Matematički modeli u nogometu

Franjo, Martina

Master's thesis / Diplomski rad

2020

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:126:342847

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

Martina Franjo

Matematički modeli u nogometu

Diplomski rad

Osijek, 2020.

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

Martina Franjo

Matematički modeli u nogometu

Diplomski rad

Mentor: doc. dr. sc. Danijel Grahovac

Osijek, 2020.

Sadržaj

Uvod

Nogomet je jedan od najpopularnijih sportova današnjice. Ljepota nogometne igre te njezina dinamičnost osnovni su uzroci zbog kojih se nogomet igra na svim kontinentima i u svim slojevima društva. Nogometni prijenosi značajnijih utakmica okupljaju oko TV prijamnika i do milijardu ljudi. U našoj zemlji nogomet se smatra jednim od nacionalnih sportova. Prisjetimo se 2018. godine i svjetskog nogometnog prvenstva u Rusiji kada je Hrvatska postala viceprvak svijeta. Finalnu utakmicu između Hrvatske i Francuske gledalo je 1,12 milijardi ljudi, a tijekom cijelog prvenstva u Hrvatskoj je vladala euforija, ulice i trgove preplavile su "kockice" - navijači, a sve to zbog nogometa. Općenito u svijetu je najgledanija engleska profesionalna nogometna liga Premier Liga, često zvana i Premiership.

Nogometna utakmica je dvoboj dviju momčadi od 11 igrača, 10 na terenu i vratar. Sastoji se od dva dijela po 45 minuta koje nazivamo poluvrijeme, a između njih je pauza od otprilike 15 minuta. Nakon svakog poluvremena igra se može produžiti da bi se nadoknadilo izgubljeno vrijeme zbog primjerice zamjene igrača, iznošenja igrača s terena ili bilo kojeg drugog uzroka.

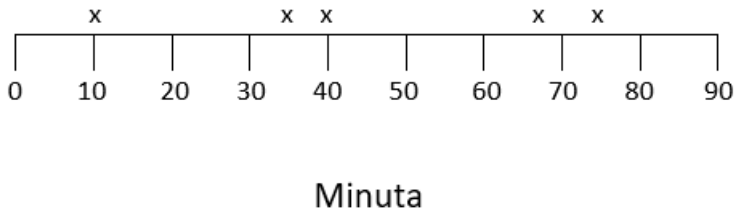
Nogomet kao što i samo ime govori, igra se nogom, a samo vratar smije rukom primiti loptu. Taj spoj noga - lopta zaslužuje posebnu pažnju, jer upravo to doprinosi nepredvidljivosti od početka do samog kraja utakmice. Igrači vježbaju satima, danima, godinama, posao im je da vladaju loptom, a ipak ne mogu dostići takvo savršenstvo i sigurnost da uvijek s loptom uspiju napraviti što su naumili. Svuda postoji jedna izreka "lopta je okrugla" i znači da je sve moguće, odnosno u nogometu ništa nije sigurno. Dakle u toj se igri ništa sa sigurnošću ne može predvidjeti, svaki napad je i nova neizvjesnost te nova mogućnost preokreta. Nepredvidivost je jedna od osnovnih čari nogometa. Možda je baš zbog te neizvjesnoti ljudima zanimljivo predviđati različite događaje na utakmici, kao što su: koja momčad će pobijediti, koliki će biti ukupan broj golova na utakmici, broj kornera, koja momčad će prva zabiti gol, koji igrač će prvi zabiti i slično. Ova predviđanja ljudi rade većinom intuitivno, bez prevelike analize, u svrhu klađenja, uzimajući u obzir rezultate pojedine momčadi u prijašnjim utakmicama.

Da ne bi sve ostalo samo na intuiciji, u ovom radu baviti ćemo se matematičkim modelima kojima je moguće predviđati ishode pojedinih utakmica, na osnovi prethod-

nih podataka. Najznačajniji modeli kojima možemo predviđati ishode utakmica su model s Poissonovom distribucijom broja golova i Dixon-Coles model. U praktičnom dijelu rada primijeniti ćemo modele na stvarnim podacima.

1 Osnovne definicije i Poissonova distribucija

"Događaji" koji se pojavljuju u sportu su nasumični. Ako gledamo događaje tijekom pojedine utakmice, možemo zamijetiti da ih možemo prikazati na vremenskom intervalu. Uzmimo za primjer nogometnu utakmicu gdje se golovi koje je postigla bilo koja momčad tijekom 90 minuta događaju nasumično u razdoblju od 90 minuta igre. Za utakmicu koja je završila 3-2, recimo da su golovi postignuti u 10.,35.,40.,67. i 84. minuti (zanemarimo koji je tim postigao pogodak), tada na jediničnom vremenskom intervalu možemo označiti kada je postignut gol. Ovaj nasumičan odabir "događaja"



Slika 1: Primjer postignutih golova na utakmici

odnosno golova može se promatrati kao postupak prebrojavanja. Uočimo da je ukupan broj golova na utakmici 5. Distribucija koja se može koristiti za modeliranje podataka koji se mogu prebrojati je Poissonova distribucija. Ako znamo koliko se puta nešto očekuje, možemo pronaći vjerojatnost da se to dogodi bilo koji broj puta. Na primjer, ako znamo da se nešto očekuje 5 puta, možemo izračunati vjerojatnosti da će se to dogoditi 0, 1, 2, ... puta. Ispada da broj golova koje momčad postigne u nogometnoj utakmici ima približno Poissonovu distribuciju, pa imamo metodu dodjeljivanja vjerojatnosti broju golova u utakmici i iz toga možemo pronaći vjerojatnosti za različite rezultate utakmice.
Sada kada je jasno da se nogometni rezultati mogu modelirati koristeći Poissonovu distribuciju prije nego počnemo kreirati model prisjetimo se osnovnih definicija, definicije očekivanja, varijance te Poissonove distribucije i njezinih bitnih svojstava.

1.1 Očekivana vrijednost i varijanca

Očekivana vrijednost može se izračunati kao zbroj vjerojatnosti svih mogućih ishoda pomnoženih s pripadajućim dobicima ili gubicima.

Definicija 1.1. ***(Očekivanje)***
Za diskretnu slučajnu varijablu X sa skupom mogućih vrijednosti x_i, raspodjelom vjerojatnosti p_i za svaki ishod $i \in \mathbb{N}$, definiramo ***očekivanje*** *kao*

$$E[X] = \sum_{i=1}^{\infty} p_i x_i. \tag{1.1}$$

Definicija 1.2. ***(Varijanca)***
Ako postoji broj $E(X - EX)^2$, onda taj pozitivan realan broj nazivamo ***varijanca*** *slučajne varijable X i označavamo sa Var(X).*

Varijanca je očekivano kvadratno odstupanje slučajne varijable od njenog očekivanja, a korijen iz varijance nazivamo standardna devijacija.

1.2 Poissonova distribucija i svojstva

Poissonovom distribucijom modeliramo slučajnu varijablu koja broji uspjehe u nekom jediničnom intervalu vremena, površine, mase i slično. Pri tome moraju biti zadovoljena sljedeća tri uvjeta:

1) Vjerojatnost pojavljivanja uspjeha ne ovisi o tome u kojem će se jediničnom intervalu on dogoditi.
2) Broj uspjeha u jednom, neovisan je o broju uspjeha u bilo kojem drugom jediničnom intervalu.
3) Očekivani broj uspjeha isti je za sve jedinične intervale i dan je pozitivnim realnim brojem λ.

Kako nas zanima broj golova, nama će uspjeh biti postignuti gol na utakmici, bez obzira koja od dvije momčadi je postigla gol. Nogometna utakmica zadovoljava sva tri uvjeta. Svaki napad je nova prilika za gol, mi ne možemo znati u kojem jediničnom intervalu će biti postignuto više golova. Očekivani broj golova je isti u svim intervalima utakmice, a također je i broj golova u jednom intervalu neovisan o broju golova u bilo kojem drugom intervalu. Iako, neke su analize pokazale da je najviše postignutih golova na utakmici krajem prvog i krajem drugog poluvremena. Uzrok tome može biti umor, gubitak koncentracije jedne momčadi, a u tom trenutku suprotna momčad iskoristi priliku i postigne gol. Ovu iskazanu činjenicu ćemo zanemariti, te ćemo gledati da je vjerojatnost pogotka jednaka u svim intervalima nogometne utakmice.

Definicija 1.3. *Kažemo da diskretna slučajna varijabla X ima* ***Poissonovu distribuciju s parametrom*** $\lambda > 0$*, ako prima vrijednosti iz skupa* $\mathbb{N}_0$ *s vjerojatnostima*

$$p_i = P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0, 1, 2, ... \tag{1.2}$$

Tada pišemo $X \sim \mathcal{P}(\lambda)$.

Provjerimo je li na ovaj način dobro definirana distribucija, tj. je li $\sum_{i=0}^{\infty} p_i = 1$. Vrijedi:

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1.$$

Jedinstveno svojstvo Poissonove distribucije je da su i očekivanje i varijanca jednaki λ, što ćemo pokazati.

Očekivanje:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k \cdot P(X = k) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k \cdot \lambda^k}{k\,(k-1)!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda \cdot e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}, \end{aligned}$$

a jer je $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$, slijedi da je

$$E[X] = \lambda.$$

Varijanca:

$Var(X)$ možemo zapisati i ovako

$$Var(X) = EX^2 - (EX)^2 = EX^2 - \lambda^2. \tag{1.3}$$

Odmah smo uvrstili očekivanje koje smo prethodno izračunali. Preostaje nam izračunati EX^2.

$$E\left[X^2\right] = \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\lambda}\lambda^k}{k!}, \tag{1.4}$$

uvrstiti ćemo $k = 0$ i $k = 1$, te ćemo k^2 zapisati kao $k^2 = k^2 + k - k = k(k-1) + k$. Tada slijedi,

$$\begin{aligned} E\left[X^2\right] &= e^{-\lambda}\left[\lambda + \sum_{k=2}^{\infty}\left(k^2 - k + k\right)\cdot\frac{\lambda^k}{k!}\right] \\ &= e^{-\lambda}\left[\lambda + \sum_{k=2}^{\infty}(k-1)\,k\cdot\frac{\lambda^k}{k!} + \sum_{k=2}^{\infty}k\cdot\frac{\lambda^k}{k!}\right], \end{aligned}$$

λ uvrstimo u drugu sumu. Na isti način kao što smo izračunali očekivanje sredimo prethodni izraz. Dobijemo sljedeće,

$$\begin{aligned} E\left[X^2\right] &= e^{-\lambda}\left[\lambda\cdot\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} + \lambda^2\cdot\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}\right] \\ &= e^{-\lambda}\left[\lambda\cdot e^{\lambda} + \lambda^2\cdot e^{\lambda}\right] \\ &= \lambda + \lambda^2. \end{aligned}$$

Sada u jednakost (1.3) uvrstimo prethodno dobiveni rezultat, te slijedi da je

$$Var(X) = \lambda.$$

Navesti ćemo još dva bitna svojstva Poissonove distribucije.
Neka su X i Y nezavisne Poissonove slučajne varijable s parametrom λ_1, odnosno λ_2, $X \sim \mathcal{P}(\lambda_1), Y \sim \mathcal{P}(\lambda_2)$. Tada

- suma nezavisnih Poissonovih slučajnih varijabli je Poissonova slučajna varijabla s parametrom koji je suma parametara

$$p_k = P\left(X + Y = k\right) = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!} \tag{1.5}$$

  Tada pišemo $X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$ (vidi dokaz [9]).

