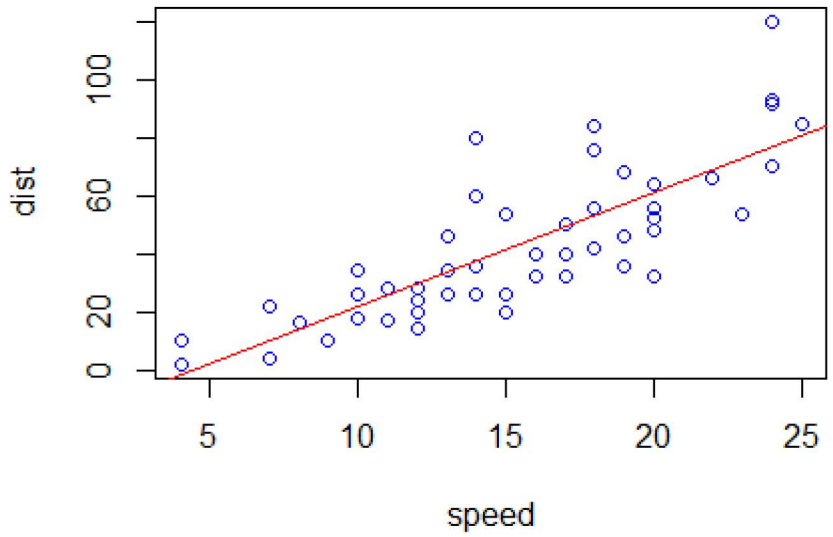
ili dvostrana:

$$H_1 : \rho_S > 0 \qquad H_1 : \rho_S < 0.$$

Sljedeća svojstva važna su za interpretaciju dobivene vrijednosti:

- $r_S \in [-1, 1]$
- vrijednost $r_S \approx 0$ sugerira da ne postoji monotona veza između varijabli X i Y
- vrijednost $r_S \approx 1$ ili $r_S \approx -1$ sugerira da postoji monotona veza između varijabli X i Y
- ako je $r_S < 0$, veza između varijabli X i Y je monotono padajuća
- ako je $r_S > 0$, veza između varijabli X i Y je monotono rastuća.

Primjer 6. *U bazi podataka* cars *iz R paketa* datasets *varijabla* speed *sadrži podatke o izmjerenoj brzini automobila u miljama na sat, a varijabla* dist *podatke o udaljenosti koja je potrebna kako bi se auto zaustavio. Želimo analizirati vezu između te dvije varijable. Pogledajmo najprije dijagram raspršenosti vrijednosti tih varijabli.*



Slika 7: Dijagram raspršenosti vrijednosti varijabli speed i dist

Iz dijagrama uočavamo kako se parovi vrijednosti varijabli speed *i* dist *grupiraju oko pravca koji ima pozitivan koeficijent smjera. Sada želimo izračunati vrijednost Spearmanovog*

koeficijenta korelacije tih dviju varijabli. Dobivena vrijednost je približno 0.8304*, što nam ukazuje na postojanje monotono rastuće veze između varijabli* speed *i* dist*. Želimo to potvrditi provođenjem korelacijskog testa s nul-hipotezom:*

$$H_0 : \rho_S = 0,$$

te alternativnom hipotezom:

$$H_1 : \rho_S > 0.$$

Dobivena p-vrijednost je 4.412×10^{-14} *pa na razini značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. možemo tvrditi kako postoji monotono rastuća veza između varijabli* speed *i* dist*, odnosno kako se porastom brzine automobila povećava i duljina zaustavnog puta.*

4.2 Kendallov koeficijent korelacije ranga

Neka je $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ realizacija jednostavnog slučajnog uzorka iz slučajnog vektora (X, Y). Pretpostavimo da je $x_i \neq x_j$ i $y_i \neq y_j$ za $i \neq j$.

Kendallov koeficijent korelacije ranga bazira se na principu usklađenih parova podataka.

Kažemo da je par podataka (x_i, y_i) i (x_j, y_j) usklađen ako vrijedi jedan od sljedeća dva uvjeta:

- $x_i < x_j$ i $y_i < y_j$
- $x_i > x_j$ i $y_i > y_j$.

