Premo sorazmernost človeških mer

Jure Repe
Mentor: prof. dr. Aleksandar Jurišić

12. september 2017

1 Uvod

Večinoma smo si ljudje zelo podobni v številu udov, vsi imamo dvoje ušes, oči, itd. Ali naši udi oz. naše telo raste v povezavi med sabo? Je možno, da ima nekdo izjemno veliko glavo, a izredno majhne roke? Je mogoče predvideti, kako veliko bomo imeli glavo iz dolžine naših dlani? Na ta vprašanja poizkuša vsaj v grobem odgovoriti ta projektna naloga in dati bralcu možnost, da spozna, v kakšnih povezavah raste njegovo telo. Prav ta vprašanja in zanimanje o zakonitostih človeške rasti so tudi glavna motivacija za izvedbo projekta.

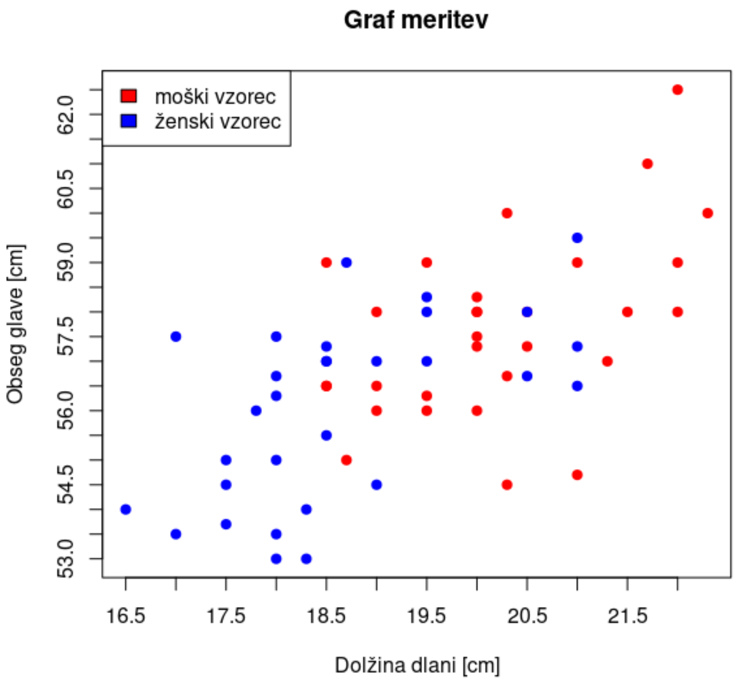
Cilj seminarske naloge je ugotoviti, ali obstaja linearna povezava med človeškimi merami, in sicer med obsegom glave in dolžino dlani. Z regresijsko analizo lahko nato poizkusimo napovedati, kako veliko glavo ima posameznik glede na podatek o velikosti oz. dolžini dlani.

O drugih telesnih merah že imamo raziskave, ki si postavljajo podobna vprašanja, npr. o velikosti človeka in razponu rok [3]. Ali o velikosti noge in velikosti človeka [4]. Z uporabo rezultatov takih raziskav lahko nato že zelo kmalu najdemo ljudi, ki so mogoče nove zvezde košarke, odbojke ali drugih športov, kjer človeške mere dajo igralcem prednost ali pa predstavljajo oviro. Prav tako bi jih lahko uporabljali v medicini in z njimi zgodaj opazili kakšne anomalije, ki bi odstopale od običajnih mer.

V projektni nalogi najprej predstavim zbrane podatke in način, na katerega sem jih pridobil. Temu sledi poglavje z izračuni o vzorčnem povprečju ter standardnem odklonu. Nato izračunam korelacijo in preverim domnevo s testno statistiko ter predstavim ugotovitve računov in v naslednjem poglavju izračunam še regresijsko premico, preverim domnevo o regresijskem koeficientu β ter regresijsko premico predstavim na grafu. Na koncu povzamem še rezultate in na njih osnovane ugotovitve.