- razlika nezavisnih Poissonovih slučajnih varijabli nije Poissonova slučajna varijabla jer poprima i negativne vrijednosti, odnosno prima vrijednosti iz skupa $\mathbb{Z}$ s vjerojatnostima

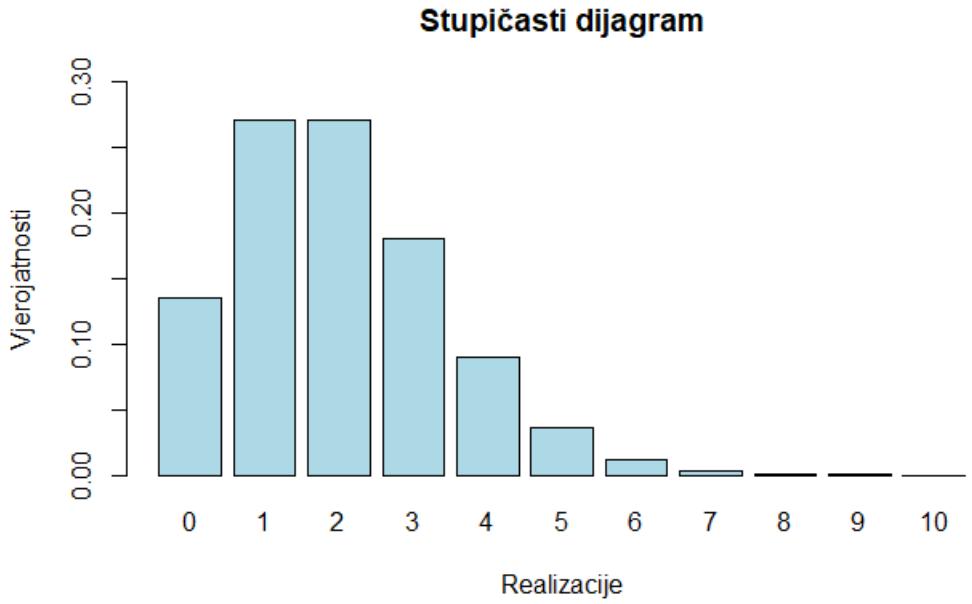
$$p_k = P\left(X - Y = k\right) = e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{k}{2}} I_k(2\sqrt{\lambda_1\lambda_2}), \tag{1.6}$$

gdje je I_k preinačena Besselova funkcija prve vrste,

$$I_k(x) = \left(\frac{x}{2}\right)^k \sum_{i=0}^{\infty} \frac{(x^2/4)^i}{i!(k+i)!},$$

(vidi dokaz [1]). Razlika dviju Poissonovih distribucija naziva se **Skellam distribucija**.

Grafički prikaz Poissonove distribucije izgledati će drugačije za različite vrijednosti λ. Na sljedećoj slici prikazan je stupičasti dijagram distribucije Poissonove slučajne varijable s parametrom 2 za skup realizacija $\{0, 1, 2, \ldots, 10\}$.



Slika 2: Stupičasti dijagram Poissonove distribucije s parametrom 2 za skup realizacija {0, 1, 2,..., 10}

1.3 Procjena parametara

Statističko zaključivanje se temelji na statističkim modelima. Statistički model je familija funkcija distribucije koja se uzima u obzir za zaključivanje o danom problemu. $\mathcal{P}$ je oznaka koju ćemo koristiti za statistički model. Poissonova distribucija

je parametarski zadana diskretna distribucija i svrstavamo je u parametarske familije distribucija, stoga ćemo se u ovom radu baviti parametarskim statističkim modelima. Definicije u ovom poglavlju prate [3]. Zanima nas procjena parametra. U ovom poglavlju parametar ćemo označavati s $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, a prostor dozvoljenih vrijednosti parametra s Θ.

Definicija 1.4. ***Statistički model*** $\mathcal{P}$ *je familija dozvoljenih funkcija distribucije slučajnog vektora za koji baza podataka čini jednu realizaciju.*

Slučajan vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ ćemo zvati **slučajan uzorak**, a njegovu realizaciju $(x_1, \ldots, x_n)$ samo **uzorak**.

Najjednostavniji statistički model je model jednostavnog slučajnog uzorka.

Definicija 1.5. *Statistički model zovemo model jednostavnog slučajnog uzorka iz funkcije distribucije F ako za slučajni vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ čija je realizacija $(x_1, \ldots, x_n)$ vrijedi:*

- *sve slučajne varijable $X_1, \ldots, X_n$ imaju istu funkciju distribucije F,*
- *slučajne varijable $X_1, \ldots, X_n$ su nezavisne.*

Promatrana veličina u ovim modelima je slučajna varijabla sa svojom funkcijom distribucije F.

Definicija 1.6. *Neka je $t : \mathbb{R}^n \to S$ izmjeriva funkcija, gdje se $S \subseteq \mathbb{R}^k$, $\boldsymbol{X} = (X_1, \ldots, X_n)$ slučajan uzorak. Kompoziciju funkcije t i slučajnog uzorka $\boldsymbol{X}$ zovemo* ***statistika*** *i označavamo s $T = t(\boldsymbol{X})$.*

Parametarski model pretpostavlja zaključivanje o parametru, što nas dovodi do problema procjene parametara modela.

Definicija 1.7. *Neka je $\mathcal{P} = \{\mathcal{F}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ parametarski statistički model, gdje je Θ parametarski prostor. Parametar $\boldsymbol{\theta}$ je* ***odrediv*** *ako za svaki*

$$\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \Longrightarrow F_{\boldsymbol{\theta}_1} \neq F_{\boldsymbol{\theta}_2}.$$

Sada kada smo spomenuli osnovne pojmove, vratimo se na Poissonovu distribuciju s parametrom λ. Pokazali smo da je jedinstveno svojstvo Poissonove distribucije da su očekivanje i varijanca jednaki λ. Kako mi želimo modelirati podatke, za

određivanje distribucije iz podataka možemo koristiti pretpostavljeni tip distribucije te procijeniti nepoznate parametre: očekivanje i varijancu. Za procjenu očekivanja slučajne varijable koristiti ćemo aritmetičku sredinu, odnosno **procjenu parametra** Poissonove distribucije procijeniti ćemo **artimetičkom sredinom** uzorka.

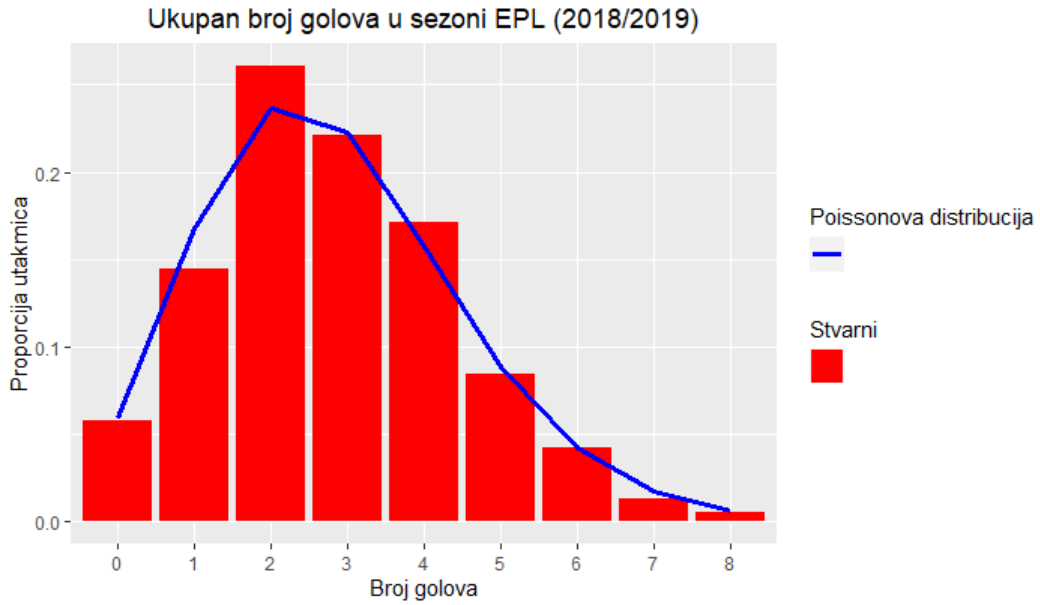
1.4 Modeliranje nogometnih rezultata Poissonovom distribucijom

Uzeti ćemo za primjer nogometnu sezonu Premier Lige (2018/2019). Baza je preuzeta sa football-data.co.uk. U Premier Ligi natječe se 20 klubova. Svaka momčad igra najviše 19 utakmica kod kuće i 19 u gostima, pa je ukupan broj utakmica u sezoni 380. Igra se ukupno 38 kola, 10 utakmica u jednom kolu. Ukupan broj postignutih golova u toj sezoni je bio 1072, dajući prosjek $\frac{1072}{380} = 2.821053$ golova po utakmici.

Broj postignutih golova	0	1	2	3	4	5	6	7	8
Frekvencija	22	55	99	84	65	32	16	5	2

Tablica 1. Ukupan broj golova po utakmici u nogometnoj sezoni EPL (2018/2019)

Na slici (3) lako je uočiti da raspodjela postignutih golova izgleda vrlo slično obliku Poissonove distribucije. Stoga, neka je X skup podataka o ukupnom broju postignutih golova u svih 380 utakmica, onda je $E[X] \approx 2.821053$, a to je zapravo procjena parametra. Na prethodno spomenutoj slici vidimo prikazan stupičasti dijagram stvarnih podataka o ukupnom broju postignutih golova u svakoj utakmici i vidimo Poissonovu distribuciju dobivenu s procijenjenim parametrom 2.821053, gdje se vjerojatnost svakog pojavljivanja izračunava pomoću procijenjenog parametra.



Slika 3: Ukupan broj postignutih golova u sezoni EPL (2018/2019) u usporedbi s Poissonovom distribucijom dobivenom s procijenjenim parametrom 2.821053

Jasno je da ukupan broj golova po utakmici u sezoni EPL (2018/2019) itekako prati Poissonovu distribuciju s očekivanjem jednakim prosječnom broju postignutih golova po utakmici.
Koliko dobro se poklapa s Poissonovom distribucijom provjeriti ćemo pomoću testa "goodness of fit". Kako su Poissonova distribucija i nogometni podaci diskretni, možemo iskoristiti "χ^2 goodness of fit" test, koji testira slijede li promatrani podaci određenu distribuciju. Nul-hipoteza je da distribucija ukupnog broja golova ima Poissonovu distribuciju, dok je alternativna da nema Poissonovu distribuciju. Već smo procijenili da je očekivani broj golova po utakmici 2.821053, te smo u tablici (1) prikazali frekvenciju broja golova po utakmici. Sada ćemo koristeći program R izračunati koliki je očekivani broj golova po utakmici dobiven s parametrom procijenjenim pomoću artimetičke sredine.

broj golova	stvarni	očekivani
0	22	22.626428
1	55	63.830345
2	99	90.034382
3	84	84.663910

4	65	59.710337
5	32	33.689200
6	16	15.839835
7	5	6.383572
8	2	2.251049

Koristeći odabrani test dobili smo p-vrijednost = 0.6590108 > 0.05, pa na razini značajnosti 0.05, ne odbacujemo nul-hipotezu, odnosno možemo tvrditi da ukupan broj golova po utakmici ima približno Poissonovu distribuciju.