U suprotnom kažemo da je par podataka (x_i, y_i) i (x_j, y_j) neusklađen.

Mogućih parova podataka ima $\binom{n}{2} = \frac{n(n-1)}{2}$.

Označimo s n_U broj usklađenih parova, a s n_N broj neusklađenih parova. Tada je procjena za τ:

$$r_K = \frac{n_U - n_N}{\frac{1}{2}n(n-1)}.$$

Uočimo kako je $r_K \in [-1, 1]$. Ako su svi parovi usklađeni, tada je $r_K = 1$. Ako su svi parovi neusklađeni, tada je $r_K = -1$.

Za testiranje hipoteze o nepostojanju monotone veze između varijabli X i Y:

$$H_0 : \tau = 0$$

može se koristiti test statistika s egzaktnom distribucijom ako uzorak nije prevelik (do 50 u R-u) te ako među podacima $(x_1, x_2, ..., x_n)$ te $(y_1, y_2, ..., y_n)$ nema jednakih. U suprotnom, koristi se asimptotski test za koji je realizacija test statistike

$$\hat{z} = \frac{r_K}{\sqrt{\frac{9n(n-1)}{2(2n+5)}}}.$$

Distribucija test statistike, u uvjetima kada je H_0 istinita, može se aproksimirati standardnom normalnom distribucijom.

Alternativna hipoteza može biti jednostrana:

$$H_1 : \tau \neq 0$$

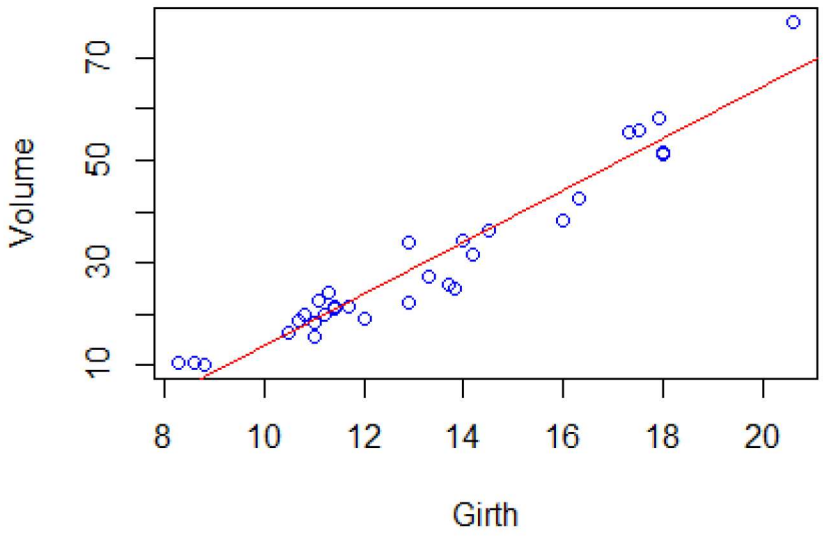
ili dvostrana:

$$H_1 : \tau > 0 \qquad H_1 : \tau < 0.$$

Kendallov koeficijent korelacije ranga $\tau \in [-1, 1]$ daje informaciju o tome u kojoj se mjeri veza između slučajnih varijabli X i Y može opisati monotonom funkcijom tako da vrijedi:

- vrijednost $\tau \approx 0$ sugerira da ne postoji monotona veza između varijabli X i Y
- vrijednost $\tau \approx 1$ ili $\tau \approx -1$ sugerira da postoji monotona veza između varijabli X i Y
- ako je $\tau < 0$, veza između varijabli X i Y je monotono padajuća
- ako je $\tau > 0$, veza između varijabli X i Y je monotono rastuća.