2 Pridobivanje podatkov

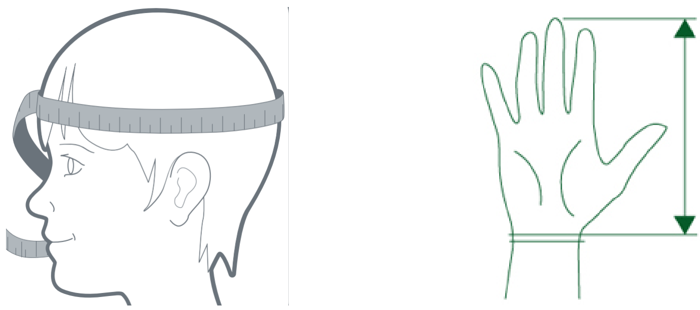
Podatke, potrebne za izvedbo naloge, sem pridobil od mojih vrstnikov na fakulteti, zboru, ki ga obiskujem ter na gimnaziji, ki sem jo obiskoval, tako da je starost udeležencev med 18 in 22 let. Udeleženci so bili izbrani ne glede na lastnosti njihove telesne konstitucije (teža, velikost, itd.). Velikost priročnega vzorca je 60, in sicer 30 za moški in 30 za ženski spol. Vsakemu udeležencu sem izmeril dolžino dlani (X) ter obseg glave (Y), predstavljeno na grafu 1.



Slika 1: Zbrane meritve, prikazane na grafu

Za merjenje dolžin sem uporabil klasičen šiviljski meter. Dlan sem meril od konca zapestja, oziroma "pregiba roke", do konca srednjega prsta. Obseg glave sem meril okrog čela nad obrvmi ter vodoravno okrog glave. Pri merjenju se pojavijo tri merske napake.

Pri dlani sem začetno točko merjenja ocenil približno, zato sem nekomu lahko izmeril malo višje, drugemu malo nižje. Pri merjenju glave se pojavi enak problem glede položaja šiviljskega traku, poleg tega pa je še dodatna napaka zaradi količine las. Gostota in dolžina las vplivata na obseg, še posebej pri moškem vzorcu, kjer se pojavlja največja razlika med posamezniki (brez las, kratki lasje, dolgi lasje).



Slika 2: Pozicija merjenja obsega glave in dolžine dlani (Vir, 3.8.2017: [6])

3 Vzorčno povprečje, standardni odklon

Naj bo $x_1, ..., x_n$ vzorec velikosti n, kjer x_i predstavlja posamezno meritev. Vzorčno povprečje se izračuna po formuli

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

in popravljeni vzorčni standardni odklon po formuli

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}.$$

Tej formuli najbolj ustrez nepristranska cenilka s za odklon, glej cenilko [1, str. 146]. Vzorčno povprečje in popravljeni vzorčni standardni odklon sem izračunal znotraj programskega jezika R s funkcijo mean() in sd(). V tabeli so predstavljeni izračuni za moški in ženski spol, kot tudi za združen vzorec obeh spolov.

spremenljivka	$\bar{x}$ [moški]	s [moški]	$\bar{x}$ [ženske]	s [ženske]	$\bar{x}$ [skupno]	s [skupno]
obseg glave	57,65	1,814	56,06	1,840	56,58	1,982
dolžina dlani	20,21	1,140	18,62	1,224	19,42	1,422

Tabela 1: Povprečje in standardni odklon za moški, ženski in skupni vzorec

Iz podatkov zgornje tabele lahko razberemo, da je povprečna velikost obsega glave moškega v priročnem vzorcu večja od ženske za 1,6 centimetra. Prav tako to velja za povprečno velikost dolžine dlani moškega v vzorcu, ki je tudi večja od povprečne dolžine ženske dlani v vzorcu za 1,6 centimetra. Razlika med povprečno velikostjo obsega glave in dolžine dlani med moškimi in ženskami v vzorcu je pričakovana, saj ima povprečen moški močneje grajeno konstitucijo, torej je večji, težji in fizično močnejši od povprečne ženske. Seveda obstajajo izjeme, nekaj se jih je pojavilo tudi med zbranimi podatki.

