Povezanost veličine populacije te bruto domaćeg proizvoda i uspjeha na Olimpijskim igrama u Riju 2016. godine

Krčmar, Marko

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Marko Krčmar

POVEZANOST VELIČINE POPULACIJE TE BRUTO DOMAĆEG PROIZVODA I USPJEHA NA OLIMPIJSKIM IGRAMA U RIJU 2016.

ZAVRŠNI RAD

Varaždin, 2019.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Marko Krčmar

Matični broj: 43248/14–R

Studij: Poslovni sustavi

POVEZANOST VELIČINE POPULACIJE TE BRUTO DOMAĆEG PROIZVODA I USPJEHA NA OLIMPIJSKIM IGRAMA U RIJU 2016.

ZAVRŠNI RAD

Mentor/Mentorica:

Jelena Gusić, mag. math.

Varaždin, srpanj 2019.

Marko Krčmar

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Tema rada je iz dostupnih demografskih i ekonomskih pokazatelja zemalja te uspjeha na Olimpijskim igrama u Riju 2016. izraditi regresijski model koji će se kasnije moći koristiti u predikcijske svrhe. Kako bi se to postiglo prvo se definiraju pojmovi korelacije i regresije te se detaljnije obrađuju jednostavna i višestruka regresija. Pokazuje se koliko transformacija varijabli može povećati točnost predikcije te se definiraju kriteriji koji će biti važni prilikom odabira najboljeg regresijskog modela. Definiraju se odabrani ekonomski i demografski pokazatelji te način na koji se mjeri uspjeh pojedine države na Olimpijskim igrama 2016. godine. Nakon raznih testiranja odabran je najreprezentativniji regresijski model na temelju kojeg je napravljena prognoza rezultata za sljedeće Olimpijske igre koje će se održati 2020. godine u Tokiju.

Ključne riječi: regresijska analiza, obrada podataka, statistika u sportu, korelacija

Sadržaj

1. Uvod

Tema ovog završnog rada je analiza rezultata s Olimpijskim igrama 2016. u Riju. temeljem linearne regresijske analize. Kroz rad se upoznaje s osnovnim podacima i primjerima linearne regresije koji olakšavaju razumijevanje istraživanje koje je obrađeno na kraju rada. Teorijski dio rada prolazi kroz jednostavni i višestruki oblik regresijske analize. Nakon tog slijedi analiza država i njihovih uspjeha na Olimpijskim igrama u Riju 2016. te kako na taj uspjeh utječu demografski i ekonomski pokazatelji tih istih država. Analiza će biti sastavljena od svih obrađenih načela regresijske analize u prvom poglavlju i sadržavati sve službeno prikupljene ulazne podatke te sve izlazne podatke koje izbacuje korišteni alat.

Cilj ovog rada je naći što reprezentativniji regresijski model. Kako bi to postigli neophodno je testiranje više regresijskih modela s različitim kombinacijama nezavisnih varijabli čije su vrijednosti pribavljene sa službene web stranice Svjetske banke (eng. The World Bank). Razrađen je utjecaj raznih faktora na olimpijski uspjeh država od kojih je jako puno teško mjerljivih faktora. Tijekom rada je testirano nekoliko modela višestruke linearne regresije, te je nakon usporedbe rezultata dana preporuka o najboljem modelu. Nakon odluke o najboljem regresijskom modelu provesti će se kratka prognoza uspjeha na sljedećim Olimpijskim igrama koje će se 2020. održati u Tokiju.

2. Regresija i korelacija

Svakodnevno se odvijaju masovne pojave među kojima postoje međusobni utjecaji tako da promjena jedne ili više promjena povlači za sobom promjene nekih drugih ovisnih pojava. U slučaju da rezultat ispitivanja pojave $\boldsymbol{A}$ ima za posljedicu promjenu pojave $\boldsymbol{B}$, tj. ako vrijedi funkcionalan izraz $A = f(B)$ govori se o regresijskom modelu te regresijskoj analizi. Isto tako, postoje slučajevi kod kojih vrijedi konstatacija da je svejedno koja se relacija napiše, tj. da je $A = f(B)$ isto što i $B = f(A)$. Kada postoji takav slučaj, on uvijek rezultira pojavom korelacijskog modela tj. korelacijske analize. (Kero K..,2003).

- **korelacija** – mjera povezanosti dvije ili više varijabli, *Karl Pearson*
- **regresija** – definira oblik povezanosti dvaju ili više varijabli, *Francis Galton*

(„Fakultet strojarstva i brodogradnje [FSB]“, 2012)

Primjenom spomenutih metoda, vrlo je važno da definiramo veze po tri različita kriterija: **oblik**, **smjer i jakost.**

- Ovisno o obliku postoje **linearne** ili **krivolinijske** veze
- Ovisno o smjeru postoje **pozitivne** ili **negativne** veze
- Ovisno o jakosti (intenzitetu) postoje **funkcionalne** ili **statističke** veze

Linearni oblik veze prepoznajemo u modelu kada promjena jedne pojave za jedinicu mjere veže za sobom promjenu druge pojave za određeni jednaki iznos. S druge strane, u slučaju kada promjena jedne pojave ne povlači promjene jednakim iznosima druge pojave kažemo da je između dvije pojave prisutna krivolinijska veza. Pozitivan smjer veze primjećujemo u slučaju kada rast jedne pojave povlači porast druge pojave ili kad pad jedne pojave prati pad druge pojave. S druge strane, negativan smjer veze između pojava primjećujemo kada jedna pojava svojom kretnjom ukazuje na porast, a druga pad ili obrnuto. (Kero K.,2003).

Moguće je provesti korelacijsku analizu i u slučaju da se radi o kvalitativnim varijablama. Koriste se koeficijent korelacije ranga kako bi se izmjerio stupanj povezanosti između pojava. Kada se radi analiza kvalitativnih varijabla potrebno je rangirati elemente od „najgoreg“ prema „najboljem“ ovisno o njegovim svojstvima koje su dio analize. Postoje dvije vrste koeficijenta ranga : Spermanov koeficijent i Kendallov koeficijent. (Dumičić K., Bahovec V.,2011).

2.1. Pearsonov koeficijent korelacije

„Pearsonov ***koeficijent korelacije*** ili ***koeficijent jednostavne linearne korelacije*** primjenjuje se u slučaju linearne povezanosti te u analizi samo dviju pojava predočenih varijablama x i y. To je brojčani pokazatelj kojim se mjeri stupanj jakosti i smjer statističke linearne povezanosti između promatranih varijabli.“ (Dumičić K., Bahovec V.,2011, str. 324).

Prije nego se definira način na koji se računa koeficijent korelacije, potrebno je izračunati kovarijancu varijabli x i y. „Kovarijanca je aritmetička sredina umnožaka odstupanja vrijednosti varijable x od njezine aritmetičke sredine i vrijednosti varijable y od njezine aritmetičke sredine. (Dumičić K., Bahovec V.,2011).

Kovarijanca varijabli x i y se time računa prema sljedećem izrazu:

$$Cov(x,y) = \frac{(\sum_{i=1}^{n} x_i - \bar{x}) \cdot (\sum_{i=1}^{n} y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{n-1}.$$

Kada je vrijednost kovarijance poznata, slijedi izračun Pearsonovog koeficijenta korelacije uz pomoć jednadžbe koja slijedi:

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

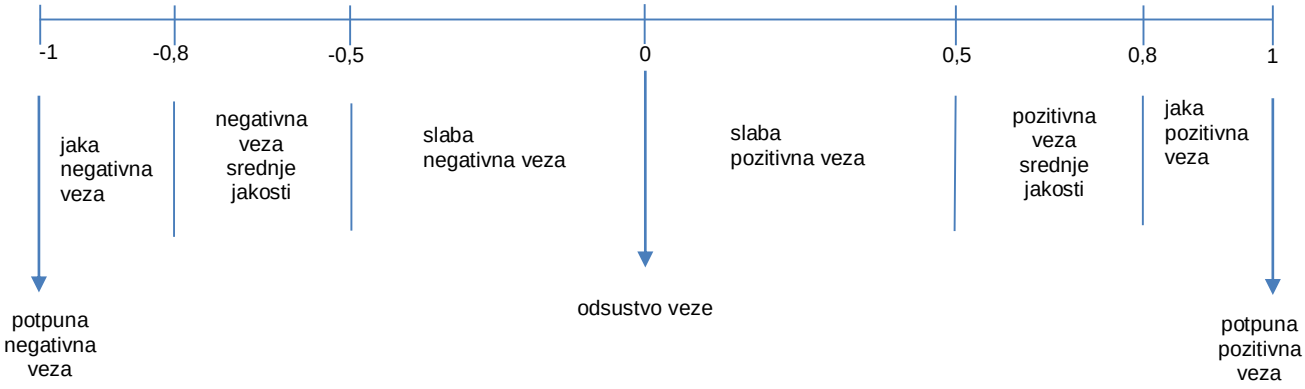
s druge strane, razvijeni oblik jednadžbe glasi:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x^2{}_i - n\overline{x^2}} \sqrt{\sum_{i=1}^{n} y^2{}_i - n\overline{y^2}}}.$$

Pearsonov koeficijent korelacije može poprimiti vrijednosti iz sljedećeg intervala:

$$-1 \leq r \leq 1.$$

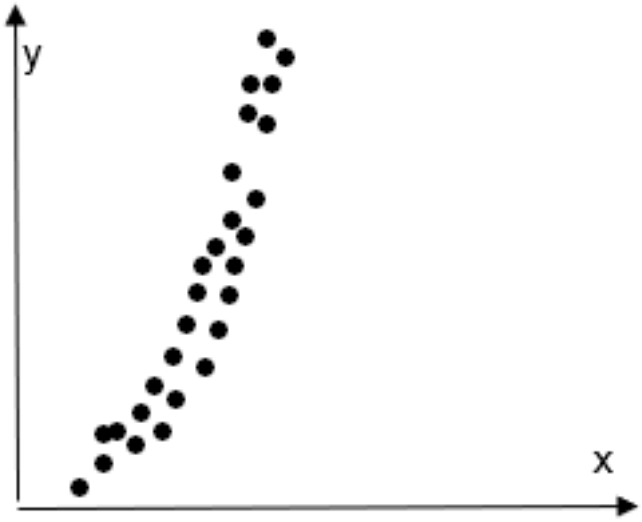
Što je koeficijent bliže graničnim vrijednostima to je veza među pojavama jača, samo što su smjerovi veze različiti. U slučaju da koeficijent iznosi +1 ili -1, to ukazuje na potpunu linearnu povezanost među pojavama. S druge strane, kada je koeficijent jednak 0 u tom linearna povezanost nije prisutna između pojava. Predznak koeficijent definira smjer veze između pozitivne i negativne veze čije su funkcije objašnjene u ranijem poglavlju. (Dumičić K., Bahovec V.,2011).



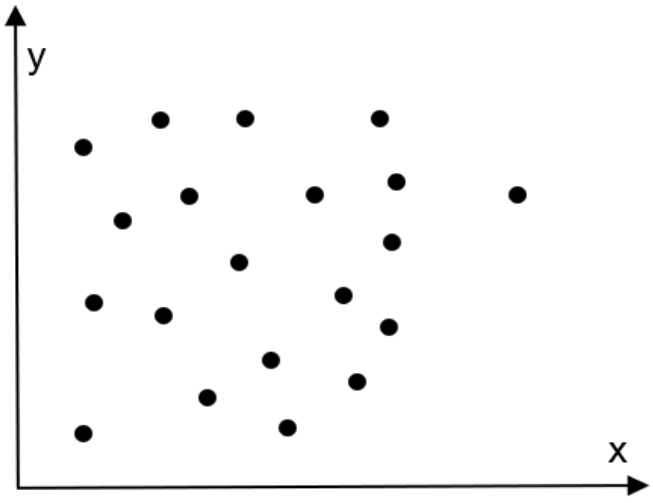
Slika 1. Grafički prikaz vrijednosti koeficijenta korelacije (Izvor: Dumičić K., Bahovec V.,2011)

2.2. Dijagram rasipanja

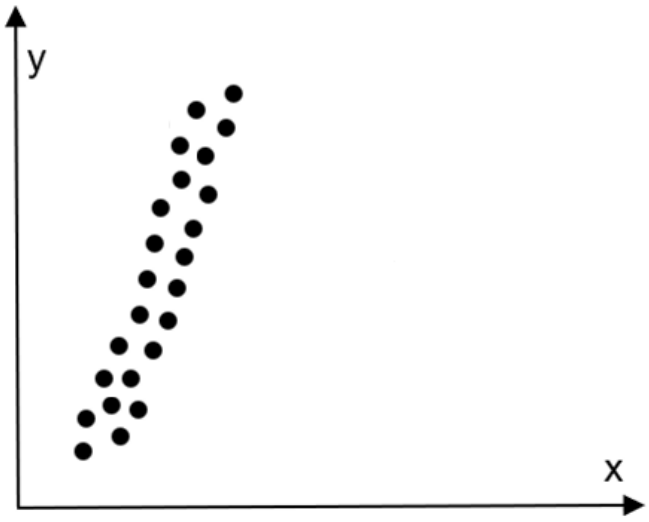
U slučaju primjene korelacijske i regresijske analize, uvijek se kreće od polazne točke koja je **dijagram rasipanja.** To je grafički prikaz točaka u koordinatnom sustavu (najčešće I. kvadrant) koje predstavljaju niz uređenih parova: $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$. Pri čemu su x_1, x_2, x_n vrijednosti varijable (X), a y_1, y_2, y_n vrijednosti varijable (Y). S obzirom na raspored točaka unutar dijagrama zaključujemo o kakvom obliku, smjeru i jakosti veze je riječ. (Veleučilište u Rijeci [VELERI], bez dat.)



