Modeliranje vrijednosti igrača u nogometu

Kalmar, Kristijan

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Mathematics / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:126:067744

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

Kristijan Kalmar

Modeliranje vrijednosti igrača u nogometu

Diplomski rad

Osijek, 2022.

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

Kristijan Kalmar

Modeliranje vrijednosti igrača u nogometu

Diplomski rad

Mentor: izv. prof. dr. sc. Danijel Grahovac

Osijek, 2022.

Sadržaj

1 Uvod

Nogomet je sport u kojemu se dvije momčadi, od kojih se svaka sastoji od 11 igrača, nadmeću na pravokutnom igralištu travnate površine. Osnovni cilj u nogometu je postizanje više pogodaka od protivničke momčadi bilo kojim dijelom tijela osim ruke. Vratar ili popularnije 'golman' je jedini igrač kojemu je dozvoljeno igrati, to jest braniti rukama, doduše samo u jasno označenom pravokutniku ispred svoga gola.

Pogledamo li dublje u povijest, igre s loptom zabilježene su i prije nove ere u drevnoj Kini te u Rimskom carstvu. Što se tiče razvoja nogometa u Hrvatskoj, prvi službeni zapis nalazi se u Rijeci na vanjskom zidu crkve sv. Roka gdje su engleski mornari odigrali utakmicu protiv Mađara.

Nogomet se razvijao i širio svijetom velikom brzinom. Danas je najmasnovniji i najpopularniji sport na svijetu. Kako u današnje vrijeme popularnost donosi zainteresiranost mnogih sponzora, nogometni klubovi u 21. stoljeću često imaju bogate vlasnike. Najpoznatiji primjer takvih klubova su Manchester City, Paris Saint Germain i Newcastle koji su u vlasništvu naftnih kompanija čije se bogastvo mjeri u milijardama dolara. Kako danas postoji na stotine programa koji prenose sport, nogometni klubovi mnogo profitiraju i od televizijskih prava. Velike svote novca su u svijetu nogometa pa tako kompanije žele da se njihove reklame prikazuju na utakmicama čime klubovi zarađuju dodatan novac od sponzora.

Budući da se sve vrti oko novca, sve prethodno navedeno uvod je za temu ovog diplomskog rada. Dakle, u ovom diplomskom radu izradit ćemo model koji će procjenjivati tržišnu vrijednost igrača u nogometu. Za taj postupak koristit ćemo višestruku linearnu regresiju o kojoj ćemo više reći u prvom poglavlju rada. Na temelju fizičkih predispozicija, statističkih podataka, područja rođenja te raznih drugih parametara procjenjivat ćemo tržišnu cijenu nogometaša.

2 Regresija

U većini istraživanja cilj je opisivanje veza između pojava koje nas okružuju. U statističkim istraživanjima takav postupak zovemo *regresija*. *Regresija* je od velikog značaja, najviše u ekonomiji, ali se često koristi i u prirodnim znanostima poput biologije, kemije itd. Regresijska analiza koristi se za donošenje zaključaka o slučajnoj varijabli Y (u radu će to biti tržišna cijena nogometaša) koje ovise o nezavisnoj varijabli x ili o nizu slučajnih varijabli $Y_1, \ldots, Y_n$ koje ovise o vektoru varijabli $\mathbf{x} = (x_0, x_1, \ldots, x_p)$ koje nazivamo prediktorima.

2.1 Linearna regresija

Osnovna ideja linearne regresije je zbrojiti učinke svih varijabli kako bi se dobila vrijednost predviđanja. Očekivanje modela linearne regresije je linearna funkcija parametara. Pomoću neovisnih varijabli $x_0, x_1, \ldots, x_p$ modeliramo očekivanje ovisne varijable Y na sljedeći način:

$$\mathrm{E}(Y) = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \tag{2.1}$$

gdje su $\beta_0, \beta_1, \ldots, \beta_p$ nepoznati parametri. Pretpostavimo da varijanca ne ovisi o $\mathbf{x}$, odnosno $Var(Y) = \sigma^2$ i da su vrijednosti $x_0, x_1, \ldots, x_p$ izmjerene bez greške. Učestalo možemo vidjeti i sljedeći zapis modela:

$$Y = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon, \tag{2.2}$$

gdje ε predstavlja slučajnu varijablu sa svojstvima $\mathrm{E}(\varepsilon) = 0$ i $Var(\varepsilon) = \sigma^2$. Osnovni cilj linearne regresije je procjena nepoznatih parametara $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ u svrhu određivanja predikcijskih vrijednosti ovisne varijable i u svrhu lakše interpretacije veza između ovisne varijable i prediktora. U jednadžbu je moguće uvesti konstantni član postavljanjem prvog elementa u $\mathbf{x}$ na 1, odnosno $x_0 = 1$. Tada za vektor $\mathbf{x} = (1, x_1, \ldots, x_p)$ vrijedi $\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. Promotrimo za početak jednostavni oblik linearne regresije gdje je $p = 1$ i $\mathbf{x} = (1, x_1)$.

Prvo ćemo opisati model jednostavne linearne regresije

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.3}$$

gdje je $\mathrm{E}(\varepsilon) = 0$ i $Var(\varepsilon) = \sigma^2$. Zatim ćemo metodom najmanjih kvadrata (*eng.* Least Square Method) procijeniti vrijednost parametara β_0 i β_1.

Pretpostavimo da su $x_0, x_1, \ldots, x_n$ realni brojevi koji predstavljaju zabilježene vrijednosti niza n nekoreliranih slučajnih varijabli oblika (2.3). Zatim pretpostavimo da za $i = 1, \ldots, n$ vrijedi

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i, \qquad Var(Y_i) = \sigma^2, \qquad Cov(Y_i, Y_j) = 0,\ i \neq j.$$

Promatrat ćemo podatke kao uređene parove $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, pri čemu su $x_1, x_2 \ldots, x_p$ vrijednosti neovisne varijable $\mathbf{x}$, a $y_1, y_2 \ldots, y_n$ odgovarajuće vrijednosti slučajnih varijabli $Y_1, Y_2 \ldots, Y_n$. Zabilježenu vrijednost slučajne varijable Y_i zapišimo u obliku $y_i = \beta_0 + \beta_1 x_i + e_i$ tako da e_i u tom slučaju nazivamo greškom odnosno e_i predstavlja razliku između teorijske vrijednosti $\mathrm{E}(Y_i)$ i zabilježene vrijednosti u i-tom koraku pokusa. Sljedeći korak jest da pokušamo povući pravac kroz $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ takav da se minimizira udaljenost tih točaka od pravca, odnosno biramo pravac koji minimizira greške modela $e_i = y_i - \beta_0 - \beta_1 x_i$, tj. želimo pronaći β_0 i β_1 takve da greške budu minimalne. Sada ćemo metodom najmanjih

kvadrata minimizirati sumu kvadrata udaljenosti od pravca. To će nam pomoći pri traženju procijenjenih vrijednosti od β_0 i β_1. Nazvat ćemo ih $\widehat{\beta_0}$ i $\widehat{\beta_1}$ koji minimiziraju sljedeću sumu

$$S = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Sljedeći korak jest da parcijalno deriviramo funkciju S po β_0 i β_1 nakon toga izjednačavanjem parcijalnih derivacija s 0 dobijemo procjenitelje $\widehat{\beta_0}$ i $\widehat{\beta_1}$ kao rješenja sljedećih jednadžbi:

$$-2 \sum_{i=1}^{n} [y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i] = 0$$

$$-2 \sum_{i=1}^{n} x_i [y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i] = 0.$$

Rješavanjem prethodnog sustava jednadžbi slijedi:

$$\begin{aligned}
\widehat{\beta_1} &= \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}, \\
\widehat{\beta_0} &= \bar{y} - \widehat{\beta_1} \bar{x}.
\end{aligned}$$

Sumu kvadrata grešaka između zabilježenih vrijednosti i točaka danog pravca minimizira pravac $y = \widehat{\beta_0} + \widehat{\beta_1} x$. Neka je

$$\text{SSE} = \sum_{i=1}^{n} \widehat{e_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2.$$

Vrijednosti $\widehat{e_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$ poznate su kao reziduali, dok je SSE (*eng.* error sum of squares) suma kvadrata reziduala. Least square metoda nam ne daje procjenitelja za σ^2, ali možemo pokazati kako je nepristran procjenitelj za σ^2 dan s

$$\widehat{\sigma}^2 = \frac{SSE}{n-2}.$$

$\widehat{\sigma}^2$ koristit ćemo za nepristranog procjenitelja od σ^2. Sumu kvadrata reziduala možemo zapisati i u sljedećem obliku:

$$\text{SSE} = \sum y_i^2 - \widehat{\beta_0} \sum y_i - \widehat{\beta_1} \sum x_i y_i.$$

$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$ ćemo koristiti za predikciju vrijednosti Y, a istom vrijednosti procjenjivat ćemo očekivanu vrijednost od Y, to jest procjenitelj za $\text{E}(Y) = \beta_0 + \beta_1 x$ koji izgleda ovako

$$\widehat{E}(Y) = \widehat{\beta_0} + \widehat{\beta_1} x.$$

Procjenitelj metodom najmanjih kvadrata linearna je funkcija slučajnih varijabli $Y_1, Y_2, \ldots, Y_n$. Dokazano je da je među ostalim linearnim nepristranim procjeniteljima, on procjenitelj s najmanjom varijancom. Iz tog razloga se često naziva linearnim nepristranim procjeniteljem s najmanjom varijancom (*eng.* best linear unbiased estimator). Jednadžbe i izvodi preuzeti su iz [12].

Teorem 2.1. *Ako je* $E(Y_i) = \beta_0 + \beta_1 x_i$, $Var(Y_i) = \sigma^2$ *i* $Cov(Y_i, Y_j) = 0$ *za* $i \neq j$ *te* $i = 1, \ldots, n$, *tada procjenitelj najmanjih kvadrata ima sljedeća svojstva:*

i) $E(\widehat{\beta_1}) = \beta_1$, $Var(\widehat{\beta_1}) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

ii) $E(\widehat{\beta_0}) = \beta_0$, $Var(\widehat{\beta_0}) = \dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

iii) $E(c_1\widehat{\beta_0} + c_2\widehat{\beta_1}) = c_1\beta_0 + c_2\beta_1$

iv) $c_1\widehat{\beta_0} + c_2\widehat{\beta_1}$ *je najbolji linearni nepristran procjenitelj za* $c_1\beta_0 + c_2\beta_1$.

Dokaz se nalazi u [2].

2.2 Višestruka linearna regresija

Pogledajmo model linearne regresije oblika (2.2). Nadalje pretpostavimo da je realizacija y_i ovisne slučajne varijable Y zabilježena za $i = 1, \ldots, n$ za koje je $n \geq p + 1$. Također pretpostavimo da vrijedi sljedeće:

$$\mathrm{E}(Y_i) = \sum_{j=0}^{p} \beta_j x_{ij}, \qquad Var(Y_i) = \sigma^2, \qquad Cov(Y_i, Y_j) = 0, \qquad \text{gdje je } i \neq j.$$

Neka je V matrica kojoj je element u i-tom retku i j-tom stupcu kovarijanca slučajnih varijabli Y_i i Y_j, odnosno $V = [Cov(Y_i, Y_j)]_{i,j=1,\ldots,n.}$ Matricu V nazivamo matricom kovarijanci od $Y_1, Y_2, \ldots, Y_n$. Model višestruke linearne regresije može se zapisati u matričnom obliku: $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $V = \sigma^2 I$, pri čemu je I jedinična matrica, a $\mathbf{Y}, \boldsymbol{\beta}$ i $\mathbf{X}$:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} x_{10} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n0} & \ldots & x_{np} \end{bmatrix}.$$

Procjenitelji za β_k dobiju se minimiziranjem sljedeće funkcije:

$$S = \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \Big)^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Koristit ćemo sličan pristup kao i kod jednostavne linearne regresije, to jest njega ćemo generalizirati. Dakle, parcijalno ćemo derivirati fukciju S te njene parcijalne derivacije po svim β_k izjednačiti s 0:

$$\frac{\partial S}{\partial \beta_k} = \sum_{i=1}^{n} 2\Big[y_i - \sum_{j=0}^{p} \beta_j x_{ij}\Big](-x_{ik}) = 0, \text{ za } k = 0, 1, \ldots, p.$$

Prethodno navedeni sustav linearan je u parametrima i možemo ga zapisati u obliku sljedeće matrične jednadžbe:

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

Za nesingularnu matricu $\mathbf{X}^T\mathbf{X}$ postoji jedinstveno rješenje matrične jednadžbe oblika:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

U nastavku ćemo pretpostaviti da je $\mathbf{X}^T\mathbf{X}$ nesingularna, ukoliko nije drugačije naglašeno.

Procjenitelji $\widehat{\beta}_j$ linearne su funkcije od $Y_1, Y_2, \ldots, Y_n$ te je lako pokazati da su oni nepristrani procjenitelji za β_j:

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Procjenitelji metodom najmanjih kvadrata za β_j su kao i kod modela jednostavne linearne regresije nepristrani procjenitelji s najmanjom varijancom.

Varijance i kovarijance nepristranog linearnog procjenitelja za β_j s najmanjom varijancom elementi su sljedeće matrice $\boldsymbol{C} = [Cov(\widehat{\beta}_i, \widehat{\beta}_j)]_{i,j=1,\ldots,n} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Također možemo pokazati da je linearan nepristran procjenitelj s najmanjom varijancom bilo koja linearna kombinacija od β_j, npr. $\boldsymbol{r}^T\widehat{\boldsymbol{\beta}} = \sum_{j=0}^{p} r_j\widehat{\beta}_j$ je linearan nepristran procjenitelj s najmanjom varijancom za $\boldsymbol{r}^T\boldsymbol{\beta} = \sum_{j=0}^{p} r_j\beta_j$. Izvodi i pojmovi preuzeti su iz [12].

3 Dijagnostika modela

U ovom ćemo poglavlju navesti i pojasniti korake koje smo kasnije koristili pri izradi modela.