4 Korelacija

S koeficientom korelacije lahko na vzorcu preverimo, kako močna je linearna povezava med dvema slučajnima spremenljivkama [2, str. 513]. Izračunamo Pearsonov koeficient korelacije, ki je definiran s formulo

$$r_{XY} = \frac{k(X,Y)}{s_X * s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Pearsonov koeficient korelacije zavzame vrednost na intervalu $[-1, 1]$ [1, str. 190]. Rezultat je predznačen glede na to ali vrednosti naših podatkov padajo ali naraščajo. Bolj ko se vrednost koeficienta bliža robovom intervala, močneje sta ustrezni slučajni spremenljivki linearno povezani. V kolikor je vrednost koeficienta korelacije blizu 0, slučajni spremenljivki nista korelirani. V programskem jeziku R korelacijski koeficient izračunamo s funkcijo cor().

Velikost koeficienta	Interpretacija
0,90 do 1,0 (−0,90 do −1,0)	Zelo močna pozitivna (negativna) korelacija
0,70 do 0,90 (−0,70 do −0,90)	Močna pozitivna (negativna) korelacija
0,50 do 0,70 (−0,50 do −0,70)	Zmerna pozitivna (negativna) korelacija
0,30 do 0,50 (−0,30 do −0,50)	Šibka pozitivna (negativna) korelacija
0,00 do 0,30 (−0,00 do −0,30)	Zanemarljiva korelacija

Tabela 2: Tabela interpretacije vrednosti korelacijskega koeficienta [5]

Za interpretacijo vrednosti Pearsonovega koeficienta korelacije sem uporabil zgornjo tabelo, ki razdeli interval od $[1, -1]$ na več manjših intervalov, s katerimi nato določimo, kako močna je korelacija med dvema slučajnima spremenljivkama.

r [moški]	r [ženske]	r [skupno]
0,4980	0,5725	0,6338

Tabela 3: Korelacijski koeficient za moški, ženski in skupni vzorec

Rezultat izračuna korelacije za vzorce se giblje okoli 0,56 (0,49 za moške ter 0,57 za ženske in 0,63 za skupni vzorec). Vrednost se nahaja blizu sredine med 0 in 1, kjer 0 pomeni, da slučajni spremenljivki nista korelirani, 1 pa pomeni, da sta zelo močno korelirani. Ker se nahaja v sredini intervala, na podlagi Pearsonovega koeficienta korelacije in zgornje tabele za interpretacijo vrednosti koeficienta, lahko rečemo, da sta slučajni spremenljivki Y (obseg glave) in X (dolžina dlani) med seboj zmerno pozitivno linearno povezani. V naslednjem poglavju bomo izračunali še testno statistiko, s katero bomo lahko bolj natančno preverili, če povezava drži glede na stopnjo značilnosti, ki si jo bomo izbrali.

4.1 Preverjanje linearne povezave

Da se bolje prepričamo o linearni povezavi med slučajnima spremenljivkama, bomo le-to preverili še s testno statistiko, ki ima formulo

$$\mathrm{TS} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

in se porazdeljuje po t porazdelitvi z $n-2$ prostorskimi stopnjami [2, str.515].

Najprej nastavimo ničelno (H_0 - nista linearno povezani) in osnovno (H_1 - sta linearno povezani) domnevo. Povezanost med obsegom glave in dolžino dlani bomo preverili pri 1% stopnji značilnosti. Stopnja značilnosti predstavlja tveganje, da zavrnemo ničelno domnevo, čeprav je ta pravilna [2, str. 372]. Torej, če imamo stopnjo značilnosti 1%, potem lahko z 99% gotovostjo trdimo, da je ničelna domneva res napačna in se nismo zmotili.

Izračun testne statistike TS za skupni vzorec

Velikost skupnega vzorca je 60, zato imamo 58 prostorskih stopenj. Izračunamo TS:

$$\text{TS} = \frac{0{,}6338 * \sqrt{58}}{\sqrt{1 - 0{,}6338^2}} = 6{,}2403.$$

Kritično območje dvostranskega testa pri 1% stopnji značilnosti je določeno z vrednostjo

$$t_{\frac{\alpha}{2}}(n-2) = \pm t_{0{,}005}(58) = \pm 2{,}663,$$

p-vrednost pa je

$$\text{p-vrednost} = 5{,}46 * 10^{-8}.$$

Izračun testne statistike TS za moški in ženski vzorec

Velikost posameznega vzorca je 30, torej imamo 28 prostorskih stopenj. Izračunajmo TS in p-vrednost še za vsak spol posebej:

$$\text{TS}_{\text{moški}} = \frac{0{,}4980 * \sqrt{28}}{\sqrt{1 - 0{,}4890^2}} = 3{,}0387, \qquad \text{TS}_{\text{ženske}} = \frac{0{,}5725 * \sqrt{28}}{\sqrt{1 - 0{,}5725^2}} = 3{,}6948.$$

