УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Јована Ђорђевић

Дводимензионални Пуасонови регресиони модели и њихова примена на моделирање резултата фудбалских утакмица

— мастер рад —

Београд, 2017.

Предговор

У последњих неколико година дошло је до експанзије интернета, тако да у овом тренутку скоро свако поседује приступ интернету и његовом садржају. Подаци Републичког завода за статистику показују да интернет у Србији свакога дана користи више од 2,85 милиона људи. Ови догађаји довели су до ширења спортских кладионица, јер се и компаније шире заједно са интернетом у циљу задовољења потребе тржишта. Ширење спортске кладионице огледа се у проширењу понуде, која сада може у реалном времену брзо да промени понуду квота у складу са сваком променом сценарија меча. Услед поменутих промена спортске кладионице нуде 24 часовно клађење на интернету на најразличитије игре, нарочито везане за фудбалске утакмице. Играчи могу да се кладе на: резултат меча, који ће играч на утакмици следећи постићи гол, време када ће гол бити постигнут, број жутих картона које ће судије на мечу доделити... Разлог велике популарности фудбала је различит, првенствено реч је о модерним клубовима који представљају профитабилне компаније, док је други разлог новац потрошен на спортске опкладе који бележи велики раст у Европи. Уз сав новац који љубитељи спортског клађења улажу на фудбалске мечеве јавља се природно питање: можемо ли користити математику за предвиђање резултата фудбалских мечева? Статистичка литература пружа најразличитије моделе као одговор на поменуто питање. Неки модели дају одговор на питање победника меча, док се други односе на тачан резултат меча. Са друге стране избор модела којим желимо да предвидимо исход неког меча зависи и од лиге односно такмичења коме меч припада. У наредним поглављима покушаћемо да предвидимо резултате фудбалских мечева који су много пута окарактерисани као непредвидљиви. Посматраћемо: резултате фудбалских мечева, како тим напада и брани се, форму тима и питање ефекта домаћег терена. Такође ћемо разматрати како су други покушали да предвиде резултате истих фудбалских мечева.

Садржај

Поглавље 1

Пуасонови модели

Претпоставимо да су догађаји на спортским мечевима случајни. Узмимо на пример фудбалски меч, где тимови постижу голове током случајних тренутака времена трајања меча. Уколико је на пример резултат меча 4:2, голови су постигнути у 18, 26, 36, 45, 52 и 85 минуту, притом не водећи рачуна о томе који је тим дао гол, меч можемо посматрати као случајни процес.

Дефиниција 1.0.1 *Нека је $N(t)$ укупан број догађаја који се десе у интервалу $[0,t]$. Стохастички процес $\{N(t), t \geq 0\}$ назива се процес бројања догађаја или процес пребрајања.*

Процес пребрајања има следеће особине:
1. За фиксирано t, $N(t)$ је случајна променљива чије су вредности из скупа N_0.
2. Функција $N(t)$ је неопадајућа, тј.

$$N(t_2) - N(t_1) \geq 0, \text{ ако је } t_2 > t_1 \geq 0,$$

штавише $N(t_2) - N(t_1)$ представља број догађаја који се догоде у интервалу $(t_1, t_2]$. [3]

Дефиниција 1.0.2 *Хомогени Пуасонов процес са стопом раста λ,где је $\lambda > 0$, је процес пребрајања $\{N(t), t \geq 0\}$ за који важи:*
1. $N(0) = 0$;
2. Процес $N(t)$ има независне прираштаје;
3. Број догађаја у произвољном интервалу дужине t има Пуасонову расподелу са параметром λt, односно

$$P\{N(t+s) - N(s) = n\} = e^{-\lambda t}\frac{(-\lambda t)^n}{n!}, \text{ за свако } t \geq 0 \text{ и } n \in N_0.$$

Из претходне једнакости можемо закључити да Пуасонов процес има стационарне прираштаје јер расподела догађаја $\{N(t+s) - N(s)\}$ не зависи од s, односно зависи само од дужине интервала, а не зависи од његовог положаја на временској оси.
Уколико посматрамо фудбалски меч као Пуасонов процес, особина да су независни прираштаји можда и није случај са фудбалским мечом. Ово можда није случај јер фудбалски меч има два полувремена, резултат другог полувремена није у потпуности независан јер од резултата првог полувремена итекако зависи игра и сценарио другог. Нпр. тренер може увести боље играче у игру јер је био незадовољан исходом првог дела меча. Уколико посматрамо фудбалски меч на самом почетку је $N(0) = 0$, док уколико је резултат на крају меча 4:2 тада је $N(90) = 6$.
Постоји мишљење да фудбалски мечеви нису независни због варирања снаге тимова, затим сваки тим има неку посебну снагу када игра на домаћем терену. Упркос свему овоме, многи ипак показују да се резултат фудбалске утакмице може моделирати коришћењем Пуасоновог процеса.[3]

Поглавље 2

Пуасонова случајна променљива и основне особине

Дефиниција 2.0.1 *Нака је X дискретна случајна величина за коју важи закон расподеле:*

$$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!} \ , \text{ где } x \text{ узима вредности } \ x=0,1,2,\ldots . \quad .$$

Ова расподела се зове Пуасонова расподела са параметром λ, у ознаци $X \sim P(\lambda)$.