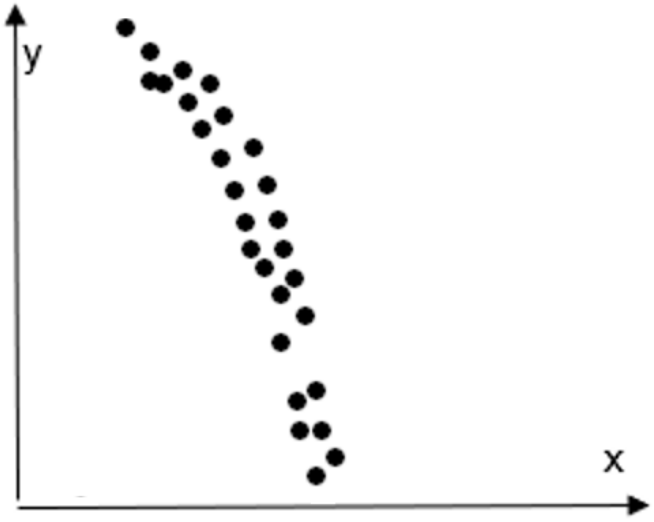
Slika 2. DR - linearna veza [1]



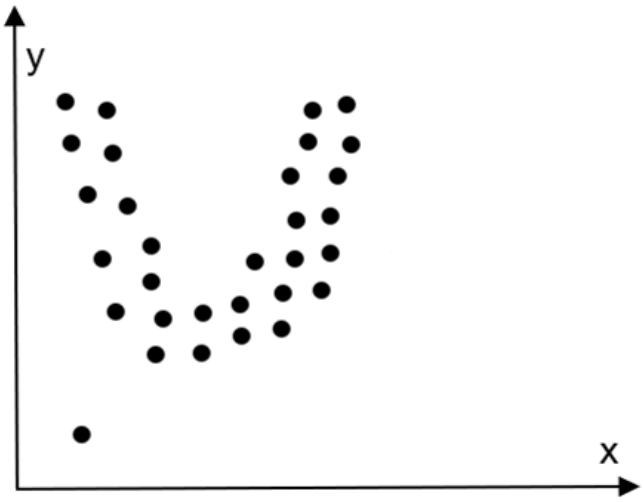
Slika 3. DR - odsutnost veze [1]



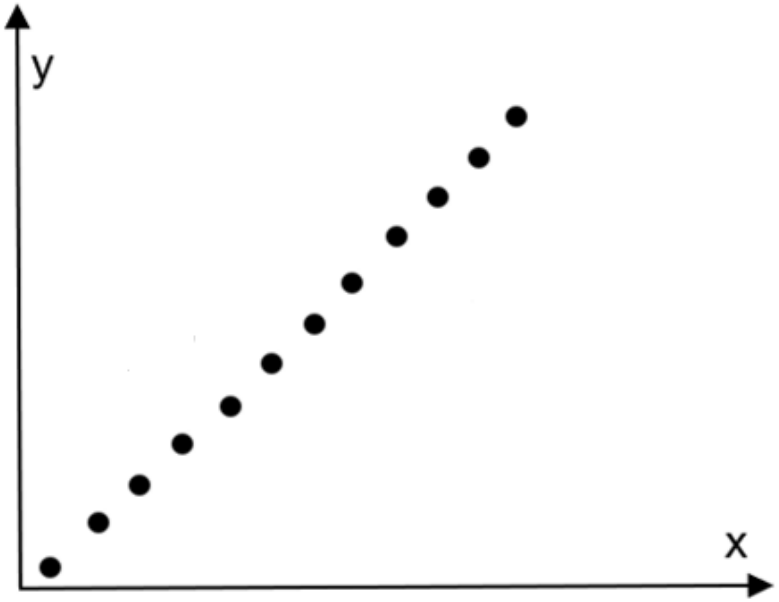
Slika 4. DR - statistička pozitivna veza [1]



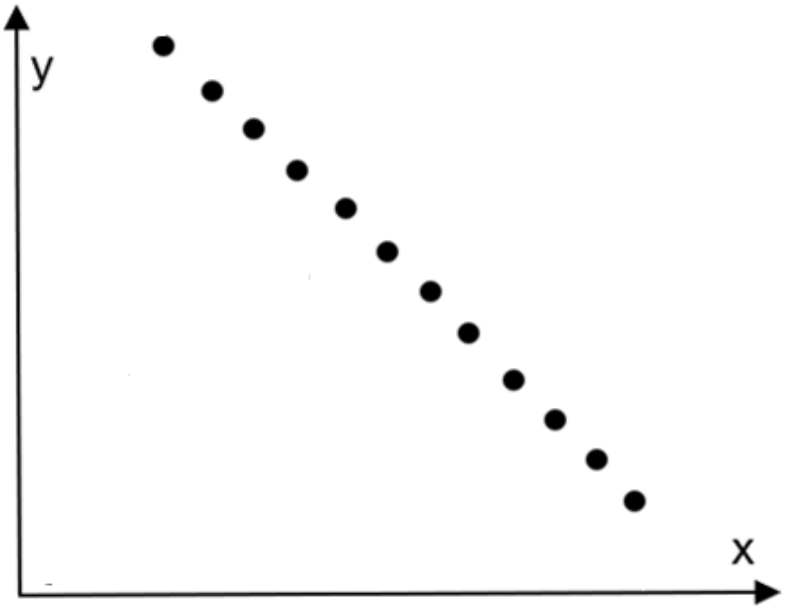
Slika 5. DR - statistička negativna veza [1]



Slika 6. DR – nelinearna veza [1]



Slika 7. DR – funkcionalna pozitivna veza [1]



Slika 8. DR - funkcionalna negativna veza [1]

3. Regresijska analiza

Regresijska analiza je matematičko-statistički postupak kojim se utvrđuje odgovarajuća veza (relacija) između jedne zavisne varijable i jedne ili više nezavisnih varijabli. Cilj analiziranja veze među varijablama je utvrditi statističku ovisnost istih varijabli. Regresijska analiza se vrlo često koristi za predviđanje i prognoziranje ishoda neke pojave. („Kineziološki fakultet [KIF]“, bez dat.)

„Regresijska analiza modela uključuje ocjenjivanje nepoznatih parametara, izračunavanje mjere disperzije i drugih statističko-analitičkih pokazatelja, te primjenu postupaka kojima se ispituje kvaliteta dobivenih rezultata s obzirom na polazne pretpostavke o modelu i svojstvima varijabli u njemu (regresijska dijagnostika).“ (Šošić I., Serdar V., 1995).

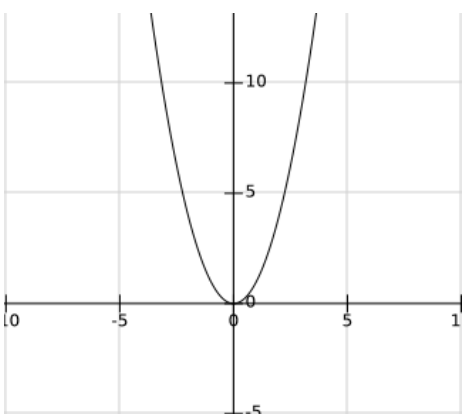
Ranije je već spomenuto da odnos među pojavama može biti funkcionalni i statistički.

3.1. Funkcionalna veza

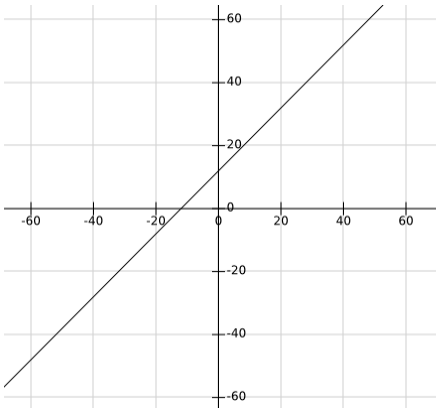
Funkcionalna veza između dvije varijable zadana je oblikom

$$y = f(x)$$

Gdje je y zavisna varijabla, x nezavisna varijabla, a $f: R \to R$ zadana funkcija. Na primjer, funkcijama $y = x + 12$ *i* $y = x^2$ zadane su funkcionalne (deterministične) veze među varijablama x i y upravo jer za svaku dopuštenu varijablu x točna vrijednost varijable y koju možemo lako izračunati. Na slici 9. prikazana su dva primjera funkcionalnih veza. (Dumičić K., Bahovec V.,2011).



$y=x^2$



$y=x+12$

Slika 9. Primjer funkcionalnih veza (Izvor: Dumičić K., Bahovec V.,2011)

3.2. Statistička veza

U statističkim analizama stvarnih pojava nije realno očekivati da veze budu funkcionalne. U praksi se javljaju statističke veze koje su slabije od funkcionalnih te jedna vrijednost jedne pojave odgovara za više različitih vrijednosti druge pojave. Prilikom takve veze javljaju se neobjašnjene varijacije varijable y zbog slučajnih utjecaja. Statistički model je oblika:

$$y = f(x) + \varepsilon,$$

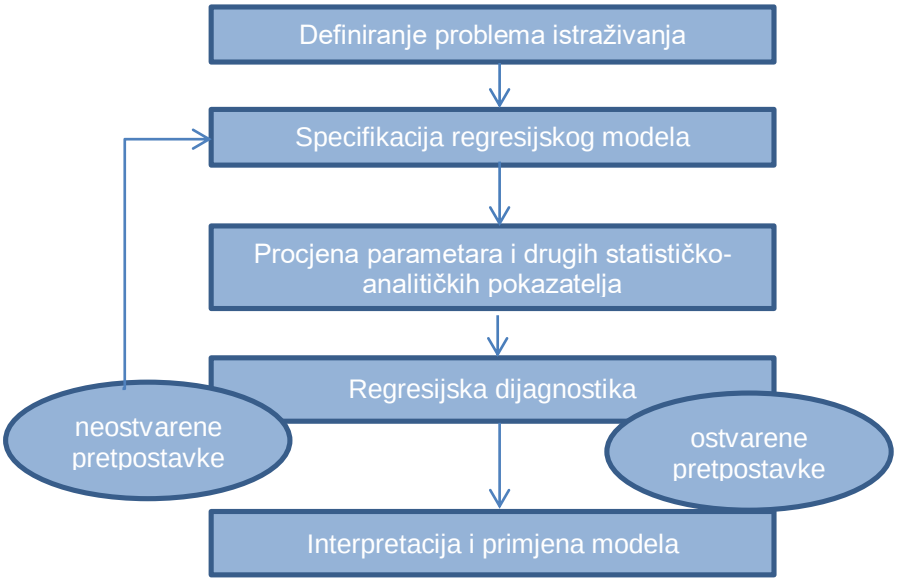
gdje se pretpostavlja da je ε slučajna varijabla koja opisuje grešku u modeliranju. Model se prihvaća adekvatnim ako je u njemu postignuta normalna distribucija grešaka ε. (Dumičić K., Bahovec V.,2011).

4. Regresijski modeli

Analitički izraz veze među pojavama (zavisnosti) zove se regresijski model**.** Kod takve vrste analize znamo što je uzrok, a što posljedica. U slučaju da naš promatrani model izražava vezu između zavisne i jedne nezavisne varijable, riječ je o jednostavnom (simple) regresijskom modelu. S druge strane, ako promatrani model izražava vezu između zavisne i dviju ili više nezavisnih varijabli, riječ je o višestrukom regresijskom modelu. („VELERI“, bez dat.).