Kritično območje je določeno z vrednostjo

$$t_{\frac{\alpha}{2}}(n-2) = \pm t_{0{,}005}(28) = \pm 2{,}76,$$

p-vrednost pa

$$\text{p-vrednost}_{\text{moški}} = 0{,}0051, \qquad \text{p-vrednost}_{\text{ženske}} = 0{,}00095.$$

P-vrednost sem za vzorce izračunal s programskim jezikom R in ukazom pt(), ki sprejme rezultat testne statistike (6,2403, 3,0387 in 3,6948) in prostorsko stopnjo (58 in 28) kot argument. Ker imamo dvostranski test, rezultat še pomnožimo z 2 in dobimo zgornje rezultate.

Izračunamo jo zato, ker p-vrednost za specifičen statistični test predstavlja verjetnost (če predpostavimo da H_0 drži), da kot rezultat testne statistike dobimo vrednost, ki je vsaj tako nasprotujoča ničelni hipotezi, kot ta, ki smo jo izračunali iz podatkov [2, str. 383]. Glede na p-vrednost ponavadi zavrnemo ničelno hipotezo takrat, ko je p-vrednost manjša od stopnje značilnosti α, ki smo si jo izbrali za test.

Ker se vrednosti testnih statistik za skupni, moški in ženski vzorec nahajajo v kritičnem območju pri 1% stopnji značilnosti, poleg tega pa je tudi izračunana p-vrednost manjša od stopnje značilnosti, lahko zavrnemo H_0 za oba spola ter skupni vzorec in rečemo, da sta vrednosti "obseg glave" ter "dolžina dlani" linearno povezani pri 1% stopnji značilnosti.

5 Regresijska premica

Regresijska premica $Y = \alpha + \beta X$ nam pokaže, kako na spremenljivko Y vpliva spremenljivka X. V našem primeru, kako na obseg glave vpliva dolžina dlani. Z regresijsko analizo lahko tudi napovemo, kako se bodo podatki obnašali pri vrednostih, ki jih trenutno še ne poznamo. Regresijsko premico sem za podatke izračunal z uporabo ukaza abline(A, B) v jeziku R, kjer je

$$B = r * \frac{s_Y}{s_X} \quad , \qquad A = \bar{y} - B * \bar{x}$$

in je r koeficient korelacije (izračunan v prejšnjem poglavju) ter A in B cenilki za koeficienta α in β [1, str. 195] in s_X ter s_Y nepristranski cenilki za popravljeni vzorčni standardni odklon. Regresijska premica je izračunana in predstavljena posamezno na grafu združenega vzorca podatkov [3] ter ločeno za moški [4] in ženski spol [5] in nazadnje še na skupnem grafu, kjer sta predstavljeni obe premici vzorcev [6].

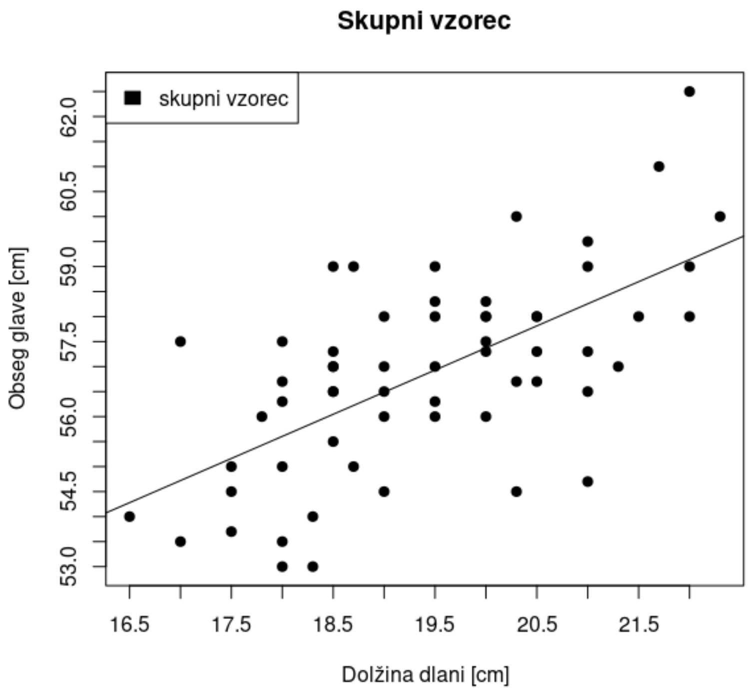
Izračun regresijske premice za skupni vzorec

