

Sveučilište J. J. Strossmayera u Osijeku  
Odjel za matematiku  
Sveučilišni diplomski studij matematike  
Financijska matematika i statistika

**Matej Petrinović**

**Multivarijatna analiza**

Seminarski rad

**Regresijski model očekivane životne dobi**

Voditelj: prof. dr. sc. Mirta Benšić

Osijek, 2019.

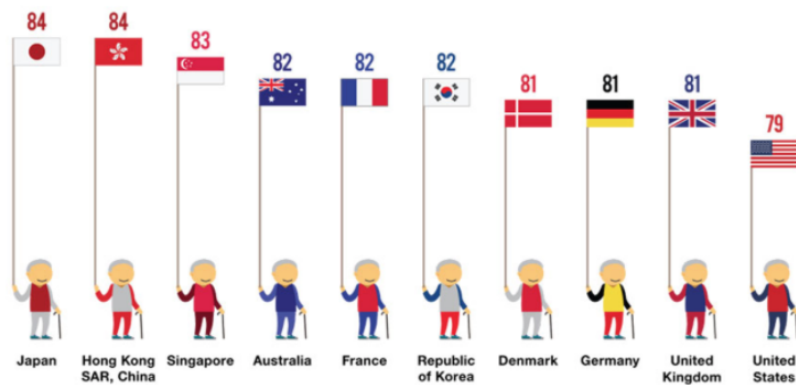
# Sadržaj

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Uvod</b>                                     | <b>1</b>  |
| <b>2</b> | <b>Analiza i priprema podataka</b>              | <b>2</b>  |
| 2.1      | Deskriptivna statistika varijabli . . . . .     | 2         |
| 2.1.1    | Stopa rođenih i umrlih . . . . .                | 2         |
| 2.1.2    | Stopa smrtnosti doječadi . . . . .              | 3         |
| 2.1.3    | Dob Muškaraca i Dob Žena . . . . .              | 4         |
| 2.1.4    | BDP . . . . .                                   | 5         |
| 2.1.5    | Grupa i Ime . . . . .                           | 6         |
| 2.2      | Priprema podataka . . . . .                     | 7         |
| <b>3</b> | <b>Modeliranje</b>                              | <b>9</b>  |
| 3.1      | Procjena koeficijenata . . . . .                | 9         |
| 3.2      | $K$ -fold kros validacija . . . . .             | 9         |
| 3.3      | Testiranje . . . . .                            | 10        |
| <b>4</b> | <b>Dijagnostika modela</b>                      | <b>11</b> |
| 4.1      | Analiza greški modela . . . . .                 | 11        |
| 4.2      | Homoskedastičnost modela . . . . .              | 12        |
| 4.3      | Multikolinearnost . . . . .                     | 13        |
| <b>5</b> | <b>Stršeće vrijednosti i utjecajna mjerenja</b> | <b>13</b> |
| 5.1      | Leverage score . . . . .                        | 13        |
| 5.2      | Cook-ova udaljenost . . . . .                   | 14        |
| <b>6</b> | <b>Zaključak</b>                                | <b>15</b> |

# 1 Uvod

Baza podataka *Poverty* sadrži podatke o 97 zemalja svijeta. Podaci su uzeti iz knjiga UNESCO 1990 Demographic Year Book (1990), New York: United Nations i Day, A. (ed.) (1992), The Annual Register 1992, 234, London: Longmans. Za ove zemlje dani su podaci za broj rođenih, umrlih, stope smrtnosti dojenčadi, očekivane životne dobi kod muškaraca i žena, te bruto domaći proizvod (BDP).

U ovom seminarskom radu analizirat ćemo opisanu bazu *Poverty*. Napraviti ćemo regresijski model u kojem ćemo pokušati modelirati životnu dob na temelju prediktora kao što su stopa rođenih i umrlih, stopa smrtnosti dojenčadi i BDP-a. Za potrebe ovoga semirana definirana je nova varijabla Očekivana Dob kao prosjek očekivane dobi muškaraca i žena. Baza podataka je preuzeta sa sljedećeg linka: [https://ww2.amstat.org/publications/jse/jse\\_data\\_archive.htm](https://ww2.amstat.org/publications/jse/jse_data_archive.htm)



Slika 1: Očekivani životni vijek

## 2 Analiza i priprema podataka

U ovom poglavlju dat ćemo pregled podataka koji se nalazi u skupu podataka POVERTY. Prikazat ćemo nekoliko redova tablice podataka kako bi dobili prvi dojam o podacima s kojima radimo. Nadalje ćemo svaku varijablu posebno analizirati i dati komentar, te grafički prikaze. Za početak pogledajmo nekoliko redova talice podataka:

|   | Natality | Mortality | Infant_Mortality | Male_Life_Exp | Female_Life_Exp | GDP  | Group | Name              |
|---|----------|-----------|------------------|---------------|-----------------|------|-------|-------------------|
| 1 | 24.70    | 5.70      | 30.80            | 69.60         | 75.50           | 600  | 1     | Albania           |
| 2 | 12.50    | 11.90     | 14.40            | 68.30         | 74.70           | 2250 | 1     | Bulgaria          |
| 3 | 13.40    | 11.70     | 11.30            | 71.80         | 77.70           | 2980 | 1     | Czechoslovakia    |
| 4 | 12.00    | 12.40     | 7.60             | 69.80         | 75.90           | 75.9 | 1     | Former_E._Germany |
| 5 | 11.60    | 13.40     | 14.80            | 65.40         | 73.80           | 2780 | 1     | Hungary           |
| 6 | 14.30    | 10.20     | 16.00            | 67.20         | 75.70           | 1690 | 1     | Poland            |

Tablica 1: Pregled podataka

### 2.1 Deskriptivna statistika varijabli

#### 2.1.1 Stopa rođenih i umrlih

Varijable *Stopa rođenih i umrlih* (varijable *Natality* i *Mortality*) su numeričkog tipa i sadrže informacije o broju rođenih, odnosno umrlih ljudi na 1000 osoba. Osnovne informacije o varijablama *Stopa rođenih* i *Stopa umrli* mogu se vidjeti na sljedećoj tablici.