Primjer 7. *U bazi podataka* trees *iz R paketa* datasets *varijabla* Girth *sadrži podatke o promjeru stabla crne višnje u inčima, a varijabla* Volume *podatke o volumenu stabla u kubičnim metrima. Želimo analizirati vezu između te dvije varijable. Pogledajmo najprije dijagram raspršenosti vrijednosti tih varijabli*



Slika 8: Dijagram raspršenosti vrijednosti varijabli Girth i Volume

Iz dijagrama uočavamo kako se parovi vrijednosti varijabli Girth *i* Volume *grupiraju oko pravca koji ima pozitivan koeficijent smjera. Sada želimo izračunati vrijednost Kendallovog koeficijenta korelacije tih dviju varijabli. Dobivena vrijednost je približno* 0.8303 *što nam*

ukazuje na postojanje monotono rastuće veze između varijabli Girth *i* Volume. *Želimo to potvrditi provođenjem korelacijskog testa s nul-hipotezom:*

$$H_0 : \tau = 0,$$

te alternativnom hipotezom:

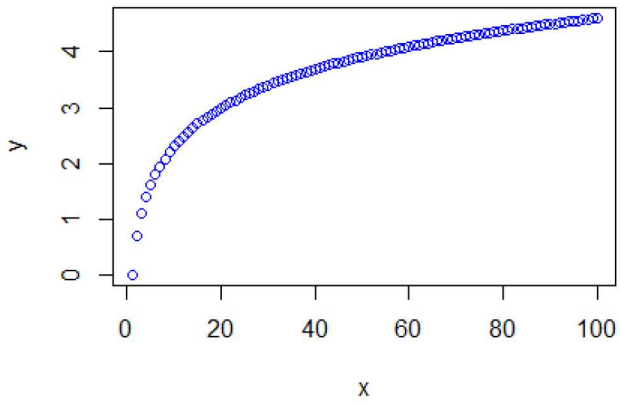
$$H_1 : \tau > 0.$$

Dobivena p-vrijednost je 3.259×10^{-11} *pa na razini značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. možemo tvrditi kako postoji monotono rastuća veza između varijabli* Girth *i* Volume, *odnosno kako se povećanjem promjera stabla povećava i volumen tog stabla.*

4.3 Nelinearne monotone veze

Kako smo ranije naveli, Spearmanov koeficijent korelacije i Kendallov τ prikladni su za korištenje i kada veza između vrijabli nije linearna. U sljedećem primjeru izračunat ćemo vrijednosti tih koeficijenata za neke nelinearne monotone veze.

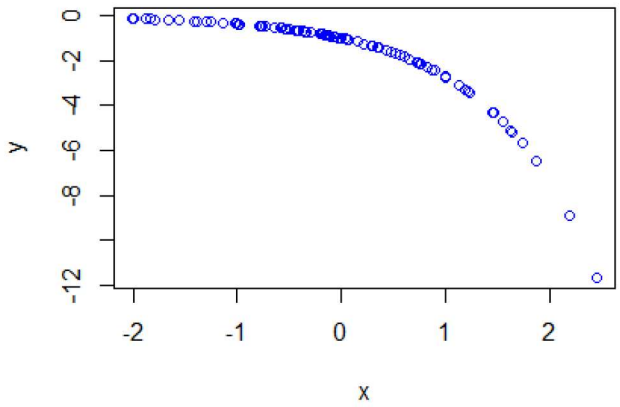
Primjer 8. *Iz dijagrama raspšenosti uočavamo kako veza između varijabli* X *i* Y *nije linearna, ali je monotono rastuća.*



Slika 9: Nelinearna monotono rastuća veza

Vrijednosti Spearmanovog i Kendallovog koeficijenta korelacije iznose 1, što nam potvrđuje postojanje monotono rastuće veze između varijabli X *i* Y.

Na sljedećem dijagramu raspršenosti vidljiva je nelinearna monotono padajuća veza između varijabli X *i* Y.