Regresijski model ima još jednu vrstu podjele, a to je podjela prema odnosu između nezavisne i zavisne varijable. Tako da model može biti linearni ili nelinearni što je već ranije spomenuto. U praksi se za nezavisnu varijablu uzima najčešće ona fizikalna veličina koja se može najtočnije mjeriti. Najjednostavniji oblik zavisnosti dviju varijabli tj. najjednostavniji regresijski model je upravo model jednostavne linearne regresije. („Prirodoslovno - matematički fakultet [PMF]“, bez dat.).

Na slici 10. prikazane su faze provođenja regresijske analize. Regresijska analiza može započeti tek nakon što se definira problem istraživanja. Nakon definiranja problema radi se specifikacija regresijskog modela nakon koje slijedi procjena nepoznatih parametara i drugih statističko-analitičkih pokazatelja. Kada je procjena nepoznatih parametara završena slijedi provođenje postupka regresijske dijagnostika. U slučaju da je regresijska dijagnostika rezultirala ostvarenjem statističko-teorijskih pretpostavki, radi se interpretacija modela koji je isto tako spreman i za primjenu. (Dumičić K., Bahovec V.,2011).



Slika 10. Dijagram tijeka regresijske analize (Izvor: Dumičić K., Bahovec V.,2011)

5. Jednostavna linearna regresijska analiza

Jednostavna linearna regresijska analiza nam služi za utvrđivanje linearne povezanosti između jedne nezavisne i jedne zavisne varijable. Ovo je najjednostavniji oblik regresijskog modela te njegova jednadžba ima sljedeći oblik:

$$Y = \alpha + \beta X + e$$

Gdje je

- Y je zavisna varijabla
- X je nezavisna varijabla
- β su regresijski koeficijenti
- α je konstantan član
- **ε** je odstupanje od funkcionalnog odnosa (Šošić I., Serdar V. ,1995)

Regresijska analiza provodi se na temelju n parova vrijednosti varijabli $X\ i\ Y$: $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$ pa se tako ovaj statistički model prikazuje sustavom od n jednadžbi:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1,2 \dots, n$$

Gdje je y_i i-ta vrijednost zavisne varijable, x_i i-ta vrijednost nezavisne varijable, α i β nepoznati parametri, a ε_i i-ta slučajna varijabla. To su nemjerljive slučajne varijable za koje pretpostavljamo da su međusobno nezavisne i da sve imaju normalnu distribuciju s očekivanjem 0 i istom varijancom σ^2. (Dumičić K., Bahovec V.,2011).

S obzirom da su odnosi među pojavama statistički, treba odrediti kriterij prema kojemu će se izabrati jednadžba pravca:

$$\hat{y}_i = a + bx_i \qquad i = 1,2 \dots, n$$

Da bi jednostavni regresijski model bio valjan, mora zadovoljiti sljedeće **preduvjete**:

Varijable

- kvantitativni podaci
- točno izmjerena nezavisna varijabla (nema slučajne varijacije)

Linearnost

- provjera linearnosti preko dijagrama rasipanja

- ako preduvjet linearnosti nije zadovoljen može pokušati transformacija varijable kako bi se postigla linearnost

Randomizacija

- uzorak treba biti reprezentativni uzorak populacije

Reziduali

- međusobno nezavisni reziduali
- reziduali su nezavisni od prediktora i zavisne varijable

Jednaka varijanca (homoskedastičnost)

- raspršenje opservacija oko pravca treba biti podjednako za sve vrijednosti varijable X

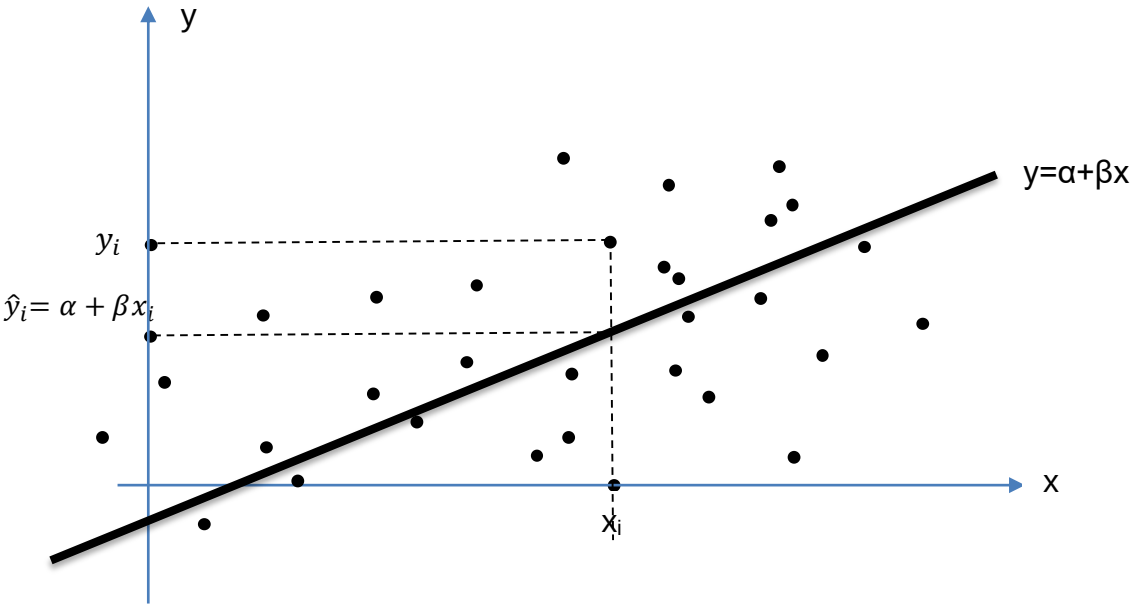
Normalna distribucija reziduala

- pretpostavlja se da je razdioba pogrešaka oko idealnog regresijskog pravca normalna
  (Šimić D., bez dat.)

5.1. Metoda najmanjih kvadrata

S obzirom da su vrijednosti parametara α i β nepoznati samim time nepoznati je i regresijski pravac $y = a + bx$. Unutar dijagrama rasipanja dodan je pravac $y = a + bx$ (slika 11).

Za svaku vrijednost x_i nezavisne varijable x, varijabla $\hat{y}_i$ dobiva neku vrijednost koju zovemo **teorijska vrijednost** zavisne varijable u x_i (eng. predicted value). Izmjerena vrijednost zavisne varijable u x_i je y_i (eng. observed value). Izmjerena vrijednost se u pravilu razlikuje od teorijske vrijednosti tako da točke $(x_i, y_i)\ i = 1, \dots, n$ uglavnom nisu smještene na regresijskom pravcu. Da bi model bio primjenjiv, razlike između izmjerenih i teorijskih vrijednosti bi trebale biti što manje. Svrha ove metode je što točnije odrediti koeficijente regresijskog pravca a i b pomoću **metode najmanjih kvadrata**. (Benšic M., Šuvak N., 2013)



Slika 11. Dijagram rasipanja s regresijskim pravcem (Izvor: Benšic M., Šuvak N. 2013)

Nas zanima razlika između izmjerenih i teorijskih vrijednosti tj. rezidualna odstupanja sljedećeg oblika:

$$\widehat{\varepsilon_i} = y_i - \hat{y}_i \quad \text{ili} \quad \widehat{\varepsilon_i} = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

pri čemu je $\hat{y}_i$ i-ta procijenjena vrijednost zavisne varijable, a $\hat{\alpha}\, i\, \hat{\beta}$ procjene parametara. Iz prethodnog izraza proizlazi da je zbroj kvadrata rezidualnih odstupanja sljedeći:

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Suma kvadrata svih reziduala predstavlja jednu mjeru kvalitete korištenog modela koja se označava SSE (eng. Sum of squares of errors):

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2$$

Kretanjem u minimiziranje zbroja kvadrata rezidualnih odstupanja s obzirom na procjene parametara $\hat{\alpha}\, i\, \hat{\beta}$, dolazi se do sustava normalnih jednadžbi:

$$n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Rješenjem prethodnog sustava jednadžbi dolazi se do sljedećih izraza za procjenu regresijskih koeficijenta $\hat{\beta}$ i konstantog člana $\hat{\alpha}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}, \qquad \alpha = \bar{y} - \hat{\beta}\bar{x}$$

Nakon dobivenih procjena parametara $\hat{\alpha}\ i\ \hat{\beta}$, regresijski model ima sljedeći oblik:

$$\hat{y} = \hat{\alpha} + \hat{\beta}.$$

(Benšic M., Šuvak N. 2013)

5.2. Pokazatelji uspješnosti modela

Nakon završetka procjene parametara regresijskog modela potrebno je ispitati sposobnost modela da objasni varijacije zavisne varijable y s obzirom na odabranu nezavisnu varijablu x. Kod mjerenja raspršenosti oko regresijskog pravca polazi se od zbroja kvadrata rezidualnih promjena SSE primjenom kojeg se radi procjena varijance. (Dumičić K., Bahovec V.,2011)

5.2.1. Procjena varijance regresije:

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

5.2.2. Procjena standardne devijacije regresije:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

Procjena standardne devijacije regresije je apsolutna mjera disperzije ili raspršenosti i izražava se u mjernim jedinicama zavisne varijable. Na temelju disperzije u apsolutnom obliku nije praktično ni lagano odlučiti o razini reprezentativnosti modela. Kako bi se taj problem zanemario, uvodi se relativna mjera disperzije tog regresijskog modela, odnosno procjena koeficijenta varijacije regresije. (Dumičić K., Bahovec V.,2011)

5.2.3. Procjena koeficijenta varijacije regresije:

$$\hat{V} = \frac{\hat{\sigma}}{\bar{y}} \cdot 100\%$$

S obzirom da se radi o relativnoj vrijednosti i ne ovisi o mjernim jedinicama zavisne varijable, rezultatom koeficijenta varijacije se zaključuje stupanj disperzije oko regresijskog

pravca. Što je procjena koeficijenta varijacije regresije bliža nuli, to je regresijski model reprezentativniji. Najčešće se uzima dogovorena granica reprezentativnosti od 30% koja odlučuje da li je model prihvatljiv ili se odbacuje. (Dumičić K., Bahovec V.,2011)

Procjena standardne devijacije i procjena koeficijenta varijacije interpretira se kao prosječno odstupanje stvarnih vrijednosti zavisne varijable y od regresijskih vrijednosti $\hat{y}_i$. (Dumičić K., Bahovec V.,2011)

5.2.4. Analiza varijance u modelu jednostavne linearne regresije

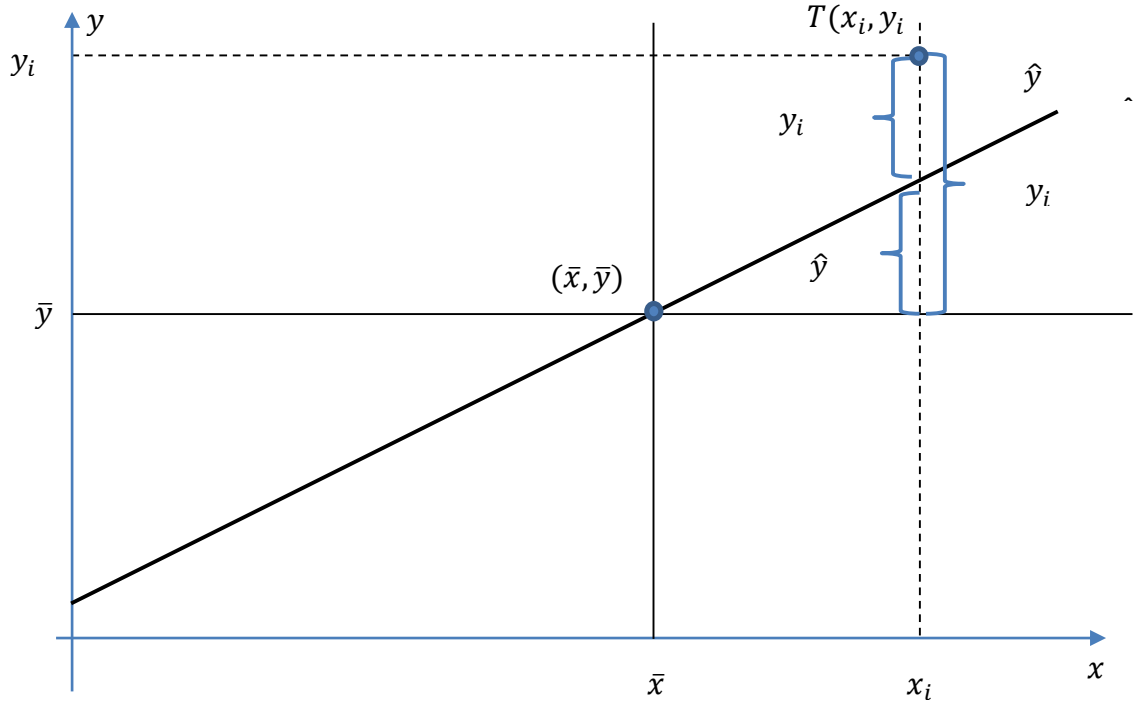
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Jednadžba analize varijance se simbolički zapisuje kao:

$$SST = SSR + SSE$$

gdje je SST zbroj kvadrata odstupanja stvarnih vrijednosti zavisne varijable od njezina prosjeka, SSR zbroj kvadrata odstupanja regresijskih vrijednosti od prosjeka, a SSE već ranije spomenuti je zbroj kvadrata odstupanja stvarnih vrijednosti zavisne varijable od pripadajućih regresijskih vrijednosti. Elementi analize varijance se prikazuju u ANOVA (eng. Analysis of variance) tablici. (Dumičić K., Bahovec V.,2011).