3.1 Homoskedastičnost

Homoskedastičnost je jedna od pretpostavki modela linearne regresije. Homoskedastičnost se odnosi na situaciju u kojoj je varijanca grešaka modela konstanta. Suprotnost homoskedastičnosti je heteroskedastičnost, odnosno to je izraz za stanje kada varijanca grešaka modela nije konstanta. Uz pretpostavku da su ispunjene polazne pretpostavke o modelu, matrica varijanci i kovarijanci vektora procijenjenih metodom najmanjih kvadrata dana je s:

$$Var(\widehat{\boldsymbol{\beta}}) = S^2(\boldsymbol{X^T X})^{-1}.$$

Međutim, ukoliko je varijanca promjenjiva, odnosno $Var(\epsilon_i) = \sigma_i, i = 1, \ldots, k$, tada je matrica varijanci i kovarijanci $D = Var(\epsilon) = diag(\sigma_i^2), i = 1, \ldots, k$. Kako je $Var(Y) = Var(\epsilon)$ matrica kovarijanci i varijanci procjenitelja parametara:

$$\begin{aligned} Var(\widehat{\beta}) &= Var((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T Var(Y) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T D X (X^T X)^{-1}. \end{aligned}$$

Iz prethodne jednadžbe vidimo da je formula za $Var(\widehat{\beta})$ drugačija, pa je i pogrešno računanje varijance. Bitno je da su podaci homoskedastični zbog toga što bez homoskedastičnosti podataka ne možemo primjenjivati F- test i ANOVA. Za provjeru homoskedastičnosti u radu koristit ćemo ncvTest (*eng.* Non - constant variance test) iz software-a R. Više detalja može se vidjeti u [6].

3.2 Multikolinearnost

Problem multikolinearnosti je prisutan ako između dvije ili više neovisnih regresorskih varijabli u modelu postoji jaka korelacija, tj. vrijednost ovisne varijable ima smisla predviđati koristeći se linearnom vezom s nekima ili svim preostalim neovisnim varijablama. Nama je cilj kasnije pokazati da ne postoji problem s multikolinearnosti, tj. da nam varijable nisu multikolinearne. Za tu svrhu u radu koristili smo VIF (*eng.* Variance Inflation Factor).

Variance Infation Factor ili faktor inflacije varijance je često korišten alat za otkrivanje multikolinearnosti u regresijskim modelima. Računa se na sljedeći način: $VIF_i = \frac{1}{1-R_i^2}$, gdje R_i^2 predstavlja nekorigirani koeficijent determinacije i - tog regresora na ostale varijable. Brojčana vrijednost za VIF govori nam za koji je postotak varijanca uvećana za svaki koeficijent. Ukoliko je faktor inflacije varijance manji od 5 smatramo da nemamo problema s multikolinearnosti. Više detalja može se vidjeti u [1].

3.3 AIC

Akaike informacijski kriterij ili *eng. Akaike Information Criterion* je broj koji mjeri koliko dobro model odgovara danom skupu podataka i pomaže pri određivanju najboljeg odgovarajućeg modela koji koristi najmanje varijabli. Njegov iznos dan je izrazom

$$AIC = -2 \ln L + 2p,$$

pri čemu L označava maksimum funkcije vjerodostojnosti promatranog modela, dok je p broj procijenjenih parametara modela. U ovom radu AIC smo koristili u statističkom software-u R (vidi [7]). Modele s manjim AIC-om smatramo boljima od modela s većim AIC-om.

3.4 Koeficijent determinacije R^2

U statistici koeficijent determinacije R^2 je mjera koja procjenjuje sposobnost modela da predvidi ishod u postavci linearne regresije. R^2 opisuje jačinu linearne veze između zavisne i nezavisnih

varijabli. Koeficijent determinacije R^2 zapisujemo u sljedećem obliku:

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

pri čemu je

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \widehat{y}_i = \sum_{j=0}^{p} \widehat{\beta}_j x_{ij}, \quad \widehat{\beta}_j \text{ je } j \text{ - ti element od } \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Koeficijent determinacije poprima vrijednosti u intervalu $0 \leq R^2 \leq 1$. Jedan od nedostataka koeficijenta determinacije je da se njegova vrijednost povećava dodavanjem nezavisnih varijabli u modelu, bez obzira na to jesu li one značajne za poboljšanje modela ili nisu. Zbog toga koristimo korigirani koeficijent determinacije $\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)}\frac{SSE}{SST} = 1 - \frac{n-1}{n-(k+1)}(1 - R^2)$. U formuli n predstavlja veličinu uzorka, a k broj korištenih nezavisnih varijabli pri izradi modela. $\bar{R}^2$ je manji ili jednak koeficijentu determinacije R^2, jer on 'kažnjava' uključivanje novih nezavisnih varijabli u model. Dakle, što su veće vrijednosti k, to će korigirani koeficijent determinacije $\bar{R}^2$ biti manji od koeficijenta determinacije R^2. Korigirani koeficijent determinacije koristi se kao jedan od mogućih kriterija za izbor modela višestruke linearne regresije. On nam govori koliko je ukupne varijabilnosti u podacima objašnjeno modelom. Prema tom kriteriju najbolji je model s najvećim $\bar{R}^2$. Više detalja može se vidjeti u [12].

4 Analiza varijance

4.1 F-test

Statistički test nazvan F-test ispituje imaju li dva skupa normalne distribucije istu varijancu. Vrlo je važan za ANOVA metodu zbog toga što uzima u omjere dvije varijance dva različita skupa podataka. Test statistika za F-test izgleda ovako:

$$F = \frac{\sigma_1^2}{\sigma_2^2},$$

gdje je σ_1^2 varijanca prvog skupa podataka, a σ_2^2 varijanca drugog skupa. Hipoteze:

$$H_0 : \sigma_1^2 = \sigma_2^2, \qquad H_1 : \sigma_1^2 \neq \sigma_2^2,$$

odnosno nul-hipoteza tvrdi da su varijance oba uzorka jednake, dok alternativna tvrdi suprotno. Formula je preuzeta iz [8].

4.2 Anova

ANOVA ili *eng. ANalysis Of VAriance* je metoda koja govori o postojanju statistički značajnih razlika između očekivanja varijabli. Zaključak se donosi na temelju F-testa. Ukoliko dodajemo

nezavisne varijable u model pomoću ANOVA metode možemo zaključiti doprinose li dodane varijable poboljšanju modela. Hipoteze u ovoj metodi izgledaju ovako:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

$$H_1 : \mu_i \neq \mu_j, \quad \text{za neki par (i, j)},$$

pri čemu je k - broj nezavisnih usporednih skupina. Hipoteze možemo interpretirati na sljedeći način:

H_0 : očekivanja svih skupina su jednake

H_1 : očekivanja se razlikuju u barem jednoj skupini

Test statistika za testiranje H_0 je:

$$F = \frac{\frac{\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2}{k-1}}{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2}{N-k}}.$$

U prethodnoj test statistici n_j predstavlja veličinu uzorka u j - toj skupini, $\bar{X}_j$ aritmetičku sredinu uzorka u j - toj skupini, $\bar{X}$ predstavlja aritmetičku sredinu svih podataka, X_{ij} predstavlja i-ti uzorak j - te skupine. i N predstavlja veličinu cijelog uzorka (oznake ćemo koristiti i u tablici 1). Izračune ANOVA metodom često prikazujemo pomoću ANOVA tablice (vidi tablicu 1). Iz

Izvor varijabilnosti	**Suma kvadrata**	**df**	**Srednje kvadratno odstupanje**
Između skupina	$SSB = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$	$k-1$	$MSB = \frac{SSB}{k-1}$
Greške (Reziduali)	$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$	$N-k$	$MSE = \frac{SSE}{N-k}$
Ukupno	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2$	$N-1$	

Tablica 1: ANOVA tablica

tablice 1 možemo odrediti test statistiku F koja ima sljedeći oblik $F = \frac{MSB}{MSE}$. Za opširnije vidi [8].

4.2.1 Primjer primjene ANOVA metode

U istraživanju je sudjelovalo 15 studenata matematike. Glavni cilj istraživanja jest usporediti gubitak tjelesne težine s obzirom na program treniranja. Studenti su slijedili program dva mjeseca. Rezultat će biti izražen u gubitku, koji ćemo definirati kao razliku između tjelesne težine na kraju istraživanja i tjelesne težine na početku istraživanja. Podijelili smo studente u tri skupine: prva skupina radila je kardio treninge, druga treninge snage i izdržljivosti, a trećoj skupini je rečeno da treniraju kako i koliko žele, tu skupinu ćemo označavati sa 'Opcionalna'. Nakon 2 mjeseca izmjerene su razlike u tjelesnoj težini studenata koje možemo vidjeti u tablici 2. Pozitivne razlike ukazuju na gubitak težine, a negativne na dobitak kilograma. Dobivene razlike zaokružene su na najbliže cijele brojeve radi jednostavnosti izračuna.

Kardio	**Snaga**	**Opcionalna**
6	3	2
5	4	1
8	4	0
7	2	1
8	3	3

Tablica 2: Rezultati nakon 2 mjeseca

Zanima nas postoji li statistički značajna razlika u očekivanom gubitku kilograma za različite programe treniranja. Tu tvrdnju ćemo provjeriti pomoću ANOVA metode. Za početak ćemo odrediti hipoteze i razinu značajnosti α.
Hipoteze:

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_i \neq \mu_j$ za neki par (i, j); gdje su $i, j = 1, 2, 3$; $\alpha = 0.05$.

Sljedeći korak je postaviti test statistiku za ANOVU, $F = \frac{MSB}{MSE}$. Nakon toga moramo odrediti vrijednost test statistike F. Kako bi odredili kritično područje moramo izračunati stupnjeve slobode kako je napisano u tablici 1, $df_1 = k - 1$ te $df_2 = N - k$. U našem primjeru, $df_1 = 3 - 1 = 2$, a $df_2 = 15 - 3 = 12$. Kritična vrijednost iznosi $F = 3.2874$ (vidi [11]). Dakle, odbacit ćemo nul-hipotezu ukoliko je $F \geq 3.2874$. Četvrti korak jest odrediti aritmetičke sredine svake skupine te cijelog uzorka (vidi tablicu 3), te nakon toga izračunati srednje kvadratno odstupanje za svaku skupinu (vidi tablicu 4, tablicu 5, tablicu 6).

	Kardio	**Snaga**	**Opcionalna**	**Ukupno**
n	5	5	5	15
Aritmetička sredina	6.8	3.2	1.4	3.8

Tablica 3: Aritmetičke sredine skupina

Sada možemo izračunati $SSB = \sum_{j=1}^{3} n_j(\bar{X}_j - \bar{X})^2$. Dakle, u našem slučaju:

$$SSB = 5(6.8 - 3.8)^2 + 5(3.2 - 3.8)^2 + 5(1.4 - 3.8)^2 = 75.6.$$

Sljedeće što ćemo izračunati je $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$ za svaku skupinu posebno. Za skupinu koja je prakticirala kardio treninge:

Kardio	$(X_{i1} - 6.8)^2$
6	0.64
5	3.24
8	1.44
7	0.04
8	1.44
Ukupno	$\sum_{i=1}^{5}(X_{i1} - \bar{X}_1)^2 = 6.8$

Tablica 4: Suma kvadrata za kardio skupinu

Za skupinu koja je prakticirala treninge snage i izdržljivosti:

Snaga	$(X_{i2} - 3.2)^2$
3	0.04
4	0.64
4	0.64
2	1.44
3	0.04
Ukupno	$\sum_{i=1}^{5}(X_{i2} - \bar{X}_2)^2 = 2.8$

Tablica 5: Suma kvadrata za skupinu treninga snage i izdržljivosti

Za kraj je ostalo izračunati sumu kvadrata za 'Opcionalnu' skupinu:

Opcionalno	$(X_{i3} - 1.4)^2$
2	0.36
1	0.16
0	1.96
1	0.16
3	2.56
Ukupno	$\sum_{i=1}^{5}(X_{i3} - \bar{X}_3)^2 = 5.2$

Tablica 6: Suma kvadrata za opcionalnu skupinu

Sada možemo izračunati $SSE = \sum_{j=1}^{3}\sum_{i=1}^{5}(X_{ij} - \bar{X}_j)^2 = 6.8 + 2.8 + 5.2 = 14.8$.
Konstruirajmo sada ANOVA tablicu.

Izvor varijabilnosti	Suma kvadrata	df	Srednje kvadratno odstupanje
Između skupina	$SSB = 75.6$	3 - 1 = 2	$MSB = \frac{75.6}{3-1} = 25.2$
Greške (Reziduali)	$SSE = 14.8$	15 - 3 = 12	$MSE = \frac{14.8}{12} = 1.19$
Ukupno	$SST = 90.4$	15 - 1 = 14	

Tablica 7: ANOVA tablica za Primjer 1.3.3.

Dakle, sada možemo izračunati $F = \frac{MSB}{MSE} = \frac{25.2}{1.19} = 21.18$. Odbacujemo H_0 zato što je $21.18 \geq 3.2874$. Na razini značajnosti $\alpha = 0.05$ možemo tvrditi da postoji statistički značajna razlika u gubitku tjelesne težine između navedene tri skupine.

5 Analiza zavisnosti

5.1 Pearsonov korelacijski test

Pearsonov koeficijent korelacije prvi je koristio Francis Galton, britanski profesor i pokretač istraživanja o statističkoj korelaciji i regresiji, a ime je dobio po britanskom matematičaru i statističaru Karlu Pearsonu. Pearsonov koeficijent korelacije jedna je od najčešće korištenih mjera linearne povezanosti između dvije numeričke varijable. Koeficijent može zaprimiti vrijednost između -1 i 1. Izračunava se pomoću kovarijance i standardnih odstupanja obje varijable prema sljedećoj formuli:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y},$$

gdje se kovarijanca računa prema sljedećoj formuli:

$$Cov(X, Y) = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \cdot \sigma_Y},$$

pri čemu su σ_X i σ_Y standardne devijacije od X i Y te $\mu_X = \mathrm{E}[X]$ i $\mu_Y = \mathrm{E}[Y]$ očekivanja slučajnih varijabli X i Y.