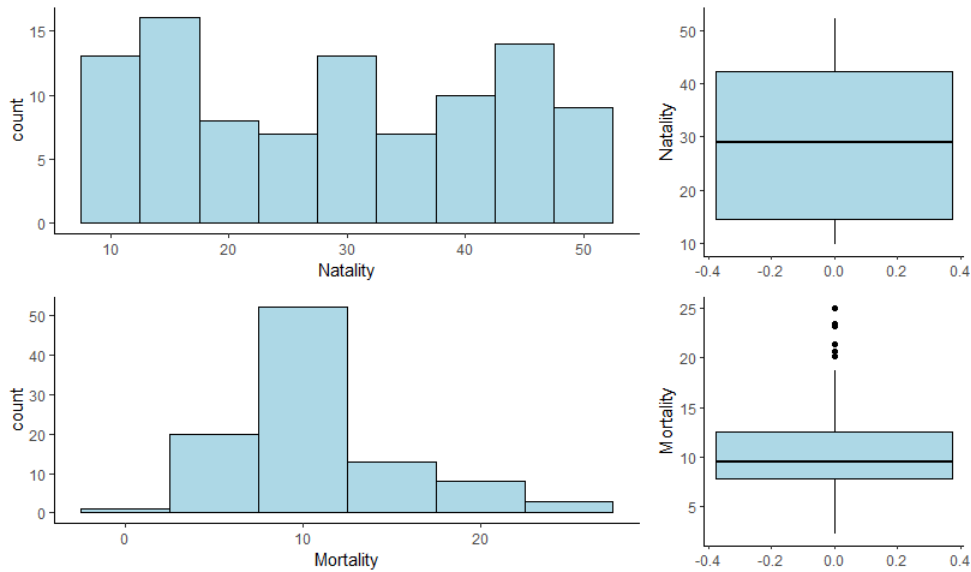
|        | Min  | Q1    | Med   | Mean  | Q3    | Max   |
|--------|------|-------|-------|-------|-------|-------|
| Rođeni | 9.70 | 14.70 | 29.00 | 29.46 | 42.55 | 52.20 |
| Umrli  | 2.20 | 7.70  | 9.50  | 10.73 | 12.30 | 25.00 |

Tablica 2: Osnovi podaci varijabli stope rođenih i umrlih

Formule za izračunavanje stope umrlih i rođenih, odnosno nataliteta i mortaliteta su sljedeće i gledaju se na 1000 stanovnika:

$$\text{NAT} = \frac{N_B}{N} \cdot 1000$$
$$\text{MORT} = \frac{N_D}{N} \cdot 1000$$

pri čemu  $N_B$  označava broj rođenih,  $N_D$  broj umrlih, te  $N$  označava ukupnu populaciju u određenoj zemlji. Najveći broj rođenih ljudi ima Uganda, dok najmanji broj ima Italija, a najveći broj umrlih ljudi ima Zimbabve, a najmanji ima Kuvajt.



Slika 2: Histogram i boxplot varijabli

Histogram i kutijasti dijagram jasno prikazuju raspon podataka. Također histogram varijable stope umrlih, bi mogao sugerirati na normalnu distribuciju pa stoga, testirajmo navedenu hipotezu. Nakon testiranja  $p$ -vrijednost iznosi 0.0003 što je svakako manje od  $\alpha = 0.05$  razine značajnosti, stoga ne možemo tvrditi postojanje normalne distribucije.

### 2.1.2 Stopa smrtnosti doječadi

Varijabla *Stopa smrtnosti doječadi* (Infant\_Mortality) je numeričkog tipa i sadrži informacije o broju umrle djece ispod dobi od godine dana na 1000 osoba. Osnovne informacije mogu se vidjeti u danoj tablici.

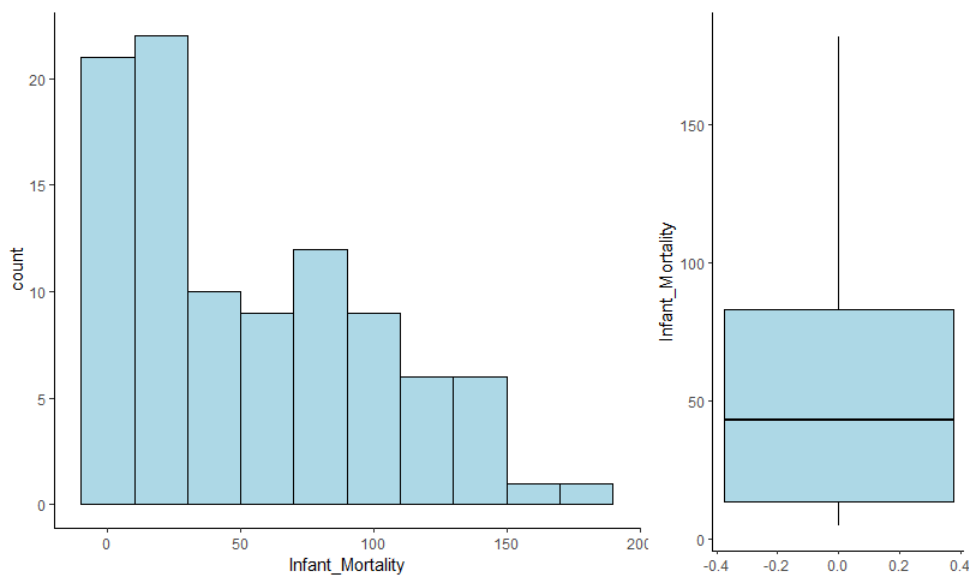
| Min  | Q1    | Med   | Mean  | Q3    | Max    |
|------|-------|-------|-------|-------|--------|
| 4.50 | 13.05 | 43.00 | 55.28 | 86.50 | 181.60 |

Tablica 3: Osnovi podaci varijable Stopa smrtnosti doječadi

Stopa smrtnosti dojenčadi izračunava se po sljedećoj formuli:

$$IM = \frac{1000}{N} \sum_{i=1}^n \mathbb{1}_{\{G_i < 1\}}$$

pri čemu  $N$  označava ukupnu populaciju neke zemlje, a navedena suma prebrojavanje ukupne umrle djece mlađe od godine dana. Pogledajmo dalje osnovne grafičke prikaze.