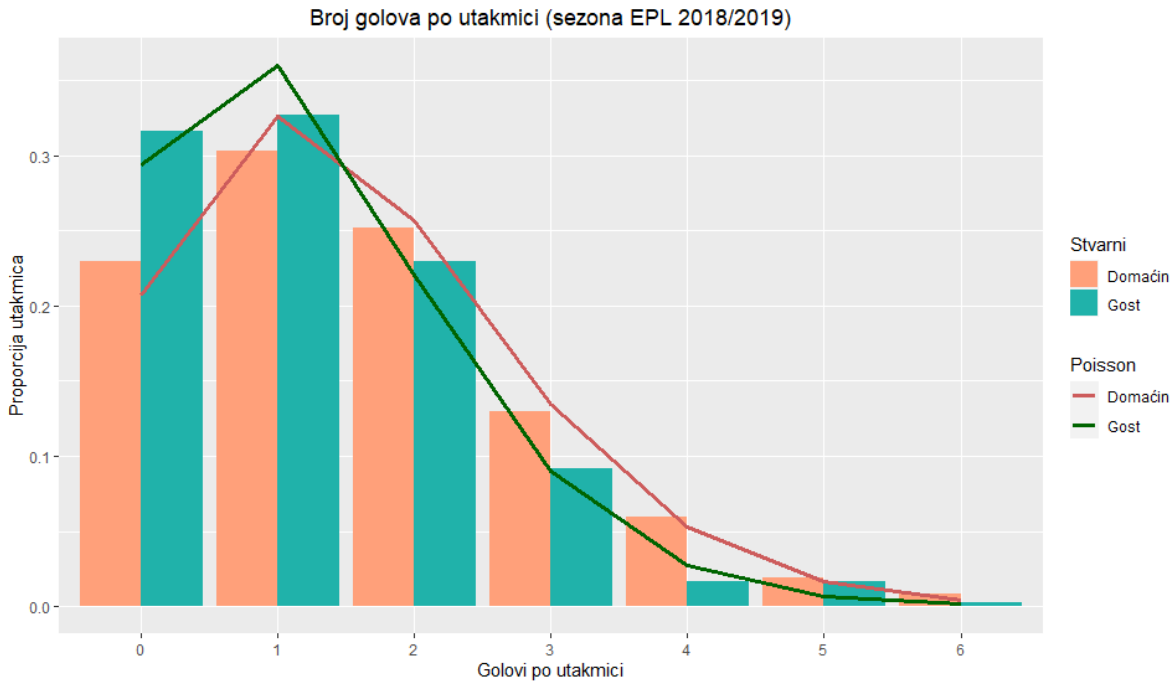
1.5 Razlika u postignutim golovima

Usporediti ćemo broj golova domaćih i gostujućih momčadi, a zatim ćemo uzeti dvije momčadi i analizirati njihove uspjehe kad igraju na domaćem i kad igraju na gostujućem terenu.
Osvrnuti ćemo se na broj golova koje momčadi postignu na domaćem terenu i gostujućem terenu. Igra na domaćem terenu je veliki plus za domaću momčad jer igra u okolini koju poznaje, ali najvažniju prednost daju navijači. Neki statistički modeli sugeriraju da domaća momčad ima 60% šanse za pobjedu. Je li to istina ili mit? Izračunati ćemo prosječan broj golova na domaćem terenu i na gostujućem terenu, te ih usporediti.

Domaćin	Gost
1.575657	1.224324

Tablica 2. Prosječan broj golova domaćih momčadi i prosječan broj golova gostujućih momčadi

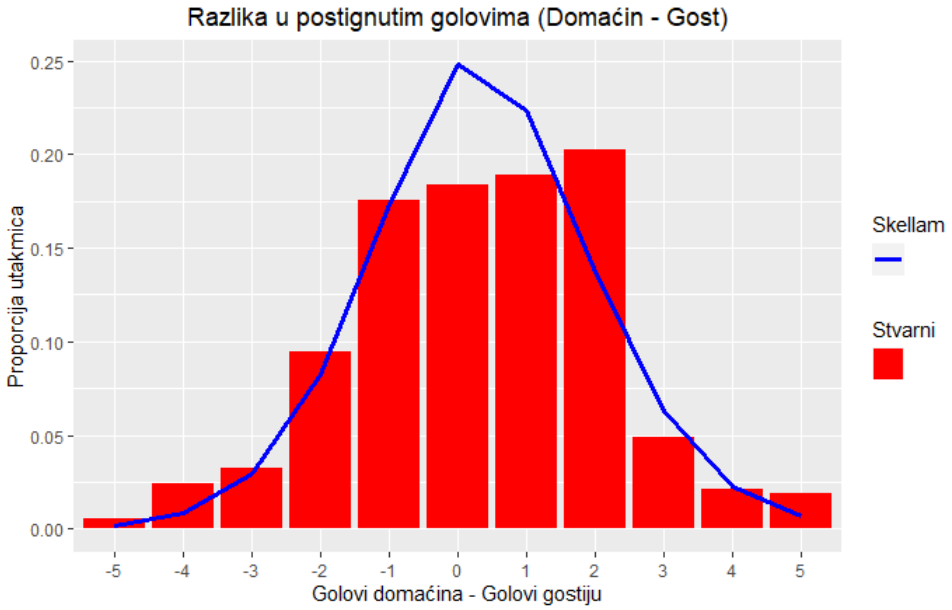
Možemo primjetiti da u prosjeku domaće momčadi postignu više golova nego gostujuće. Kao što smo rekli broj golova možemo modelirati Poissonovom distribucijom s procjenjenim parametrom. Dakle broj golova domaćih i gostujućih momčadi možemo gledati kao dvije Poissonove varijable.
Na sljedećoj slici prikazan je omjer postignutih golova u usporedbi s distribucijom dobivenom s procijenjenim parametrima koji se nalaze u tablici (2).



Slika 4: Broj golova domaćih i gostujućih momčadi po utakmici

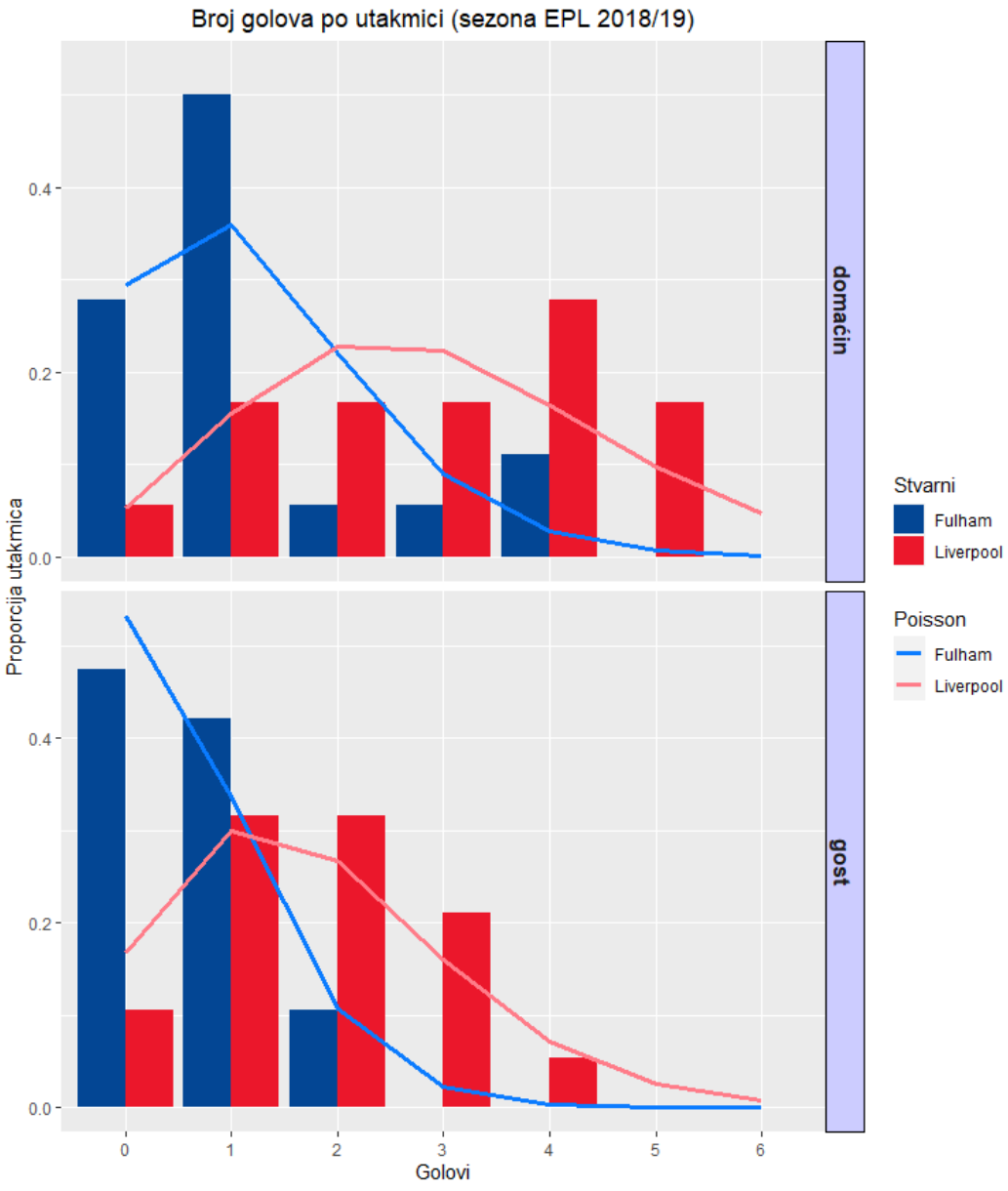
Iz grafa vidimo da gostujuće momčadi češće ne postignu nijedan gol na utakmici, a također primjećujemo da domaćini češće postižu veći broj golova. Također možemo uočiti da broj postignutih golova domaćih i gostujućih momčadi prati Poissonovu distribuciju dobivenu s procijenjenim parametrima. Broj golova koje je postigla svaka momčad smatramo nezavisnim događajima.

Želimo analizirati razliku dviju nezavisnih Poissonovih slučajnih varijabli, iskoristiti ćemo prethodno definiranu Skellam distribuciju. Neka je X slučajna varijabla koja modelira broj golova domaćih momčadi, a Y slučajna varijabla koja modelira broj golova gostujućih momčadi. U Skellam distribuciju uvrstili smo vrijednosti iz tablice (2) te smo izračunali da je vjerojatnost da utakmica završi neriješeno 0.2484069, a vjerojatnost da gostujuća momčad (nije određeno koja momčad) pobjedi s jednim pogotkom razlike 0.1734483. Na sljedećoj slici možemo vidjeti vjerojatnosti ostalih ishoda.