Neka je dan uzorak (x_i, y_i), $i = 1, \ldots, n$. Procjena za ρ je uzoračka korelacijska koeficijent

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\widehat{\sigma_x}^2\widehat{\sigma_y}^2},$$

gdje su

$$\widehat{\sigma_x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}), \quad \widehat{\sigma_y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}), \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

U statističkom zaključivanju o korelaciji postavljamo nultu i alternativnu hipotezu

$$H_0 : \rho = 0,$$

$$H_1 : \rho \neq 0,$$

pri tome koristimo test statistiku

$$t = \frac{\sqrt{n-2} \cdot r}{\sqrt{1-r^2}},$$

koja ima Studentovu distribuciju s $n-2$ stupnja slobode. Pozitivna korelacija postoji kada su vrijednosti prve (X) i druge (Y) varijable obje visoke ili niske. Tada je koeficijent pozitivan i blizu 1 (npr. 0.91). Negativna korelacija postoji kada su vrijednosti prve varijable visoke, a vrijednost druge niske ili obrnuto. Tada je koeficijent negativan i blizu -1 (npr -0.81). Ukoliko je $\rho = 0$ kažemo da varijable nisu korelirane. Više detalja može se vidjeti u [8].

5.2 Kendallov koeficijent korelacije ranga

Kendallov koeficijent korelacije ranga ili jednostavnije Kendallov τ bazira se na principu usklađenih parova podataka. Kažemo da je par podataka (x_i, y_i) i (x_j, y_j) usklađen ukoliko vrijedi jedan od uvjeta $x_i < x_j$ i $y_i < y_j$ ili $x_i > x_j$ i $y_i > y_j$, u suprotnom kažemo da je par podataka neusklađen. Procjena za τ izgleda ovako:

$$\tau = \frac{n_U - n_N}{\frac{1}{2}n(n-1)},$$

pri čemu je n_U broj usklađenih parova, a n_N broj neusklađenih parova, te n veličina uzorka. Uočima kako je $\tau \in [-1, 1]$. Dakle, ukoliko su svi parovi usklađeni tada je $\tau = 1$. Ako su svi parovi neusklađeni tada je $\tau = -1$.

Hipoteze za testiranje o nepostojanju monotone veze između dviju varijabli:

$$H_0 : \tau = 0, \quad \text{ne postoji monotona veza}$$

$$H_1 : \tau \neq 0 \quad \text{(jednostrana)}, \quad H_1 : \tau > 0 \qquad H_1 : \tau < 0 \quad \text{(dvostrana)}$$

Kendallov koeficijent korelacije ranga $\tau \in [-1, 1]$ daje informaciju o tome u kojoj se mjeri veza između dvije slučajne varijable može opisati monotonom funkcijom tako da vrijedi:

- $\tau \approx 0$, sugerira da ne postoji monotona veza između varijabli,
- $\tau \approx 1$ ili $\tau \approx -1$, sugerira da postoji monotona veza između varijabli,
- $\tau < 0 \quad (\tau > 0)$, veza između varijabli je monotono padajuća (rastuća).

U uvjetima kada je H_0 istinita, test statistika

$$\widehat{z} = \frac{\tau}{\sqrt{\frac{9n(n-1)}{2(2n+5)}}},$$

se može aproksimirati standardnom normalnom distribucijom.

6 Empirijsko istraživanje: Modeliranje vrijednosti igrača u nogometu

6.1 Kratki uvod

U današnje vrijeme gdje je nogomet više biznis nego sport, zanimljiv je način na koji se određuju tržišne cijene nogometaša. U ovom poglavlju bit će napravljeno istraživanje na temelju baze preuzete sa interneta. Baza se sastoji od 2644 podatka i 15 varijabli te je preuzeta sa stranice *Kaggle.com* (vidi [10]).

6.2 Varijable i podaci

U ovom poglavlju bit će stavljen naglasak na bazu te podatke koje sadrži. Ovisna varijabla Cijena predstavlja tržišnu cijenu nogometaša. Osim Cijene imat ćemo 8 neovisnih varijabli koje opisuju svakog nogometaša. U nastavku ćemo nešto više reći o svakoj varijabli:

- Područje - kategorijalna varijabla koja opisuje geografsko područje s kojeg igrač potječe,
- Pozicija - kategorijalna varijabla koja opisuje koju poziciju igrač pokriva u momčadi,
- Broj_godina - numerička varijabla koja opisuje koliko igrač ima godina,
- Visina - numerička varijabla koja opisuje koliko je igrač visok,
- Noga - kategorijalna varijabla koja opisuje koja je 'jača' noga igrača, to jest kojom nogom se igrač bolje koristi,
- Liga - kategorijalna varijabla koja opisuje u kojoj nogometnoj ligi igrač nastupa,
- povjerenje - numerička varijabla koja opisuje omjer između dvije originalne varijable Utakmice i Započete_utakmice (Utakmice - broj utakmica igrača, Započete_utakmice - broj utakmica u kojima je igrač bio u početnih 11 svoje momčadi), tj. koliko je puta igrač zaslužio povjerenje svoga trenera da započne utakmicu,
- Minuta - numerička varijabla koja opisuje broj odigranih minuta na utakmicama,
- omjerga - numerička varijabla koja opisuje omjer između zbroja 2 originalne, varijable Golovi (postignuti golovi) i Asistencije (pružene asistencije) te varijable Utakmica.

6.3 Statističko proučavanje varijabli i odnos cijene prema neovisnim varijablama

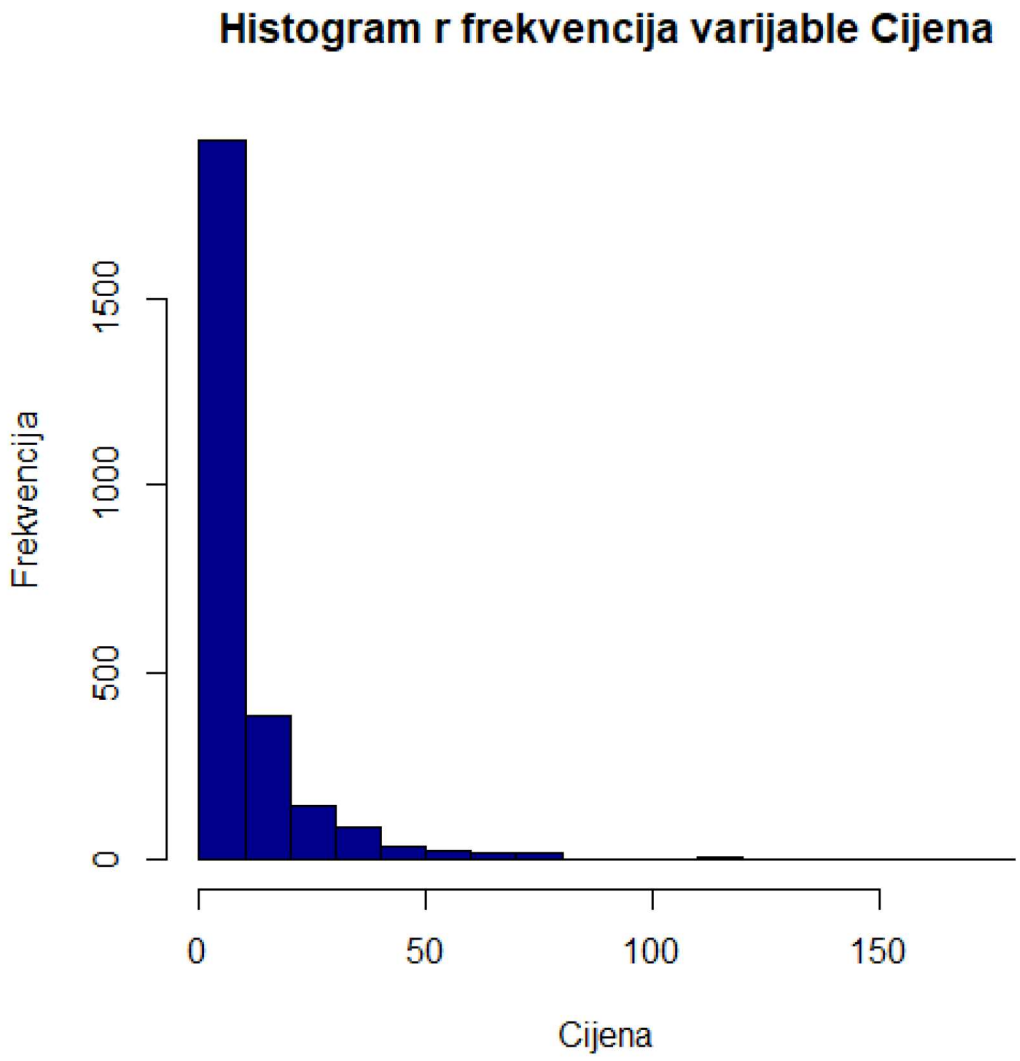
6.3.1 Cijena

Varijabla Cijena je numerička varijabla koja sadrži informaciju o tržišnoj cijeni igrača u eurima. Zbog lakšeg praćenja rezultata podijelili smo izvornu tržišnu cijenu igrača sa 10000. Deskriptivnu statistiku varijable Cijena možemo vidjeti u tablici 8.

Minimum	Donji kvantil	Medijan	Prosjek	Gornji kvantil	Maksimum	Sd
0.0005	1.00	4.00	9.57	12.00	180.00	14.90

Tablica 8: Deskriptivna statistika varijable Cijena

Iz prethodne tablice vidimo kako prosječna tržišna cijena nogometaša iznosi 9.57 milijuna eura. Najskuplji nogometaš u korištenoj bazi, ali i na svijetu u trenutku skidanja baze je Kylian Mbappe kojemu je cijena 2020. godine iznosila 180 milijuna eura. Na slici 1 možemo vidjeti histogram frekvencija varijable Cijena.



Slika 1: Histogram relativnih frekvencija varijable Cijena

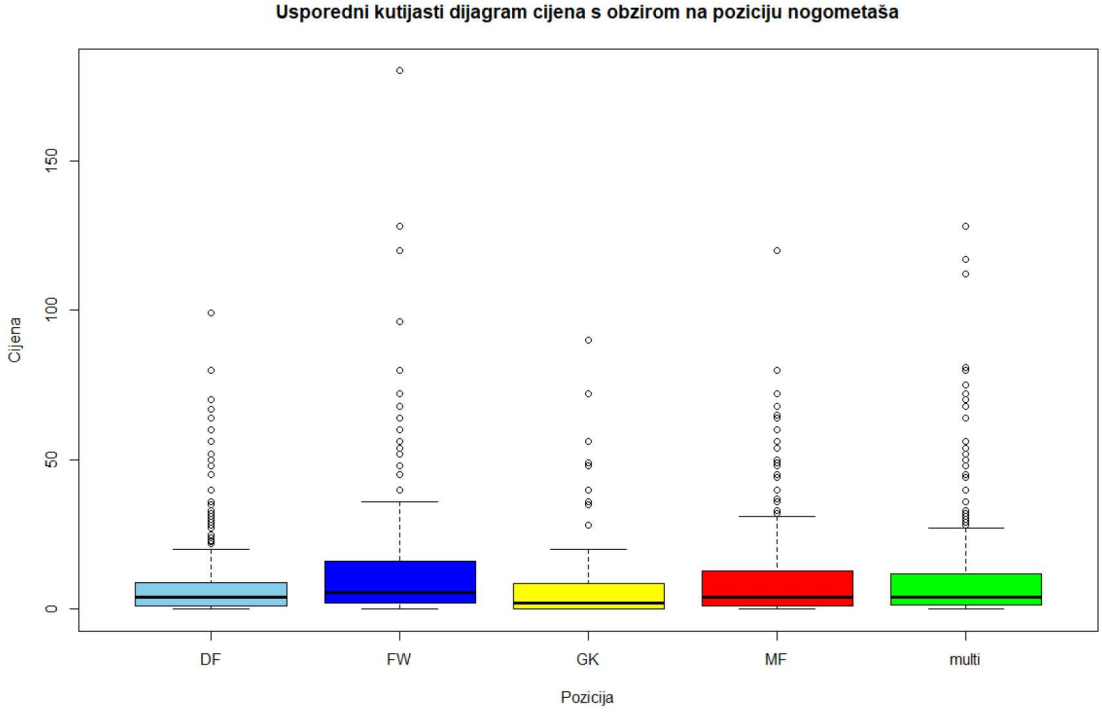
6.3.2 Pozicija

Varijabla Pozicija je kategorijalna varijabla koja opisuje poziciju koju u nogometnoj momčadi igrač zauzima. Sastoji se od 5 kategorija: GK (vratar), DF (obrambeni), MF (vezni), FW (napadač) i multi (igrači koji pokrivaju više pozicija). U tablici 9 možemo vidjeti frekvencije i relativne frekvencije varijable Pozicija.

Pozicija	**Frekvencija**	**Relativna frekvencija**
GK	195	0.0738
DF	852	0.3222
MF	570	0.2156
FW	368	0.1392
multi	659	0.2492

Tablica 9: Frekvencije i relativne frekvencije varijable Pozicija

Iz tablice 9 uočavamo da je u bazi sadržano najviše obrambenih i veznih igrača, što objašnjava činjenica da su najčešće postave u nogometu 1-4-4-2, 1-5-3-2 te 1-3-5-2. Najmanje ima vratara što je očekivano jer ih u igri tj. među vratnicama može biti samo jedan u ekipi od 11 igrača. Na slici 2 možemo vidjeti kako se kreću cijene s obzirom na poziciju igrača.



Slika 2: Usporedni kutijasti dijagram cijena s obzirom na varijablu Pozicija

Iz prethodne slike vidimo kako su napadačima cijene u prosjeku nešto više od ostalih pozicija, te vidimo kako su vratari na tržištu nešto podcjenjeniji. Provedbom ANOVA procedure možemo

zaključiti da postoji statistički značajna razlika u očekivanim cijenama s obzirom na poziciju igrača (p-vrijednost = 1.152e-08).

6.3.3 Broj_godina

Varijabla Broj_godina je numerička varijabla, čija je deskriptivna statistika prikazana u tablici 10. Spomenutu varijablu ćemo transformirati u kategorijalnu te ju podijeliti u tri skupine. Prvu skupinu će činiti mladi igrači od 14 do 23 godine, drugu skupinu zreli igrači, odnosno popularnijim nazivom, igrači u najboljim godinama tj. od 24 do 32 godine, te treću skupinu stariji igrači stariji s 32 ili više godina.