Дефиниција 2.0.2 *Функција генератрисе вероватноће случајне величине $X \sim P(\lambda)$ једнака је :*

$$g_X(t)=E\left(t^X\right)=\sum_{k=0}^{\infty} t^k \ \frac{\lambda^k e^{-\lambda}}{k!}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!} \ =e^{-\lambda}e^{\lambda t}=e^{\lambda(t-1)} \ .$$

Диференцирањем претходно дате функције генератрисе вероватноће добијамо:

$$g_X^{'}(t)=\lambda e^{\lambda(t-1)} \ ,$$
$$g_X^{''}(t)=\lambda^2 e^{\lambda(t-1)} \ ,$$

Заменом $t=1$ у претходну једнакост добијамо:

$$g_X^{'}(1)=\lambda \ ,$$
$$g_X^{''}(1)=\lambda^2 .$$

Очекивање случајне промељиве X може се дефинисати користећи претходне једнакости као:

$$E(X) = g_X^{'}(1) = \lambda,$$

односно дисперзија случајне променљиве X :

$$D(X) = g_X^{''}(1) + g_X^{'}(1) - (g_X^{'}(1))^2 = \lambda \ .$$

Дакле уколико $X \sim P(\lambda), \lambda > 0$ тада важи : $E(X) = D(X) = \lambda$. Очекивање и дисперзија Пуасонове случајне променљиве једнаки су λ, тако да нема потребе оцењивати посебно сваки од ова два параметра. Као што ћемо видети касније, Пуасонова расподела је погодна за моделирање пребројивих података. Реални подаци који не могу бити добро моделирани помоћу Пуасонове расподеле обично имају већу дисперзију од средње вредности и тада имамо проблем прераспршености података. У том случају модел мора бити прилагођен овој особини.[4]

Поглавље 3

Уопштени линеарни регресиони модели

Уопштени линеарни модели представљају значајну генерализацију линеарне регресије у уопштенију регресију, која користи експоненцијалну фамилију расподела. Уопштен линеаран модел модел је заснован на следећем:

- регистроване вредности се укључују у модел путем линеарне функције ($X^T\beta$)

- условно очекивање зависне променљиве се представља као функција линеарне комбинације

$$E(Y|X) = \mu = f(X^T\beta)$$

- добијене вредности се изводе из експоненцијалне фамилије расподела са средином μ .

Уопштени линеарни модели се састоје из три компоненте:

- Компонента случајности или зависна променљива
дефинише условну расподелу обележја Y_i (за i-ту од n независинсих вредности), за дате вредности независних променљивих у моделу. У оригиналној формулацији расподела, Y_i је члан експоненцијалне фамилије расподела коју чине: Нормална, Пуасонова, Биномна, Гама и инверзна Гаусова расподела.

- Компонента систематичности или линеарни предиктор је линеарна функција параметара регресије

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik}.$$

Линеарни предиктор може бити сачињен од: независних променљивих, трансформација независних променљивих и нелинеарних функција променљивих (као у полиномијалној регресији). У случају полиномијалне регресије модел остаје линеаран све док је линеаран вектор параметараβ.[4]

- Глатка и инвертибилна функција везе која трансформише очекивање обележја,

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik} \ .$$

Како је функција везе инвертибилна,можемо такође да напишемо

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik}) \ .$$

Стога се уопштени линеарни модели могу посматрати и као линеарни модели трансформација очекивања обележја или као нелинеарни регресиони модели обележја. Инверзна веза се назива и функција средње вредности. Најчешће коришћене функције везе и њихове инверзне вредности дате су следећом табелом 3.1, где μ_i представља очекивану вредност резултата, док η_i представља линеарно предвиђање.[4]

Табела 3.1: Најчешће коришћене функције везе и њихове инверзне вредности

Веза	$g(\mu_i) = \eta_i$	$\mu_i = g^{-1}(\eta_i)$
Идентитет	μ_i	η_i
Лог	$\ln \mu_i$	e^{η_i}
Инверзна	${\mu_i}^{-1}$	${\eta_i}^{-1}$
Инверзна квадратна	${\mu_i}^{-2}$	${\eta_i}^{-1/2}$
Квадратни корен	$\sqrt{\mu_i}$	${\eta_i}^2$
Логит	$\ln \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$

Табелом 3.2 дате су расподеле из експоненцијале фамилије, односно њихове каноничке везе, домен резултата и условне функције дисперзије. Параметар σ_i је параметар дисперзије, η_i линеарно предвиђање, а μ_i представља очекивану вредност Y_i .