Vrednosti A in B za skupni vzorec

$$B = 0{,}884, \qquad A = 39{,}701,$$

enačba regresijske premice pa je

$$Y = 0{,}884x + 39{,}701.$$



Slika 3: Regresijska premica na grafu skupnega vzorca

Izračun regresijske premice za moški in ženski vzorec

Vrednosti A in B za moški spol

$$B_{\text{moški}} = 0{,}792, \qquad A_{\text{moški}} = 41{,}640$$

ter za ženski spol

$$B_{\text{ženske}} = 0{,}861, \qquad A_{\text{ženske}} = 40{,}031.$$

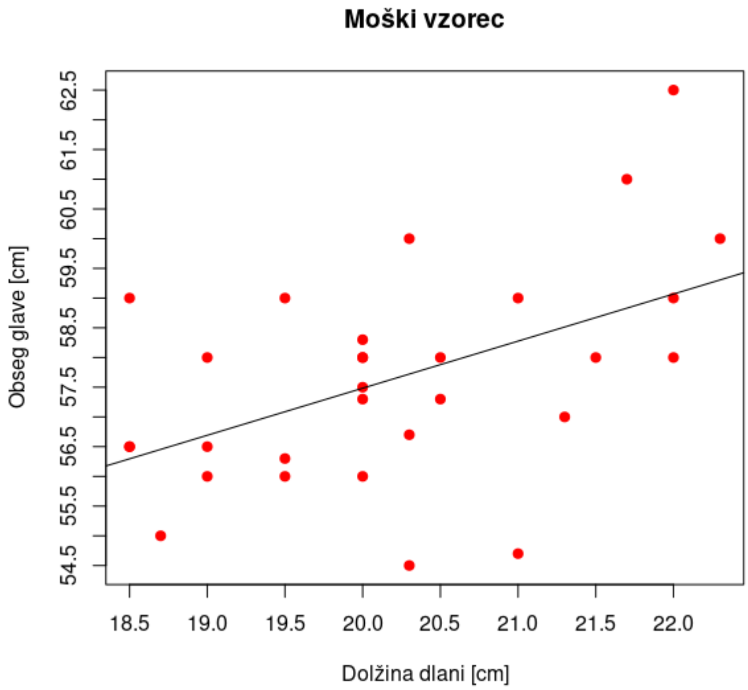
Zapišimo še enačbo regresijske premice

$$Y_{\text{moški}} = 0{,}798x + 41{,}640, \qquad Y_{\text{ženske}} = 0{,}861x + 40{,}031$$