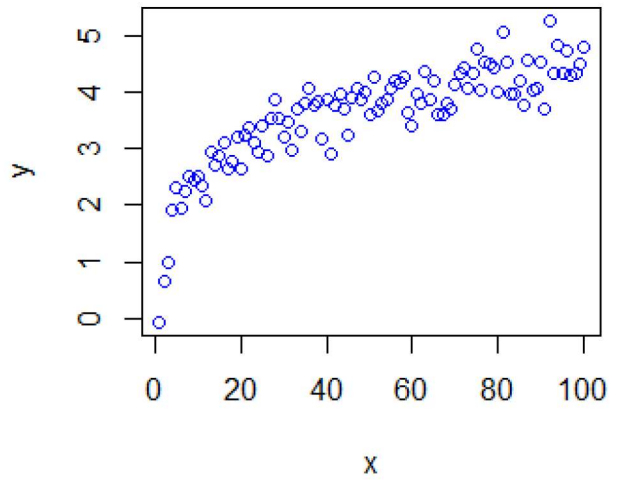


Slika 10: Nelinearna monotono padajuća veza

Vrijednosti Spearmanovog i Kendallovog koeficijenta korelacije iznose -1, što nam potvrđuje postojanje monotono padajuće veze između varijabli X i Y.

U statističkim analizama nije realno očekivati determinističke monotone veze između slučajnih varijabli bez određenih odstupanja. U sljedećem primjeru izračunat ćemo vrijednosti Spearmanovog i Kendallovog koeficijenta korelacije za varijable između kojih je sugerirano postojanje monotone veze.

Primjer 9. *Iz dijagrama raspršenosti možemo naslutiti kako postoji monotono rastuća veza između varijabli X i Y.*



Slika 11: Sugerirana monotono rastuća veza

Vrijednost Spearmanovog koeficijenta korelacije iznosi približno 0.871*, dok je vrijednost Kendallovog koeficijenta korelacije približno* 0.699*, što nam ukazuje postojanje monotono rastuće veze između varijabli* X *i* Y*. Želimo to potvrditi provođenjem korelacijskog testa s nul-hipotezom:*

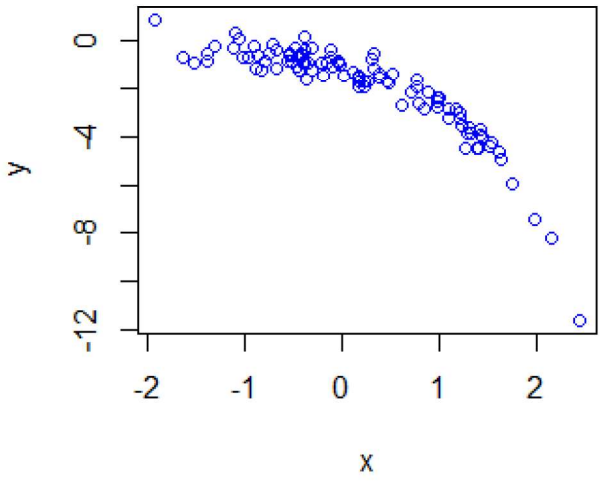
$$H_0 : \rho_S = 0,$$

te alternativnom hipotezom:

$$H_1 : \rho_S > 0.$$

Dobivena p-vrijednost je manja od 2.2×10^{-16} *pa na razini značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. možemo tvrditi kako postoji monotono rastuća veza između varijabli* X *i* Y*.*

Na sljedećem dijagramu raspršenosti naslućujemo postojanje monotono padajuće veze između varijabli X *i* Y*.*



Slika 12: Sugerirana monotono padajuća veza

Vrijednost Spearmanovog koeficijenta korelacije iznosi približno -0.891*, dok je vrijednost Kendallovog koeficijenta korelacije približno* -0.739*, što nam ukazuje na postojanje monotono padajuće veze između varijabli* X *i* Y*. Želimo to potvrditi provođenjem korelacijskog testa s nul-hipotezom:*

$$H_0 : \tau = 0,$$

te alternativnom hipotezom:

$$H_1 : \tau < 0.$$

Dobivena p-vrijednost je manja od 2.2×10^{-16} *pa na razini značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. možemo tvrditi kako postoji monotono padajuća veza između varijabli* X *i* Y*.*

Literatura

[1] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2014.

[2] M. Benšić, N. Šuvak, *Primijenjena statistika*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2013.

[3] D.J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman & Hall/CRC, 2004.

[4] C.J. Kowalski, *On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient*, Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 21, No. 1, pp. 1-12, 1972.

[5] M.G. Kendall, A. Stuart, *The Advanced Theory of Statistics*, Charles Griffin ltd., 1961.

[6] Materijali s predavanja i vježbi iz kolegija Statistički praktikum