Minimum	**Donji kvantil**	**Medijan**	**Prosjek**	**Gornji kvantil**	**Maksimum**	**Sd**
14	22	25	25.35	28	41	4.44

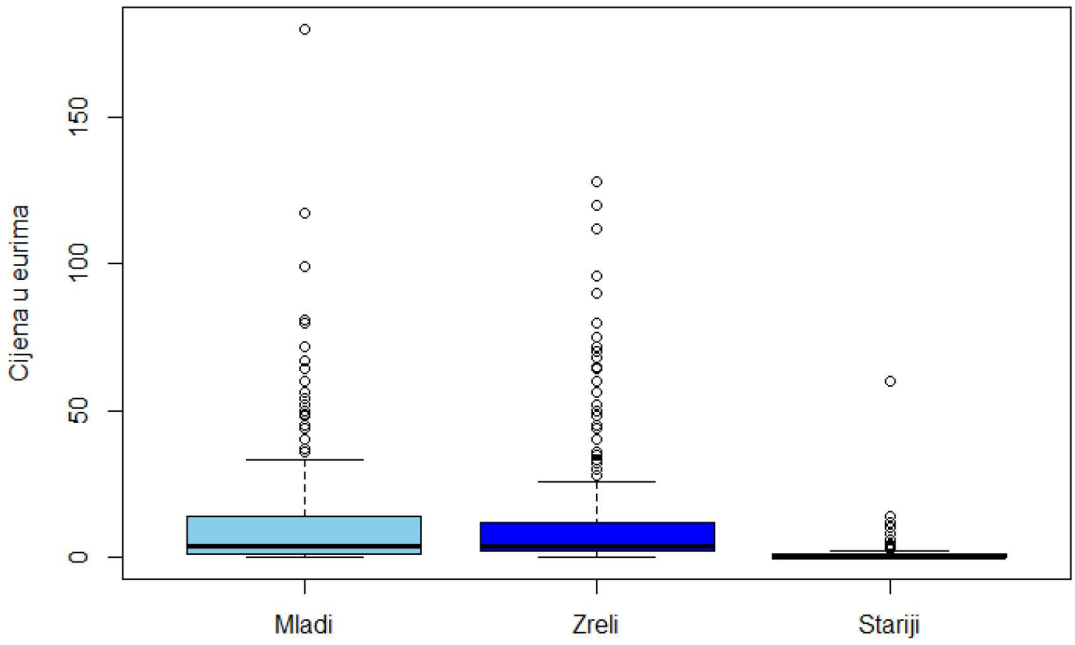
Tablica 10: Deskriptivna statistika varijable Broj_godina

Iz prethodne tablice vidimo kako najmađi igrač ima 14 godina dok najstariji ima 41 godinu. Također vidimo kako više od 50% igrača ima 25 ili više godina.

Broj_godina	**Frekvencija**	**Relativna frekvencija**
mladi	979	0.3703
zreli	1515	0.5730
stariji	150	0.0567

Tablica 11: Frekvencije i relativne frekvencije varijable Broj_godina

U tablici 11 možemo vidjeti frekvencije i relativne frekvencije transformirane kategorijalne varijable Broj_godina. Iz tablice je vidljivo kako najveći postotak u bazi čine igrači u 'najboljim' godinama odnosno 57.30% te da je najmanje igrača u posnim godinama, tj. pri kraju karijere, odnosno samo njih 5.67%. Nakon tablice ćemo na slici 3 vidjeti kako se kreću cijene igrača s obzirom na kategorijalnu podjelu po godinama.



Slika 3: Usporedni kutijasti dijagram cijena s obzirom na transformiranu varijablu Broj_godina

S prethodne slike vidljivo je da su mlađi igrači u prosjeku skuplji te imaju veći gornji kvantil od igrača u najboljim godinama. To možemo opravdati činjenicom da klubovi sve više imaju 'projekt' igrače u koje ulažu puno novca kako bi mladi igrač ostao u klubu što više godina. Kao što bi bilo očekivano, stariji igrači su najjeftiniji na tržištu što smo samo potvrdili ovim usporednim kutijastim dijagramom. Provedbom ANOVA metode možemo tvrditi da postoji statistički značajna razlika u očekivanim cijenama s obzirom na transformiranu varijablu Broj_godina (p-vrijednost = 1.678e-10).

6.3.4 Visina

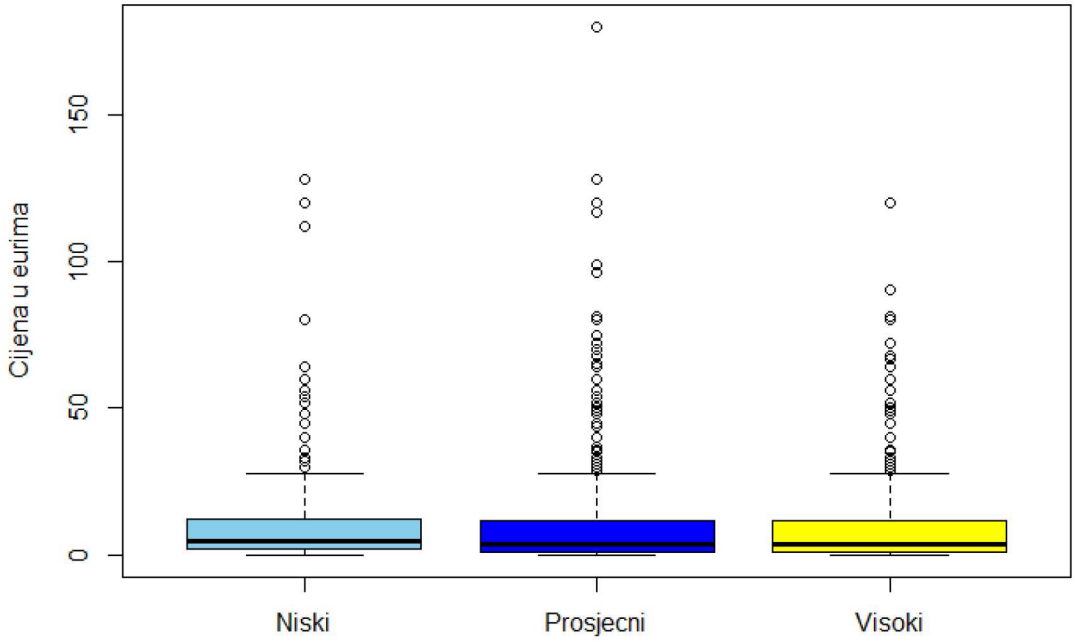
Varijabla Visina, kao što je ranije navedeno, govori o visini igrača u centimetrima. Slično kao i varijablu Broj_godina raspodijelit ćemo varijablu Visina u tri skupine. Prvu skupinu će činiti 'niski' igrači odnosno igrači koji imaju visinu od 162 do 174 centimetra, druga skupina bit će igrači 'prosječne' visine, tj. igrači visine između 175 te 187 centimetra, posljednju skupinu čine 'visoki' igrači odnosno igrači visine iznad 187 centimetara.

U tablici 12 možemo vidjeti frekvencije i relativne frekvencije preoblikovane varijable Visina.

Visina	Frekvencija	Relativna frekvencija
niski	336	0.1270
prosjecni	1581	0.5980
visoki	727	0.2750

Tablica 12: Frekvencije i relativne frekvencije varijable Visina

Iz prethodne tablice vidimo kako je najviše igrača koji su prosječno visoki, njih čak 59.80% što je bilo i za očekivati. Također vidimo kako najmanje ima niskih igrača, tj. samo 12.70%. Sada ćemo na slici 4 vidjeti kako se kreću cijene igrača s obzirom na visinu.



Slika 4: Usporedni kutijasti dijagram cijena s obzirom na transformiranu varijablu Visina

Kao što vidimo sa prethodne slike, ne postoji vidljiva razlika između cijena igrača s obzirom na ove tri kategorije njihovih visina. Provjerit ćemo to i statistički, provedbom ANOVA metode (p-vrijednost = 0.7036) ne možemo tvrditi da postoji statistički značajna razlika u očekivanim cijenama igrača s obzirom na njihovu visinu.

6.3.5 Noga

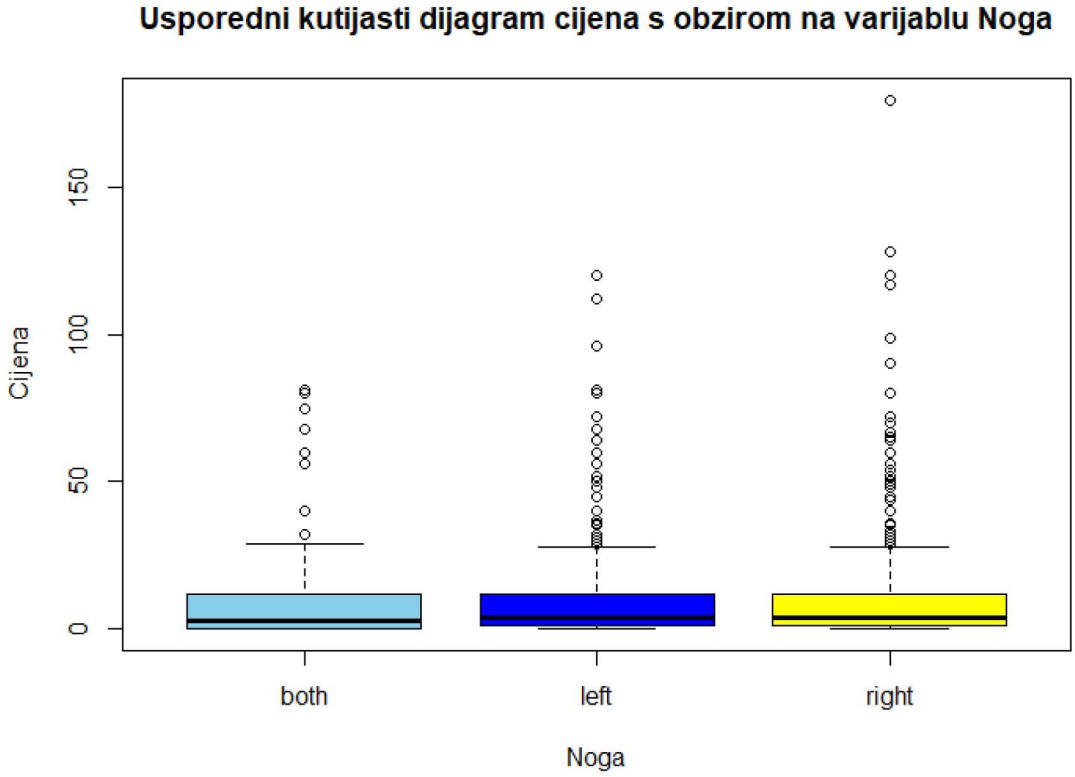
Varijabla Noga, kao što je ranije objašnjeno, govori koju nogu igrač više koristi ili popularnije koja mu je 'jača' noga. Postoje tri kategorije: prva su dešnjaci, druga su ljevaci, a treća su igrači

koji koriste obje noge podjednako dobro. U tablici 13 možemo vidjeti frekvencije i relativne frekvencije varijable Noga.

Noga	**Frekvencija**	**Relativna frekvencija**
dešnjaci	1870	0.7073
ljevaci	656	0.2481
obje	118	0.0446

Tablica 13: Frekvencije i relativne frekvencije varijable Noga

Bilo je za očekivati da će biti najviše dešnjaka jer ljevake čini samo 10% svjetske populacije. U našoj bazi čak 70.73% čine dešnjaci, 24.82% ljevaci, a igrači koji podjednako dobro barataju loptom sa obje noge čine tek 4.46% naše baze. Sada ćemo pogledati sliku 5 da bi vidjeli kako se kreću cijene u odnosu na varijablu Noga.



Slika 5: Usporedni kutijasti dijagram cijena s obzirom na transformiranu varijablu Noga

S prethodne slike teško je zaključiti da postoji statistički značajna razlika u cijeni s obzirom na varijablu Noga. Provedbom ANOVA metode ne možemo tvrditi da postoji statistički značajna razlika u očekivanoj cijeni s obzirom na to s kojom nogom nogometaš preferira igrati (p-vrijednost = 0.6455).

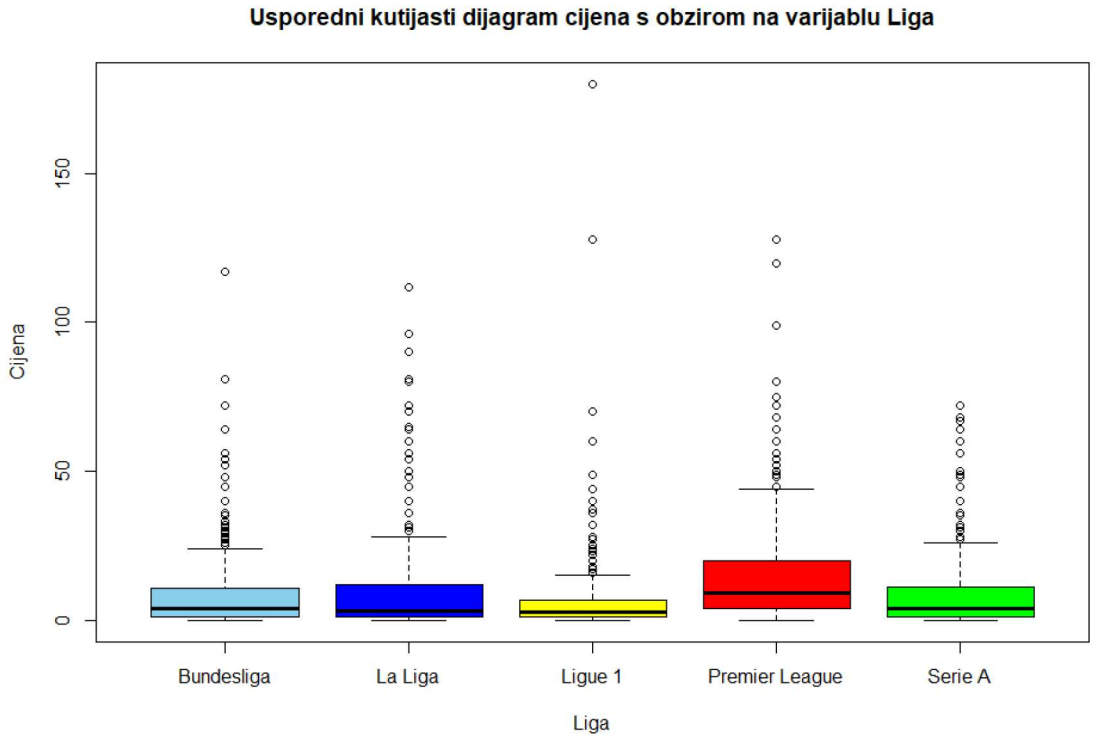
6.3.6 Liga

Varijabla Liga sastoji se od pet najboljih europskih liga, a to su: Premier Liga (Engleska), Bundesliga (Njemačka), La Liga (Španjolska), Seria A (Italija) te Ligue 1 (Francuska). U tablici 14 možemo vidjeti tablicu frekvencija i relativnih frekvencija varijable Liga.