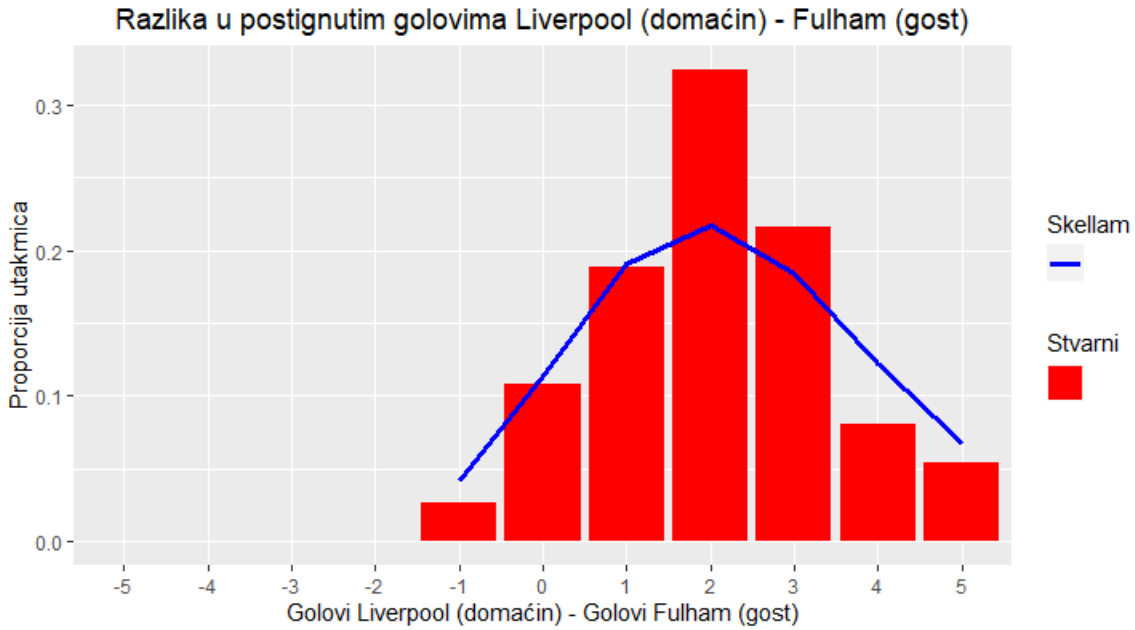
Slika 5: Razlika golova između domaćih i gostujućih momčadi

Sada kada smo analizirali domaće i gostujuće momčadi uzeti ćemo za primjer dvije momčadi te ćemo analizirati njihove rezultate kada su bili u ulogama domaćina i gosta. Pogledajmo raspodjelu golova koje su postigli Liverpool i Fulham, momčadi koje su završile na 2. i pretposljednjem mjestu na kraju sezone. Na sljedećoj slici su prikazani grafovi iz kojih možemo isčitati sljedeće zanimljive činjenice.



Slika 6: Broj golova Fulhama i Liverpoola

Fulham je na domaćem terenu najčešće postizao jedan ili nijedan gol, a u dvije utakmice na domaćem terenu su postigli čak 4 pogotka. Do odigravanja zadnjeg kola Liverpool je samo jednu utakmicu na domaćem terenu završio bez postignutog gola. Što se tiče utakmica u gostima, Fulham je najviše utakmica završio bez postignutog pogotka i s jednim postignutim golom, a u nijednoj utakmici nisu postigli više od 2 gola. Liverpool je imao odlične rezultate i u gostima, najčešće su postizali jedan ili dva pogotka. Ako gledamo svih 37. kola, u samo tri utakmice Liverpool nije postigao nijedan pogodak.
Uzmimo sada konkretan primjer, utakmicu između Liverpoola i Fulhama, gdje je Liverpool domaćin, a Fulham gost. Prosječan broj golova koji Liverpool postiže kao domaćin je 2.894737, dok je prosječan broj golova koji Fulham postiže kao gost je 0.6315789. Sada ćemo te procjene uvrstiti u Skellam distribuciju. Vjerojatnost da utakmica završi neriješeno je 0.1133734, a vjerojatnost da Liverpool (domaćin) pobjedi s dva gola razlike je 0.217757. Na sljedećoj slici možemo vidjeti vjerojatnosti ostalih ishoda.



Slika 7: Razlika golova između Liverpoola i Fulhama

Nerealno je prepostaviti nezavisan broj golova u međusobnom susretu jer kao što vidimo na ovom primjeru Liverpool je momčad koja postiže veliki broj pogodaka,

dok Fulham postiže mali broj. Nije jednako gledati međusobni susret dviju momčadi koje su podjednake ili različite kvalitete.

2 Poissonova regresija

Poissonova regresija pripada generaliziranim linearnim modelima. U ovom poglavlju definirat ćemo što su to generalizirani linearni modeli te reći neka svojstva, a nakon toga ćemo se fokusirati na Poissonovu linearnu regresiju. Definicije i primjeri u ovom poglavlju prate [4] i [7].

2.1 Općenito o generaliziranim linearnim modelima

Neka su dani podaci oblika $(y_i, x_{1i}, \ldots, x_{ki})$ za $i = 1, \ldots, n, k \in \mathbb{N}$, pri čemu je y_i realizacija neke slučajne varijable Y_i čija distribucija ovisi o vrijednostima $x_{1i}, \ldots, x_{ik}$. Odabir modela nije jednostavan, a jedan od najpoznatijih modela koji nam daje rješenje traženog problema je linearna regresija. Neka su uz dane podatke dani vektor koeficijenta $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$ te vektor grešaka $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Linearni regresijski model pretpostavlja da vrijedi

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

Pretpostavke koje slučajna varijabla greške ε_i, za svaki $i = 1, \ldots, n$ mora zadovoljiti su $E[\varepsilon_i] = 0$ i $Var(\varepsilon_i) = \sigma^2$. Linearni regresijski model u matričnoj notaciji možemo zapisati na sljedeći način:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

gdje je $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ vektor, a $\mathbf{X} = (\mathbf{1}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ matrica dimenzije $n \times k+1$, čiji su stupci vektori

$$\mathbf{1} = (1, \ldots, 1)^T, \text{ dimenzije n}, \qquad \boldsymbol{x_i} = (x_{1i}, \ldots, x_{ni})^T, \quad i = 1, \ldots, k. \tag{2.3}$$

Nadalje, pretpostavljamo da vrijedi

$$E[Y_i] = \sum_{j=0}^{k} = \beta_j x_{ji}, \quad Var(Y_i) = \sigma^2, \quad Cov(Y_i, Y_j) = 0, \ i \neq j.$$

Iz čega slijedi

$$\mu = E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{2.4}$$

Linearna kombinacija $x_1, \ldots, x_k$ i parametara $\beta_0, \ldots, \beta_k$ definira parametar η tj.

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.5}$$

Veza između slučajne varijable Y i prediktorskih varijabli $x_1, \ldots, x_k$ je

$$\mu = \eta.$$

Gornji zapis je generalizacija izraza (2.4). Zadnji izraz iz generalizacije možemo zapisati kao

$$\eta = g(\mu)$$

gdje je g funkcija identitete u slučaju linearne regresije. Funkciju g zovemo funkcija veze (eng. *link function*).
Generalizaciju možemo proširiti. Slučajne komponente mogu imati bilo koju razdiobu iz eksponencijalne familije razdioba, a ne samo normalnu razdiobu te vezna funkcija može biti bilo koja monotona diferencijabilna funkcija. Time smo dobili cijelu skupinu generaliziranih linearnih modela (GLM).

Definicija 2.1. *Kažemo da slučajna varijabla Y pripada nekoj eksponencijalnoj familiji ako je njezina fukcija gustoće oblika*

$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{2.6}$$

za neke funkcije a, b i c. Parametar θ zovemo prirodni parametar, dok je ϕ parametar disperzije ili skaliranja.

Lako se pokaže da je

$$E[Y] = \mu = b'(\theta) \tag{2.7}$$

$$Var[Y] = a(\phi)b''(\theta) \tag{2.8}$$

gdje su b' i b'' prva i druga derivacija funkcije b. Vidimo da očekivanje od Y ovisi isključivo od parametru θ, dok varijanca općenito ovisi o oba parametra. Može se pokazati da je funkcija b neprekidna i invertibilna, osim u nekim specijalnim slučajevima pa možemo definirati funkciju varijance

$$V(\mu) = b''(\theta).$$

Sada varijancu od Y možemo pisati i kao

$$Var[Y] = a(\phi)V(\mu).$$

Za GLM imamo da vrijedi

$$g(\mu_i) = g(b'(\theta_i)) = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki}.$$

Iz gornje jednakosti možemo definirati **prirodnu funkciju veze** kao

$$g = (b')^{-1} \tag{2.9}$$

$$\Rightarrow g(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{1i} + \dots \beta_k x_{ki}. \tag{2.10}$$

Primjer 2.1. *Neka je Y slučajna varijabla s normalnom distribucijom s parametrima μ i σ^2. Funkcija gustoće slučajne varijable Y je*

$$\begin{aligned} f_Y(y;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma)\right)\right] \end{aligned} \tag{2.11}$$

Iz (2.6) i (2.11) slijedi da je

$$\theta = \mu, \quad \phi = \sigma^2, \quad a(\phi) = \phi, \quad b(\theta) = \frac{\theta^2}{2},$$
$$c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma)\right).$$

Nadalje, iz izraza (2.7) i (2.8) možemo odrediti očekivanje i varijancu slučajne varijable Y. Slijedi

$$b(\theta) = \frac{\theta^2}{2} \Rightarrow E[Y] = b'(\theta) = \theta = \mu$$
$$a(\phi) = \phi \Rightarrow Var[Y] = a(\phi)b''(\theta) = \phi = \sigma^2$$

što smo već znali.

2.2 Poissonova linearna regresija

Sada ćemo se fokusirati na jedan GLM, a to je Poissonova regresija. Komponente Y_i zavisne varijable Y su međusobno nezavisne i imaju Poissonovu distribuciju, koja pripada eksponencijalnoj familiji. Prema (2.5) moramo još pronaći funkciju veze i time ćemo točno odrediti model Poissonove regresije. Već smo ranije iskazali definiciju Poissonove distribucije, izraz (1.2) možemo zapisati i na sljedeći način

$$f_Y(y) = \exp(y\log\mu - \mu - \log y!) \tag{2.12}$$

Iz (2.6) i (2.12) slijedi da je

$$\theta = \log\mu, \quad \phi = 1, \quad a(\phi) = 1, \quad b(\theta) = e^\theta,$$
$$c(y,\phi) = -\log y!$$

Stoga vrijedi sljedeće:

- prirodni parametar Poissonove distribucije je $\log \mu$
- iz (2.7) slijedi da je $E[Y] = \mu$
- iz (2.8) slijedi da je $Var[Y] = \mu$
- iz (2.9) slijedi da je prirodna funkcija veze $g(\mu) = \log \mu$, zbog $\mu = g^{-1}(\eta)$

$$\Rightarrow \mu = e^{\eta} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

Uzimajući u obzir prethodne tvrdnje, model Poissonove regresije možemo opisati na sljedeći način. Neka su $y_1, \ldots, y_n$ gdje je $n \in \mathbb{N}$ opservacije koje su realizacije nezavisnih slučajnih varijabli $Y_1, \ldots, Y_n$ takvih da vrijedi $Y_i \sim \mathcal{P}(\mu_i)$. Neka su $x_1, \ldots, x_k$ prediktorske varijable te $\beta_0, \beta_1, \ldots, \beta_k \in \mathbb{R}, \forall k \in \mathbb{N}$ te neka je

$$\eta = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

gdje je X matrica dizajna definirana kao u (2.3). Slučajna varijabla Y je s kovarijantama $x_1, \ldots, x_k$ povezana na sljedeći način

$$\mu = e^{\eta} = e^{\mathbf{X}\boldsymbol{\beta}} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \tag{2.13}$$

gdje je $\mu = (\mu_1, \ldots, \mu_n)$. Preostaje nam još procijeniti koeficijente $\beta_0, \beta_1, \ldots, \beta_k$, a to činimo metodom maksimalne vjerodostojnosti (eng. *maximum likelihood estimation*).
Vrijedi da je

$$f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\beta}) = \frac{\mu^y}{y!} e^{-\mu} = \frac{e^{y\beta x}}{y!} e^{-e^{\beta x}} = \prod_{i=1}^{n} \frac{e^{y_i \beta x_i}}{y_i!} e^{-e^{\beta x_i}}. \tag{2.14}$$

Sada želimo pronaći vektor koeficijenata $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ tako da je gornja vjerojatnost maksimalna.