Slika 3: Histogram i boxplot varijabli

Iz histograma možemo vidjeti da veliki broj država ima mali broj umrle djece ispod godinu dana, dok mali broj država ima više od 150 umrle djece ispod godine dana. Zemlje koje imaju preko 150 umrle djece ispod godine dana su Afganistan i Sierra Leone.

### 2.1.3 Dob Muškaraca i Dob Žena

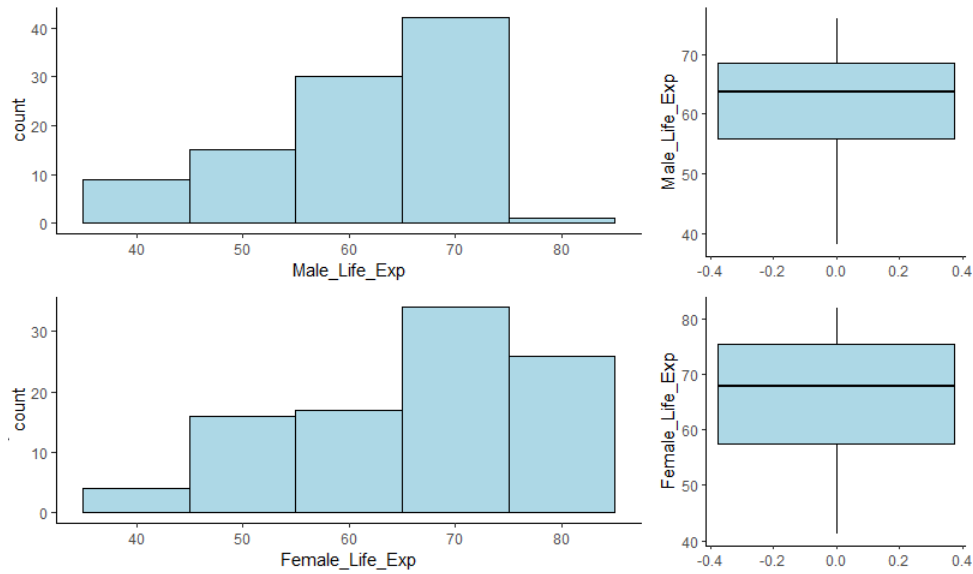
Ove dvije varijable *Dob Muškaraca* i *Dob žena* (Male\_Life\_Exp, Female\_Life\_Exp) su numeričke varijable koje prikazuju očekivanu dob muškaraca odnosno žena. Očekivano trajanje života statistička je mjera prosječnog vremena za koje se očekuje da će organizam živjeti na temelju godine rođenja, njezine trenutne dobi i drugih demografskih čimbenika uključujući spol. Takva mjera često se koristi u aktuaristici kao dodatna mjera izračuna premije na neku vrstu životnog osiguranja. Nova varijabla Očekivani životni vijek (koju ćemo kasnije u modeliranju koristiti kao zavisnu varijablu) definirana je na sljedeći način

$$\text{Očekivani vijek} = \frac{\text{Dob Muškaraca} + \text{Dob žena}}{2}$$

Osnove podatke možemo vidjeti u danoj tablici.

|               | Min   | Q1    | Med   | Mean  | Q3    | Max   |
|---------------|-------|-------|-------|-------|-------|-------|
| Žene          | 41.20 | 56.75 | 67.60 | 66.03 | 75.45 | 81.80 |
| Muškarci      | 38.10 | 55.40 | 63.40 | 61.38 | 68.50 | 75.90 |
| Očekivana dob | 39.65 | 56.12 | 65.65 | 63.71 | 71.55 | 78.85 |

Tablica 4: Deskriptivna statistika varijabli Dob Muškaraca i Žena



Slika 4: Histogram i boxplot varijabli

Iz slike kutijastih dijagrama možemo uočiti veće vrijednosti za žene nego kod muškaraca. Također iz histograma vidimo da više žena doživi oko ili više od 80 godina. Najveće očekivane dobi imaju stanovnici Japana, gdje je očekivana dob za muškarce 75.9, te za žene 81.8 godina.

#### 2.1.4 BDP

Varijabla *BDP* opisuje bruto domaći proizvod (BDP) po osobi. Bruto domaći proizvod je makroekonomski indikator koji pokazuje vrijednost finalnih dobara i usluga proizvedenih u zemlji tijekom dane godine, izraženo u novčanim jedinicama. U našem slučaju BDP je izražen u američkim dolarima. Često se u makroekonomskim i ekonometrijskim modelima koristi logaritam BDP-a. Osnovne podatke možemo vidjeti u sljedećoj tablici.

|          | Min   | Q1    | Med   | Mean  | Q3    | Max    |
|----------|-------|-------|-------|-------|-------|--------|
| GDP      | 80    | 475   | 1690  | 5741  | 7325  | 34064  |
| log(GDP) | 4.382 | 6.174 | 7.432 | 7.487 | 8.755 | 10.436 |

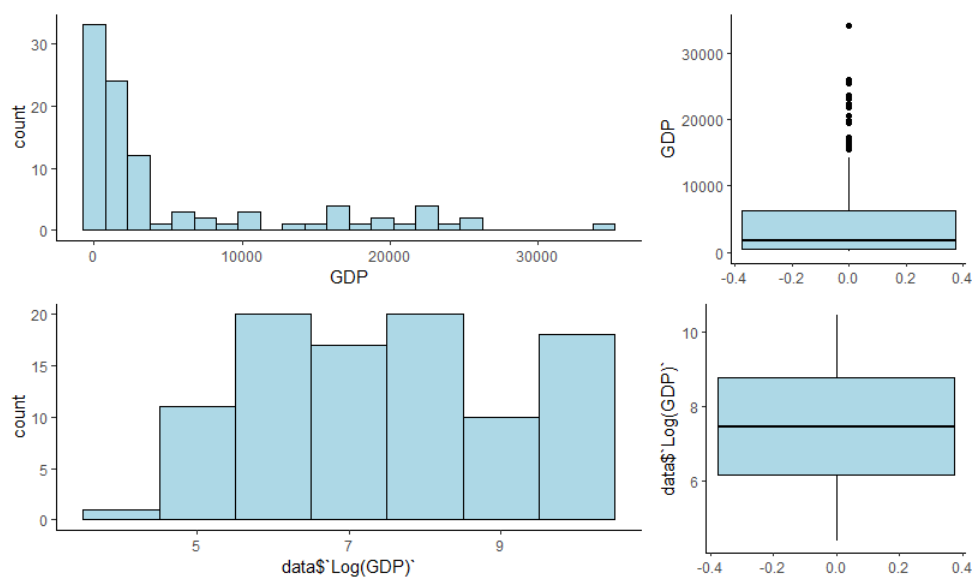
Tablica 5: Deskriptivna statistika varijable BDP