Slika 12. Grafički prikaz analize varijance jednostavne linearne regresije (Izvor: Dumičić K., Bahovec V.,2011)

Tablica 1: ANOVA tablica jednostavne regresije

Izvor varijacije	Stupnjevi slobode	Zbroj kvadrata	Sredina kvadrata	*F*-omjer
Protumačen modelom linearne regresije	1	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	SSR	$\frac{\frac{SSR}{1}}{\frac{SSE}{n-2}}$
Neprotumačen modelom linearne regresije	$n-2$	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$\frac{SSE}{n-2}$	
Ukupno	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$		

5.2.5. Koeficijent determinacije R^2

Koeficijent determinacije je mjera reprezentativnosti regresijskog modela koji se zasniva na analizi varijance. Definira se kao omjer zbroja kvadrata odstupanja regresijskih vrijednosti od prosjeka (SSR) i zbroja kvadrata odstupanja stvarnih vrijednosti zavisne varijable od prosjeka (SST). (Dumičić K., Bahovec V.,2011).

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

Koeficijent determinacije R^2 nam pokazuje u koliko mjeri je rasipanje eksperimentalnih vrijednosti zavisne varijable objašnjeno linearnom funkcijom $f(x) = \alpha + \beta x$, a u kolikoj mjeri se radi o rezidualnom rasipanju. (Dumičić K., Bahovec V.,2011)

Koeficijent poprima vrijednost iz sljedećeg intervala:

$$0 \leq R^2 \leq 1$$

Kada je vrijednost koeficijenta R^2 vrlo blizu 1 to ukazuje da linearan model objašnjava velik dio rasipanja u eksperimentalnim vrijednostima zavisne varijable te da rezidualima pripada mali dio rasipanja vrijednosti. Regresijski model je reprezentativniji kada je vrijednost koeficijenta bliže 1. (Dumičić K., Bahovec V.,2011)

5.2.6. Korigirani koeficijent determinacije $\bar{R}^2$

Mjera reprezentativnosti regresijskog modela koja za razliku od R^2 uzima u obzir veličinu uzorka i broj nezavisnih varijabli naziva se korigirani koeficijent determinacije te se računa po sljedećoj formuli:

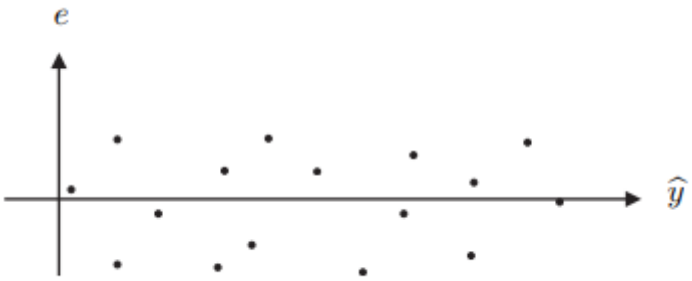
$$\bar{R}^2 = 1 - \frac{n-1}{n-2}(1 - R^2).$$

Vrijednost korigiranog koeficijenta determinacije može biti samo manja ili jednaka vrijednosti R^2. Nedostatak takve mjere reprezentativnosti je da $\bar{R}^2$ može poprimiti i negativnu vrijednost u slučaju slabe reprezentativnosti regresijskog modela. Važnost korigiranog koeficijenta determinacije više dolazi do izražaja kod višestruke linearne regresije zbog većeg broja nezavisnih varijabli. (Dumičić K., Bahovec V.,2011)

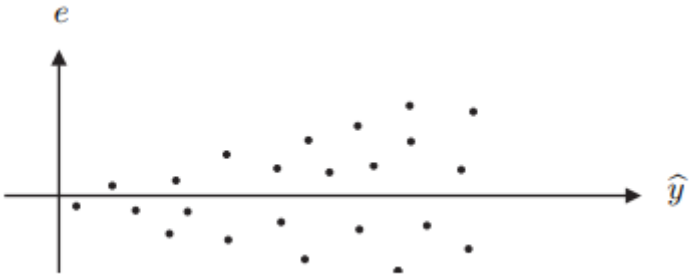
5.2.7. Analiza reziduala

Važno je da reziduali zadovoljavaju sljedeće potrebne preduvjete:

1. Varijance reziduala moraju biti jednake. Često se govori o homogenosti reziduala. Dijagram rasipanja uvelike pomaže uočavanje nehomogenosti reziduala. U slučaju da je na dijagramu uočljivo sustavno povećanje ili smanjenje rasipanja vezano uz vrijednosti $\hat{y}$, to je znak da varijance reziduala nisu homogene. Sljedeći primjeri demonstriraju homogenost, odnosno nehomogenost varijanci.



Slika 13. Homogenost varijanci reziduala (Izvor: Benšic M., Šuvak N. 2013)



Slika 14. Nehomogenost varijanci reziduala (Izvor: Benšic M., Šuvak N. 2013)

2. Slučajne varijable $\varepsilon_1 \dots \epsilon_n$ moraju biti normalno distribuirane s očekivanjem 0 i varijancom σ^2.
3. Slučajne greške modela su nezavisne. Nezavisnost reziduala provjeravamo preko dijagrama rasipanja.
4. Ako reziduali $\hat{\varepsilon}_i$ zadovoljavaju prethodno navedene uvjete smatraju se dobrim procjenama stvarnih normalnih grešaka ε. (Benšic M., Šuvak N. 2013)

5.3. Testovi značajnosti jednostavnog regresijskog modela

Osnova testa su hipoteze i oblik distribucije procjenitelja parametra. Test značajnosti parametra β u modelu jednostavne linearne regresije provodi se primjenom dvije vrste testova, a to su **F**-test i **t**-test. F-test provjerava da li je cijeli regresijski model značajan, dok t-test provjerava značajnost nezavisne varijable x u modelu. S obzirom da se radi o jednostavnoj regresiji oba test će biti donositi isti zaključak te sadržavati iste hipoteze. (Benšic M., Šuvak N. 2013).

Hipoteze testova jednostavne regresijske analize su:

$$H_0 \dots \beta = 0$$

$$H_1 \dots \beta \neq 0,$$

gdje H_0 sadrži tvrdnju koja ukazuje na suvišnost varijable x u modelu, dok H_1 sadrži suprotnu tvrdnju koja opravdava važnost varijable x te da njezina prisutnost objašnjava varijacije zavisne varijable y. F-omjer je empirijska vrijednost test veličine sljedećeg oblika:

$$F = \frac{SSR}{\frac{SSE}{n-2}}$$

Vrijednost F-omjera se može očitati iz ranije spomenute ANOVA tablice u prošlom poglavlju zajedno s pojašnjenjem SSR i SSE varijabli. Test veličina t-testa je:

$$t = \frac{\beta}{\sigma_\beta}$$

(Benšic M., Šuvak N. 2013)

6. Višestruka linearna regresijska analiza

Nakon formiranja jednostavnog regresijskog modela, javila se potreba za ubacivanjem više prediktorskih (nezavisnih) varijabli jer sa jednostavnim modelom predviđanja su bila u nekim situacijama vrlo loša i neiskoristiva. Istraživanje provedeno među niskom djecom na način da je zavisna varijabla bila maksimum plazme hormona rasta, a dok su nezavisne (prediktorske) varijable bile godine, spol, razna tjelesna mjerenja i tako čak 14 nezavisnih varijabli koje su bile dio regresijske analize. Zaključak je takav da kad bi to bio jednostavni linearni model sa samo jednom prediktorskom varijablom opis tog rezultata analize ne bi uopće reprezentativan s obzirom na zavisnu varijablu. Upravo zato jer ne možemo do tako kompleksnog zaključaka doći samo uz procjenu jedne nezavisne varijable kada znamo da puno više čimbenika utječe na rezultat u ovom slučaju na niski rast djece. (Kutner M.H., Nachtsheim C.J.,Neter J.,Li W. ,2005)

Višestruka linearna regresijska analiza nam služi za utvrđivanje linearne povezanosti između dviju ili više nezavisnih (prediktorskih) i jedne zavisne varijable pri čemu regresijska jednadžba ima sljedeći oblik:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

gdje je

- y je zavisna (regresand) varijabla
- $x_1, x_2, \ldots x_k$ su nezavisne varijable (prediktori)
- $\beta_0, \beta_1, \ldots, \beta_k$ su regresijski koeficijenti (nepoznati parametri)
- ε je slučajna varijabla (normalno distribuirana s očekivanjem 0) (Šošić I., Serdar V. ,1995).

Preduvjeti su jednaki kao i kod modela jednostavne linearne regresije uz proširenje vezano za nezavisne varijable. Preduvjeti su sljedeći:

- Veza između zavisne varijable i nezavisnih varijabli je linearna.
- Varijanca reziduala je nezavisna od prediktora i zavisne varijable.
- Reziduali su normalno distribuirani.
- Nezavisne varijable su međusobno linearno nezavisne
- Reziduali su međusobno nezavisni. (Šimić D., bez dat.)

U prošlom poglavlju je spomenuto kako se jednostavna linearna regresija grafički prikazuje pomoću pravca, a s obzirom da je ovdje riječ o višestrukoj linearnoj regresiji nju prikazujemo pomoću hiperravnine u višedimenzionalnom prostoru. Isto kao i kod jednostavne linearne regresije za pronalazak regresijskih koeficijenata se koristi metoda najmanjih kvadrata. Kod višestruke linearne regresije za pronalazak regresijskih koeficijenata često se koristi matrični zapis jednadžbe. (Šimić D., bez dat.) (Montgomery D.C., Runger G.C.,2010, str. 417)

Kako bi iskazali višestruki regresijski model u matričnom obliku, potrebno je standardizirani oblik regresijske jednadžbe koja ima sljedeći izgled prebaciti u matrice:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1,2,\ldots n$$

Ovaj model je sustav od n jednadžbi koje se izražavaju pomoću sljedećeg matričnog zapisa:

$$Y = X\beta + \varepsilon$$

gdje je:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & \mathrm{X}_{11} & \mathrm{X}_{12} & \ldots & \mathrm{X}_{1,\mathrm{k}} \\ 1 & \mathrm{X}_{21} & \mathrm{X}_{22} & \ldots & \mathrm{X}_{2,\mathrm{k}} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & \mathrm{X}_{\mathrm{n}1} & \mathrm{X}_{\mathrm{n}2} & \ldots & \mathrm{X}_{\mathrm{n,k}} \end{bmatrix}$$

gdje:

- Y je vektor standardiziranih rezultata entiteta
- β je vektor parametara
- X je matrica konstanti
- **Ɛ** je vektor rezidualnih vrijednosti

Uvrste li se u zadnju jednadžbu umjesto parametara β njihove procjene tada slijedi:

$$y = X\hat{\beta} + \hat{\varepsilon}$$

$$\hat{\varepsilon} = y - X\hat{\beta}$$

pri čemu je $\hat{\varepsilon}$ procjena slučajne varijable ε na bazi uzorka i zovu se rezidualna ili neprotumačena odstupanja. (Montgomery D.C., Runger G.C.,2010, str. 417)