Табела 3.2: Расподеле из експоненцијале фамилије

Фамилија	Канонička веза	Домен Y_i	$D(Y_i\|\eta_i)$
Гаусова	Идентитет	$(-\infty, +\infty)$	σ_i
Биномна	Логит	$\frac{0,1,2,\ldots n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{\eta_i}$
Пуасонова	Лог	0,1,2..	μ_i
Гама	Инверзна	$(0, +\infty)$	$\sigma_i {\mu_i}^2$
Инверзна Гаусова	Инверзна квадратна	$(0, +\infty)$	$\sigma_i {\mu_i}^3$

Користећи статистички софтвер R уопштене линеарне моделе можемо моделирати уз помоћ стандардних расподела датих табелом 3.2 . Приликом избора модела, читав низ регресионих модела узимамо у разматрање. Модели које разматрамо су: комплетан или потпун модел (има онолико параметара колико и регистрованих вредности), максимални модел (највећ модел који смо спремни разматрати), минимални модел (садржи минималан скуп параметара који морају бити присутни) и тренутни модел (модел који је тренутно предмет истраживања, а налази између минималног и максималног по броју предиктора). Потпуни модел описује регистроване вредности тачно, али баш због тога има врло мале шансе да буде погодан за понављање истраживања уз коришћење истих метода, али других регистрованих вредности. Он не наглашава важне особине података. Насупрот томе, минимални модел има добре шансе да одговара и подацима из поновљених истраживања. Међутим, битне карактеристике података су код минималног модела обично испуштене. Дакле, мора се пронаћи баланс између успешности уклапања података и једноставности. [4]

Поглавље 4

Пуасонова регресија и дефинисање Пуасоновог регресионог модела

4.1 Основне особине модела

Пуасонова регресија је врста уопштених линералних модела, где зависну случајну променљиву моделирамо претпостављајући да има Пуасонову расподелу. Пуасонова расподела подразумева случајне променљиве са ненегативним целобројним вредностима, као што су, пребројиви подаци. Зависна променљива представља број догађаја у одређеном временском интервалу.
Као што смо већ напоменули, код линеарних модела процене средњих вредности могу да буду негативне, међутим када посматрамо пребројиве податке, средине морају бити ненегативне. Даље, пребројиви подаци често испољавају хетероскедастичност, где већа дисперзија прати већу средњу вредност. Најједноставнији уопштени линеарни модел за податке добијене пребројавањем подразумева Пуасонову расподелу компоненте случајности.
Моделовање Пуасонове регресије се одвија у четири корака:

- постављање модела
- оцењивање параметара модела
- провера адекватности модела или другачије речено колико модел

добро уклапа податке

- закључак

Претпоставимо да имамо узорак обима n, дат са y_1,y_2 , $\ldots, y_n$ које можемо посматрати као реализацију независних Пуасонових случајних променљивих односно $Y_i \sim P(\mu_i)$, дефинишемо модел облика:

$$\log(\mu_i) = {x_i}^T \beta \ ,$$

односно

$$\mu_i = E(y_i|x_i) = e^{{x_i}^T \beta} \ .$$

Затим, уколико диференцирамо претходну једнакост, добијамо следеће:

$$\frac{\partial \mu_i}{\partial x_i} = \frac{\partial E(y_i|x_i)}{\partial x_i} = \frac{\partial E\left(e^{{x_i}^T \beta}\right)}{\partial x_i} = e^{{x_i}^T \beta}\beta_i = \mu_i \beta_i \ .$$

Главна особина Пуасоновог модела је да су математичко очекивање и дисперзија једнаке, односно да важи:

$$E(y_i|x_i) = \mu_i = e^{{x_i}^T \beta} = \ D(y_i|x_i) \ .$$

Уколико се, пак догоди да је $E(y_i|x_i) < \ D(y_i|x_i)$, тада су посматрани подаци прераспрштени, и Пуасонов модел мора бити модификован, да бисмо добили добро слагање модела са подацима.

4.2 Оцењивање параметара модела методом максималне веродостојности

Посматрајмо n независних случајних променљивих Y_i где i узима вредности $i = 1, 2, \ldots, n$, односно y_1,y_2 , $\ldots, y_n$ њихове реализоване вредности. Функција веродостојности за променљиве Y_i представља вероватноћу да дати узорак буде изабран и дата је следећом једнакошћу:

$$l_i(\theta_i, y_i) = P(Y_i = y_i) = e^{(y_i b(\theta_i) + c(\theta_i) + d(y_i))}$$

где $\theta_i, i = 1, 2, \ldots, n$ представља параметар расподеле.
Функције $l_i(\theta_i)$ и $\ln l_i(\theta_i)$ постижу максимум за исту вредност θ_i , у пракси је често лакше наћи максимум природног логаритма функције веродостојности. У том случају важи следећа једнакост

$$\ln l_i(\theta_i, y_i) = y_i b(\theta_i) + c(\theta_i) + d(y_i).$$

Уколико диференцирамо функцију $\ln l_i(\theta_i, y_i)$ по $\partial\theta_i$ добијамо следеће:

$$U_i(\theta_i, y_i) = \frac{\partial l n l_i(\theta_i, y_i)}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) \ .$$