Realni BDP (oznaka  $Y$ ) možemo izračunati koristeći matematičku formulu

$$Y = C + I + G + E - U$$

pri čemu je  $C$  osobna potrošnja,  $I$  nacionalne investicije,  $G$  državna potrošnja,  $E$  izvoz i  $U$  uvoz. U našem slučaju se radi o BDP-u per capita kojeg dobijemo djeljenjem realnog BDP-a sa ukupnim brojem stanovništva (ozn.  $S$ )

$$Y_{\text{per capita}} = \frac{Y}{S}$$



Slika 5: Histogram i boxplot BDPa

Iz histograma vidimo da s povećanjem BDP-a broj zemalja se smanjuje. Na kutijastom dijagramu uočavamo stršeće vrijednosti. Najveća stršeća vrijednost je Švicarska čiji BDP iznosi čak 34064 \$, dok zemlja s najmanjim BDP-om je Monzabik koji iznosi svega 80 \$.

### 2.1.5 Grupa i Ime

Varijabla *Ime* sadrži ime države, a varijabla *Grupa* je kategorijalna varijabla koja opisuje grupu kojoj država pripada i to na način:

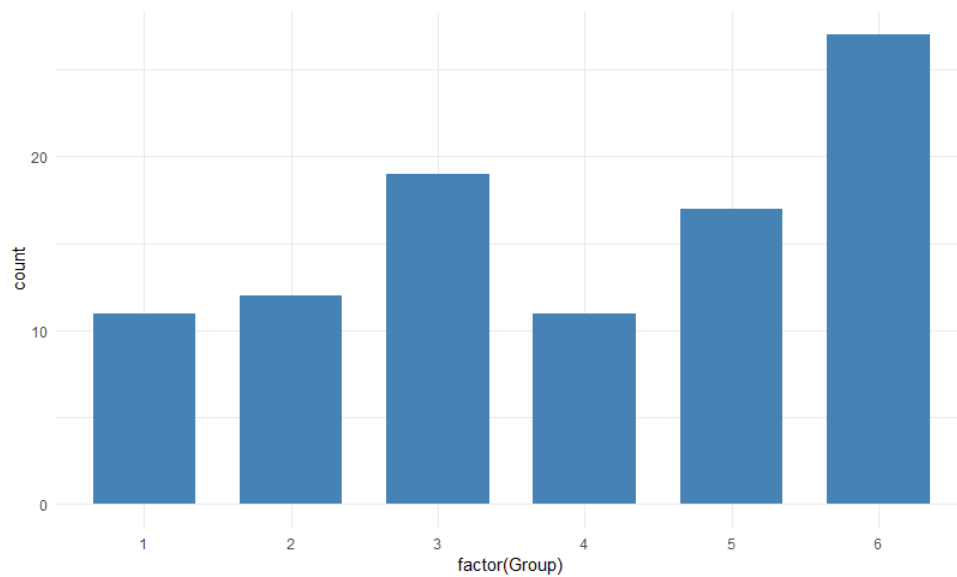
1. Istočna Europa
2. Južna Amerika i Meksiko
3. Zapadna Europa, Sjeverna Amerika, Japan, Australija, Novi Zeland
4. Bliski Istok
5. Azija
6. Afrika

U sljedećoj tablici možemo vidjeti broj zemalja po grupama.

| Grupa        | 1 | 2  | 3  | 4  | 5  | 6  |
|--------------|---|----|----|----|----|----|
| Broj zemalja | 9 | 12 | 19 | 10 | 14 | 27 |

Tablica 6: Broj zemalja po grupama

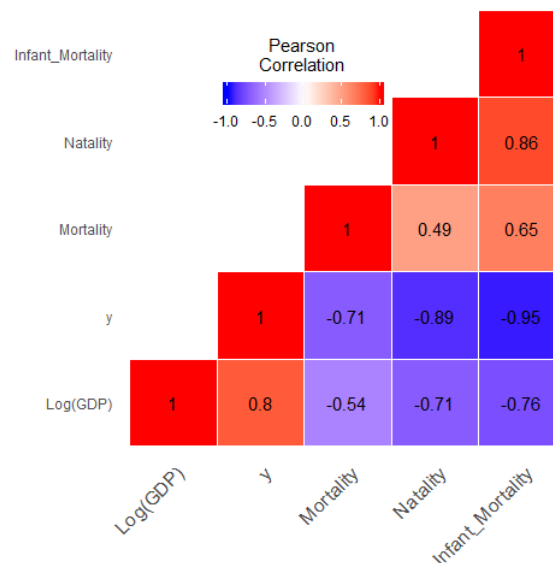
Također možemo prikazati i stupčasti dijagram koji će jasnije predložiti razmjer zemalja u pojedinim grupama.