Procjena vektora β metodom najmanjih kvadrata dobit će se iz zahtjeva koji glasi da zbroj kvadrata rezidualnih odstupanja bude minimalan. Normalne jednadžbe metode najmanjih

kvadrata u matričnom obliku imaju sljedeći izgled, a ekvivalentne su skalarnom zapisu normalnih jednadžbi metode najmanjih kvadrata. (Dumičić K., Bahovec V.,2011)

$$\mathrm{X}'\mathrm{X}\hat{\beta} = \mathrm{X}'\mathrm{y}$$

Iz tih jednadžbi množenjem sa $(\mathrm{X}'\mathrm{X})^{-1}$ dobivaju se procjenitelji najmanjih kvadrata varijable β:

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

(Montgomery D.C., Runger G.C.,2010, str. 417)

Transformirani oblici varijabli također mogu biti dio regresijskog modela, neovisno radilo se o jednostavnoj ili višestrukoj regresijskoj analizi. Puno puta se nezavisne varijable transformiraju kako bi se došlo do što boljeg regresijskog modela, tako da je sljedeći izraz u potpunosti prihvatljiv (Kutner M.H., Nachtsheim C.J.,Neter J.,Li W. ,2005):

$$y_i = \beta_0 + \beta_1 x^2{}_{i1} + \beta_2 x_{i1} + \beta_3 \log x_{i2} + \varepsilon_i$$

Kod takvih slučajeva se preporuča da se definiraju nove nezavisne varijable koje će preglednije predstavljati ulogu tih transformiranih varijabli. Definirajmo transformirane varijable:

$$z_{i1} = x^2{}_{i1} \qquad z_{i2} = x_{i1} \qquad z_{i3} = \log x_{i1}$$

Nakon toga, formirani oblik regresijskog modela izgleda:

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \varepsilon_i$$

(Kutner M.H., Nachtsheim C.J.,Neter J.,Li W. ,2005)

6.1. Analiza varijance u modelu višestruke linearne regresije

Jednadžba analize varijance ima sljedeći oblik:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

Iz jednadžbe je vidljivo da se zbroj kvadrata odstupanja empirijskih vrijednosti varijable y od prosjeka (SST) rastavlja na protumačeni zbroj kvadrata SSR i rezidualni zbroj kvadrata SSE. Kako bi došli do sredine kvadrata, tj. nezavisnih procjena komponenti varijance potrebno je podijeliti zbrojeve kvadrata s odgovarajućim stupnjevima slobode. Svi ti podaci vezani uz varijancu prikazani su unutar ANOVA tablice. (Dumičić K., Bahovec V.,2011)

Tablica 2: ANOVA tablica višestruke regresije

Izvor varijacije	Stupnjevi slobode	Zbroj kvadrata SS	Sredina kvadrata MS	*F*-omjer
Protumačen modelom	k	SSR	$\frac{SSR}{k}$	$\frac{\frac{SSR}{k}}{\frac{SSE}{(n-(k+1))}}$
Neprotumačen modelom	$n-(k+1)$	SSE	$\frac{SSE}{(n-(k+1))}$	
Ukupno	$n-1$	SST		

Rezidualna suma kvadrata podijeljena s $[n-(k+1)]$ stupnjeva slobode je procijenjena varijanca regresije:

$$\hat{\sigma}^2 = \frac{SSE}{n-(k+1)}$$

(Dumičić K., Bahovec V.,2011)

6.2. Koeficijent determinacije R^2 i korigirani koeficijent determinacije $\bar{R}^2$

Koeficijent determinacije je omjer:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{[n-(k+1)]\hat{\sigma}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 može poprimiti vrijednost u intervalu [0,1] dok je promatrani model reprezentativniji što je koeficijent bliže 1. S obzirom na broj regresorskih varijabli (k) koeficijent raste bez obzira da li su varijable bitne za objašnjenje varijacija zavisne varijable ili ne. (Dumičić K., Bahovec V.,2011)

Puno bolji kriterij za ocjenjivanje modela je korigirani koeficijent determinacije $\bar{R}^2$ definiran na sljedeći način:

$$\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2) = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SSE}{SST} = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{SST}{n-1}} = 1 - \frac{\hat{\sigma}^2}{\frac{SST}{n-1}}$$

Korigirani koeficijent determinacije je veći kada je procijenjena varijanca regresije manja, odnosno te dvije veličine su međusobno obrnuto proporcionalne. (Dumičić K., Bahovec V.,2011)

6.3. Testovi hipoteza kod višestruke linearne regresije

Kod modela višestruke linearne regresije testiranje hipoteza se dijeli u tri grupe:

- Test o značajnosti jedne nezavisne varijable (pojedinačni test)
- Test o značajnosti svih nezavisnih varijabli (skupni test, test o značajnosti regresije)
- Test o značajnosti podskupa nezavisne varijabli (Dumičić K., Bahovec V.,2011)

Test o značajnosti jedne nezavisne varijable (pojedinačni test)

Pojedinačni test: Kako bi se saznao smjer veze između varijabli y i x_j provode se jednosmjerni testovi o značajnosti pojedine varijable x_j. Takva vrsta testova je najčešća kod regresijske analize. Kod ovog test hipoteza, tvrdnja istraživača (radna hipoteza) se formulira kao alternativna hipoteza. (Dumičić K., Bahovec V.,2011).

$$H_0: \beta_j = 0 \qquad H_0: \beta_j = 0$$
$$H_1: \beta_j > 0 \qquad H_1: \beta_j < 0$$

Uz razinu signifikantnosti α hipoteza H_0 se odbacuje ako je t_j veći od t_α u testu na gornju granicu, odnosno ako je t_j manji od $-t_\alpha$ u testu na donju granicu. Rezultat testa pokazuje da li je model prihvatljiv ili je neprihvatljiv te se mora zamijeniti novim modelom. (Dumičić K., Bahovec V.,2011)

Test o značajnosti svih nezavisnih varijabli (skupni test, test o značajnosti regresije)

Nulta hipoteza sadrži tvrdnju da niti jedna od k nezavisnih varijabli nema utjecaja na varijacije zavisne varijable. Alternativna hipoteza je suprotna nultoj, pa tako ona glasi da barem jedna od nezavisnih varijabli značajna u modelu. (Dumičić K., Bahovec V.,2011). Formulacija hipoteza je sljedeća:

$$H_0: \beta_1 = \beta_2 = \cdots \beta_k = 0$$
$$H_1: \exists \beta_j \neq 0, \quad j = 1,2, \ldots, k.$$

Test veličina za skupni test je omjer protumačene i neprotumačene sredine kvadrata:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-(k+1))}}$$

(Dumičić K., Bahovec V.,2011)

7. Analiza podataka

7.1. Uvod

Cilj ovog istraživanja je pokazati utjecaj broja stanovnika te bogatstva pojedine države na osvojene olimpijske medalje tj. ukupan uspjeh država na Olimpijskim igrama 2016. Prema informacijama na navedenim Olimpijskim igrama je sudjelovalo 207 država. („Rio 2016: The greatest show on Earth in stats“, 2016). Carter i sur. (2014) su pokazali da je asimetrija podataka jako velika ako se sve države uključe u analizu rezultata. Većina država nije osvojila medalje pa su prouzrokovali navedenu asimetriju podataka. U ovo istraživanje su uključene države koje su osvojile barem jednu zlatnu medalju i time je u istraživanje obuhvaćeno 58 država. Podaci o broju osvojenih medalja po državama su preuzeti s web stranice http://olympics.sporting99.com/medal-tally-2016.html, a podaci o ukupnom BDP-u države, o BDP po stanovniku države te o broju stanovnika preuzeti su s web stranice Svjetske banke. (http://databank.worldbank.org.) Prikupljeni demografski i ekonomski podaci referentni su za 2016. godinu jer se istraživanje bazira na uspjesima pojedinih država na Olimpijskim igrama 2016. u Riju. U istraživanju će se primjenjivati modeli jednostruke i/ili višestruke regresijske analize ovisno o broju nezavisnih varijabli u modelu.

Uspjeh na Olimpijskim igrama se najčešće mjeri na tri načina: prvi način uzima u obzir ukupan broj zlatnih medalja neke države, ZlatneMedalje=Y_1; drugi način gleda ukupan broj medalja (zlato+srebro+bronca) neke države, UkupnoMedalja=Y_2; treći način broji ukupan broj Borda bodova prema Borda metodi koja za svaku zlatnu medalju dodaje 3 boda, za svaku srebrnu dodaje 2 boda i za svaku brončanu dodaje 1 bod, BordaBodovi=Y_3. Navedene tri zavisne varijable su jako međusobno linearno povezane jer su korelacijski koeficijenti između svake dvije varijable veći od 0.9 (između Y_1 i Y_2 korelacijski koeficijent iznosi 0.972 , između Y_1 i Y_3 korelacijski koeficijent iznosi 0.986, između Y_2 i Y_3 korelacijski koeficijent iznosi 0.996). Zbog navedenog promatrana zavisna varijabla u modelu može biti bilo koja od tri navedene. U ovom istraživanju odabrana zavisna varijabla će biti Y_2 koja predstavlja ukupan broj medalja koje je pojedina država osvojila na Olimpijskim igrama.

7.2. Utjecaj broja stanovnika

Za očekivati je da ako pojedina država ima više stanovnika da je time i veći broj potencijalnih sportaša pa i veća vjerojatnost osvajanja medalje na Olimpijskim igrama ili bilo kojem drugom sportskom internacionalnom natjecanju. Time je pretpostavka da će veličina populacije utjecati na uspjeh države na Olimpijskim igrama 2016. i to je razlog zašto je nezavisna varijabla X_2 u modelu. Zbog prevelikih razlika u ukupnom broju stanovnika među državama, možda će modeli bolje prognozirati uspjehe ako navedena varijabla bude transformirana da se smanji utjecaj razlike u broju stanovnika.

Postoje i geografske karakteristike pojedinih država (npr. geografska pozicija alpskih država olakšava im bavljenje zimskim sportovima i veću uspješnost) kao i iskustva u natjecanju države na velikim svjetskim natjecanjima, tradicije prakticiranja određenih disciplina (npr. azijske zemlje imaju tradiciju bavljenja raznim borilačkim disciplinama) pa i to sigurno utječe na ukupni broj medalja. U ovom radu će se promatrati samo utjecaj ukupnog broja stanovnika pojedine države na ukupan broj medalja jer je prethodno navedene varijable jako teško izmjeriti pa samim time i uključiti u model.

7.3. Utjecaj bogatstva

Za očekivati je da države koje ulažu više u infrastrukturu, u kampove za proizvodnju novog kadra, osvajaju više medalja na Olimpijskim igrama nego države u kojima su sportaši prepušteni sami sebi. Kako bi država mogla ulagati u sport potrebna su financijska sredstva i samim time je za očekivati da će bogatijim državama biti lakše uložiti sredstva nego državama koje su siromašnije.

U istraživanju su kao nezavisne varijable vezane uz bogatstvo pojedine države uzeti bruto domaći proizvod (kratica BDP) te bruto domaći proizvod po stanovniku. Bruto domaći proizvod je vrijednost svih dovršenih roba i usluga, koji su bili proizvedeni unutar jedne države u određenom vremenskom razdoblju („Bruto domaći proizvod – BDP“, bez dat.) dok se bruto domaći proizvod po stanovniku dobije tako da se ukupni BDP podijeli s ukupnim brojem stanovnika.

7.4. Deskriptivna statistika varijabli

Kao što je prethodno navedeno zavisne varijable u modelu su Y_1=broj zlatnih medalja, Y_2=ukupan broj medalja, Y_3=borda bodovi dok su nezavisne varijable X_1=BDP po stanovniku u 10.000 (prihod), X_2=broj stanovnika u milijardama i X_3=BDP u 10 bilijuna.

U sljedećoj tablici se nalazi deskriptivna statistika svih zavisnih i nezavisnih varijabli za odabranih 58 država. Vidljivo je da je 50% država osvojilo 1 ili 2 zlatne medalje, a 50% 2 ili više dok je maksimalan broj zlatnih medalja koje je neka država osvojila 46. Iako su izbrisane države koje nisu osvojile nijednu zlatnu medalju i dalje je distribucija broja osvojenih medalja asimetrična. Slični zaključci su vidljivi iz deskriptive druge dvije zavisne varijable. Kod ukupnog broja medalja 75% država je osvojilo od 1 do 18 medalja, dok je 25% država osvojilo od 18 do 121 medalju. Ako se promatraju borda bodovi, 50% država je osvojilo od 3 do 16 borda bodova, a samo 25% država je osvojilo od 33.75 do 250 borda bodova. Slične asimetrije podataka su vidljive i kod nezavisnih varijabli jer je poznato da broj stanovnika, kao ni svjetsko bogatstvo nisu ravnomjerno raspoređeni. Kod broja stanovnika se uočava najveća asimetrija podataka, a i kod ukupnog BDP tako da bi BDP po stanovniku možda bio bolja nezavisna varijabla nego preostale dvije. Ranije u teorijskom dijelu rada spomenuto je da se dozvoljava transformacija oblika nezavisnih varijabli tako da će se unutar nekih modela naći prirodni logaritam ili korijen od pojedinih varijabli kako bi se smanjila asimetrija podataka.