Definicija 2.2. *Za danu realizaciju $\boldsymbol{x} = (x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$ s gustoćom $f(\boldsymbol{x}; \boldsymbol{\theta})$ gdje je $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, n, $k \in \mathbb{N}$. Funkcija vjerodostojnosti parametra $\boldsymbol{\theta}$ je*

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}). \tag{2.15}$$

Iz prethodne definicije i (2.14) slijedi da je funkcija parametra β jednaka

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{y_i \beta x_i}}{y_i!} e^{-e^{\beta x_i}}. \tag{2.16}$$

Prethodnu jednakost možemo logaritmirati pa dobivamo log-vjerodostojnost

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \beta x_i - e^{\beta x_i} - \log(y_i!)). \tag{2.17}$$

Funkcija $f(x) = \log x$ je monotona te će se zbog toga maksimum vjerodostojnosti postići u istim točkama kao i maksimum log-vjerodostojnosti. Iz sume možemo izbaciti $-\log(y_i!)$ jer ne sadrži β pa neće utjecati na rezultat. Stoga, možemo promatrati

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \beta x_i - e^{\beta x_i}). \tag{2.18}$$

Da bismo pronašli maksimum još moramo riješiti sustav jednadžbi

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0,$$

i time ćemo dobiti procijenjene parametre Poissonove regresije.
Sada kada smo definirali generalizirane linearne modele i Poissonovu regresiju možemo se vratiti na stvarne podatke i izgraditi model.

3 Poissonov model

Sljedeći korak je primjena Poissonovog regresijskog modela na stvarne nogometne podatke. Za izradu modela uzeti ćemo nogometnu sezonu Premier Lige (2018/2019) koju smo koristili i prethodnim poglavljima, a analizu podataka raditi ćemo u programu RStudio. Kao što smo već rekli, u ligi se natječe 20 klubova, ukupan broj utakmica je 380, svaka momčad igra najviše 19 utakmica kod kuće i 19 u gostima. Iz baze smo izbacili posljednje kolo jer ćemo njega predvidjeti i usporediti sa stvarnim podacima.

3.1 Osnovni Poissonov model

U ovom dijelu formulirati ćemo osnovni Poissonov model u matematičkim terminima koji se još naziva Maherov model. Kao što smo već naveli, ključna pretpostavka modela je da broj postignutih golova domaćina i gostujuće momčadi u svakoj

utakmici su zapravo dvije nezavisne Poissonove varijable, čije je očekivanje određeno napadačkim i obrambenim kvalitetama svake strane.
Neka je sa i označena domaća momčad, a sa j gostujuća momčad, te neka je $X_{i,j}$ i $Y_{j,i}$ broj postignutih golova domaćina odnosno gostiju. Tada pretpostavljamo da je

$$X_{i,j} \sim \mathcal{P}(\alpha_i \beta_j \gamma), \tag{3.1}$$
$$Y_{j,i} \sim \mathcal{P}(\alpha_j \beta_i), \tag{3.2}$$

gdje su $X_{i,j}$ i $Y_{j,i}$ nezavisne slučajne varijable, $\alpha_i, \beta_i > 0$, $\forall i$. α_i je mjerilo snage napada domaće momčadi, β_j je mjerilo slabosti obrane gostujuće momčadi, dok γ predstavlja faktor domaćeg terena.
Model poprima sljedeći izgled:

$$P(X_{i,j} = x, Y_{j,i} = y) = \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\mu}\mu^y}{y!} \tag{3.3}$$

gdje je $\lambda = \alpha_i \beta_j \gamma$, a $\mu = \alpha_j \beta_i$. Sada trebamo procijeniti α, β i γ za svaki tim. Umjesto da svaku utakmicu gledamo zasebno, izgraditi ćemo općenitiji Poissonov linearni regresijski model.

3.2 Izgradnja modela

Za izgradnju modela pratili smo blog *Predicting Football Results With Statistical Modelling* [12].
Model za predviđanje vjerojatnosti rezultata u nogometnoj utakmici koristi samo golove postignute u prethodnim utakmicama. Povijesni rezultati se smatraju mjerom napadačke i obrambene kvalitete timova, jer se momčad koja postigne puno golova smatra napadački jakom, a momčad koja primi puno golova smatra se obrambeno slabom. Za model ćemo koristiti samo 4 varijable, a to su 'home - domaćin', 'away - gost', 'homeGoals - golovi domaćina' i 'awayGoals - golovi gostiju'.

```
        home             away         homeGoals      awayGoals
1     Man United      Leicester             2              1
2    Bournemouth        Cardiff             2              0
3         Fulham Crystal Palace             0              2
4   Huddersfield        Chelsea             0              3

⋮
```

366	Wolves	Fulham	1	0
367	Arsenal	Brighton	1	1
368	Chelsea	Watford	3	0
369	Huddersfield	Man United	1	1
370	Man City	Leicester	1	0

Tablica prikazuje bazu podataka do posljednjeg kola.
Vidljivo je da je su momčadi kategorijalne varijable pa uvodimo indikator varijable. Neka je $teams = (\text{"}Arsenal\text{"}, \text{"}Bournemouth\text{"}, \ldots, \text{"}West\ Ham\text{"}, \text{"}Wolves\text{"})$ vektor svih momčadi. Tada generalizirani linearni model s Poissonovom funkcijom veze možemo zapisati kao:

$$\mu = m(team, opponent, home) = \exp(\beta_0 + \sum_{i=1}^{20} \beta_i \, I_{\{team=teams_i\}} + \sum_{i=1}^{20} \beta_{20+i} \, I_{\{opponent=teams_i\}} + \beta_{41} \, home),$$

gdje je za svaki $i = 1, \ldots, 20$,

$$I_{\{team\,=\,teams_i\}} = \begin{cases} 1, & \text{ako je } team \text{ jednak } teams_i \\ 0, & \text{inače} \end{cases}$$

$$I_{\{opponent\,=\,teams_i\}} = \begin{cases} 1, & \text{ako je } opponent \text{ jednak } teams_i \\ 0, & \text{inače} \end{cases}$$

$$home = \begin{cases} 1, & \text{ako je } team \text{ domaćin} \\ 0, & \text{inače} \end{cases}$$

Primjerice, za utakmicu "Arsenal - Wolves" gde je Arsenal domaćin, broj golova momčadi "Arsenal" procjenjujemo s $m(\text{"}Arsenal\text{"}, \text{"}Wolves\text{"}, 1) = \exp(\beta_0 + \beta_1 + \beta_{40} + \beta_{41})$, dok broj golova ekipe "Wolves" procjenjujemo s $m(\text{"}Wolves\text{"}, \text{"}Arsenal\text{"}, 0) = \exp(\beta_0 + \beta_1 + \beta_{40})$.
Za procjenu parametara $\beta_0, \beta_1, \ldots, \beta_{41}$ koristili smo sljedeći algoritam u R-u te smo dobili sljedeće rezultate.

```
poisson model <- rbind(
```

```
    data.frame(goals=epl_data$homeGoals,
               team=epl_data$home,
               opponent=epl_data$away,
               home=1),
    data.frame(goals=epl_data$awayGoals,
               team=epl_data$away,
               opponent=epl_data$home,
               home=0))
  glm(goals ~ home + team +opponent, family=poisson(link=log),data=.)
summary(poisson_model)

 Call:
glm(formula = goals ~ home + team + opponent, family = poisson(link = log),
    data = .)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.17511  -1.02684  -0.05616   0.48462   2.98913

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.49255    0.19153   2.572 0.010120 *
home                  0.25265    0.06268   4.031 5.56e-05 ***
teamBournemouth      -0.27027    0.18251  -1.481 0.138637
teamBrighton         -0.73344    0.20938  -3.503 0.000460 ***
teamBurnley          -0.44926    0.19248  -2.334 0.019596 *
teamCardiff          -0.77017    0.21376  -3.603 0.000315 ***
teamChelsea          -0.12488    0.17402  -0.718 0.472976
teamCrystal Palace   -0.41252    0.19021  -2.169 0.030099 *
teamEverton          -0.31651    0.18342  -1.726 0.084416 .
teamFulham           -0.69864    0.20946  -3.335 0.000852 ***
teamHuddersfield     -1.18201    0.24914  -4.744 2.09e-06 ***
teamLeicester        -0.32612    0.18448  -1.768 0.077101 .
teamLiverpool         0.18490    0.16091   1.149 0.250511
teamMan City          0.22907    0.15931   1.438 0.150462
teamMan United       -0.06212    0.17271  -0.360 0.719084
```

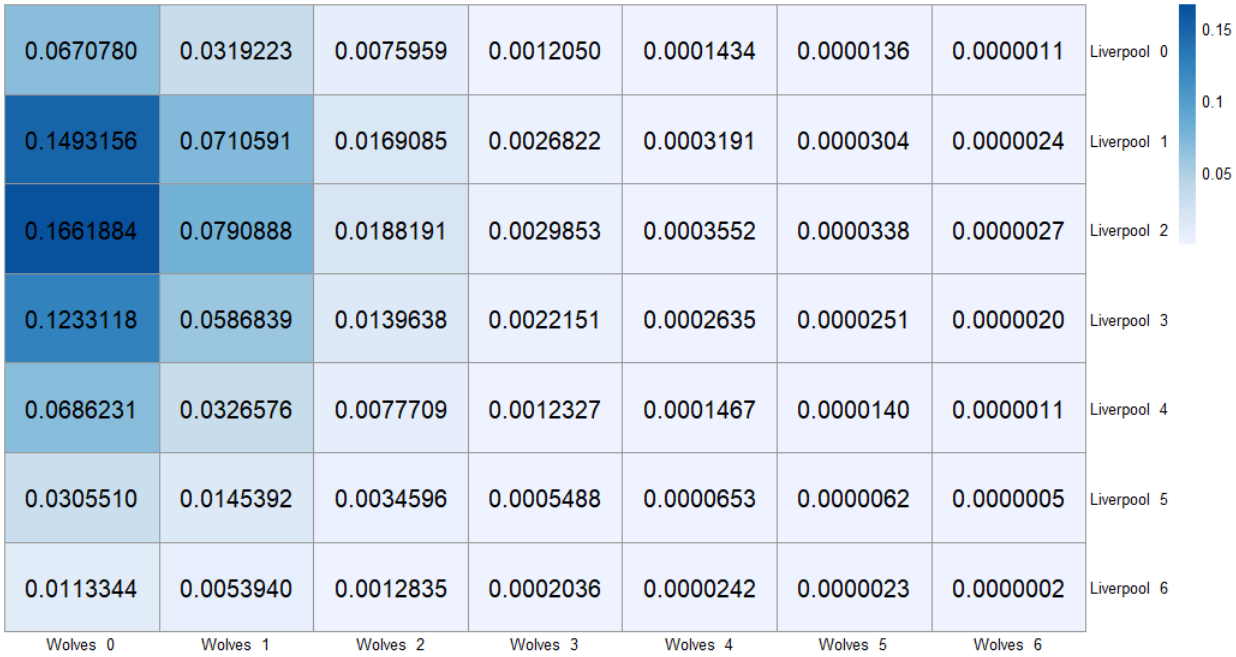
```
teamNewcastle          -0.61039    0.20184  -3.024 0.002493 **
teamSouthampton        -0.43869    0.19284  -2.275 0.022912 *
teamTottenham          -0.09173    0.17263  -0.531 0.595136
teamWatford            -0.31059    0.18452  -1.683 0.092338 .
teamWest Ham           -0.37878    0.18778  -2.017 0.043679 *
teamWolves             -0.42390    0.18891  -2.244 0.024834 *
opponentBournemouth     0.24578    0.18857   1.303 0.192441
opponentBrighton        0.08898    0.19497   0.456 0.648101
opponentBurnley         0.24133    0.18821   1.282 0.199748
opponentCardiff         0.29496    0.18614   1.585 0.113057
opponentChelsea        -0.25371    0.21407  -1.185 0.235938
opponentCrystal Palace -0.02685    0.20040  -0.134 0.893410
opponentEverton        -0.13655    0.20713  -0.659 0.509733
opponentFulham          0.38609    0.18201   2.121 0.033901 *
opponentHuddersfield    0.35513    0.18294   1.941 0.052224 .
opponentLeicester      -0.05948    0.20249  -0.294 0.768947
opponentLiverpool      -0.81119    0.25624  -3.166 0.001547 **
opponentMan City       -0.80828    0.25625  -3.154 0.001609 **
opponentMan United      0.02311    0.19852   0.116 0.907343
opponentNewcastle      -0.08018    0.20244  -0.396 0.692067
opponentSouthampton     0.20434    0.18914   1.080 0.279984
opponentTottenham      -0.31066    0.21727  -1.430 0.152772
opponentWatford         0.07114    0.19582   0.363 0.716397
opponentWest Ham        0.05754    0.19669   0.293 0.769854
opponentWolves         -0.12989    0.20714  -0.627 0.530607
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 942.63  on 739  degrees of freedom
Residual deviance: 732.30  on 700  degrees of freedom
AIC: 2140.2

Number of Fisher Scoring iterations: 5
```