На претходно наведени начин дефинисана функција U_i назива се скор статистика и она представља оцену непознатог параметра β_i. Како можемо приметити, функција U_i зависи од реализованих вредности y_i. Можемо је посматрати и као случајну променљиву на следећи начин:

$$U_i = Y_i b'(\theta_i) + c'(\theta_i) \ .$$

Тада је очекивана вредност овако дефинисане случајне променљиве U_i дата као:

$$E(U_i) = E(Y_i) b'(\theta_i) + c'(\theta_i) \ ,$$

односно

$$E(U_i) = \left(-\frac{c'(\theta_i)}{b'(\theta_i)}\right) b'(\theta_i) + c'(\theta_i) = 0 \ .$$

Дисперзија случајне величине U_i назива се информациона матрица и означићемо је са I_i . На основу особина дисперзије о линеарним трансформацијама случајне променљиве и дефиницији случајне променљиве U_i добићемо следећу једнакост:

$$I_i = D(U_i) = b'^2(\theta_i) D(Y_i) \ .$$

Дисперзија случајне променљиве U_i дата је са:

$$D(U_i) = \frac{b''(\theta_i) c'(\theta_i)}{b'(\theta_i)} - c''(\theta_i)$$

Наведена скор статистика има примену код статистичког оцењивања и тестирања параметара уопштених линеарних модела.
Пошто је $E(U_i) = 0$ за дисперзија случајне променљиве U_i важи:

$$D(U_i) = E\left(U_i^2\right) - E^2(U_i) = E\left(U_i^2\right) \ .$$

Функција веродостојности за све Y_i , $i = 1, 2, , n$ је:

$$L = \sum_{i=1}^{n} l n l_i(\theta_i) = \sum_{i=1}^{n} y_i b(\theta_i) + c(\theta_i) + d(y_i).$$

Случајна величина U_j задовољава следећу једнакост :

$$U_j = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{D(Y_i)} x_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i}\right).$$

Елементи коваријационе матрице случајне величине U_i имају облик:

$$I_{jk} = E\left(U_j\, U_k\right) .$$

Уколико искористимо $E\left(\left(Y_i - \mu_i\right)\left(Y_l - \mu_l\right)\right) = 0$ за $i \neq j$, јер како је раније речено случајне величине Y_i , $i = 1, 2, ..., n$ су међусобно независне, елементи коваријационе матрице случајне величине U_i добијају следећи облик:

$$I_{jk} = \sum_{i=1}^{n} \frac{x_{ij}\, x_{ik}}{D(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 .$$

На основу свега написаног добијамо следећу једнакост:

$$b^{(m)} = b^{(m-1)} + \left(I^{(m-1)}\right)^{-1} U^{(m-1)} ,$$

где је $b^{(m)}$ вектор оцена параметара $\beta_1, \ldots, \beta_p$ у m-тој итерацији и $\left(I^{(m-1)}\right)^{-1}$ инверзна информациона матрица са елементима I_{jk} . Уколико помножимо обе стране претходне једнакости са $I^{(m-1)}$ добијамо:

$$b^{(m)} I^{(m-1)} = I^{(m-1)} b^{(m-1)} + U^{(m-1)} ,$$

односно

$$I = X^T W X ,$$

где је W дијагонална матрица димензије $n \times n$ са елементима:

$$w_{ii} = \frac{1}{D(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Код уопштених линеарних модела оцене добијене методом максималне веродостојности користе алгоритам итеративних тежинских најмањих квадрата.

4.3 Тест количника веродостојности

Статистичку проверу дефинисаног модела вршимо уз помоћ интервала поверења и тестирања хипотеза.
Тестирање хипотезе врши се тако што се пореде два модела и колико се добро поклапају са подацима. Када је реч о уопштеним линеарним моделима о којима је у овом раду реч, два модела требало би да имају