Liga	**Frekvencija**	**Relativna frekvencija**
Premier Liga	511	0.1933
Bundesliga	487	0.1841
La Liga	538	0.2034
Serie A	585	0.2212
Ligue 1	523	0.1978

Tablica 14: Frekvencije i relativne frekvencije varijable Liga

Iz prethodne tablice vidimo kako su igrači gotovo pravilno raspoređeni po ligama, to jest svaka liga sadrži otprilike 1/5 podataka baze. Sada ćemo na slici 6 pogledati kako se kreću cijene nogometaša s obzirom na to u kojoj ligi igraju.



Slika 6: Usporedni kutijasti dijagram cijena s obzirom na varijablu Liga

S prethodne slike možemo utvrditi kako su najviše cijenjeni igrači iz Premier Lige, a iz Ligue 1 najmanje. Ostale tri lige su podjednake što se tiče tržišnih cijena igrača. Sada ćemo statistički provjeriti postoji li značajna razlika u cijenama s obzirom na to u kojoj se ligi igrač natječe. Provedbom ANOVA metode možemo tvrditi kako postoji statistički značajna razlika u očekivanim cijenama igrača s obzirom na to u kojoj ligi nastupaju (p-vrijednost = 2.2e-16).

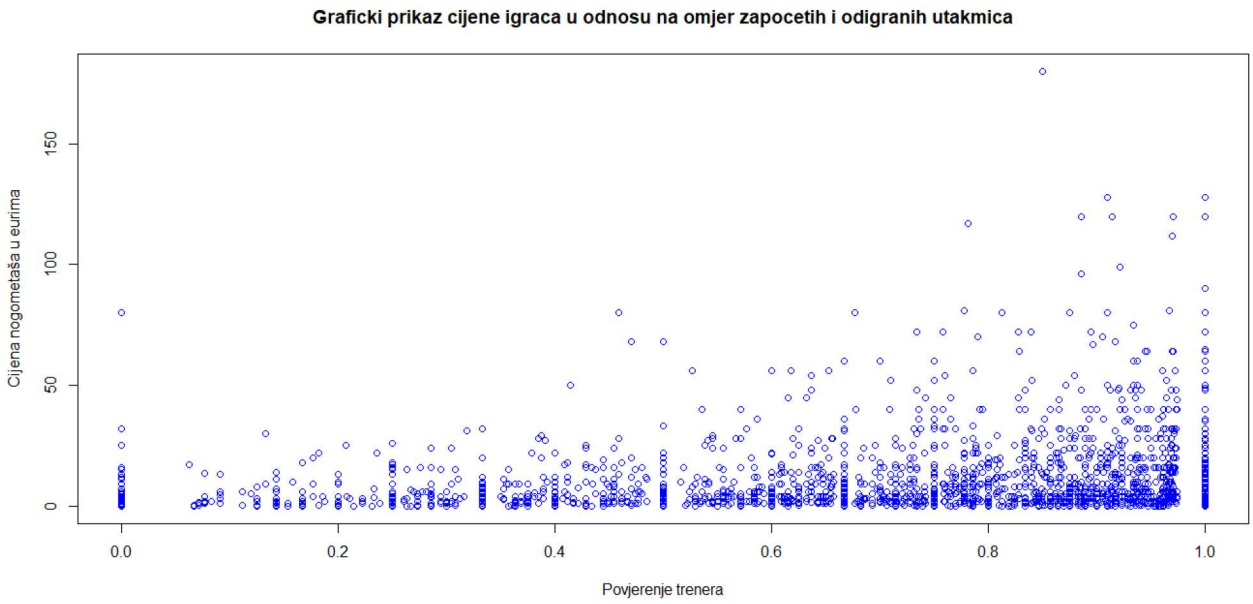
6.3.7 Povjerenje

Varijabla povjerenje govori o omjeru u kojem nogometaš ima povjerenje svog trenera. Ukratko, varijabla povjerenje nastala je od varijabli Utakmice i Započete_utakmice, odnosno

$$povjerenje := \frac{Započete\_utakmice}{Utakmice}.$$

Dakle, ukoliko je omjer 1/1 znači da je igrač započeo svaku od utakmica. Na slici 7 nalazi se grafički prikaz cijene nogometaša u odnosu na varijablu povjerenje.

Iz prethodne slike vidimo kako postoji trend rasta cijene u odnosu na povjerenje trenera. To bi značilo da što više utakmica igrač započne to ima veću cijenu. Sada ćemo to statistički provjeriti pomoću Kendallovog koeficijenta korelacije ranga. Nakon provedbe testa na razini značajnosti 0.05 odbacujemo nul-hipotezu (p-vrijednost = 2.2e-16 $<$ 0.05). Dobivena vrijednost za τ iznosi 0.19 što je veće od 0, te možemo tvrditi da je veza između varijabli monotono rastuća. Dakle, igrači koji imaju veće povjerenje trenera imaju veću tržišnu cijenu.



Slika 7: Grafički prikaz cijene igrača u odnosu na varijablu povjerenje

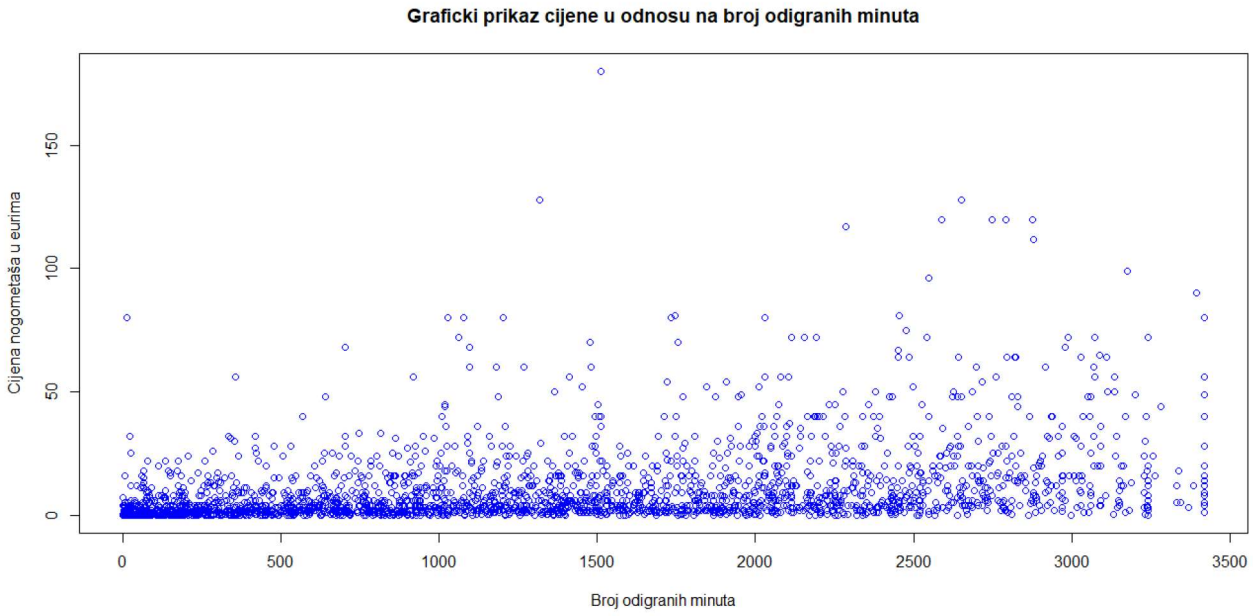
6.3.8 Minuta

Varijabla Minuta govori o tome koliko nogometaš provede vremena na terenu za vrijeme utakmica. U tablici 15 možemo vidjeti deskriptivnu statistiku varijable Minuta.

Minimum	Donji kvantil	Medijan	Prosjek	Gornji kvantil	Maksimum	Sd
1	424.5	1181.5	1285.9	2050.2	3420	947.35

Tablica 15: Deskriptivna statistika varijable Minuta

Iz prethodne tablice možemo vidjeti kako je prosječno vrijeme nogometaša iz baze provedeno na terenu 1285.9 minuta. Dok je zanimljivo vidjeti kako je neki od igrača odigrao samo jednu minutu, a 3420 minuta je najviše vremena što je neki od igrača proveo na terenu. Sada ćemo pogledati sliku 8, odnosno grafički prikaz kretanja cijena igrača u odnosu na broj odigranih minuta.



Slika 8: Grafički prikaz cijene igrača u odnosu na varijablu Minuta

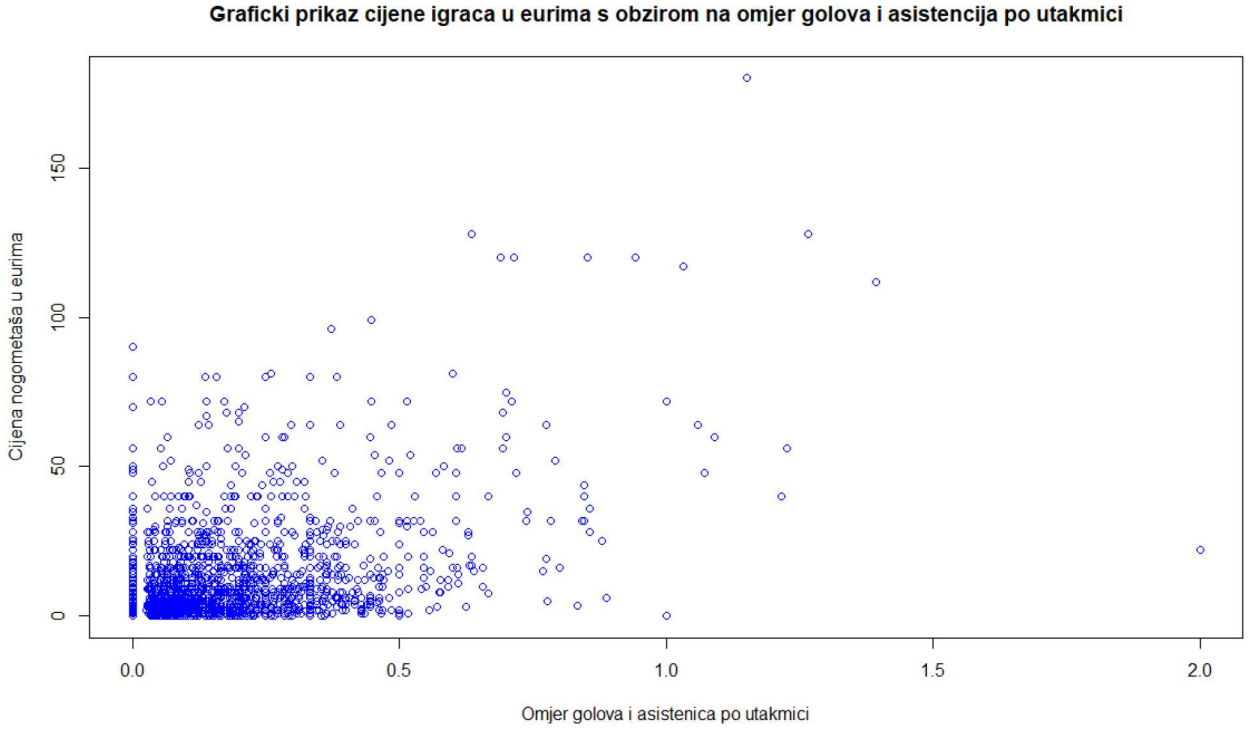
Zanima nas utječe li na tržišnu cijenu nogometaša broj odigranih minuta, što ćemo statistički provjeriti Kendallovim koeficijentom korelacije ranga. Nakon provedbe testa na razini značajnosti 0.05 odbacujemo nul-hipotezu (p-vrijednost = 2.2e-16 $<$ 0.05). Vrijednost za $\tau = 0.351 > 0$, to jest veza između varijabli je monotono rastuća. Dakle, igrači koji imaju veću minutažu će imati veću tržišnu cijenu.

6.3.9 Omjerga

Varijabla omjerga govori o igračevom učinku po utakmici, odnosno koliko zabija i asistira na utakmicama:

$$omjerga := \frac{Golovi + Asistencije}{Utakmice}.$$

Na slici 9 možemo vidjeti kretanje cijena nogometaša s obzirom na njihov učinak po utakmici.



Slika 9: Grafički prikaz cijene igrača u odnosu na varijablu omjerga

S prethodne slike možemo vidjeti trend rasta cijene igrača s obzirom na porast učinka, što ćemo u nastavku statistički provjeriti Kendallovim koeficijentom korelacije ranga. Nakon provedbe testa, na razini značajnosti 0.05 odbacujemo nul-hipotezu (p-vrijednost = 2.2e-16 $<$ 0.05). Dobivena vrijednost za τ iznosi 0.32 što je veće od 0, odnosno možemo tvrditi da je veza između varijabli monotono rastuća. Dakle, igrači koji imaju veći učinak na utakmicama će imati veću tržišnu cijenu.