Zanimaju nas vrijednosti u stupcu 'Estimate' (procjena).
Uzeti ćemo eksponencijalnu vrijednost parametra. Pozitivna vrijednost implicira veći broj golova ($e^x > 1, \forall x > 0$), dok vrijednosti bliže nuli predstavljaju neutralni efekat ($e^0 = 1$). Ako je eksponencijalna vrijednost veća od 1, to znači da se parametar Poissonove distribucije povećava (množimo ga s nečim većim od 1) u suprotnom se smanjuje.
Prvo što smo uočili u tablici je procjena 0.25265 za 'home'. 'home' je zapravo prednost domaćeg terena, što znači da domaće momčadi općenito postignu veći broj golova nego gostujuće, a to smo već ranije spomenuli. Također, možemo primijetiti da nisu sve ekipe ravnopravne. Primjerice napad Liverpoola je 0.18490, dok odgovarajuća vrijednost za Fulham iznosi -0.69864. Što bismo interpretirali ovako, nogometaši Liverpoola su bolji strijelci od prosjeka, a nogometaši Fulhama lošiji.
Nadalje, prefiks 'opponent' predstavlja protivnika, te njegova vrijednost ovisi o kvaliteti obrane svake momčadi. Primjerice obrana Liverpoola iznosi -0.81119, a Fulhama 0.38609, drugim riječima, vjerojatnije je da ćete postići gol protiv Fulhama.
Zaključujemo da model ima statistički i intuitivni smisao te možemo započeti s predviđanjem posljednjeg kola.

3.3 Predviđanje posljednjeg kola

U ovom dijelu predviđati ćemo vjerojatnosti ishoda posljednjeg kola. Prvo ćemo uzeti direktan primjer, utakmicu između Liverpoola i Wolverhampton Wanderers F.C. koji su poznati pod nazivom Wolves, te ćemo izračunati vjerojatnosti mogućih ishoda. Zatim ćemo na analogan način, izračunati vjerojatnosti ishoda za ostale utakmice i usporediti vjerojatnosti dobivene pomoću modela s vjerojatnostima koje je nudila kladionica Bet365.
U model proslijedimo momčadi koje želimo usporediti. Za svaku momčad zasebno izračunavamo očekivani prosječni broj golova. Pogledajmo koliki je očekivani broj golova na utakmici između Liverpoola i Wolvesa, gdje je Liverpool domaćin, a Wolves gost. Očekujemo da će Liverpool postići 2.226001 pogodaka kod kuće, s druge strane, očekujemo da će Wolves u gostima postići 0.475899 pogodaka kada igra protiv Liverpoola.
Kao ranije, imamo dvije Poissonove distribucije, te pomoću njih možemo izračunati vjerojatnosti različitih događaja. Simulirati ćemo utakmice, uzeti ćemo da je mak-

simalan broj golova koje momčadi mogu postići 6, jer je to bio maksimalan broj golova koji je neka momčad postigla tijekom sezone u jednoj utakmici. Matrica prikazana na slici (8) pokazuje vjerojatnosti da će Liverpool (reci matrice) i Wolves (stupci matrice) postići određeni broj golova.



Slika 8: Matrica vjerojatnosti

Duž dijagonale oba tima postižu isti broj golova(npr. vjerojatnost da oba tima postignu jedan gol jednaka je 0.0670780)). Stoga, možemo izračunati izgled neriješenog rezultata tako da sumiramo sve vjerojatnosti na dijagonali. Sve ispod dijagonale predstavlja pobjedu Liverpoola, dok sve iznad predstavlja pobjedu Wolvesa. Možemo uočiti da je za rezultat 2-0 (Liverpool-Wolves) vjerojatnost 0.1661884 najveća što odgovara očekivanom broju golova na utakmici između Liverpoola i Wolvesa.
Sljedeća tablica prikazuje vjerojatnosti mogućih ishoda.

Utakmica	**1**	**x**	**2**
Liverpool - Wolves	0.7682426	0.1593244	0.06450702

1 - pobjeda domaćina, x - neriješeno, 2 - pobjeda gostiju

Tablica 3. Vjerojatnosti ishoda utakmice između Liverpoola i Wolvesa

Kao što vidimo naš model daje Wolvesu 6.45% šanse za pobjedu. Da bismo provjerili točnost predviđanja, usporeditit ćemo vjerojatnosti koje je vratio naš model s vjerojatnostima koje nudi kladionica Bet365.
U kladionicama su vjerojatnosti prikazane kao koeficijenti, a ne kao postotci. Koeficijent predstavlja nivo vjerojatnosti da će se nešto dogoditi ili da se nešto neće dogoditi. Evo primjera kako se od koeficijenata izračunavaju vjerojatnosti:

- Koeficijent 10.00 predstavlja $\frac{1}{10} = 10\%$ šanse da će se nešto dogoditi,
- Koeficijent 5.00 predstavlja $\frac{1}{5} = 20\%$ šanse da će se nešto dogoditi, dok
- Koeficijent 1.25 predstavlja $\frac{1}{1.25} = 80\%$ šanse da će se nešto dogoditi,

Na sljedećoj tablici prikazani su koeficijenti koje je ponudila kladionica Bet365 za posljednje kolo EPL 2018/2019.

Utakmica	**1**	**x**	**2**
Brighton - Man City	19.00	8.50	1.16
Burnley - Arsenal	3.25	3.80	2.20
Crystal Palace - Bournemouth	1.90	4.20	3.80
Fulham - Newcastle	2.50	3.60	2.90
Leicester - Chelsea	2.40	3.75	2.90
Liverpool - Wolves	1.30	6.00	11.00
Man United - Cardiff	1.28	6.50	11.00
Southampton - Huddersfield	1.44	4.75	8.50
Tottenham - Everton	2.20	3.50	3.50
Watford - West Ham	2.25	3.75	3.20

1 - pobjeda domaćina, x - neriješeno, 2 - pobjeda gostiju

Tablica 4. Decimalni koeficijenti (Bet365)

Uzmimo direktan primjer utakmicu Manchester United - Cardiff, koeficijent 1.28 predstavlja $\frac{1}{1.28} = 78.125\%$ šanse da će Manchester United pobijediti, koeficijent 6.5 predstavlja 15.38% šanse da će utakmica završiti neriješeno. Šanse koje je vratio naš model usporediti ćemo s koeficijentima iz prethodne tablice. Koeficijente ćemo pretvoriti u vjerojatnosti.

Utakmica (stvarni rezultat)		1	x	2
Brighton - Man City (1-4)	Bet365	0.053	0.118	0.862
	Predviđeno	0.059	0.155	0.778
	Razlika	0.006	0.037	0.084
Burnley – Arsenal (1-3)	Bet365	0.308	0.263	0.455
	Predviđeno	0.244	0.211	0.539
	Razlika	0.064	0.052	0.084
Crystal Palace – Bournemouth (5-3)	Bet365	0.526	0.238	0.263
	Predviđeno	0.504	0.233	0.260
	Razlika	0.022	0.005	0.003
Fulham – Newcastle (0-4)	Bet365	0.400	0.278	0.345
	Predviđeno	0.276	0.280	0.444
	Razlika	0.124	0.002	0.099
Leicester – Chelsea (0-0)	Bet365	0.417	0.267	0.345
	Predviđeno	0.324	0.266	0.409
	Razlika	0.093	0.001	0.064
Liverpool – Wolves (2-0)	Bet365	0.769	0.167	0.091
	Predviđeno	0.768	0.159	0.065
	Razlika	0.001	0.008	0.026
Man United – Cardiff (0-2)	Bet365	0.781	0.154	0.091
	Predviđeno	0.754	0.142	0.085
	Razlika	0.027	0.012	0.006
Southampton – Huddersfield (1-1)	Bet365	0.694	0.211	0.118
	Predviđeno	0.685	0.202	0.109
	Razlika	0.009	0.009	0.009
Tottenham – Everton (2-2)	Bet365	0.455	0.286	0.286
	Predviđeno	0.562	0.242	0.195
	Razlika	0.107	0.044	0.091
Watford - West Ham (1-4)	Bet365	0.444	0.267	0.313
	Predviđeno	0.473	0.244	0.281
	Razlika	0.029	0.023	0.032

1 - pobjeda domaćina, x – neriješeno, 2 - pobjeda gostiju

Slika 9: Apsolutna razlika vjerojatnosti između Bet365 i modela

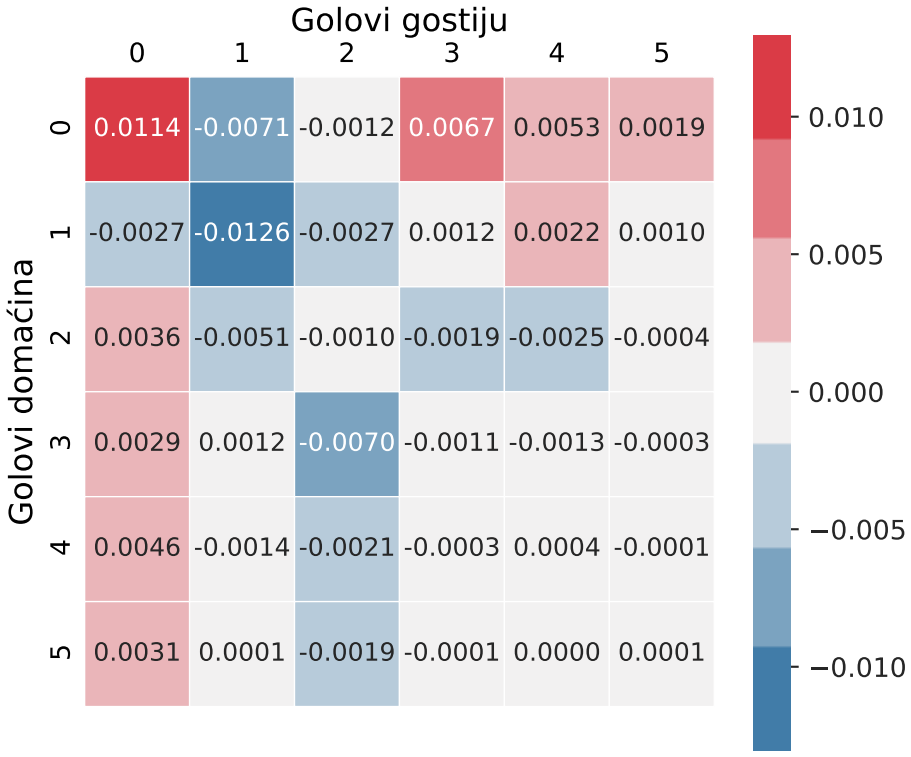
Razlika između Bet365 i našeg modela prikazana je u aposlutnom smislu. Plave ćelije prikazuju da naš model daje veću vjerojatnost određenom ishodu od Bet365. Dok crvene ćelije prikazuju da je Bet365 dao veću vjerojatnost. Možemo uočiti da prevladavaju slabe nijanse ćelija što ukazuje da ne postoje velike razlike između vjerojatnosti koje je ponudio Bet365 i koje smo dobili pomoću modela.
Samo su dvije ćelije obojane jakim bojama, a one impliciraju da je naš model pogrešan ili da su pogrešne vjerojatnosti koje je ponudio Bet365. Ovakve razlike mogu nastati kad kladionice odluče korigirati pojedine koeficijente uzimajući u obzir neke druge čimbenike. Primjerice, nekoj momčadi treba pobjeda da bi se kvalificirala u Ligu Prvaka, znamo da će ta momčad iskoristiti zadnje atome snage da pobjedi, dok protivničkoj momčadi pobjeda ne donosi ništa. U takvim situacijama kladionice ponude jako mali koeficijent na pobjedu momčadi kojoj je ta pobjeda iznimno važna, a jako veliki koeficijent na pobjedu momčadi kojoj ta pobjeda nije nužno potrebna. Takvi veliki koeficijenti privlače kladioničare koji su spremni na njih staviti ulog i riskirati, te je to jedan od načina pomoću kojih kladionice dobro zarađuju.
Osnovni Poissonov model koji je Maher objavio 1982. (vidi [10]) predstavlja dobro polazište na kojeg možemo dodati značajke koje približe odražavaju stvarnost, što nas dovodi do Dixon - Coles modela.