исту расподелу вероватноћа и исту функцију везе, али линеарни предиктор једног модела треба да садржи више параметара од другог. Једноставнији модел, који одговара H_0 хипотези мора бити специјалан случај другог модела који је општији. Уколико пак једноставнији модел једнако добро описује податке као и општији тада ћемо наравно користити једноставнији модел.
Како бисмо упоредили два модела, морамо поставити статистике које описују добро модел, односно колико се исти добро поклапа са подацима. Такве статистике могу бити базиране на функцији веродостојности, максималној вредности логаритма функције веродостојности, критеријуму минималне вредности суме квадрата или пак разлици статистика за одступање резидуала.
Уколико је S статистика коју посматрамо, основна идеја је да под одређеним условима апрокцимација задовољава следеће:

$$\frac{S-E(S)}{\sqrt{D(S)}} \sim N(0,1) ,$$

односно

$$\frac{(S-E(S))^2}{D(S)} \sim \chi^2(1) .$$

$E(S)$ и $D(S)$ редом представљају очекивање односно дисперзију статистике S.
Овде ће бити речи о логаритму функције веродостојности као о начину процене адекватности модела, на тај начин што га упоређујемо са општијим моделом који садржи максималан број параметара који се могу оценити.
Претпоставићемо да имамо n независних случајних променљивих Y_i, $i = 1, 2, ., n$. Означићемо са β_{max} вектор параметара потпуног модела, односно b_{max} оцену за β_{max} добијену методом максималне веродостојности. У том случају функција веродостојности за потпуни модел у тачки b_{max} односно $l(b_{max}; y)$ биће већа од било које друге функције веродостојности за дате регистроване вредности. Означимо са $l(b; y)$ максималну вредност функције веродостојности за модел који посматрамо. Тада важи следећа једнакост:

$$\Lambda = \frac{l(b_{max};y)}{l(b;y)} ,$$

односно

$$\log(\Lambda) = \log(l(b_{max}; y)\ ) - \log(l(b; y)) .$$

Уколико је вредност $\log(\Lambda)$ велика, то указује да посматрани модел слабо описује податке у односу на потпуни модел.
Како бисмо одредили критичну област за $\log(\Lambda)$, потребно је да знамо његову узорачку расподелу. Претпоставимо да су посматраних n независних случајних променљивих Y_i , случајне величине са Пуасоновом расподелом $Y_i \sim P(\mu_i)$. Тада логаритам функције веродостојности задовољава следећу једнакост:

$$\ln l(\beta; y) = \sum y_i \ln(\mu_i) - \sum \mu_i - \sum lny_i! \quad .$$

Уколико је модел комплетан, тада су μ_i различите за све $, i = 1, 2, \ldots, n$ односно $\beta = [\mu_1, \mu_2, \ldots., \mu_n]^T$.
Претпоставимо да модел који тестирамо односно поредимо са потпуним моделом, има p параметара где је $p < \ n$. Оцену добијену методом максималне веродостојности b можемо користити како бисмо израчунали $\widehat{\mu_i}$, јер је $E(y_i) = \mu_i$. Тада је максимална вредност логаритма функције веродостојности једнака:

$$\ln l(\beta; y) = \sum y_i \ln(\widehat{y_i}) - \sum \widehat{y_i} - \sum lny_i! \quad .$$

Узорачко одступање резидуала статистике D_r дато је са:

$$D_r = 2(\ln(l(\beta_{max}; y)\ ) - \ln(l(b; y)) = 2\left(\sum y_i \ln\left(\frac{y_i}{\widehat{y_i}}\right) - \sum(y_i - \widehat{y_i}\right).$$

Међутим, за већину модела се може показати да важи $\sum y_i = \sum \widehat{y_i}$ и тада узорачко одступање резидуала статистике D_r задовољава следећу једнакост:

$$D_r = 2\sum \sigma_i ln\frac{\sigma_i}{e_i} \ .$$

У претходној једнакости σ_i је регистрована вредност за y_i, док је e_i оцена очекиване вредности за $\widehat{y_i}$. Вредност D_r можемо упоредити са ${\chi^2}_{n-p}$, тако што проверимо да ли је испод репа наведене расподеле и затим уколико јесте можемо рећи да се модел добро поклапа са подацима. Хипотезе о вектору параметара β дужине p могу се тестирати уз помоћ Валдовог теста чија статистика:

$$(\widehat{\beta} - \beta)^T I(\widehat{\beta} - \beta) \ ,$$

има приближно ${\chi^2}_p$ расподелу.

4.4 Прераспрштеност или прекорачење дисперзије модела

Уколико је систематички део Пуасоновог модела одговарајући, што би значило да ниједан важан параметар није изостављен, и да су функције добро дефинисане, а ипак постоји повећање варијације око оцењених вредности, тада је полазна претпоставка о Пуасоновој расподели нетачна. Прераспрштеност представља дефинисање стохастичке компоненте, при чему је систематичка структура модела тачна. Потенцијално решење наведеног проблема може бити замена Пуасоновог регресионог модела негативним Биномним или Квазипуасоновим моделом. Постоје специфични тестови који региструју прераспрштеност модела али врло често је довољна стандардна статистика χ^2 . Присуство прераспрштености модела никада не треба игнорисати јер чак иако је форма модела тачна, занемаривање прераспрштености доводи до оцена дисперзија процењених коефицијената које су превише мале, чиме настају превише уски интервали поверења и сувише мале p вредности тестова. [4]