Slika 6: Stupčasti dijagram Grupe

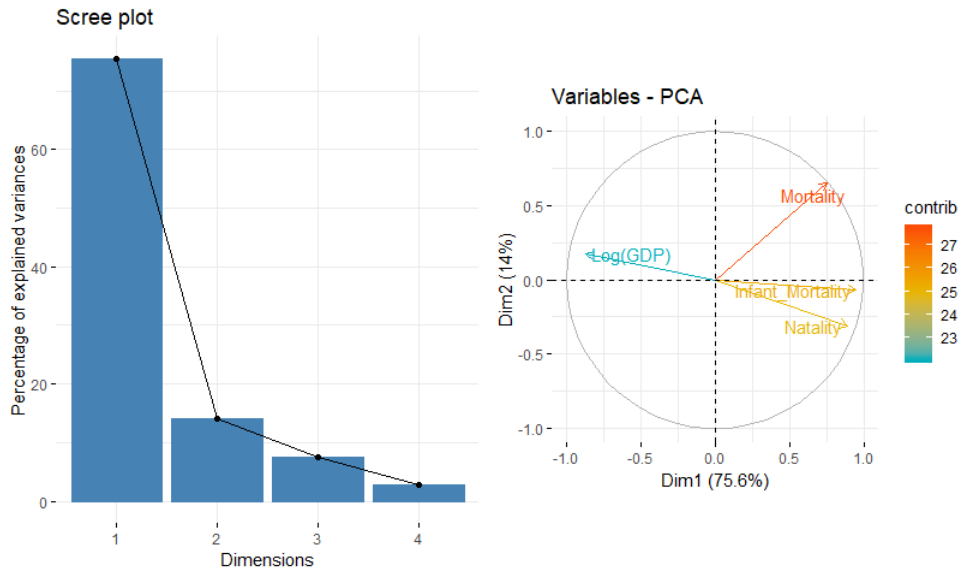
## 2.2 Priprema podataka

Prije samog modeliranja podatke moramo na neki način pripremiti. Kako smo ranije definirali zavisnu varijablu, samim time nam varijable za očekivanu životnu dob muškarca i žena ne trebaju. Nadalje, ekonometrijska teorija kaže da se financijski podaci uvijek uzimaju u logaritamskom obliku, stoga u modelu ćemo koristiti logaritam BDP-a. Najprije ćemo pogledat korelacijsku matricu zbog uvida u multikolinearnost među prediktorima i korelaciju sa zavisnom varijablom. To ćemo pokazati korelacijskom heatmapom:



Slika 7: Korelacijski prikaz

Iz slike vidimo kako je zavisna varijabla u jakim korelacijama s svim ostalim prediktorima. Naime, također se vide znatne korelacije i među prediktorima. Tako nešto bi moglo utjecati na kvalitetu modela. Najprije ćemo u model upotrijebiti sve varijable, te kasnije kroz dijagnostiku po potrebi uklanjati radi povećanja kvalitete. Pogledajmo još grafički prikaz objašnjene varijance pomoću PCA algoritma:



Slika 8: Objašnjena varijanca

Lijevi graf nam govori kako najveći udio u varijanci ima prva dimenzija, dok su sve ostale na neki način zanemarive. Desni prikaz govori vezu između prediktora i objašnjene varijance u prvoj i drugoj dimenziji. Isto tako što su strelice bliže to su prediktori međusobno jače korelirani. Na samom kraju analize i pripreme podataka ćemo dane podatke skalirati, kako nebi došlo do problema zbog međusobno drukčijih raspona. Podatke ćemo skalirati koristeći  $Z$ -score način, tako da podatke centriramo, tj. da je prosjek podataka 0, a standardna devijacija jednaka 1. Formula za  $Z$ -score je sljedeća:

$$Z_i = \frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i}$$

Pri čemu je  $X_i$   $i$ -ti prediktor, a  $\hat{\mu}_i$  i  $\hat{\sigma}_i$  su redom procjena očekivanja i standardne devijacije. Još ostaje podijeliti podatke na train i test dio. To ćemo podijeliti u omjeru 80 : 20 %.

### 3 Modeliranje

#### 3.1 Procjena koeficijenata

U ovom dijelu seminara bavit ćemo se modeliranjem zavisne varijable. Kako je zavisna varijabla numeričkog i neprekidnog tipa, model moramo tražiti u klasi regresijskih modela. Za modeliranje ćemo koristiti model linearne regresije, tj. tražimo model oblika

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

gdje su  $X_i$ ,  $i = 1, 2, 3, 4$  prediktori, a  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$  greška modela. Nakon kreiranja modela, rezultati, tj. procjena koeficijenata, pripadnim izračunatim greškama,  $p$ -vrijednostima  $T$ -statistike, i pouzdanim intervalom.

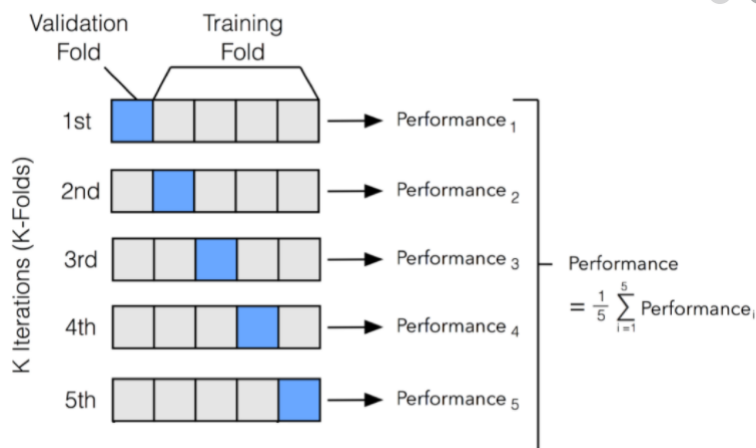
|                  | Estimate | Std. Error | $t$ value | $P(> t )$ | 2.5%  | 97.5% |
|------------------|----------|------------|-----------|-----------|-------|-------|
| (Intercept)      | 57.95    | 1.76       | 32.98     | 9.3e-53   | 54.46 | 61.44 |
| Natality         | -2.97    | 0.48       | -6.18     | 1.7e-08   | -3.92 | -2.01 |
| Mortality        | -1.97    | 0.32       | -6.15     | 1.9e-08   | -2.61 | -1.34 |
| Infant_Mortality | -4.96    | 0.57       | -8.70     | 1.2e-13   | -6.10 | -3.83 |
| logGDP           | 0.78     | 0.23       | 3.37      | 1.09e-03  | 0.32  | 1.25  |

Tablica 7: Procjenjene vrijednosti modela

Ako pogledamo tablicu, vidit ćemo kako su greške koeficijenata relativno male. Isto tako  $p$ -vrijednosti su daleko manje od zadane razine značajnosti  $\alpha = 0.05$ , te niti jedan pouzdani interval ne sadrži 0. Stoga daljnje promatranje metoda odabira koeficijenata neće biti potrebna. Nadalje,  $R^2 = 0.9518$ , što kaže da je čak 95% podataka objašnjeno ovim modelom, dok srednja kvadratna greška modela iznosi  $MSE = 5.28$  i apsolutna greška  $MAE = 1.72$ , što možemo interpretirati da je u projesku greška modela za predviđenu vrijednost otprilike jednaka  $y \pm 2$ , što je u konačnici jako dobro.