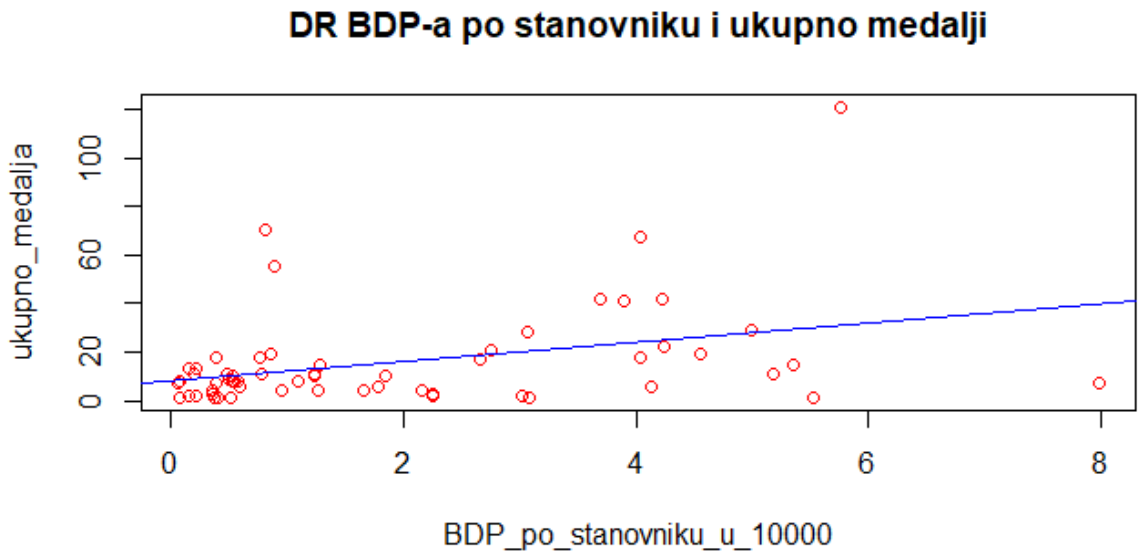
Tablica 3: Deskriptivna statistika

varijabla	prosjek	sd	min	Q1	medijan	Q3	Max	raspon	skewness
Zlato	5.276	7.862	1.0	1.0	2.0	6.0	46.0	45.0	3.172
Ukupno medalja	15.793	20.858	1.0	4.0	9.5	18.0	121.0	120.0	2.895
Borda bodovi	31.259	42.614	3.0	9.25	16.0	33.75	250.0	247.0	3.048
Broj stanovnika	0.067	0.186	0.0004	0.007	0.022	0.059	1.379	1.37	6.097
BDP po stanovniku	1.949	1.895	0.0662	0.491	1.158	3.079	7.987	7.920	1.031
Ukupni BDP	0.110	0.288	0.0005	0.005	0.020	0.084	1.862	1.862	4.663

7.5. Regresijski modeli

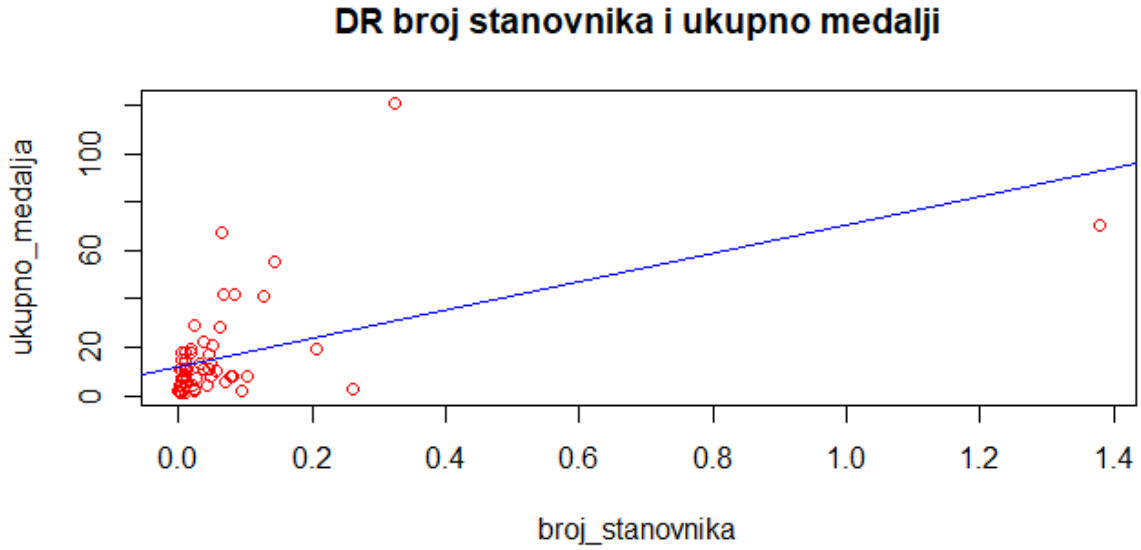
Kao što je bilo prethodno rečeno uspjeh na Olimpijskim igrama u ovom radu će se mjeriti nezavisnom varijablom Y_2 koja predstavlja ukupan broj medalja.

Kako je nezavisna varijabla X_1 nastala dijeljenjem preostale dvije nezavisne varijable, odlučeno je u nastavku rada promatrati dva tipa regresije: jednostavna s jednom nezavisnom varijablom X_1 te višestruka s dvije nezavisne varijable X_2 i X_3 te njihovim transformacijama.

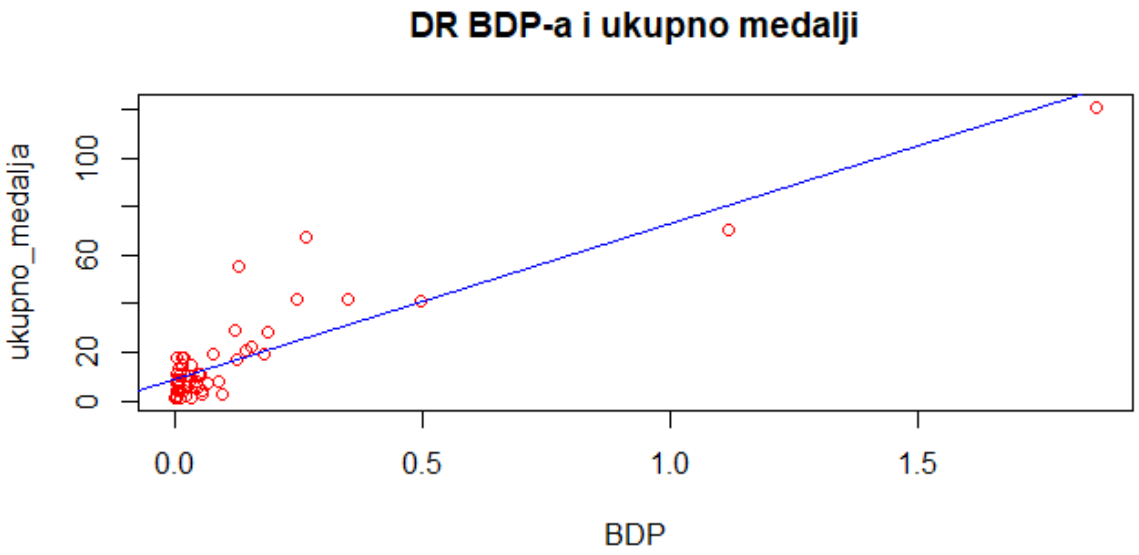


Slika 15. Dijagram rasipanja ukupnih medalja i prihoda po stanovniku

Iz slike 15. je vidljivo da postoji blaga pozitivna korelacija između varijabli X_1 *i* Y_2, korelacijski koeficijent je 0.36.



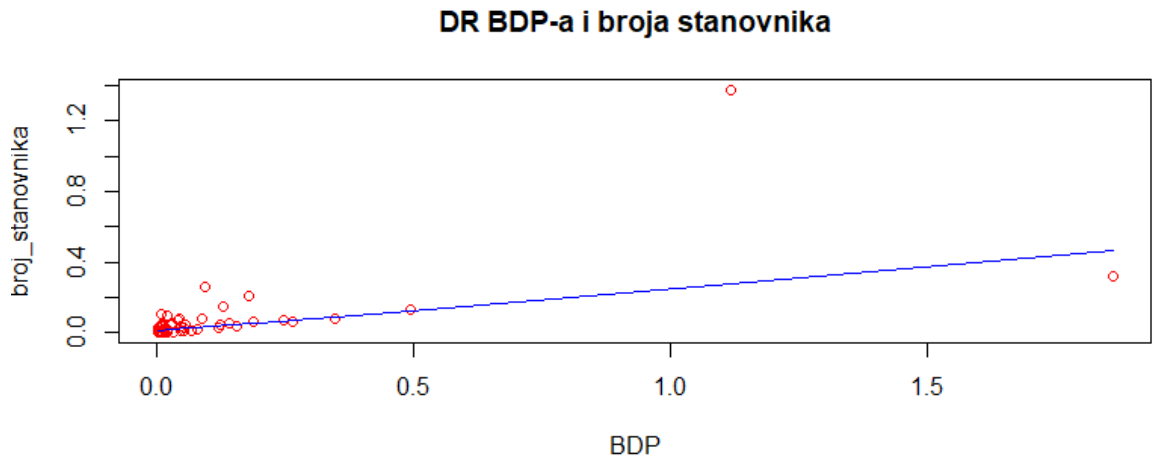
Slika 16. Dijagram rasipanja ukupnih medalja i broja stanovnika u milijardama



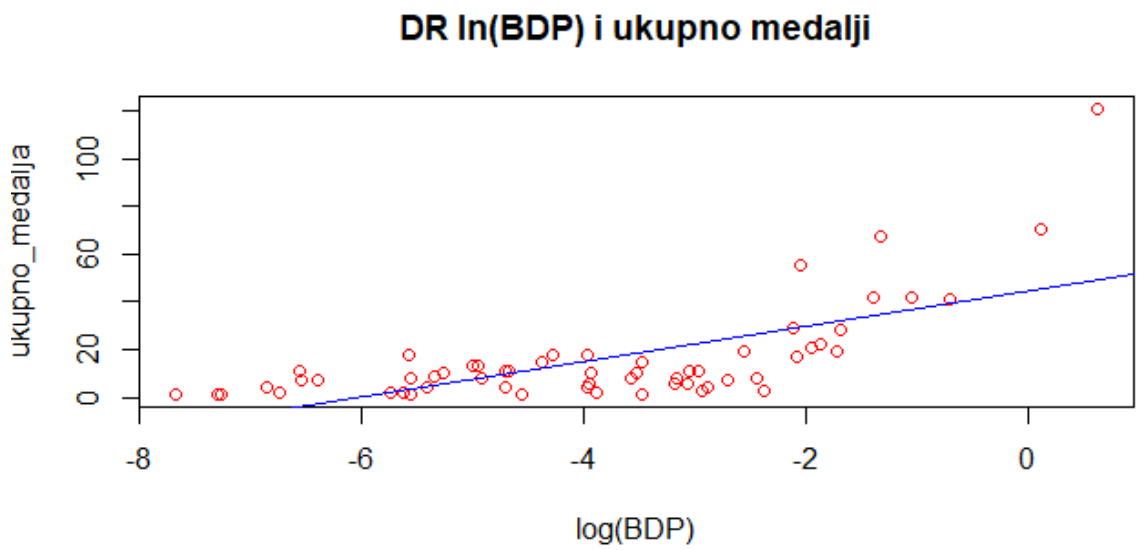
Slika 17. Dijagram rasipanja ukupnih medalja i BDP-a u 10 bilijuna

Iz slike 16. i slike 17. je vidljiva asimetrija koja je bila spomenuta prilikom deskriptivne statistike. Kod varijabli X_2 i X_3 izračunati su veći koeficijenti korelacije s zavisnom varijablom Y_2 (0.523 i 0.882), ali na iznose tih koeficijenata su utjecale ekstremne vrijednosti. Kako bi se smanjio utjecaj ekstremnih vrijednosti prije izrade modela višestruke regresije nezavisne varijable će se transformirati.