4.5 Квазипуасонов модел

Када је дисперзија пребројивих података већа него што је она моделирана Пуасоновим регресионим моделом, један од начина да то превазиђемо је да уведемо параметар распршења који ће дозволити прекорачење дисперзије.
Претпоставимо да су посматраних n независних случајних величина, величине са Пуасоновом расподелом $Y_i \sim P(\mu_i)$ и уколико уведемо параметар θ из претходно наведеног разлога на следећи начин:

$$D(Y_i) = \theta\mu_i \ .$$

Уколико је вредност уведеног параметра $\theta > 1$ тада је дисперзија већа од очекиване вредности, а уколико је $\theta < 1$ тада је дисперзија мања у односу на очекивану вредност у Пуасоновом моделу. Наведено прилагођавање Пуасоновог регресионог модела уз помоћ параметра распршења, који линеарно зависи од функције средине назива се Квазипуасонов регресиони модел.
Увођење параметра распршења за прераспрштене податке одразиће се на оцене стандардних грешака, које ће све бити помножене са $\sqrt{\theta}$ у односу на Пуасонов регреиони модел.

4.6 Негативни биномни модел

Уколико код Пуасоновог модела препознамо шум приликом мерења пребројивих података, можемо дефинисати модификацију код које на стандардни модел додајемо стохастички део ε_i и тада модел добија следећи облик:

$$\log(\lambda_i) = {x_i}^T\beta + \varepsilon_i \quad ,$$

притом је $e^{\varepsilon_i} : \Gamma(\theta,\theta)(E(e^{\varepsilon_i}) = 1, D(e^{\varepsilon_i}) = 1/\theta)$ за свако i и $cov\,(\varepsilon_i\ ,\varepsilon_j\ ) = 0$ за све $i \neq j$. [4] Последица додавања стохастичког дела ε_i је λ_i који представља модификацију μ_i за шум ε_i.
Овако дефинисан модел може се сматрати и Пуасоновим моделом са двоструком случајношћу, јер поред случајности која је укључена у Пуасонов модел имамо и други део случајности ε_i . Увођењем случајности ε_i , Пуасонова формулација сада добија следећи облик:

$$f\,(y_i|x_i,\lambda_i) = \tfrac{e^{-\lambda_i\mu_i}(\lambda_i\mu_i)^{y_i}}{y_i!} \ .$$

Последица је да условна расподела сада не зависи само од x_i већ и од λ_i, међутим ипак остаје и даље Пуасонова. Сада се поставља питање како да одредимо расподелу за y_i, која зависи само од x_i јер су оне заправо независне променљиве које посматрамо. Функција расподеле за y_i која зависи само од x_i дата је следећом једнакошћу:

$$f\,(y_i|x_i) = \tfrac{(\theta+y_i)}{(1+y_i)(\theta)}{r_i}^{y_i}(1-r_i)^{\theta} \ ,$$

где је r_i једнако

$$r_i = \tfrac{\mu_i}{\mu_i+\theta} \ .$$

Дисперзија за условну средњу вредност μ_i није μ_i већ:

$$\mu_i\left(1+\left(\tfrac{1}{\theta}\right)\mu_i\right) = \mu_i + \left(\tfrac{1}{\theta}\right){\mu_i}^2 \ .$$

Вредности параметара β и θ можемо оценити методом максималне веродостојности. Такође можемо добити и оцене стандардних грешака за оба параметра.
Један од начина да проверимо да ли постоји прераспрштеност података је да то урадимо помоћу оцена из негативног Биномног модела. Како овај модел даје оцену параметара дисперзије параметра θ потребно је да поставимо нулту хипотезу $H_0 \ : \ \theta = 0$. Уколико добијемо да је $\theta = 0$, користићемо Пуасонов модел, док уколико је $\theta > 0$ постоји прераспрштеност, случај $\theta < 0$ значи да је дисперзија мања од средње вредности.

Поглавље 5

Дводимензионалан Пуасонов регресиони модел

Дефиниција 5.0.1 *Нека су X_1, X_2, X_3 три независне случајне величине са Пуасоновм расподелом, редом са параметрима λ_1, λ_2, λ_3 . Случајан вектор (X , Y) дефинисан на следећи начин:*

$$X = X_1 + X_3 \ ,$$

$$Y = X_2 + X_3 \ ,$$

има дводимензионалну Пуасонову расподелу у ознаци $BP(\lambda_1, \lambda_2, \lambda_3)$. Закон расподеле овако дефинисане дводимензионалне Пуасонове расподеле задат је следећом једнакошћу:

$$f_{BP}(x,y) = P_{X,Y}(x,y) = P(X = x, Y = y) =$$
$$exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \frac{{\lambda_1}^x}{x!} \frac{{\lambda_2}^y}{y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k \quad .$$

Нека је i-та опсервација модела представљена као:

$$(X_i, Y_i) \ \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$$
$$\log(\lambda_{ki}) = w_{ki}\beta_k \ \ k = 1, 2, 3$$