4 Dixon - Coles model

Mark Dixon i Stuart Coles primijetili su nedostatke Maherovog modela koje su željeli popraviti te su predložili sljedeća poboljšanja. Prvo je uvođenje parametra koji bi ispravio podcijenjenu frekvenciju utakmica s malim brojem golova jer je osnovni Poissonov model predviđao rezultate (0-0,1-0,0-1,1-1) rjeđe nego što je trebao. U [6] su prikazali analize koje podupiru njihovu tezu. Drugo poboljšanje je primjena vremenske komponente kako bi utakmice koje su posljednje odigrane imale veću važnost, čime bi se modelirala kratoročna forma momčadi.

4.1 Dodavanje parametra ovisnosti

Da bismo mogli uočiti gdje model najviše griješi treba nam jako puno podataka, pa ćemo uzeti podatke od nogometne sezone Premier Lige 2006/2007 pa sve do sezone 2018/2019 (podaci su preuzeti s football-data.co.uk). Na sljedećoj slici nalazi se matrica koja prikazuje prosječnu razliku vjerojatnosti između stvarnih rezultata i rezultata predviđenih pomoću modela.
Crvene ćelije ukazuju da je model podcijenio rezultate, dok plave ukazuju da je precijenio što znači da model rekao da će biti puno češći rezultat (1-1) nego što je zapravo bio.
Vidimo da postoji problem kod neriješenog rezultata (ćelije za rezultate (0-0) i (1-1) obojane su jakim nijansama plave i crvene boje), dok je kod rezultata (0-1) i (1-0) problem manje očit, ali postoji.



Slika 10: Razlika vjerojatnosti između stvarnih rezultata i predviđenih rezultata

Prema [6] broj postignutih golova dvije momčadi nije potpuno neovisan. Utvrđena je ovisnost između golova domaće i gostujuće momčadi za utakmice s malim brojem pogodaka. Stoga su Dixon i Coles napravili korekciju osnovnog Poissonog modela.. Neka vrijede oznake kao u poglavlju (3.1). Tada Dixon-Coles (DC) model poprima sljedeći izgled:

$$P\left(X_{i,j} = x, Y_{j,i} = y\right) = \tau_{\lambda,\mu}(x,y)\frac{e^{-\lambda}\lambda^{x}}{x!}\frac{e^{-\mu}\mu^{y}}{y!}, \tag{4.1}$$

gdje je $\lambda = \alpha_i \beta_j \gamma,\ \mu = \alpha_j \beta_i$, te

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1 - \lambda\mu\rho, & \text{ako je } x = y = 0, \\ 1 + \lambda\rho, & \text{ako je } x = 0,\ y = 1, \\ 1 + \mu\rho, & \text{ako je } x = 1,\ y = 0, \\ 1 - \rho, & \text{ako je } x = y = 1, \\ 1, & \text{inače.} \end{cases}$$

Ključna razlika u odnosu na osnovni Poissonov model je uvođenje funkcije τ koja ovisi o parametru ρ koji kontrolira jačinu korekcije modela. ρ je parametar ovisnosti. Primjetimo, da ako je $\rho = 0$ onda je DC model jednak osnovnom Poissonovom modelu.
Kada je $x \leq 1$ i $y \leq 1$ parametar ovisnosti mijenja vrijednost τ i povećava ili smanjuje vjerojatnost rezultata te više ne vrijedi pretpostavka nezavisnosti između pogodaka.

4.1.1 Procjena parametara

Da bi se stvorile procjene za danu utakmicu, moraju se odrediti parametri.
S ukupno n timova, postoje parametri napada $\alpha_1, \ldots, \alpha_n$ i parametri obrane $\beta_1, \ldots, \beta_n$, te parametar domaćeg terena γ i parametar ovisnosti ρ koje treba procijeniti. Da model ne bi bio preparametriziran, postavljen je uvjet

$$\frac{1}{n} \sum_{i=1}^{n} \alpha_i = 1. \tag{4.2}$$

U Premier Ligi postoji 20 momčadi, što daje ukupno 42 parametra koje treba procijeniti.
Osnovni alat zaključivanja je funkcija vjerodostojnosti. Za utakmice označene s indeksom $k = 1, \ldots, N$, s odgovarajućim rezultatima (x_k, y_k) funkcija vjerodostojnosti ima sljedeći izgled:

$$L(\alpha_i, \beta_i, \rho, \gamma, i = 1, \ldots, n) = \prod_{k=1}^{N} \tau_{\lambda_k,\mu_k}(x_k, y_k)\, e^{-\lambda_k} {\lambda_k}^{x_k}\, e^{-\mu_k} {\mu_k}^{y_k} \tag{4.3}$$

gdje je

$$\lambda_k = \alpha_{i(k)} \beta_{j(k)} \gamma,$$
$$\mu_k = \alpha_{j(k)} \beta_{i(k)}.$$

$i(k)$ odnosno $j(k)$ označavaju indekse domaće i gostujuće momčadi na utakmici k. Da bismo pronašli parametre zapravo smo maksimizirali funkciju log-vjerodostojnosti. Iskoristili smo uvjet (4.2) kako bismo dobili jedinstveno rješenje i sljedeću tablicu parametara.

Klub	α	β
Arsenal	1.3602	-0.8579
Bournemouth	1.1145	-0.5551
Brighton	0.6323	-0.7333
Burnley	0.8971	-0.5924
Cardiff	0.6147	-0.5911
Chelsea	1.2055	-1.1300
Crystal Palace	1.0066	-0.8369
Everton	1.0542	-0.9774
Fulham	0.6247	-0.4328
Huddersfield	0.1805	-0.5072
Leicester	0.9964	-0.9398
Liverpool	1.5338	-1.6793
Man City	1.5987	-1.6379
Man United	1.2511	-0.8067
Newcastle	0.8064	-0.9460
Southampton	0.8966	-0.6365
Tottenham	1.2590	-1.1369
Watford	1.0306	-0.7294
West Ham	1.0234	-0.8031
Wolves	0.9139	-0.9884
ρ	-0.0410	
γ	0.2246	

Slika 11: Parametri koji maksimiziraju funkciju log-vjerodostojnosti za sezonu Premier Lige 2018/2019

Nakon što smo procijenili parametre napraviti ćemo predviđanje ishoda za konkretnu utakmicu između Liverpoola i Wolvesa, gdje je Liverpool domaćin. Usporediti ćemo vjerojatnosti ishoda dobivene pomoću osnovnog Poissonovog modela i DC modela.

Liverpool - Wolves	1	x	2
Osnovni Poissonov model	0.76824	0.15932	0.06451
Dixon - Coles model	0.76504	0.17165	0.06329

1 - pobjeda domaćina, x - neriješeno, 2 - pobjeda gostiju

Tablica 5. Vjerojatnosti ishoda utakmice između Liverpoola i Wolvesa

Možemo primjetiti da DC model daje veću vjerojatnost neriješenom rezultatu, dok je za ostala dva ishoda, razlika u dobivenim vjerojatnostima jako mala.
U ovom dijelu procjena parametara je bila puno zahtjevnija, nego li u prethodnom poglavlju, a nismo dobili puno bolja predviđanja. Stoga ćemo se sada osvrnuti i na drugo poboljšanje osnovnog Poissonov modela, a to je primjena vremenske komponente.

4.2 Dodavanje vremenske komponente

U nogometu se momčadi s vremenom mijenjaju. Momčad može biti u odličnoj formi i postići niz pobjeda, a s druge strane može biti u lošoj i nakupiti niz poraza. Puno faktora utječe na kvalitetu momčadi, ali glavni faktor je trenutna forma. Kada smo u DC modelu uveli ovisni parametar ρ nismo uzeli u obzir prethodnu činjenicu, te smo rekli da su tijekom sezone sve utakmice jednako važne. Stoga, sada pretpostavljamo da su parametri lokalno konstantni tijekom vremena (ne potpuno), da su povijesni rezultati utakmica manje značajni nego nedavni rezultati i određujemo procjene parametara za svaki trenutak t, koje se temelje na povijesnim rezultatima utakmica do trenutka t.
Izmjenjujemo jednadžbu (4.3), uvodimo faktor koji smanjuje važnost starijih utakmica:

$$L(\alpha_i, \beta_i, \rho, \gamma, i = 1, \ldots, n) = \prod_{k \in A_t} \left(\tau_{\lambda_k, \mu_k}(x_k, y_k)\, e^{-\lambda_k} {\lambda_k}^{x_k}\, e^{-\mu_k} {\mu_k}^{y_k} \right)^{\phi(t-t_k)} \tag{4.4}$$

gdje je t trenutak u kojem se vrši procjena, t_k je trenutak kada je odigrana utakmica k, $A_t = \{k : t_k < t\}$(skup utakmica odigranih prije trenutka t), λ_k i μ_k su definirani kao u jednadžbi (4.3), a ϕ je nerastuća funkcija vremena. Funkcija ϕ bi trebala dati manju vrijednost, što su trenutci t i t_k udaljeniji. Na taj način starije utakmice