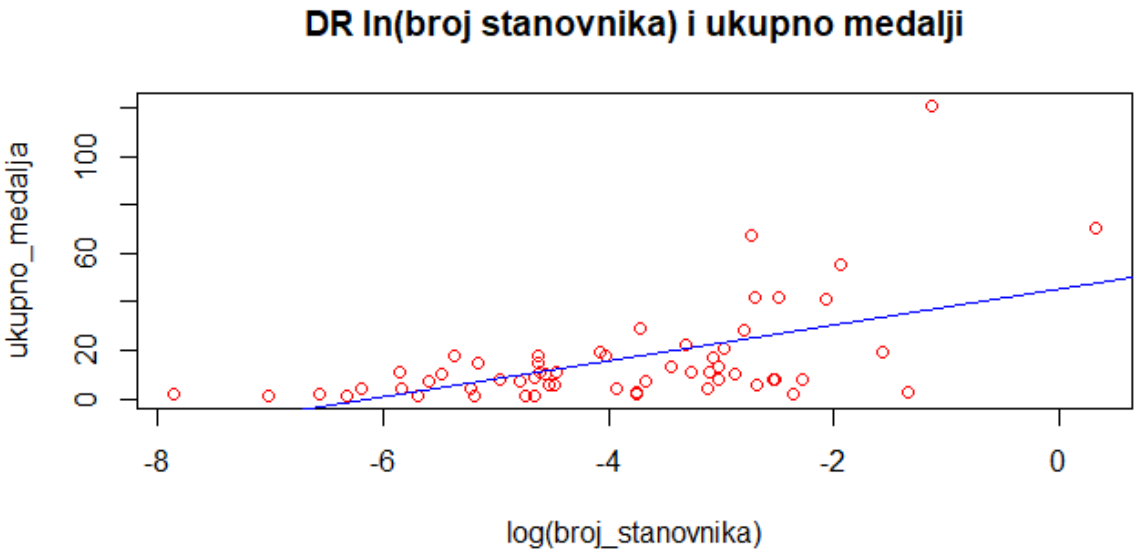
Slika 18. pokazuje da ne postoji jaka linearna povezanost između ukupnog BDP-a i ukupnog broja stanovnika pa će u daljnjoj analizi biti prikazan model višestruke regresije s 2 nezavisne varijable koje će biti transformirane.



Slika 18. Dijagram rasipanja BDP-a i broja stanovnika



Slika 19. Dijagram rasipanja ln(BDP) i ukupnih medalja



Slika 20. Dijagram rasipanja ln(broja stanovnika) i ukupnih medalja

Nakon transformacije nezavisnih X_2 i X_3 varijabli u $\ln(X_2)$ i $\ln(X_3)$, razina asimetrije se znatno smanjila. Novi koeficijenti korelacije $\ln(X_2)$ i $\ln(X_3)$ s zavisnom varijablom iznose 0.560 (povećanje za 0.037) i 0.678 (smanjenje za 0.204).

MODEL 1:

```
Call:
lm(formula = ukupno_medalja ~ BDP + broj_stanovnika)

Residuals:
    Min      1Q  Median      3Q     Max
-11.484  -6.351  -1.273   2.925  40.635

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.977      1.407   6.378 3.92e-08 ***
BDP               68.009      5.999  11.337 5.16e-16 ***
broj_stanovnika   -9.756      9.284  -1.051    0.298
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.905 on 55 degrees of freedom
Multiple R-squared:  0.7824,   Adjusted R-squared:  0.7745
F-statistic: 98.88 on 2 and 55 DF,  p-value: < 2.2e-16
```

MODEL 2:

```
 Call:lm(formula = ukupno_medalja ~ ln(BDP) + broj_stanovnika)

Residuals:
    Min      1Q  Median      3Q     Max
-27.637  -8.825  -0.967   5.471  70.429

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       37.029      5.210   7.108 2.51e-09 ***
log(BDP)           6.027      1.145   5.264 2.40e-06 ***
broj_stanovnika   30.282     11.811   2.564   0.0131 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.75 on 55 degrees of freedom
Multiple R-squared:  0.5172,   Adjusted R-squared:  0.4996
F-statistic: 29.46 on 2 and 55 DF,  p-value: 2.011e-09
```

MODEL 3:

```
lm(formula = ukupno_medalja ~ BDP + ln(broj_stanovnika))

Residuals:
    Min      1Q  Median      3Q     Max
-16.326  -5.087  -1.858   2.533  39.964

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          16.1783     4.3163   3.748 0.000429
BDP                  58.8910     5.2620  11.192 8.48e-16
log(broj_stanovnika)  1.7447     0.9642   1.809 0.075847

(Intercept)          ***
BDP                  ***
log(broj_stanovnika) .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.719 on 55 degrees of freedom
Multiple R-squared:  0.7905,   Adjusted R-squared:  0.7829
F-statistic: 103.8 on 2 and 55 DF,  p-value: < 2.2e-16
```

MODEL 4:

```
lm(formula = ukupno_medalja ~ ln(BDP) + ln(broj_stanovnika))

Residuals:
    Min      1Q  Median      3Q     Max
-25.909 -10.578  -1.105   5.835  72.188

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           46.202      5.552   8.322 2.61e-11
log(BDP)               6.553      1.671   3.922 0.000246
log(broj_stanovnika)   1.298      2.041   0.636 0.527495

(Intercept)          ***
log(BDP)             ***
log(broj_stanovnika)
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.55 on 55 degrees of freedom
Multiple R-squared:  0.4634,   Adjusted R-squared:  0.4439
F-statistic: 23.75 on 2 and 55 DF,  p-value: 3.668e-08
```

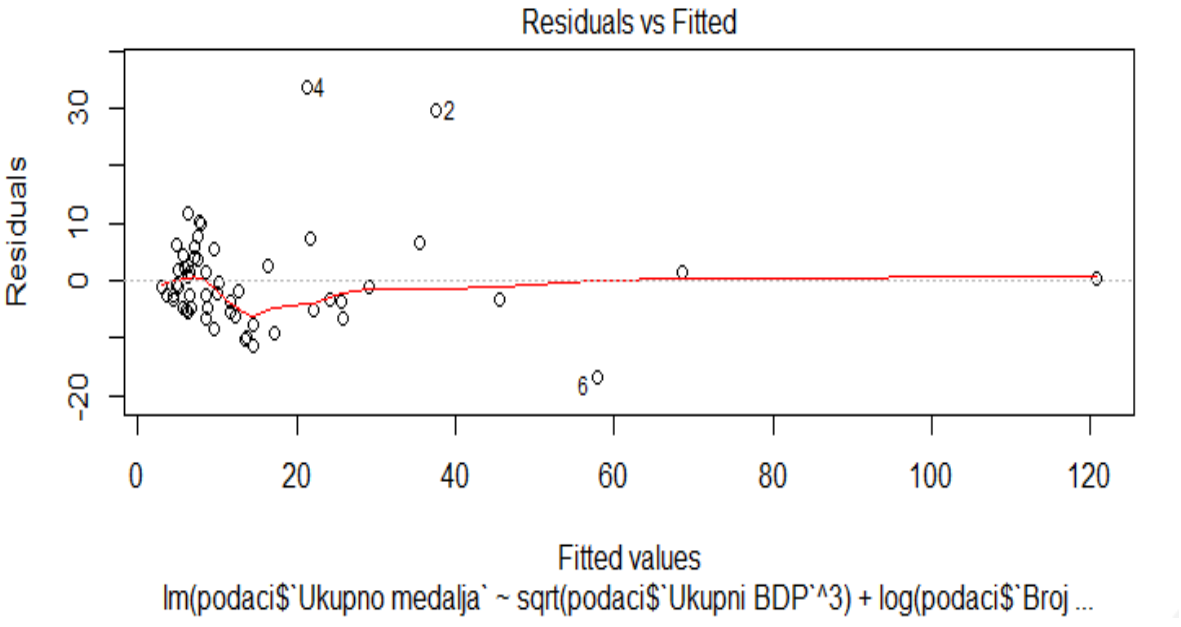
MODEL 5:

```
Call:
lm(formula = ukupno_medalja ~ sqrt(BDP^3) + ln(broj_stanovnika) +
    broj_stanovnika + BDP)

Residuals:
    Min      1Q  Median      3Q     Max
-16.871  -4.711  -1.412   3.285  33.784

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          10.2511     4.9519   2.070 0.043327
sqrt(BDP^3)         -64.1083    18.3776  -3.488 0.000986
log(broj_stanovnika)  0.9413     1.0190   0.924 0.359793
broj_stanovnika     -26.2194     9.0013  -2.913 0.005233
BDP                 151.9691    25.6400   5.927 2.37e-07

(Intercept)          *
sqrt(BDP^3)          ***
log(broj_stanovnika)
broj_stanovnika      **
BDP                  ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.641 on 53 degrees of freedom
Multiple R-squared:  0.8404,   Adjusted R-squared:  0.8284
F-statistic: 69.78 on 4 and 53 DF,  p-value: < 2.2e-16
```

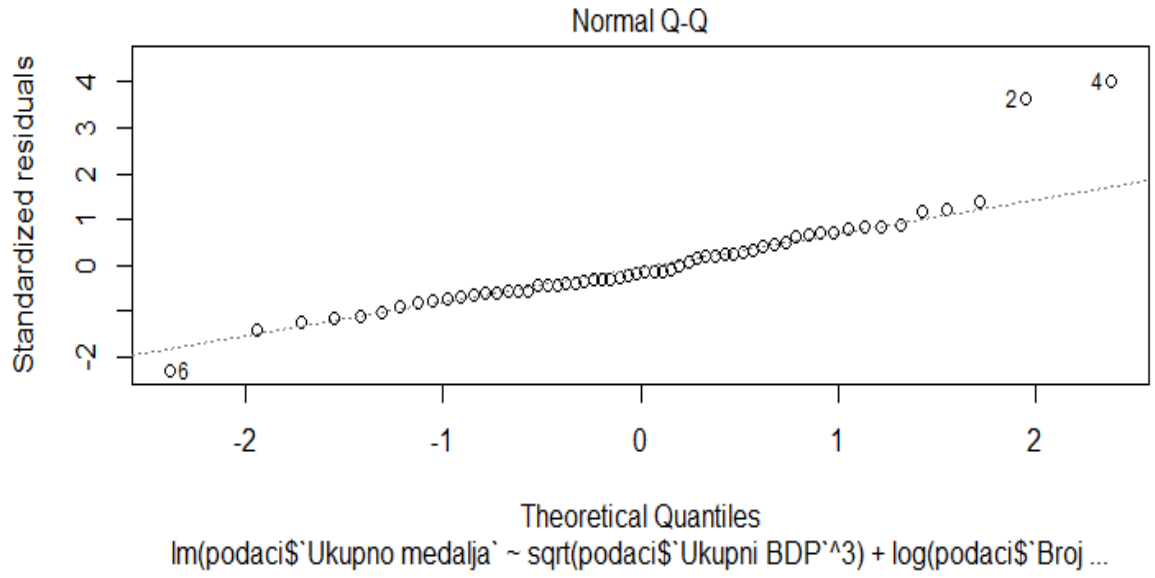
MODEL 6:

```
Call:
lm(formula = ukupno_medalja ~ sqrt(BDP^5) + ln(broj_stanovnika))

Residuals:
    Min      1Q  Median      3Q     Max
-22.268  -7.420  -1.001   5.410  47.301

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           31.307      4.620   6.777 8.75e-09
sqrt(BDP^5)           20.898      2.637   7.925 1.15e-10
log(broj_stanovnika)   4.539      1.077   4.213 9.44e-05

(Intercept)          ***
sqrt(BDP^5)          ***
log(broj_stanovnika) ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.02 on 55 degrees of freedom
Multiple R-squared:  0.6795,   Adjusted R-squared:  0.6678
F-statistic: 58.29 on 2 and 55 DF,  p-value: 2.579e-14
```