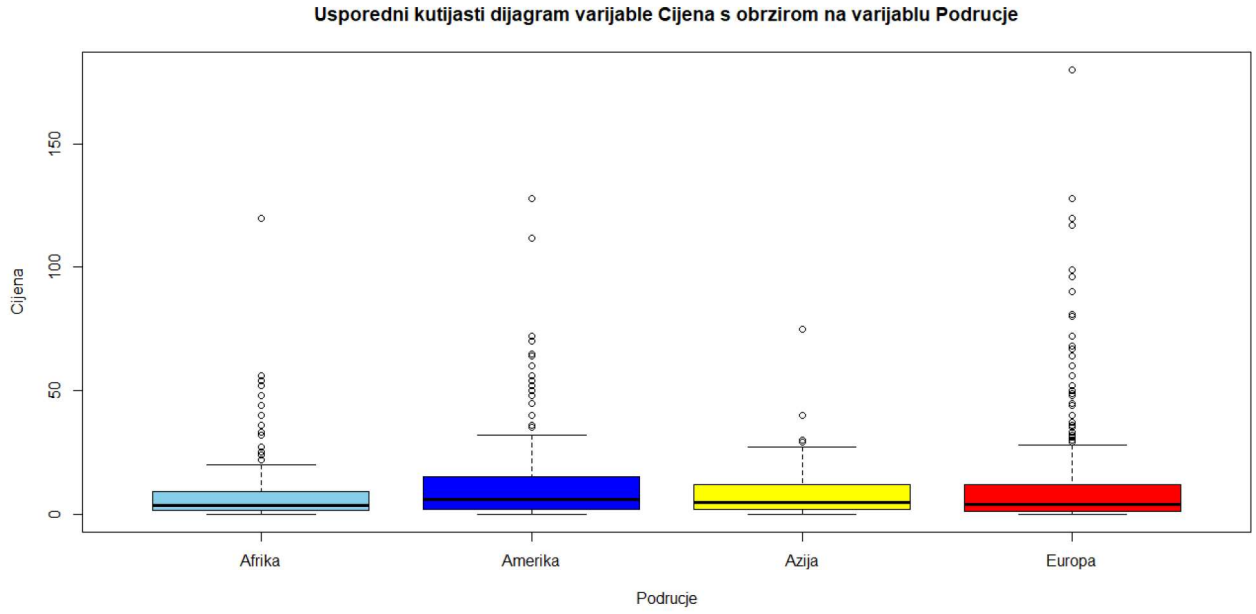
6.3.10 Područje

Varijabla Područje govori o tome s kojeg geografskog dijela Zemlje dolazi igrač. Varijabla je podijeljena u četiri kategorije: Amerika (Sjeverna i Južna), Europa, Azija te Afrika. U tablici 16 možemo vidjeti frekvencije i relativne frekvencije varijable Područje.

Područje	**Frekvencija**	**Relativna frekvencija**
Amerika	309	0.1168
Europa	2004	0.7580
Azija	50	0.0190
Afrika	281	0.1062

Tablica 16: Frekvencije i relativne frekvencije varijable Područje

Iz prethodne tablice vidimo da 75.80% igrača dolazi s područja Europe što je i očekivano jer se baza sastoji od pet najboljih europskih nogometnih liga. Također vidimo kako tek 1.9% nogometaša dolazi iz Azije zbog toga što nogomet još nije dosegao veliku popularnost u Aziji. Na slici 10 možemo pogledati usporedni kutijasti dijagram varijable Cijena u odnosu na varijablu Područje.



Slika 10: Usporedni kutijasti dijagram varijable cijena s obzirom na varijablu Područje

S prethodne slike možemo vidjeti kako su igrači iz područja Sjeverne i Južne Amerike u prosjeku nešto cjenjeniji od ostalih područja. To možemo objasniti povijesno, jer iako je u engleskoj kulturi prvobitno nastao nogomet smatra se da su Južnoamerikanci doveli nogomet na višu razinu. Posebno to vrijedi za Brazilce koji slove za jedne od najvećih 'majstora' nogometne igre. Provjerimo sada postoji li statistički značajna razlika u cijeni nogometaša s obzirom na to iz kojeg područja dolazi. Provedbom ANOVA metode možemo tvrditi kako postoji statistički značajna razlika u očekivanim cijenama s obzirom na to s kojeg područja dolaze (p-vrijednost = 7.575e-05).

7 Modeliranje vrijednosti igrača u nogometu

Metoda koju smo odabrali za modeliranje vrijednosti igrača u nogometu uz dane vrijednosti regresora je višestruka linearna regresija. Naš je zadatak naći najefikasniji takav model. Koristeći statističke procedure koje su već ugrađene u software-u R i svojih logičkih pretpostavki pokušat ćemo pronaći najbolji model za opisivanje varijable Cijena. Kriteriji po kojima ćemo odlučivati pridonosi li varijabla značajno modelu su korigirani $\bar{R}^2$ i AIC (Akaike informacijski kriterij).

7.1 Pretpostavke modela

Kako bi model bio zadovoljavajući mora zadovoljavati sljedeće pretpostavke: homoskedastičnost grešaka modela, normalnost reziduala i moramo provjeriti ima li model problem s multikolinearnosti.

7.2 Selekcija modela

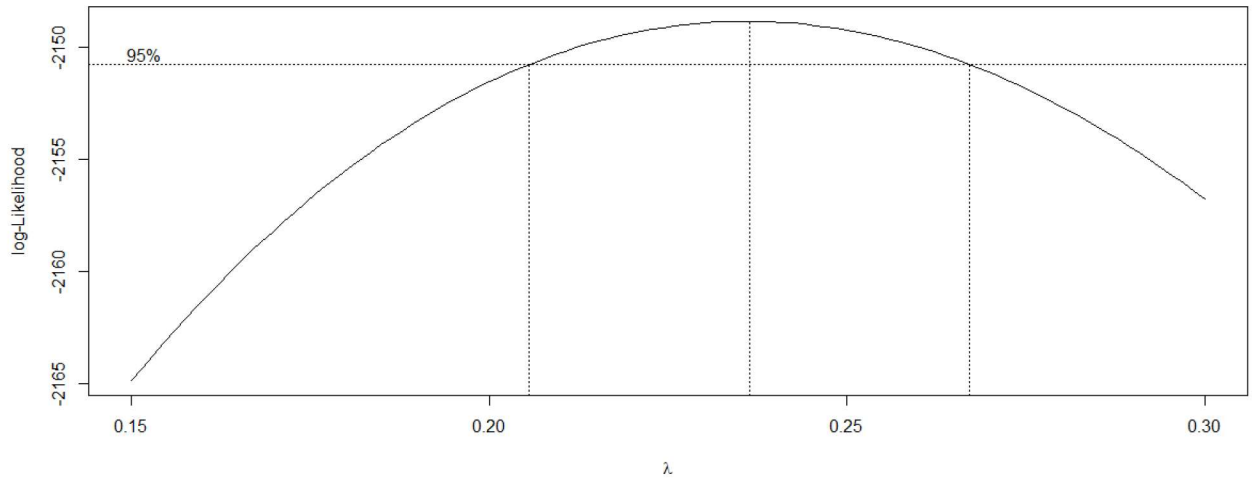
U ovom poglavlju pokušat ćemo odrediti najbolji model za naš problem. Od mnogih pokušaja kreiranja modela za našu bazu izdvojit ćemo dva najbolja.

7.2.1 Model 1

Na poduzorku od 600 vrijednosti kreirat ćemo model za određivanje vrijednosti igrača u nogometu. Istraživanjem smo zaključili kako nam višestruka linearna regresija varijable Cijena daje heteroskedastičnu varijancu, primjenom ncvTesta zaključujemo kako moramo transformirati varijablu Cijena. Primjenom Box-Cox transformacije pokušat ćemo izgraditi model. Box-cox transformacija transformira ovisne slučajne varijable koje nisu normalno distribuirane u normalno distribuirane.
Box-Cox transformacija ovisne varijable Cijena izgleda ovako :

$$y(\lambda) = \begin{cases} \frac{Cijena^{\lambda}-1}{\lambda} & , \text{ ako je } \lambda \neq 0 \\ \log(Cijena) & , \text{ ako je } \lambda = 0 \end{cases}$$

Primjenom $boxcox()$ funkcije iz paketa $MASS$ u software-u R odredili smo λ za naš model. Na slici 11 možemo vidjeti pouzdani interval za λ. Za naš poduzorak najbolji λ iznosi 0.2222222.



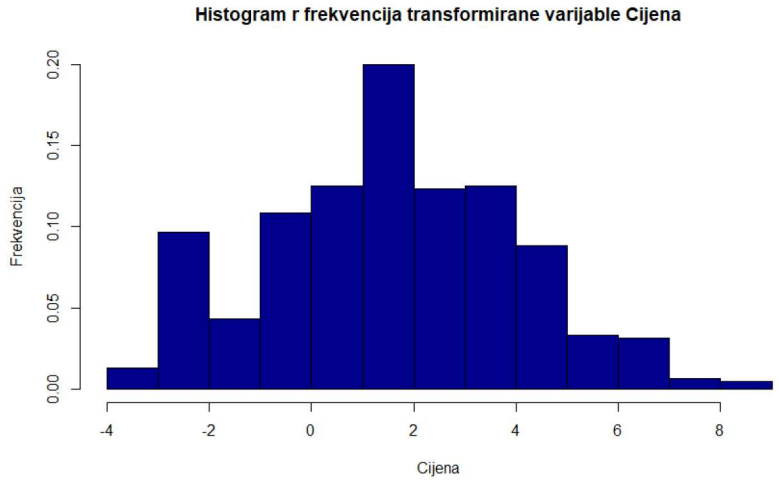
Slika 11: 95% puzdani interval za λ

Pogledajmo u tablici 17 deskriptivnu statistiku transformirane varijable Cijena.

Minimum	Donji kvantil	Medijan	Prosjek	Gornji kvantil	Maksimum	Sd
-3.709	0.00	1.624	1.666	3.317	8.539	2.4225

Tablica 17: Deskriptivna statistika transformirane varijable Cijena

Na slici 12 možemo vidjeti histogram relativnih frekvencija transformirane varijable Cijena.

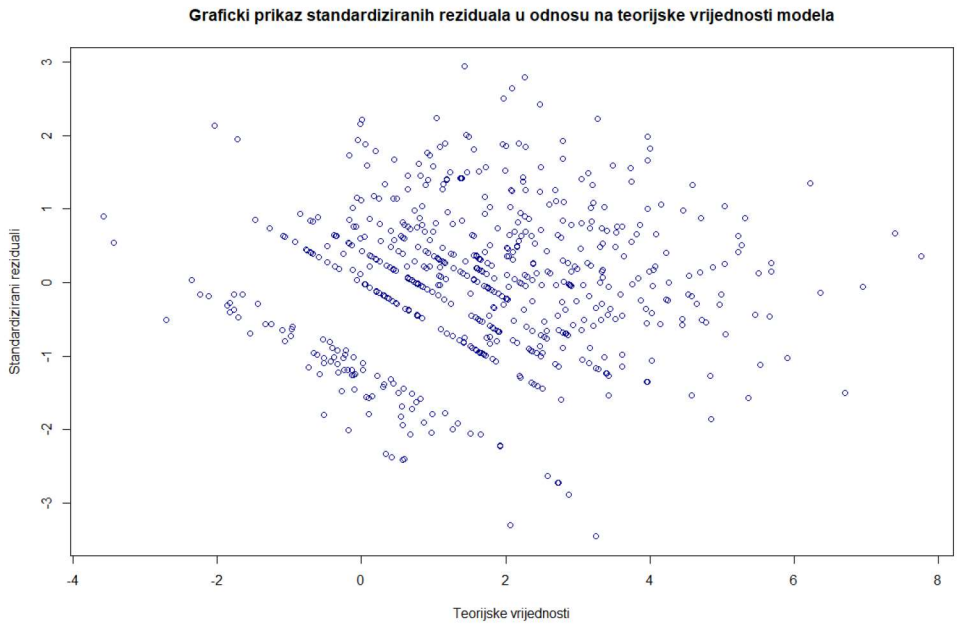


Slika 12: Histogram relativnih frekvencija transformirane varijable Cijena

7.2.2 Izgled modela 1 i provjera pretpostavki

Kako smo u poglavlju 6.3 pokazali da varijable Noga i Visina nisu statistički značajne, njih ne upotrebljavamo u izgradnji ovog modela. Za izradu modela koristit ćemo varijable Područje, Pozicija, Broj_godina, Liga, Minuta, povjerenje i omjerga (vidi tablicu 21). Model koji ima najmanji AIC ujedno je i model koji ima najveći korigirani $\bar{R}^2$.

Provjerimo sada pretpostavke. Prvo ćemo provjeriti homoskedastičnost grešaka modela. Znamo da su greške nemjerljive veličine, zato za njihovu procjenu koristimo standardizirane reziduale. Na slici 13 prikazan je grafički prikaz standardiziranih reziduala u odnosu na teorijske vrijednosti modela.