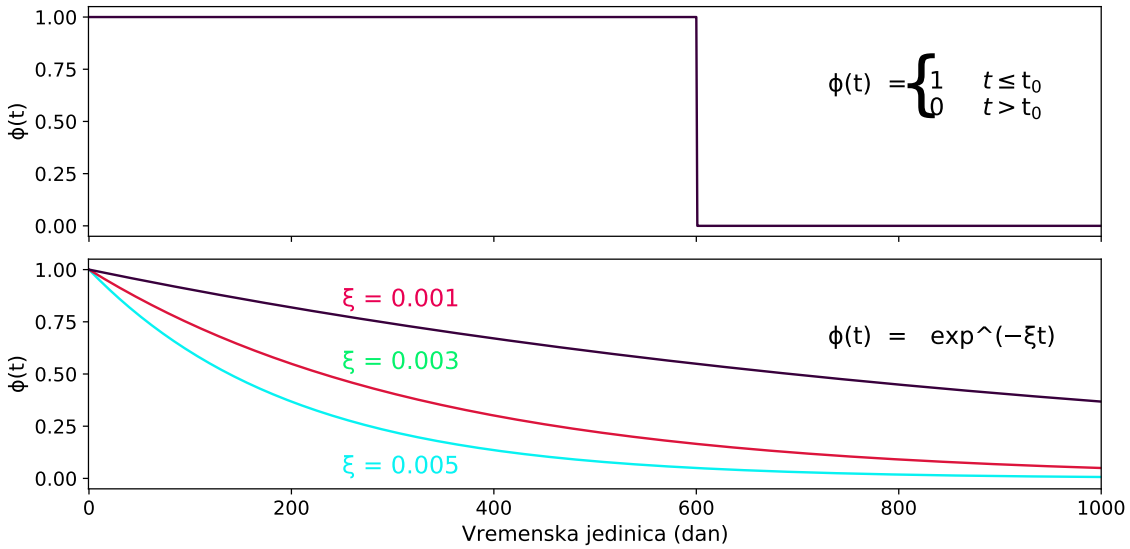
dobivaju manje značenje, dok novije dobivaju veće.
Funkcija ϕ može se odabrati na više načina. Jedan od mogućih izbora je

$$\phi(t) = \begin{cases} 1, & t \leq t_0, \\ 0, & t > t_0, \end{cases}$$

u čijem slučaju, u trenutku t, svim rezultatima unutar zadnjih t_0 trenutaka biti će dana jednaka važnost kod zaključivanja.
Dixon i Coles ispitali su i neke druge mogućnosti te predlažu izbor

$$\phi(t) = e^{-\xi t}, \quad \xi > 0. \tag{4.5}$$

Upotreba ove funkcije eksponencijalno će umanjivati važnost starijih utakmica. Ako je $\xi = 0$, sve utakmice će se gledati kao jednako važne, dok će povećanje te vrijednosti dati veću važnost nedavnim utakmicama.



Slika 12: Funkcija vremena

Kao i ranije, moramo odrediti parametre koji maksimiziraju jednadžbu, u ovom slučaju jednadžbu (4.4). Optimizacija izbora ξ je problematična. Procjenom vjerojatnosti svih rezultata mogu se pronaći vjerojatnosti pobjede domaćina, neriješenog rezultata i pobjede gostiju.

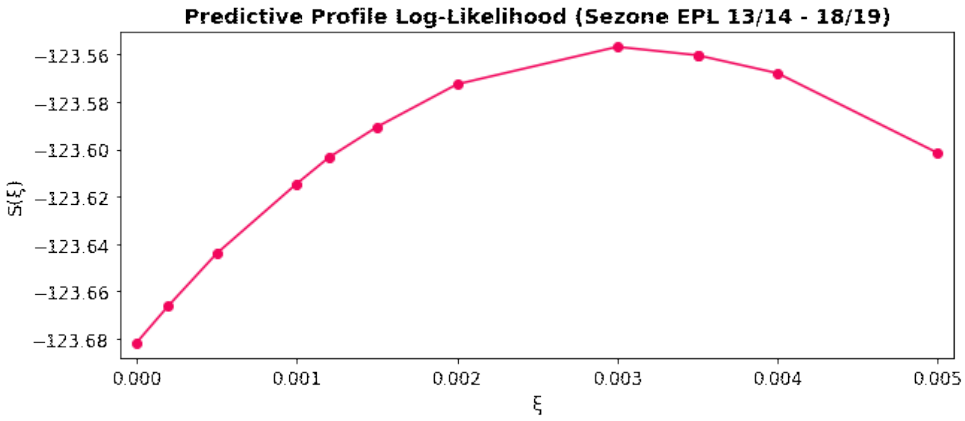
Na primjer, vjerojatnost pobjede domaćina u utakmici k procjenjuje se kao:

$$p_k^H = \sum_{l,m \in B_H} P(X_k = l, Y_k = m) \tag{4.6}$$

gdje je $B_H = \{(l, m) : l > m\}$, a vjerojatnost rezultata određuje se iz maksimizacije modela (4.4) u trenutku t_k (trenutak kada se igra utakmica k). Slični izrazi vrijede za p_k^A i p_k^D, vjerojatnost pobjede gostiju u utakmici k odnosno vjerojatnost neriješenog rezultata u utakmici k.
Sada definiramo

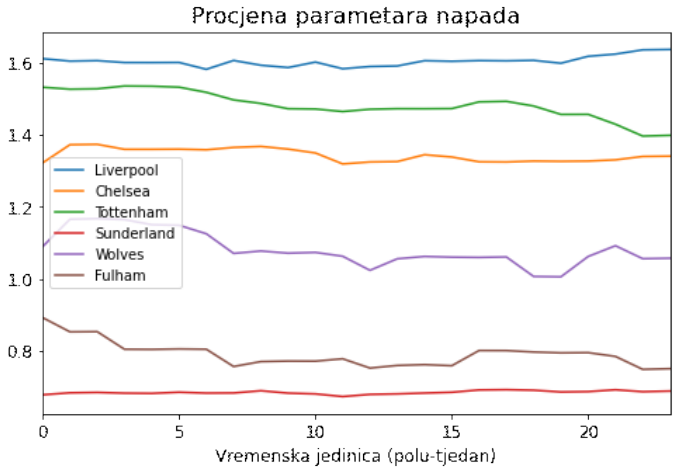
$$S(\xi) = \sum_{k=1}^{N} (\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D) \tag{4.7}$$

gdje δ bilježi ishod utakmice. Primjerice, $\delta_k^H = 1$, ako je utakmica k završila pobjedom domaćina, u suprotnom je $\delta_k^H = 0$. p_k^H, p_k^A i p_k^D su procjene maksimalne vjerodostojnosti iz modela (4.4), s težinskom parametrom postavljenim na ξ.
U radu [6] kao vremenska jedinica koristi se pola tjedna, stoga ćemo i mi uzeti istu vremensku jedinicu. Određivanje optimalnog parametra ξ je eksperimentalno. Za tu potrebu gledali smo sezone EPL (2013/2014 do 2018/2019) i radili procjene za zadnjih 100 dana sezone EPL 2018/2019, odnosno 24 polu-tjedna (izbačeno je nekoliko dana kada nije bilo utakmica). Proces računanja $S(\xi)$ je jako računalno zahtjevan pa smo isprobavali sljedeće varijante parametra ξ:
0, 0.0002, 0.0005, 0.001, 0.0012, 0.0015, 0.002, 0.003, 0.0035, 0.004, 0.005.

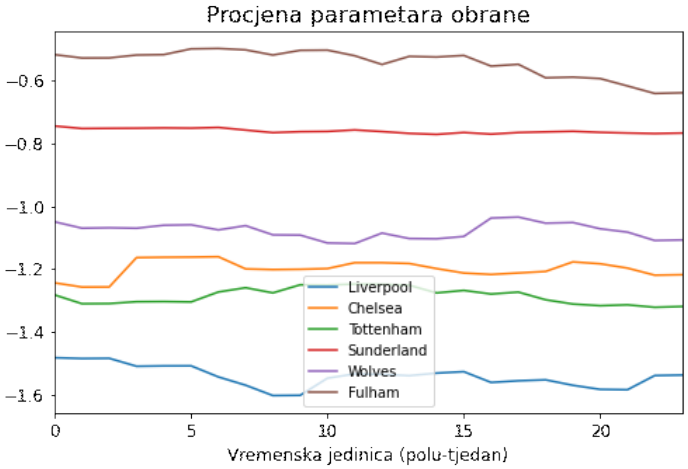


Slika 13: Funkcija postiže maksimum za $\xi = 0.003$

Sada smo za fiksni $\xi = 0.003$ napravili procjenu parametara. Pošto se procijena parametara mijenja svaki polutjedan na sljedećim slikama prikazano je kako su se mijenjale procjene parametra napada i obrane za nekoliko momčadi kroz vrijeme.



Slika 14: Procjena parametara napada



Slika 15: Procjena parametara obrane

Literatura

[1] A. A. Alzaid, M. A. Omair, *On the Poisson difference distribution inference and applications*, Bull. Malays. Math. Sci. Soc. (2), Vol. 33(2010), 17–45

[2] B. Basrak, *Aktuarska matematika II, 4.dio*, Sveučilište u Zagrebu PMF - Matematički odjel, Zagreb, 2016.

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

[4] M. Benšić, *Predavanja za kolegij Multivarijatna analiza*

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

[6] M. J. Dixon, S. G. Coles, *Modelling Association Football Scores and Inefficiencies in the Football Betting Market*, Applied Statistics, Vol. 46(1997), 265-280

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

[8] J. A. Gardner, *Modeling and simulating football results*, https://www1.maths.leeds.ac.uk/~voss/projects/2010-sports/JamesGardner.pdf

[9] O. C. Ibe, *Fundamentals of applied probability and random processes*, Academic Press, Boston, 2014.

[10] M. J. Maher, *Modelling association football scores*, Statistica Neerlandica, Vol. 36(1982), 109-118

[11] J. H. McDonald, *Handbook of Biological Statistics*, 3rd ed. Sparky House Publishing, Baltimore, Maryland, 2014.

[12] *Predicting Football Results With Statistical Modelling*, https://dashee87.github.io/data%20science/football/r/predicting-football-results-with-statistical-modelling/

[13] *Predicting Football Results With Statistical Modelling: Dixon-Coles and Time-Weighting*, https://shorturl.at/iuDN1

Sažetak

Nogomet je jedan od najpopularnijih sportova današnjice. Zbog toga je predviđanje nogometnih ishoda vrlo zanimljivo. U ovom radu predstavljeni su matematički modeli kojima je moguće predviđati ishode pojedinih utakmica na osnovi prethodnih podataka. Najznačajniji modeli kojima možemo predviđati ishode utakmica su model s Poissonovom distribucijom broja golova i Dixon-Coles model. U praktičnom dijelu rada primijenli smo modele na stvarnim podacima.

Ključne riječi: očekivanje, Poissonova distribucija, procjena parametra, Poissonov model, Dixon - Coles model

Summary

Football is one of the most popular sports today. This is why predicting football outcomes is very interesting. This paper presents mathematical models that can predict the outcomes of individual matches based on previous data. The most significant models by which we can predict the outcomes of a match are the Poisson distribution of number goals and the Dixon-Coles model. In the practical part of the paper, we applied the models to real data.

Keywords: mean, Poisson distribution, parameter estimate, Poisson model, Dixon - Coles model

Životopis

Rođena sam 10. studenog 1994. godine u Koprivnici. Obrazovanje sam započela u Osnovnoj školi Molve, a nakon toga upisala sam Opću gimnaziju u Đurđevcu koju završavam 2013. godine. Iste godine upisujem Preddiplomski studij matematike na Odjelu za matematiku Sveučilišta J.J. Stossmayera u Osijeku. Akademski naziv prvostupnice matematike stječem 2017. godine s temom završnog rada *Ravnoteža napete žice* pod mentorstvom izv.prof.dr.sc. Krešimira Burazina. U jesen 2017. godine upisujem diplomski studij matematike, smjer Financijska matematika i statistika. Tijekom diplomskog studija odradila sam stručnu praksu u Hrvatskoj agenciji za hranu gdje sam se bavila statističkom analizom podataka. Od svibnja 2019. godine zaposlena sam u BE-terna d.o.o (od srpnja 2020. godine Adacta d.o.o posluje pod imenom BE-terna d.o.o) na poziciji *aplikacijski savjetnik*.