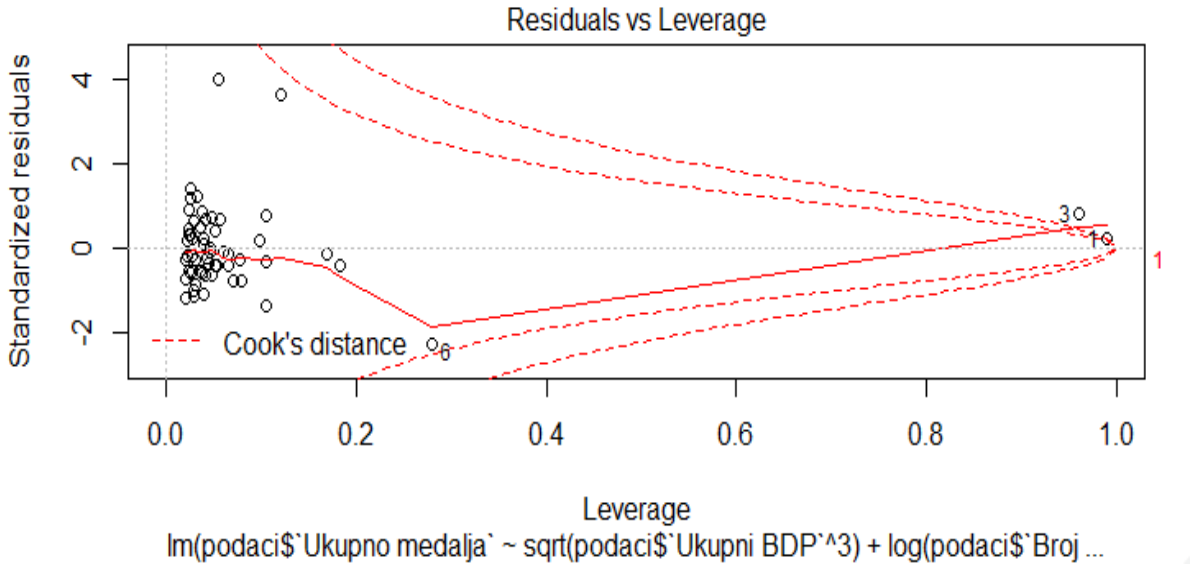
U **MODELU 1** korigirani koeficijent determinacije iznosi 0.77 dok u modelima koji uključuju transformirane nezavisne varijable koeficijent raste i do 0.83. Kao model koji najbolje opisuje promatrane podatke je **MODEL 5** (ima najveći R^2 kao i $\bar{R}^2$)**.** Iz rezultata višestruke linearne regresije slijedi da je formula regresijskog modela $\mathbf{ukupno\ medalja} = \mathbf{10.2511} - \sqrt{\boldsymbol{BDP^3}} * \mathbf{64.1083} + \mathbf{ln(broj\ stanovnika)} * \mathbf{0.9413} + \mathbf{BDP} * \mathbf{151.9691} - \mathbf{broj\_stanovnika} * \mathbf{26.2194}$.



Slika 21. Dijagram raspršenja reziduala MODELA 5



Slika 22. QQ dijagram MODELA 5



Slika 23. Dijagram utjecajnosti MODELA 5

Iz dijagrama raspršenja reziduala (slika 21.) je vidljivo postoji ravnomjerna raspršenost oko horizontalne osi. Opažena su i 3 stršila, a to su točke 2,4 i 6. Iz QQ dijagrama (slika 22.) vidimo da su svi reziduali normalno distribuirani osim ranije spomenutih stršila 2,4 i 6. Iz dijagrama utjecajnosti (slika 23.) se može primijetiti da opservacije 1 i 3 jako utječu na regresijsku analizu. To su države: Kina i SAD. Već na početku ove analize je bilo očito da se dvije države razlikuju od drugih. Kada bi izostavili te dvije države iz ulaznih podataka pokazatelji MODELA 5 bi se znatno razlikovali.

U Tablici 4. se vidi usporedba ukupno osvojenih medalja te ukupno osvojenih medalja opisano modelom 5 koji je stavljen kao najreprezentativniji u ovom radu. Tablica sadrži samo prvih 24 promatranih država jer su upravo te države pri vrhu zanimljive u ovakvom istraživanju. Sitne pogreške su uvijek prisutne kod takvih analiza, ali pojavile su se i znatne pogreške koje su istaknute plavom bojom u tablici. Primjećujemo da su Velika Britanija i Rusija osvojile poprilično više medalja nego što je to prognozirano našim modelom. S druge strane, Japan je osvojio čak 16 medalja manje nego što je predviđeno modelom 5. Sada nakon analize se jasno vidi da olimpijski uspjeh države ne ovisi samo o broju stanovnika i bogatstvu te države. Važno je koliko država ulaže u svoje sportaše i resurse kojima se oni koriste kako bi napredovali. S druge strane, same Olimpijske igre su posložene tako da sportaši poput plivača i atletičara mogu osvojiti čak i nekoliko medalja zbog više utrka unutar njihovih disciplina.

Tablica 4: Usporedba rezultata

Država	Ukupno medalja Y_2	Dobiveno modelom 5	Reziduali
1. SAD	121	120.7986931	0.2013069
2. Velika Britanija	67	37.5018423	29.4981577
3. Kina	70	68.5784693	1.4215307
4. Rusija	55	21.2163094	33.7836906
5. Njemačka	42	45.445073	-3.4450730
6. Japan	41	57.8708298	-16.8708298
7. Francuska	42	35.5675508	6.4324492
8. Južna Koreja	21	24.1996874	-3.1996874
9. Italija	28	29.1397408	-1.1397408
10. Australija	29	21.7803352	7.2196648
11. Nizozemska	19	16,39308	2.6069200
12. Mađarska	15	7.4626413	7.5373587
13. Brazil	19	25.7187041	-6.7187041
14. Španjolska	17	22.1561474	-5.1561474
15. Kenja	13	7.1698931	5.8301069
16. Jamajka	11	4.879442	6.1205580
17. Hrvatska	10	5.7408555	4.2591445
18. Kuba	11	7.0534821	3.9465179
19. Novi Zeland	18	7.7905395	10.2094605
20. Kanada	22	25.6584951	-3.6584951
21. Uzbekistan	13	7.1554934	5.8445066
22. Kazahstan	18	7.9750986	10.0249014
23. Kolumbija	8	10.0857531	-2.0857531
24. Švicarska	7	14.5835995	-7.5835995

7.6. Prognoza rezultata

Bliže se Olimpijske igre 2020. u Tokiju pa će se tako isprobati odabrani model 5 kako bi došli do prognoze uspjeha SAD-a, Kine i Italije. Te države su ušle u prognoziranje zbog jako malih rezidualnih odstupanja u prošlim Olimpijskim igrama 2016. Kako bi prognoza bila što vjerodostojna važno je i sve nezavisne varijable isto prebaci na 2020. godinu. S obzirom da su potrebni podaci za sljedeću godinu što je samo po sebi nemoguće, potrebno je prikupiti prognozirane podatke o BDP-u i broju stanovnika s reputabilne web stranice (npr. www.statista.com). Nakon unosa prognoziranih demografskih i ekonomskih podataka u jednadžbu modela dobiveni su sljedeći rezultati:

SAD 2020.

Prognoza ukupnih medalja dobivena regresijskim modelom 5 za SAD iznosi **125.698 medalja**.

KINA 2020.

Prognoza ukupnih medalja dobivena regresijskim modelom 5 za Kinu iznosi **85.484 medalja**.

ITALIJA 2020.

Prognoza ukupnih medalja dobivena regresijskim modelom 5 za Italiju iznosi **31.236 medalja**.

8. Zaključak

Cilj ovog rada je analizirati do koje mjere na olimpijski uspjeh pojedine države utječu demografski i ekonomski čimbenici. Istraživanje je dalo saznanje da se dosta dobro može prognozirati uspjeh država na Olimpijskim igrama samo ako se prikupe podaci o broju stanovnika i BDP-u.

Nakon što su teorijski obrađeni svi preduvjeti za testiranje i provedbu regresije, definirani su kriteriji koji će se koristiti prilikom odabira najboljeg modela. Tijekom istraživanja isprobano je nekoliko modela u svrhu pronalaska modela koji će najbolje predicirati uspjeh pojedine države na nekim od sljedećih Olimpijskih igara. Nakon analize više modela, odabran je regresijski model 5 u radu. Upravo taj model sadrži najveći koeficijent determinacije R^2 kao i najveći korigirani koeficijent determinacije što je bitno za naglasiti jer on uzima u obzir i broj nezavisnih varijabli u modelu. Regresijska jednadžba tog modela ima sljedeći izgled: $\text{ukupno medalja} = 10.2511 - \sqrt{(\text{BDP\^3})} * 64.1083 + \ln(\text{broj stanovnika}) * 0.9413 + \text{BDP} * 151.9691 - \text{broj\_stanovnika} * 26.2194.$

Prije samog eksperimentiranja s regresijskim modelima, izračunati su korelacijski koeficijenti između pojedine nezavisne varijable (X_1,X_2,X_3) s odabranom zavisnom varijablom Y_2. Takav test je rezultirao istaknućem visine korelacije između BDP-a i ukupnog broja medalja. Stršila u velikoj mjeri utječu na korelaciju tako da bi se rezultati znatno promijenili kada SAD i Kina ne bi bile dio analize.

Pojava slučajnih grešaka (reziduala) bila je pretpostavljena tako da tu nije bilo nikakvih iznenađenja. Bitno je napomenuti da su se pojavile tri poprilično znatne pogreške kod analize, ali sve zajedno gledano odabrani model je dosta dobro opisao rezultate.

Za ovakav rad baziran na statističkoj obradi podatka, vrlo je važno na početku prikupiti valjane ulazne podatke. Ti podaci su se prikupljali putem web servisa. Regresijske analize su se obavljale pomoću RStudio alata koji je primarno namijenjen statističkoj obradi podataka.

Zaključak ovog rada je da na olimpijski uspjeh država utječu stotine faktora od kojih je većina nemjerljiva. Pristupom ovim osnovnim ekonomskim i demografskim podacima moguće je u grubo prognozirati ljestvicu uspjeha država na Olimpijskim igrama. S druge strane, prognoziranje točnog broja medalja je iznimno teško, skoro pa nemoguće jer ima mnogo trenutno nemjerljivih faktora, a i demografski i ekonomski pokazatelji koji se koriste za prognoziranje su podložni godišnjim promjenama.

Popis literature

1. Dumičić K., Bahovec V. (2011) *Poslovna statistika*, Zagreb: Element
2. Kero, K. (2003). *Statistika u primjerima*. Varaždin: Fakultet organizacije i informatike.
3. Benšic M., Šuvak N. (2013) *Primijenjena statistika,* Osijek: Sveucilište J. J. Strossmayera
4. Šošić I., Serdar V. (1995) *Uvod u statistiku*. Zagreb: Školska knjiga
5. Carter N., Felton N., Schwertman N. (2014) *A Classroom Investigation of the Effect of Population Size and Income on Success in the London 2012 Olympics*, Chico: California State University.
6. Montgomery D.C., Runger G.C.(2010) *Applied Statistics and Probability for Engineers*, New York: John Wiley & Sons
7. Kutner M.H., Nachtsheim C.J.,Neter J.,Li W. (2005) *Applied Linear Statistical Model,* New York: McGraw-Hill/Irwin
8. Veleučilište u Rijeci [VELERI] (bez dat.) *Korelacijska i regresijska analiza*. Preuzeto: 05.07.2019. s https://www.veleri.hr/files/datotekep/nastavni_materijali/k_poduzetnistvo_s1/Kvantitativne_za_poduzetnike_Pr2_Izv.pdf
9. Prirodoslovno - matematički fakultet [PMF] (bez dat.) *Regresijska analiza.* Preuzeto 05.07.2019. s https://www.pmf.unizg.hr/_download/repository/PREDAVANJE11.pdf
10. Fakultet strojarstva i brodogradnje [FSB] (2012) *Korelacija i regresija.* Preuzeto 05.07.2019. s https://www.fsb.unizg.hr/atlantis/upload/newsboard/30_03_2012__16701_KRegress_OPTIPLAPOK_2012_Compatibility_Mode.pdf
11. Kineziološki fakultet [KIF] (bez dat.) *Regresijska analiza* Preuzeto 05.07.2019. s http://km.com.hr/wp-content/uploads/2018/04/11.-Regresijska-analiza.pdf
12. Šimić, D. (bez dat.). Regresija inferencijalno,. Statistika [Moodle]. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin
13. Polazni podaci preuzeti 27.6.2019. s https://data.worldbank.org/

14. Bruto domaći proizvod – BDP (bez dat.). U Ekonomski rječnik. Preuzeto 5.7.2019. s http://www.ekonomskirjecnik.com/definicije/bruto-domaci-proizvod-bdp.html
15. *Rio 2016: The greatest show on Earth in stats*, (2016). Preuzeto 5.7.2019. s https://www.bbc.com/sport/olympics/37148372
16. Prognozirani podaci preuzeti 27.6.2019. s www.statista.com
17. Rstudio software za statističke analize 27.6.2019. preuzeto s https://www.rstudio.com/
18. Upute za korištenje alata Rstudio https://www.statmethods.net

Popis slika

Popis tablica