Slika 13: Grafički prikaz standardiziranih reziduala u odnosu na teorijske vrijednosti modela 1

S prethodne slike možemo naslutiti da u modelu ne postoji problem s homoskedastičnosti varijance, ali to ćemo statistički potvrditi (vidi tablicu 18). Za provjeru homoskedastičnosti varijance koristit ćemo dva testa iz software-a R, a to su: ncvTest i Goldfeld-Quandt test.
Hipoteze:

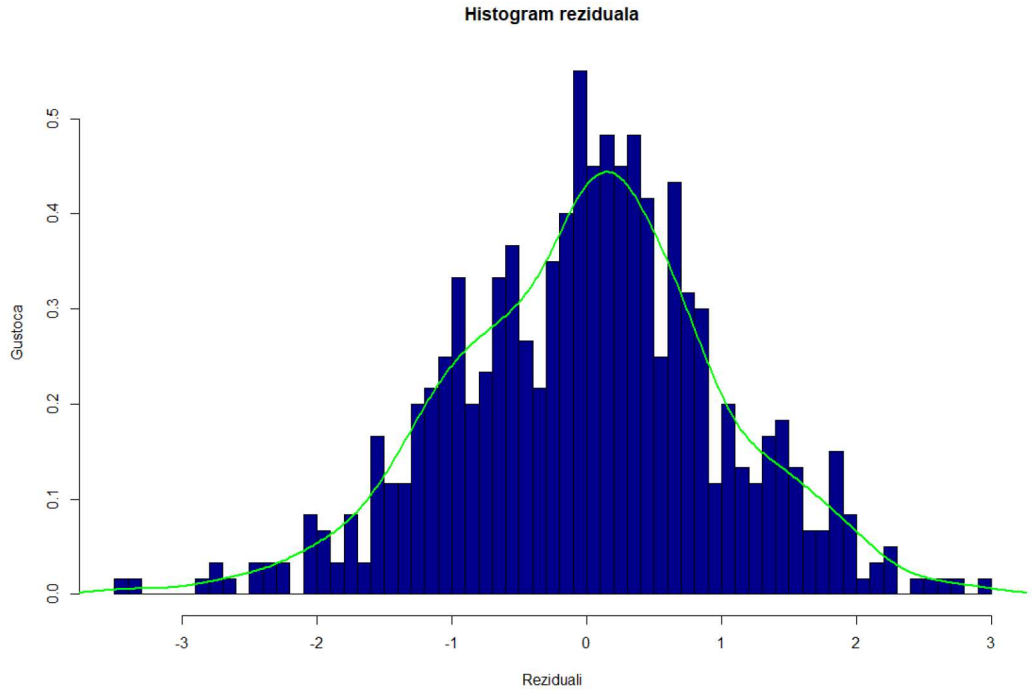
- H_0 **: Homoskedastičnost varijance modela** (Varijanca je konstanta)
- H_1 **: Heteroskedastičnost varijance modela** (Varijanca nije konstanta)

Test	**p-vrijednost**
ncvTest	0.68587
Goldfeld-Quandt test	0.875

Tablica 18: Provjera homoskedastičnosti modela 1

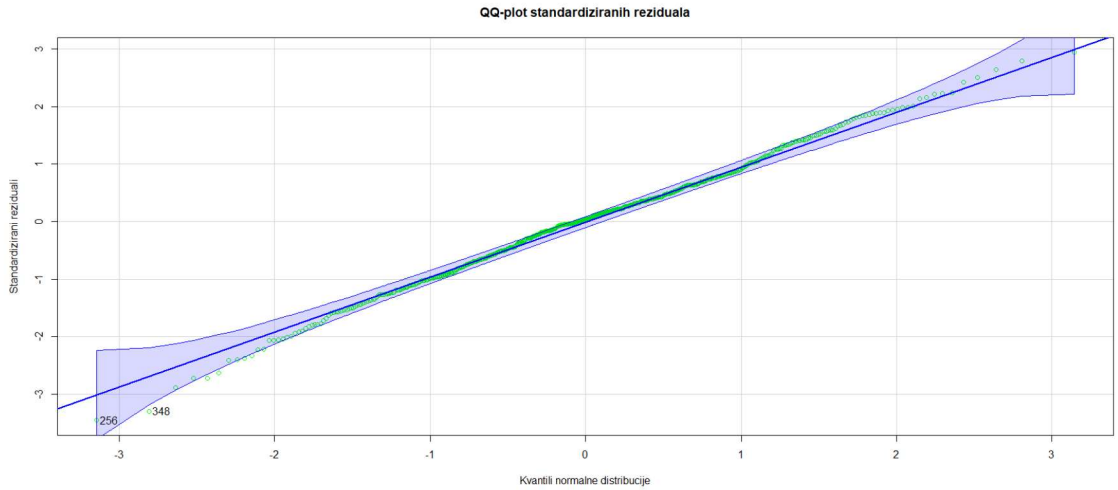
Oba testa imaju p-vrijednost veću od 0.05 stoga na razini značajnosti 0.05 nemamo razloga sumnjati u homoskedastičnost grešaka modela.

Sljedeće što ćemo provjeriti je jesu li greške modela normalno distribuirane. Kao i u prethodnoj provjeri za procjenu grešaka modela koristit ćemo standardizirane reziduale. Za početak pogledajmo sliku 14 na kojoj se nalazi histogram reziduala s obzirom na danu funkciju gustoće.



Slika 14: Histogram reziduala s obzirom na funkciju gustoće

Koristan alat za provjeru normalnosti u software-u R jest funkcija *qqplot*() (vidi sliku 15).



Slika 15: QQ plot standardiziranih reziduala

S prethodne slike naslućujemo da standardizirani reziduali modela dobro prate normalnu distribuciju, što ćemo u nastavku statistički provjeriti (vidi tablicu 19). Za provjeru normalnosti standardiziranih reziduala koristit ćemo dva testa iz software-a R, a to su: Shapiro-Wilkov test i Kolmogorov-Smirnovljev test.

Hipoteze:

- H_0 : **Reziduali su normalno distribuirani**
- H_1 : **Reziduali nisu normalno distribuirani**

Test	**p-vrijednost**
Shapiro-Wilkov test	0.2144
Kolmogorov-Smirnovljev test	0.3063

Tablica 19: Provjera normalnosti reziduala za model 1

Kako je kod oba testa p-vrijednost veća od 0.05 ne odbacujemo hipotezu H_0 te nemamo razloga sumnjati u normalnost standardiziranih reziduala modela.

Još nam je ostalo za provjeriti ima li naš model problema s multikolinearnosti. Za provjeru multikolinearnosti koristit ćemo faktor inflacije varijance. Pogledajmo sada rezultate u tablici 20.

	VIF
PodrucjeAmerika	1.6778
PodrucjeAzija	1.1983
PodrucjeEuropa	1.8875
PozicijaFW	1.6940
PozicijaGK	1.2124
PozicijaMF	1.4174
Pozicijamulti	1.6781
Broj_godinastariji	1.1867
Broj_godinazreli	1.2052
Minuta	2.0546
LigaLa Liga	1.8119
LigaLigue 1	1.8602
LigaPremier League	1.8288
LigaSerie A	1.8197
povjerenje	2.4124
omjerga	1.4041

Tablica 20: Faktor inflacije varijance za model 1

Iz prethodne tablice vidimo kako su svi faktori inflacije varijance manji od 5 što znači da nemamo razloga sumnjati u multikolinearnost modela.

Sada kada imamo model koji zadovoljava sve pretpostavke na redu je da ga interpretiramo. Rezultate procijenjenih vrijednosti koeficijenata možemo vidjeti u tablici 21.

	Koeficijenti	**Standardna greška**	**t-vrijednost**	**p-vrijednost**
Slobodni član	-0.7841	0.3789	-2.069	0.0389
PodrucjeAmerika	0.6198	0.3211	1.931	0.0540
PodrucjeAzija	-0.3975	0.5592	-0.711	0.4775
PodrucjeEuropa	0.0757	0.2329	0.325	0.7453
PozicijaFW	0.4373	0.2848	1.536	0.1252
PozicijaGK	-0.8322	0.3021	-2.755	0.0061
PozicijaMF	0.7061	0.1975	3.575	0.0004
Pozicijamulti	0.4417	0.2231	1.980	0.0481
Broj_godinastariji	-2.8471	0.3239	-8.790	$<$ 2e-16
Broj_godinazreli	-0.4160	0.1590	-2.615	0.0092
Minuta	0.0009	0.0001	8.457	$<$ 2e-16
LigaLa Liga	0.4112	0.2438	1.687	0.0922
LigaLigue 1	-0.0702	0.2409	-0.292	0.7707
LigaPremier League	1.8224	0.2375	7.674	7.04e-14
LigaSerie A	0.1191	0.2355	0.506	0.6131
povjerenje	0.6084	0.3652	1.666	0.0963
omjerga	3.0506	0.4518	6.752	3.53e-11

Tablica 21: Koeficijenti modela 1

Interpretacija koeficijenata:

- Vrijednost procijenjenog parametra uz varijablu PodručjeAmerika iznosi 0.6198. Odnosno, igrač s područja Sjeverne ili Južne Amerike prosječno ima transformaciju cijene veću za 0.6198 od igrača koji dolaze s područja Afrike, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu PodručjeAzija iznosi -0.3975. Odnosno, igrač s područja Azije prosječno ima transformaciju cijene manju za 0.3975 od igrača koji dolaze s područja Afrike, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu PodručjeEuropa iznosi 0.0757. Odnosno, igrač s područja Europe prosječno ima transformaciju cijene veću za 0.3975 od igrača koji dolaze s područja Afrike, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu PozicijaFW iznosi 0.4373. Odnosno, napadači prosječno imaju transformaciju cijene veću za 0.4373 od obrambenih igrača, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu PozicijaGK iznosi -0.8322. Odnosno, vratari prosječno imaju transformaciju cijene manju za 0.8322 od obrambenih igrača, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu PozicijaMF iznosi -0.7061. Odnosno, vezni igrači prosječno imaju transformaciju cijene manju za 0.7061 od obrambenih igrača, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu Pozicijamulti iznosi 0.4418. Odnosno, igrači koji mogu igrati više pozicija prosječno imaju transformaciju cijene veću za 0.4418 od obrambenih igrača, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu Broj_godinastariji iznosi -2.8471. Odnosno, igrači stariji od 32. godine prosječno imaju transformaciju cijene manju za 2.8471 od igrača do 23. godine, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu Broj_godinazreli iznosi -0.4160. Odnosno, igrači stariji od 23. godine prosječno imaju transformaciju cijene manju za 0.4160 od igrača do 23. godine, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu Minuta iznosi 0.0009. Odnosno, jedinično povećanje minuta provedenih na terenu se prosječno reflektira povećanjem transformacije cijene za 0.0009, uz sve ostale parametre fiksne.

- Vrijednost procijenjenog parametra uz varijablu LigaLa Liga iznosi 0.4112. Odnosno, igrači iz La Lige prosječno imaju transformaciju cijene veću za 0.4122 od igrača iz Bundeslige, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu LigaLigue 1 iznosi -0.0702. Odnosno, igrači iz Ligue 1 prosječno imaju transformaciju cijene manju za 0.0702 od igrača iz Bundeslige, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu LigaPremier League iznosi 1.8224. Odnosno, igrači iz Premier League prosječno imaju transformaciju cijene veću za 1.8224 od igrača iz Bundeslige, uz sve ostale varijable fiksne.

- Vrijednost procijenjenog parametra uz varijablu LigaSerie A iznosi 0.1191. Odnosno, igrači iz Serie A prosječno imaju transformaciju cijene manju za 0.1191 od igrača iz Bundeslige, uz sve ostale varijable fiksne

- Vrijednost procijenjenog parametra uz varijablu povjerenje iznosi 0.6084. Odnosno, jedinično povećanje omjera započetih i odigranih utakmica se prosječno reflektira povećanjem transformacije cijene za 0.6084, uz sve ostale parametre fiksne.

- Vrijednost procijenjenog parametra uz varijablu omjerga iznosi 3.0506. Odnosno, jedinično povećanje omjera golova i asistencija po utakmici se prosječno reflektira povećanjem transformacije cijene za 3.0506, uz sve ostale parametre fiksne.

Dobivena vrijednost AIC-a za ovaj model iznosi 727.08. Vrijednost $\bar{R}^2$ iznosi 0.477 što znači da je 47.7% ukupne varijabilnosti u podacima objašnjeno modelom.

Sada ćemo u tablici 22 nasumično odabrati 15 procijenjenih podataka te usporediti s originalnim podacima.

U tablici 22 upotrebljavat ćemo sljedeće skraćenice:

- **PTV** := Procijenjena transformirana vrijednost nogometaša
- **TV** := Transformirana vrijednost nogometaša
- **PV** := Procijenjena vrijednost nogometaša
- **OV** := Originalna vrijednost nogometaša

Ime i prezime	**PTV**	**PV**	**TV**	**OV**
Lionel Messi	7.757537	90.857	8.340865	112
Kevin de Bruyne	7.397453	79.4487	8.539255	120
Casemiro	4.588978	23.6517	6.878425	65
Mateo Kovačić	4.151182755	18.9401	5.985595	45
Marcus Rashford	6.946471273	66.7676	6.839289	64
Callum Hudson-Odoi	3.368676065	12.3624	5.152198	31
Jan Morávek	0.51522990	1.6287	-2.0918111	0.06
Gianmarco Cangiano	-0.12818702	0.8781	-2.1874309	0.05
Federico Valverde	3.737116113	15.1891	6.419152	54
Kalidou Koulibaly	1.434552201	3.4737	6.507754	56
Edin Džeko	10.3871	0.1590	3.006452	10
Filip Bradarić	0.944991112	2.3579	1.3273092	3.4
Júlio Tavares	2.509684427	7.3483	0.8617485	2.2
Cauly Oliveira Souza	1.779140493	4.4781	0.3493710	1.4
Darko Churlinov	-0.25181420	0.7717	-2.0279469	0.0675

Tablica 22: Usporedba podataka procijenjenih modelom 1 s originalnim podacima

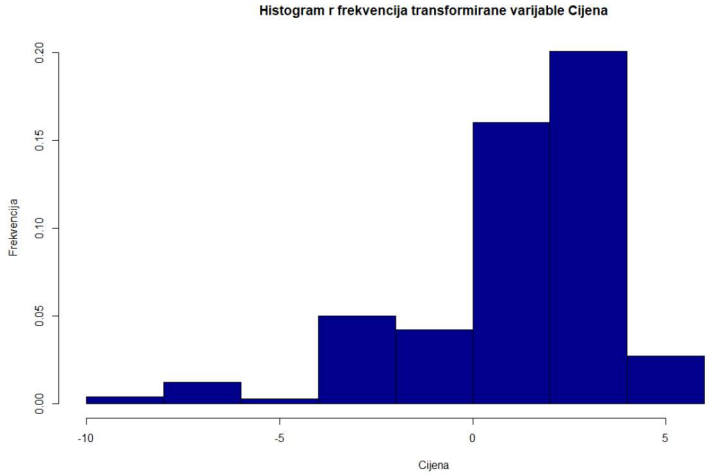
7.2.3 Model 2

Na uzorku od 368 vrijednosti kreirat ćemo model za određivanje vrijednosti igrača koji igraju na poziciji napadača. Kao i u prethodnom modelu primjenom ncvTesta zaključili smo kako višestruka linearna regresija varijable Cijena daje heteroskedastičnu varijancu, te da moramo transformirati varijablu Cijena. U ovom slučaju korištenjem logaritamske transformacije varijable Cijena pokušat ćemo izgraditi model. Pogledajmo u tablici 23 deskriptivnu statistiku transformirane varijable Cijena.