#### 3.2 $K$ -fold kros validacija

Prije testiranja modela na test dijelu podataka, provest ćemo  $K$ -fold kros validaciju za  $K = 10$ , radi pregleda prosječnih perfomansi modela. Postupak kros validacije objašnjen je na slici ispod



Slika 9: 5-fold kros validacija

Rezultati 10–fold kros validacije se nalaze u tablici ispod.

|    | RMSE | $R^2$ | MAE  | Resample |
|----|------|-------|------|----------|
| 1  | 3.31 | 0.92  | 2.60 | Fold01   |
| 2  | 3.71 | 0.90  | 2.86 | Fold02   |
| 3  | 1.29 | 0.98  | 1.15 | Fold03   |
| 4  | 2.85 | 0.90  | 1.87 | Fold04   |
| 5  | 1.88 | 0.98  | 1.30 | Fold05   |
| 6  | 1.84 | 0.98  | 1.40 | Fold06   |
| 7  | 1.97 | 0.97  | 1.84 | Fold07   |
| 8  | 2.88 | 0.93  | 2.31 | Fold08   |
| 9  | 2.23 | 0.95  | 1.98 | Fold09   |
| 10 | 1.18 | 0.99  | 0.90 | Fold10   |

Tablica 8: Rezultatiti 10-fold kros validacije

Iz tablice se izračuna prosječni  $R^2$  kao  $\bar{R}^2 = \frac{1}{10} \sum_{i=1}^{10} R_i^2 = 0.9499016$ . Slično se izračuna i  $\overline{RMSE} = 2.312698$ , te  $\overline{MAE} = 1.820664$ . Sama kros validacija daje jako dobre rezultate za svako reurzokovanje training skupa.

### 3.3 Testiranje

Kada smo pogledali rezultate kros-validacije, testirat ćemo model na test dijelu podataka. Ovo služi za uvid kako se model ponaša na podacima koje nikad nije vidio. Rezultate je najbolje pogledati u nekoliko redaka tablice:

|                  | Natality | Mortality | Infant_Mortality | logGDP | y     | pred  |
|------------------|----------|-----------|------------------|--------|-------|-------|
| Romania          | -1.15    | -0.03     | -0.61            | 7.40   | 69.45 | 70.25 |
| Byelorussian_SSR | -1.04    | -0.29     | -0.91            | 7.54   | 71.15 | 72.01 |
| Ukrainian_SSR    | -1.17    | 0.16      | -0.91            | 7.19   | 70.60 | 71.25 |
| Uruguay          | -0.83    | -0.27     | -0.72            | 7.85   | 71.65 | 70.65 |
| Denmark          | -1.24    | 0.23      | -1.03            | 10.00  | 74.75 | 74.14 |
| Netherlands      | -1.18    | -0.48     | -1.04            | 9.76   | 76.60 | 75.22 |

Tablica 9: Dobivene vrijednosti na test dijelu podataka

Ako pogledamo rezultate možemo vidjeti da su razlike između stvarnih vrijednosti (y) i prediktiranih vrijednosti (pred) jako male. Valja napomenuti da tablica prikazuje prethodno obrađene podatke, tj. skalirane i logaritmirano.  $MSE$  na test dijelu iznosi 4.11, dok je  $MAE$  1.5. U konačnici možemo reći da model pokazuje izvrsne rezultate. Koliko je model zapravo valjan i matematički točan obrađujemo u idućem poglavlju.

## 4 Dijagnostika modela

U ovom poglavlju seminara bavit ćemo se dijagnostikom kreiranog modela. Najprije da bi model bio valjan i uzeo se u razmatranje treba zadovoljavati sljedeće pretpostavke:

1. Greške su međusobno nezavisne i distribucija greške je normalna
2. Model je homoskedastičan, tj. varijance greške je konstantna.

Nakon provjere pretpostavki napraviti ćemo analizu multikolinearnosti među prediktorima, te ispitati postoji li u podacima stršaćih i utjecajnih mjerenja, tj. podataka.

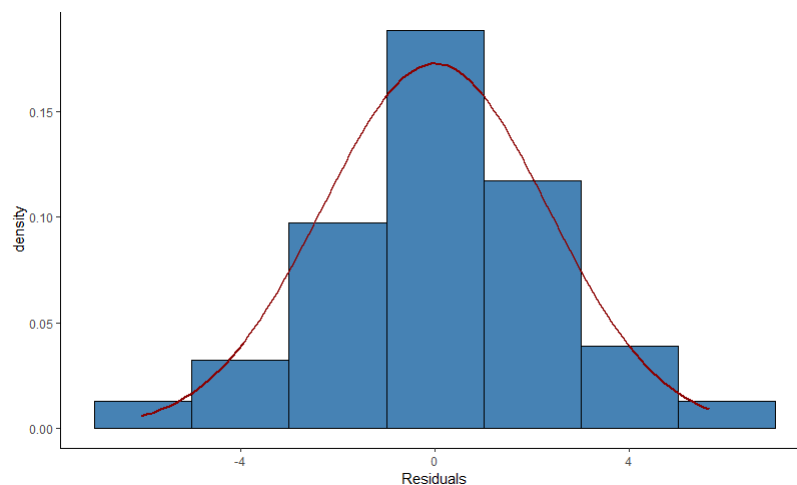
### 4.1 Analiza greški modela

Kako smo ranije rekli, greška modela  $\varepsilon$  mora slijediti normalnu distribuciju  $\mathcal{N}(0, \sigma^2)$ . Greška modela je nemjerljiva veličina, te se procjenjuje rezidualima,  $\hat{\varepsilon}$  koji se definiraju na sljedeći način kao  $\hat{\varepsilon}_i = y_i - y_{pred}$ . Pogledajmo tablicu s osnovnim karakteristikama reziduala:

|     | Min.  | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|-----|-------|---------|--------|------|---------|------|
| rez | -6.04 | -1.02   | -0.15  | 0.00 | 1.36    | 5.63 |