где i представља број итерације, w_{ki} означава вектор промењљивих за посматрање i-те опсевације коришћен за моделирање λ_{ki} и β_k означава одговарајући вектор коефицијената регресије.
Јасно је да на сваки параметар дводимензионалне Пуасонове расподеле могу утицати различите карактеристике и промењљиве. Из тог разлога објашњавајуће промењљиве које се користе за моделирање параметара λ_{ki} могу бити различите. Процена параметара за такав модел није

једноставна, користићемо алгоритам итеративних тежинских најмањих квадрата, који је ефикасна итеративна процедура за процену параметара, на основу критеријума максималне веродостојности .
Конструкцију алгоритма за дводимензионални Пуасонов регресиони модел можемо искористити за извођење дводимензионалне Пуасонове расподеле. Из поменутог разлога за сваку опсервацију i, увешћемо помоћне случајне величине X_{1i}, X_{2i} и S_i које имају Пуасонову расподелу, редом са параметрима λ_{1i}, λ_{2i} и λ_{3i} . Затим се дефинишу случајне величине X_i и Y_i на следећи начин:

$$\begin{aligned} X_i &= X_{1i} + S_i \\ Y_i &= X_{2i} + S_i \ . \end{aligned}$$

Алгоритам итеративних тежинских најмањих квадрата процењује непосматране податке преко њихових условних очекивања и затим максимизује вероватноћу комплетног скупа података. Он је намењен проблемима у којима су неке променљиве скривене (нису видљиве у опсервацијама). Дакле, добијамо очекивање случајних величина X_{1i} , X_{2i} и S_i уколико су познати подаци и тренутне вредности параметара , и затим максимизујемо вероватноћу комплетног скупа података уклапањем три Пуасонова регресиона модела.
Сада је циљ проценити регресионе коефицијенте β_k за $k = 1, 2, 3$. Наведени алгоритам је веома флексибилан и многе варијације дводимензионалног Пуасоновог модела могу се лако извести уз мале модификације. Означимо са ϕ следећи вектор параметара $\phi = (\beta'_1, \beta'_2, \beta'_3)$ тада је логвероватноћа читавог скупа података дата једнакошћу:

$$l(\phi) = -\sum_{i=1}^{n}\sum_{k=1}^{3}\lambda_{ki} + \sum_{i=1}^{n}\sum_{k=1}^{3} x_{ki}\log(\lambda_{ki}) - \sum_{i=1}^{n}\sum_{k=1}^{3}\log(x_{ki}!) \ ,$$

где је λ_{ki} одређена са $\log(\lambda_{ki}) = w'_{ki}\beta_k \ \ k = 1, 2, 3$.
Алгоритам је одређен следећим:

- Користећи вредности актуелних параметара за итерацију k добијене су следеће вредности: $\phi^{(k)}, {\lambda_{1i}}^{(k)}, {\lambda_{2i}}^{(k)}$ и ${\lambda_{3i}}^{(k)}$. Наведене вредности користе се за одређивање очекиване вредности случајне величине X_{3i} за $i = 1, \ldots, n$:

$$s_i = E\left(S_i | X_i, Y_i, \phi^{(k)}\right) = \begin{cases} {\lambda_{3i}}^{(k)} \frac{f_{BP}\left(x_i - 1, y_i - 1 | {\lambda_{1i}}^{(k)}, {\lambda_{2i}}^{(k)}, {\lambda_{3i}}^{(k)}\right)}{f_{BP}\left(x_i, y_i | {\lambda_{1i}}^{(k)}, {\lambda_{2i}}^{(k)}, {\lambda_{3i}}^{(k)}\right)} & , \ \min(x_i, y_i) > 0 \\ 0 & , \ \min(x_i, y_i) = 0 \end{cases}$$

где је $f_{BP}(x, y|\lambda_1, \lambda_2, \lambda_3)$ закон расподеле случајног вектора (X_i, Y_i). Условна очекивања случајних променљивих X_{1i} и X_{2i} задовољавају следеће [6]:

$$E\left(X_{1i}|X_i, Y_i, \phi^{(k)}\right) = X_i - s_i$$
$$E\left(X_{2i}|X_i, Y_i, \phi^{(k)}\right) = Y_i - s_i \quad .$$

- Прогноза је дата следећим једнакостима:

$$\beta_1^{(k+1)} = \hat{\beta}\,(x - s, W_1),$$
$$\beta_2^{(k+1)} = \hat{\beta}\,(y - s, W_2),$$
$$\beta_3^{(k+1)} = \hat{\beta}\,(s, W_3),$$

за $k = 1, 2, 3$ где је $s = (s_1, \ldots, s_n)^T$ вектор димензије $n \times 1$ добијен у претходном кораку, $\hat{\beta}\,(x, W)$ је оцена методом максималне веродостојности вектора x и матрице W је матрица података. Параметар ${\lambda_l}^{(k+1)}$, $l = 1, 2, 3$ се рачуна директно из дефиниције модела. [6]