Minimum	Donji kvantil	Medijan	Prosjek	Gornji kvantil	Maksimum	Sd
-9.9035	0.6931	1.7483	1.1600	2.7726	5.1930	2.5709

Tablica 23: Deskriptivna statistika transformirane varijable Cijena modela 2

Na slici 16 možemo vidjeti histogram relativnih frekvencija transformirane varijable Cijena.

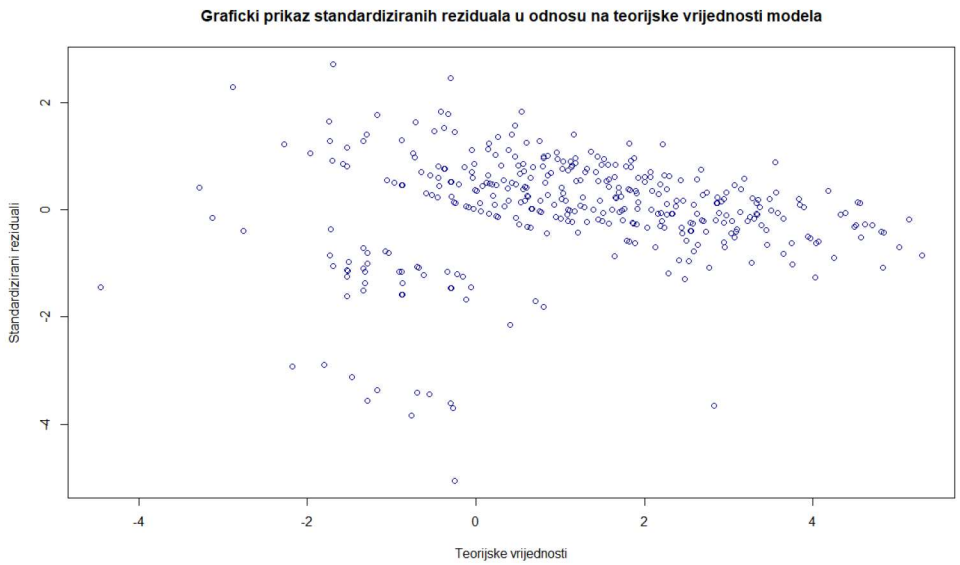


Slika 16: Histogram relativnih frekvencija transformirane varijable Cijena modela 2

7.2.4 Izgled modela 2 i provjera pretpostavki

Kako smo ranije pokazali da varijable Noga i Visina nisu statistički značajne, te varijabla Pozicija koju smo fiksirali na napadače, njih ne upotrebljavamo u izgradnji ovog modela. Pri izradi ovog modela koristit ćemo varijable Područje, Broj_godina, Liga, Minuta, povjerenje i omjerga (vidi tablicu 27).

Sljedeće što ćemo učiniti jest provjeriti pretpostavke modela. Prvo ćemo provjeriti homoskedastičnost grešaka modela. Na slici 17 prikazan je grafički prikaz standardiziranih reziduala u odnosu na teorijske vrijednosti modela.



Slika 17: Grafički prikaz standardiziranih reziduala u odnosu na teorijske vrijednosti modela 2

Za provjeru homoskedastičnosti varijance koristit ćemo ncvTest i Goldfeld-Quandt test (vidi tablicu 24).

Hipoteze:

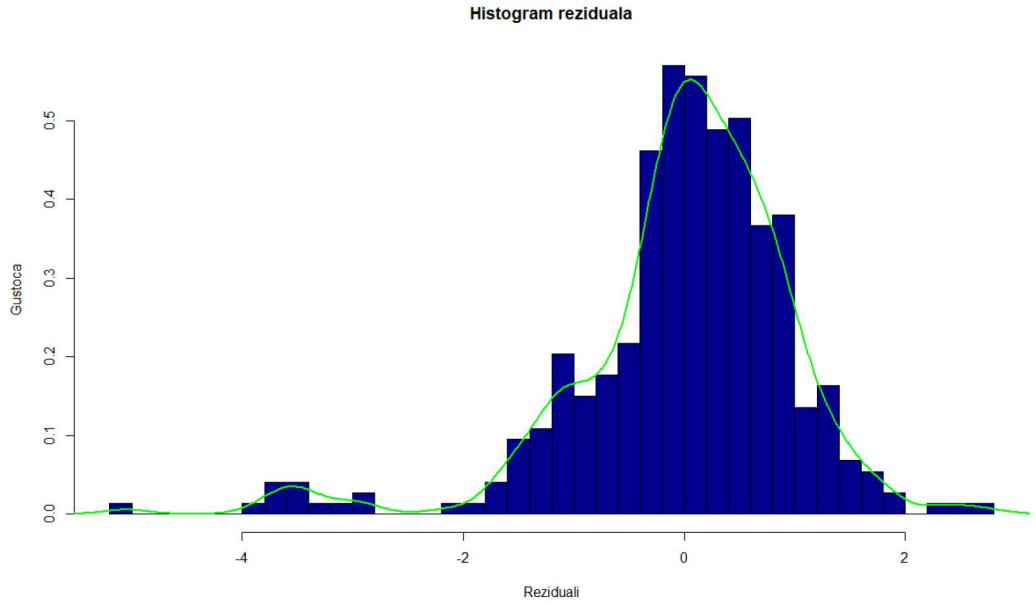
- H_0 : **Homoskedastičnost varijance modela** (Varijanca je konstanta)
- H_1 : **Heteroskedastičnost varijance modela** (Varijanca nije konstanta)

Test	**p-vrijednost**
ncvTest	2.22e-16
Goldfeld-Quandt test	0.1682

Tablica 24: Provjera homoskedastičnosti modela 2

Budući da ncvTest ima p-vrijednost manju od 0.05, na razini značajnosti 0.05 odbacujemo H_0, tj. možemo smatrati da varijanca nije konstanta.

Zatim ćemo provjeriti jesu li greške modela normalno distribuirane. Za početak pogledajmo sliku 18 na kojoj se nalazi histogram reziduala s obzirom na danu funkciju gustoće.



Slika 18: Histogram reziduala s obzirom na funkciju gustoće za model 2

Za provjeru normalnosti standardiziranih reziduala koristit ćemo Shapiro-Wilkov test i Kolmogorov-Smirnovljev test (vidi tablicu 25).

Hipoteze:

- H_0 **: Reziduali su normalno distribuirani**
- H_1 **: Reziduali nisu normalno distribuirani**

Test	**p-vrijednost**
Shapiro-Wilkov test	2.373e-14
Kolmogorov-Smirnovljev test	0.0001

Tablica 25: Provjera normalnosti reziduala za model 2

Kako je kod oba testa p-vrijednost manja od 0.05 odbacujemo hipotezu H_0 te ne možemo tvrditi da su standardizirani reziduali modela normalno distribuirani.

Još nam je ostalo za provjeriti ima li naš model problema s multikolinearnosti. Za provjeru multikolinearnosti koristit ćemo faktor inflacije varijance. Pogledajmo sada rezultate u tablici 26.

	VIF
PodrucjeAmerika	1.7808
PodrucjeAzija	1.1879
PodrucjeEuropa	1.8979
Broj_godinastariji	1.1293
Broj_godinazreli	1.2282
Minuta	3.7433
LigaLa Liga	1.9869
LigaLigue 1	1.9665
LigaPremier League	1.9685
LigaSerie A	1.9929
povjerenje	3.2721
omjerga	1.9761

Tablica 26: Faktor inflacije varijance za model 2

Iz prethodne tablice vidimo kako su svi faktori inflacije varijance manji od 5, znači da nemamo razloga sumnjati u multikolinearnost modela.

Iako nam model ne zadovoljava sve pretpostavke, rezultate procijenjenih vrijednosti koeficijenata možemo vidjeti u tablici 27.

	Koeficijenti	**Standardna greška**	**t-vrijednost**	**p-vrijednost**
Slobodni član	-1.1178	0.4286	-2.608	0.0095
PodrucjeAmerika	0.3975	0.4150	0.958	0.3388
PodrucjeAzija	0.3129	0.7119	0.440	0.6606
PodrucjeEuropa	-0.2203	0.3069	-0.718	0.4734
Broj_godinastariji	-3.1366	0.4971	-6.309	8.36e-10
Broj_godinazreli	0.0392	0.2244	0.175	0.8614
Minuta	0.0002	0.0002	1.000	0.2763
LigaLa Liga	-0.1958	0.3480	-0.563	0.5739
LigaLigue 1	-0.4005	0.3646	-1.099	0.2727
LigaPremier League	1.0384	0.3448	3.012	0.0028
LigaSerie A	0.4558	0.3366	1.354	0.1765
povjerenje	2.5140	0.5814	4.324	1.99e-05
omjerga	2.4390	0.5658	4.311	2.11e-05

Tablica 27: Koeficijenti modela 2

Dobivena vrijednost AIC-a za ovaj model iznosi 496.78. Vrijednost korigiranog $\bar{R}^2$ iznosi 0.4332 što znači da je 43.32% ukupne varijabilnosti u podacima objašnjeno modelom.

Sada ćemo u tablici 28 kao i u prethodnom modelu nasumično odabrati 15 procijenjenih podataka te usporediti s originalnim podacima.

Ime i prezime	**PTV**	**PV**	**TV**	**OV**
Raheem Sterling	4.18274584	65.5455	4.852030	128
Aritz Aduriz	-4.448806007	0.0117	-7.1308988	0.0008
Opoku Ampomah	0.043909145	1.0449	0.2623643	1.3
Lukáš Haraslín	0.341266131	1.4067	0.4700036	1.6
Zlatan Ibrahimović	0.559476521	1.7498	1.2527630	3.5
Bas Dost	2.619230707	5.6425	1.7047481	5.5
Nikola Kalinić	2.126100044	3.5556	1.7047481	5.5
Pedro	2.980384495	8.6916	2.0149030	7.5
Erling Haaland	4.38051011	79.8788	4.276666	72
Christian Pulisic	3.89051213	48.9359	3.988984	54
Darko Churlinov	-1.333355604	0.2636	-2.6956277	0.0675
Wout Weghorst	3.13188412	22.9171	3.044522	21
Lois Diony	0.535211665	1.7078	0.7884574	2.2
Aimar Oroz	-1.531476744	0.2162	-4.6051702	0.01
Romelu Lukaku	3.83570869	46.3263	4.219508	68

Tablica 28: Usporedba podataka procijenjenih modelom 2 s originalnim podacima

Popis slika

Popis tablica

Literatura

[1] V. Bahovec, N. Erjavec, *Uvod u ekonometrijsku analizu*, Element, Zagreb, 2009.

[2] L. J. Bain, M. Engelhardt, *Introduction to Probability and Mathematical Statistics*, Pacific Grove, Duxbury/Thomson Learning, 1991.

[3] A. Basilevsky, *Statistical Factor Analysisi and Related Models: Theory and Applications*, Wiley-Interscience, New York, 1994.

[4] M. Benšić, *Predavanja za kolegij Statistika*, https://www.mathos.unios.hr/images/homepages/mirta/statistika/sve1.pdf

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

[6] T. S. Breusch, A. R. Pagan, *A simple test for heteroscedasticity and random coefficient variation*, Econometrica 47, 1979.

[7] A. J. Dobson, A. G. Barnett, *An introduction to generalized linear models*, CRC Press, Boca Raton, 2018.

[8] F.E. Harrell, *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer, New York, 2001.

[9] I. Hendriks, *Modelling the trasfer prices of football players*, Tilnburg University, Tilnburg, 2017.

[10] $https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt\_fbref\_201920.csv$

[11] $https://www.statology.org/how-to-read-the-f-distribution-table/$

[12] A. C. Rencher, G. B. Schaalje, *Linear models in statistics*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2008

Sažetak

Sve vezano za nogomet dobiva ogromnu medijsku popularnost. Nogometnim zaljubljenicima vrlo je zanimljivo promatrati događanja van nogometnih terena, kao što su razne statistike, špekulacije oko transfera, vrijednosti klubova i igrača. Cilj ovog diplomskog rada je modelirati vrijednost igrača u nogometu pomoću višestruke linearne regresije. U prvom dijelu rada opisana je višestruka linearna regresija, te dijagnostika modela. U drugom dijelu rada statistički smo obradili sve varijable korištene u kreiranju modela. U posljednjem poglavlju kreirali smo dva modela višestruke linearne regresije. Jedan je uključivao sve podatke dok je drugi model uključivao samo napadače.

Ključne riječi: linearna regresija, nogomet, vrijednost igrača u nogometu

Modelling the value of football players

Summary

Everything related to football is gaining massive media popularity. It is very interesting for football enthusiasts to observe events outside football field, such as various statistics, transfer rumours and the value of clubs and players. The aim of this master thesis is to model the value of football players using multiple linear regression. The first part of thesis describes multiple linear regression and model diagnostics. In the second part of the thesis, we analysed all the variables used in the model. In the last chapter, we created two models of multiple linear regression. One model included all the data while the other model included only strikers.

Keywords: linear regression, football, value of football player

Životopis

Rođen sam 6. siječnja 1997. godine u Osijeku. Pohađao sam osnovnu školu "Hrvatski sokol" u Podravskim Podgajcima. 2011. godine upisujem Prirodoslovno-matematičku gimnaziju u Osijeku koju završavam 2015. godine. Iste godine upisujem se na Odjelu za matematiku Sveučilišta Josipa Jurja Strossmayera u Osijeku na preddiplomskom studiju matematike kojeg završavam 2019. godine i stječem naziv prvostupnika matematike s temom završnog rada *Primjena konformnih preslikavanja u aerodinamici* pod mentorstvom izv. prof. dr. sc. Snježane Majstorović. Nakon toga, 2019. godine upisujem diplomski studij matematike, smjer Financijska matematika i statistika. Trenutno sam zaposlen kao mlađi specijalist za poslovnu inteligenciju u Zagrebačkoj banci.