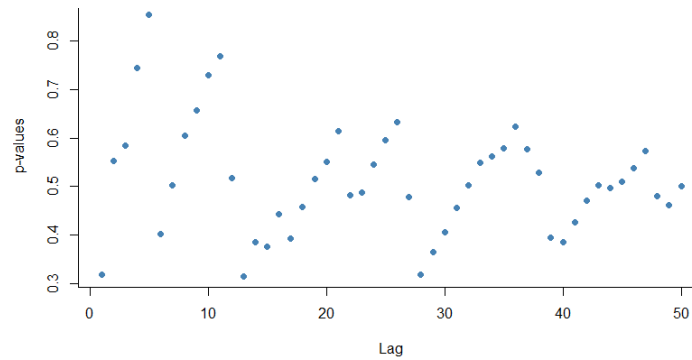
Tablica 10: Osnovna statistika reziduala

Iz tablice vidimo kako je procjena očekivanja za reziduale jednaka 0, što je jedan korak bliže zadovoljavanju pretpostavke. Dalje ćemo pogledati histogram reziduala i vidjeti postoji li sugeracija na normalnu distribuciju. Procjenjena varijanca reziduala iznosi 5.35, pa grešku modeliramo kao  $\varepsilon \sim \mathcal{N}(0, 5.35)$ .



Slika 10: Histogram reziduala

Histogram vidno sugerira na normalu distribuciju, stoga testirajmo tu hipotezu. Provođenjem Shapiro-Wilk testa dobivamo  $p$ -vrijednost = 0.3464, što je svakako veće od  $\alpha = 0.05$  stoga nema razloga sumnjati u normalnost. Još ostaje testirati međusobnu nekoreliranost reziduala. Navedeno ćemo napraviti koristeći Breusch-Godfrey Test za testiranje koreliranosti na koracima. Testirat ćemo do koraka 50, a rezultate, tj.  $p$ -vrijednosti, prikazati grafički.

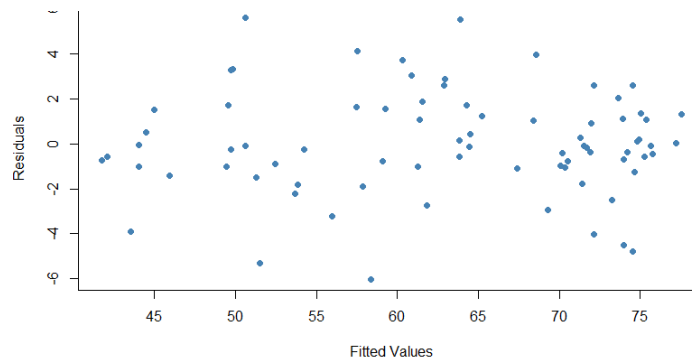


Slika 11:  $p$ -vrijednosti BG Testa

Iz slike vidimo kako je najmanja  $p$ -vrijednost 0.3, te stoga možemo tvrditi kako nema razloga sumnjati u korelaciju reziduala. Ovom analizom uspostavili smo da greške možemo smatrati normalno distribucijom, te da su reziduali međusobno ne korelirani čime je pretpostavka 1. ispunjena.

## 4.2 Homoskedastičnost modela

Homogenost varijanci reziduala provjerit ćemo analizom grafičkog prikaza ovisnosti reziduala o teorijskim vrijednostima, te iz toga ustvrditi postoji li nekakva pravilnost ili zavisnost. Prikaz je ispod na slici.



Slika 12: Odnos reziduala i teorijskih vrijednosti

Grafički prikaz ne sugerirana na postojanje neke pravilnosti, no svakako ćemo navedenu tvrdnju testirati koristeći `ncvTest` i Breusch-Pagan Test. a osnovu ovoga testirajmo hipotezu o homogenosti varijance na razini značajnosti  $\alpha = 0.05$ .

$$H_0 : \text{Varijanca je konstanta}$$

$$H_1 : \text{Varijanca nije konstanta}$$

Testiranjem `ncvTest`-om dobijamo  $p$ -vrijednost 0.26536, a Breusch-Pagan Test 0.3185 što je veće od 0.05 pa ne odbacujemo nul hipotezu, tj. nemamo razloga sumnjati u konstantnost varijance. Nakon ove analize zaključujemo kako kreirani model zadovoljana obje nužne pretpostavke i matematički je korektan, te se može koristiti za daljnju potrebu.

### 4.3 Multikolinearnost

U ovom dijelu seminara ispitat ćemo postoji li korelacija između prediktora. Izračunat ćemo faktor inflacije varijance, VIF. Sumnju na postojanje multikolinearnosti vidjeli smo na korelacijskoj matrici gdje varijabla Infant\_Mortality ima povećane korelacije sa svim ostalima. Pogledajmo tablicu VIF-ova za svaku varijablu u modelu:

| Varijabla        | VIF  |
|------------------|------|
| Nativity         | 3.84 |
| Mortality        | 1.90 |
| Infant_Mortality | 5.61 |
| logGDP           | 2.69 |

Tablica 11: VIF vrijednosti

Uočavamo kako prediktor Infant\_Mortality ima VIF vrijednost veću od 5, stoga taj prediktor stvara problem kolinearnosti s ostalim prediktorima. Idejno bi bilo taj prediktor ukloniti, no pitanje je gubimo li tada na točnosti modela? Takvu tvrdnju ćemo testirati Waldovim testom, gdje ćemo usporediti puni model i model bez navedenog prediktora. Rezultat je sljedeći:

|   | Res.Df | Df | Chisq | $P(>Chisq)$ |
|---|--------|----|-------|-------------|
| 1 | 72     |    |       |             |
| 2 | 73     | -1 | 66.35 | 3.777e-16   |

Tablica 12: Rezultati Waldovog testa

Rezultat testa, tj.  $p$ -vrijednost, kaže da bi uklanjanjem tog prediktora izgubili na kvaliteti. Stoga ćemo taj prediktor ostaviti, a svu daljnju upotrebu uzimati s oprezom.