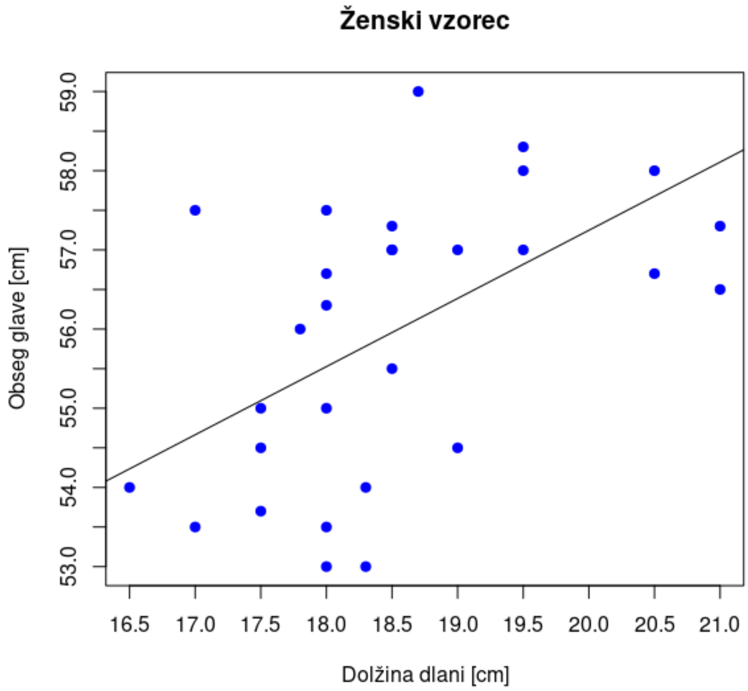
Ker imamo zapisano enačbo premice, lahko sedaj tudi poizkusimo napovedati, kako velik obseg glave ima naprimer ženska, ki ima dlan dolgo 19 cm. Podatek o dlani vstavimo v enačbo

$$Y = 0{,}861 * 19 + 40{,}031$$

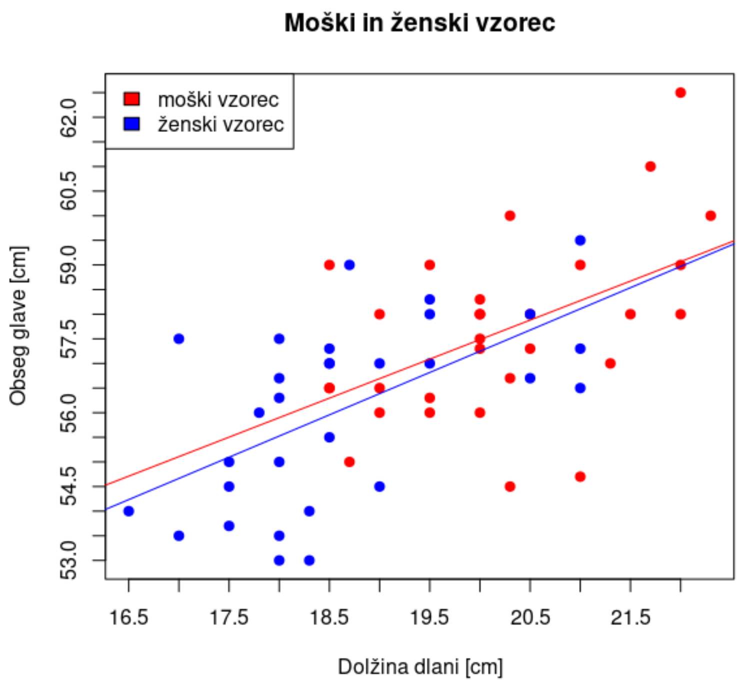
in dobimo rezultat, da je obseg njene glave enak 56,4 cm, ki se tudi sklada z ostalimi zbranimi podatki (npr. dva druga obsega glave za dlan dolžine 19 cm sta 54,5 cm in 57 cm).



Slika 4: Regresijska premica na grafu moškega vzorca



Slika 5: Regresijska premica na grafu ženskega vzorca



Slika 6: Skupen graf regresijskih premic za moški in ženski vzorec

5.1 Preverjanje regresijskega koeficienta

Preverimo še domnevo o regresijskem koeficientu β. Formula za testno statistiko je

$$\mathrm{TS} = \frac{s_Y\sqrt{n-2}}{s_X\sqrt{1-r^2}}(b-\beta_0)$$

in se porazdeljuje po t porazdelitvi z $n-2$ prostorskimi stopnjami [1, str. 197].

Izračun testne statistike TS za skupni vzorec

Postavimo ničelno hipotezo $H_0 : \beta = 0{,}9$ in osnovno domnevo $H_1 : \beta \neq 0{,}9$.

$$\mathrm{TS} = \frac{1{,}982 * \sqrt{58}}{1{,}422 * \sqrt{1-0{,}6338^2}}(0{,}884-0{,}9) = -0{,}22.$$

Pri 5% stopnji značilnosti je kritično območje dvostranskega testa določeno z vrednostjo

$$\pm t_{0{,}025}(58) = \pm 2{,}002,$$

p-vrednost pa znaša

$$\text{p-vrednost} = 0{,}827.$$

Ker se rezultat testne statistike ne nahaja v kritičnem območju in je tudi p-vrednost večja od stopnje značilnosti α, ne moremo zavrniti H_0.