## 5 Stršeće vrijednosti i utjecajna mjerenja

U ovom dijelu bavit ćemo se detekcijom stršećih i utjecajnih mjerenja. Prikazat ćemo rezultate kao što su leverage score i Cookova udaljenost, te analizu outliera.

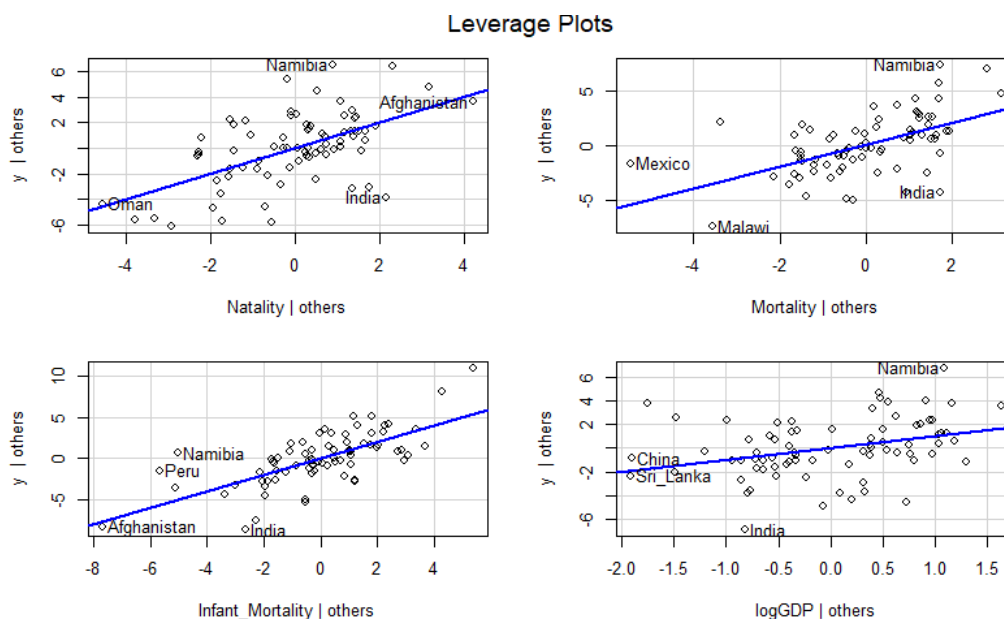
### 5.1 Leverage score

Leverage score zapravo predstavlja vrijednost utjecaja podatka na model. Što je naravno taj score veći, veći je i utjecaj. Drugim riječima taj podatak može odvući regresijski pravac. Pogledajmo najprije tablicu s osnovnim podacima leverage scorea.

|      | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|------|---------|--------|------|---------|------|
| Lev. | 0.02 | 0.04    | 0.06   | 0.06 | 0.07    | 0.21 |

Tablica 13: Osnovna statistika leverage scorea

Zemlja koja ima najveći leverage score je Afganistan, dok zemlja s najmanjim je Vijetnam. Na idućem grafičkom prikazu možemo vidjeti posebno vidjeli leverage score-ove za svaki prediktor posebno, u odnosu za zavisnu varijablu.

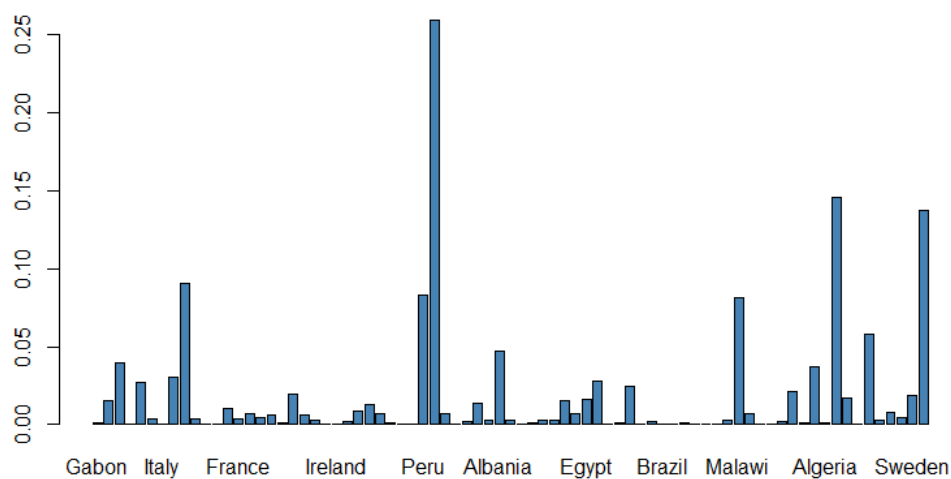


Slika 13: Leverage score

Vidimo iz grafičkih prikaza da se Afganistan ističe na čak dva grafa. Vidimo da Nambia također se posebno ističe na svim grafovima, no nema najveći score na ukupnom modelu.

## 5.2 Cook-ova udaljenost

Cook-ova udaljenost je posebna mjera za detekciju utjecaja. Definira se kao suma svih promjena u regresijskom modelu kada se ukloni promatranje  $i$ -to promatranje. Najveću vrijednost Cook-ove udaljesti ima Koreja, što znači da bi se njezin uklanjanjem dogodila najveća promjena.



Slika 14: Cook-ova udaljenost

## 6 Zaključak

Na osnovi promatranja baze podataka POVERTY modelirali smo očekivani životni vijek stanovnika na osnovu sljedećih prediktora: stopa rođenih, stopa umrlih, stopa smrtnosti dojenčadi i BDP. Podaci su prvobitno očišćeni i dovedeni u prihvatljivu formu za modeliranje. Model je pokazao dobra prediktivna svojstva, unatoč problemu sa multikolinearnosti. Također model je pokazao ispunjenost svih pretpostavki modela linearne regresije, te se kao takav može dalje koristiti. Pokazao je visoku točnost na test dijelu podatak, te jako visoku vrijednost koeficijenta determinacije. Model bi se dalje mogao nastaviti poboljšavati uporabom drugih metoda, kao što je  $L_1$  regresija, koja nije osjetljiva na utjecjana mjerenja. Također mogao bi se tražiti drugi model u klasi algoritama strojnog učenja ali to nije u sastavu ovoga kolegija.