Izračun testne statistike TS za moški in ženski vzorec

Tudi tokrat postavimo ničelno hipotezo $H_0 : \beta = 0{,}9$ in osnovno domnevo $H_1 : \beta \neq 0{,}9$.

$$\mathrm{TS}_{\text{moški}} = \frac{1{,}814 * \sqrt{28}}{1{,}140 * \sqrt{1-0{,}4980^2}}(0{,}792-0{,}9) = -1{,}05,$$

$$\mathrm{TS}_{\text{ženske}} = \frac{1{,}840 * \sqrt{28}}{1{,}224 * \sqrt{1-0{,}5725^2}}(0{,}861-0{,}9) = -0{,}38.$$

Pri 5% stopnji značilnosti je kritično območje dvostranskega testa določeno z vrednostjo

$$\pm t_{0{,}025}(28) = \pm 2{,}05.$$

P-vrednost za testno statistiko je

$$\text{p-vrednost}_{\text{moški}} = 0{,}303, \qquad \text{p-vrednost}_{\text{ženske}} = 0{,}707.$$

Tudi tu se rezultati testne statistike ne nahajajo v kritičnem območju in je p-vrednost prav tako večja od stopnje značilnosti α, zato ne moremo zavrniti H_0.

6 Zaključek

Z vrednostjo Pearsonovega koeficienta korelacije in s tabelo, s katero smo to vrednost interpretirali, smo ugotovili, da sta slučajni spremenljivki X(dolžina dlani) in Y(obseg glave) izbranega priročnega vzorca zmerno linearno povezani, torej med njima obstaja linearna povezava. To ugotovitev smo tudi podprli z izračunom testne statistike TS, saj smo zavrnili ničelno hipotezo H_0, ker je bila vrednost TS v kritičnem območju, prav tako pa je bila tudi p-vrednost manjša od stopnje značilnosti. Te ugotovitve veljajo za vse tri primere, ki smo jih obravnavali, torej skupen vzorec ($n = 60$) ter posamezno moški in ženski vzorec ($n = 30$). Izračunali smo tudi enačbo regresijske premice in z njo poskušali napovedati velikost obsega glave glede na dolžino roke. Hkrati smo regresijsko premico predstavili na grafih vzorcev.

Potrebno je omeniti, da je izbran vzorec priročen, zato te ugotovitve ne moremo posplošiti na celotno populacijo, saj priročen vzorec ne poda dobre predstavitve širše populacije, ker so bili podatki izbrani glede na dostopnost in ne na poseben način, ki bi dobro opredelil širšo populacijo.

Za izdelavo seminarske naloge in obdelave podatkov sem uporabil programski jezik R in se spoznal z metodami zbiranja, analize in obdelave podatkov.

Literatura

[1] A. Jurišić. Verjetnostni račun in statistika, skripta, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2015.

[2] W. M. Mendenhall, T. L. Sincich. Statistics for Engineering and the Sciences, 6th edition, Taylor & Francis Group, LLC, 2016.

[3] SP. Mohanty, S. Suresh Babu, N. Sreekumaran Nair. The Use of Arm Span as a Predictor of Height: A Study of South Indian Women, Kasturba Medical College and Hospital, Manipal, Karnataka, India, 2001.

[4] A. K. Agnihotri, B. Purwar, K. Googoolye, S. Agnihotri, N. Jeebun. Estimation of stature by foot length, SSR Medical College, Department of Forensic Medicine and Toxicology, Belle-Rive, Mauritius, 2007.

[5] M. M. Mukaka. A Guide to Appropriate Use of Correlation Coefficient in Medical Research, MMJ, 2012.

[6] Slike:
Kates Kaps, 2015, Source: KatesKapsBlog,
ErgoVancouver, 2015, Source: ErgoVancouver.