Поглавље 6

Примена модела на предвиђање броја голова фудбалског меча

6.1 Фудбалски подаци (Шпанска Примера лига сезона 2016/2017 године)

Шпанска примера лига први пут је одиграна у сезони 1928-1929, када је фудбалски клуб Барселона постала први шампион. Тада шпански фудбалски савез доноси одлуку којих десет тимова улази у такмичење. До тада шпански фудбал био је организован као шампионат Шпаније. Од 1950-те године Реал Мадрид и Барселона су доминирали шампионатом, Реал Мадрид је чак 33 пута био шампион док је Барселона то остварила 24 пута. Поред осталих тимова који учествују у шпанској Примера лиги, Реал Мадрид, Барселона и Атлетико Мадрид су три најуспешнија тима. Шпанска Примера лига је тренутно друга у УЕФА лествици европских лига на основу њихових наступа у последњих пет година, иза енглеске Премијер лиге, али испред Италијанске серије А.

Током сезоне, која траје од августа до маја, сваки клуб игра два пута са сваким од осталих 19 тимова, једном у гостима а једном као домаћин. Укупно се игра 38 кола. За победу тим добија три поена, један поен добија у случају нерешеног резултата док се за изгубљени меч не добијају поени. Тимови се рангирају према бодовима, и тим који на крају сезоне има највише бодова постаје првак. Уколико тимови имају исти број бодова, тада њихов међусобни дуел одлучује прво место на табели. На крају сваке сезоне, три најлошије рангирана тима испадају у другу лигу,

док прва три рангирана тима друге лиге добијају шансу да се пласирају у виши ранг односно Примера лигу.
Даљу анализу и дефиницију модела вршићемо над завршених 36 недеља шпанске Примера лиге сезона 2016-2017 године,чији резултати током одиграних 359 мечева сезоне налазе се у следећој табели:

Табела 6.1: Примера лига сезона - 2016/2017 година

Тим	БР	БРПК	БРНК	БРИК	БРГДК	БРГПК	БРПОГ	БРНЕГ	БРИЗГ	БРГДГ	БРГПГ	П
Barcelona	36	14	3	1	60	15	12	3	3	48	19	84
Real Madrid	35	13	4	1	44	19	13	2	2	52	20	84
Atletico Madrid	36	13	2	3	37	13	9	6	3	29	12	74
Sevilla	36	13	4	1	34	16	7	5	6	29	29	69
Villarreal	36	11	3	4	35	18	7	6	5	18	14	63
Athletic Club	36	13	3	2	35	17	6	2	10	16	22	62
Real Sociedad	36	10	4	4	28	22	9	1	8	27	27	62
Eibar	36	10	3	5	29	20	5	6	7	25	26	54
Espanyol	36	8	5	5	28	23	6	6	6	19	25	53
Alaves	36	6	8	4	16	20	7	4	7	21	21	51
Malaga	36	11	1	6	32	22	2	7	9	15	28	47
Celta Vigo	35	9	1	7	27	26	4	4	10	22	34	44
Valencia	36	8	4	6	31	29	4	3	11	23	33	43
Las Palmas	36	9	6	3	32	21	1	3	14	20	46	39
Betis	36	6	6	6	21	23	4	1	13	17	38	37
Leganes	36	5	5	8	21	22	3	4	11	13	31	33
Deportivo La Coruna	36	6	5	7	24	23	1	6	11	16	38	32
Sporting Gijon	36	5	3	10	24	36	1	6	11	15	34	27
Granada	36	4	4	10	16	30	0	4	14	12	48	20
Osasuna	36	1	7	10	21	38	2	2	14	16	50	18

Синтакса колона у приказаној табели је следећа :
-БР (Број одиграних мечева тима)
-БРПК (Број победа код куће тима)
-БРНК (Број нерешених мечева код куће тима)
-БРИК (Број изгубљених мечева код куће тима)
-БРГДК (Број датих голова код куће тима)
-БРГПК (Број голова примљених код куће)
-БРПОГ (Број победа у гостима тима)
-БРНЕГ (Број нерешених у гостима тима)
-БРИЗГ (Број изгубљених у гостима тима)
-БРГДГ (Број голова у гостима тима)
-БРГПГ (Број примљених голова у гостима)
-П (Број поена тима) .

6.2 Моделовање фудбалских резултата користећи Пуасонову расподелу

Да ли је број голова меча Пуасонова расподела тестирали смо уз помоћ χ^2 теста, јер је Пуасонова расподела дискретна расподела. Скуп по-

датака који смо надаље у раду користили садржи 36 недеља шпанске Примера лиге сезоне 2016/2017 године односно резултате 359 мечева.
Користећи се χ^2 тестом тестирали смо нулту хипотеза H_0: расподела укупног броја голова по мечу има Пуасонову расподелу.
Тест статистика χ_t^2 задовољава следећу једнакост:

$$\chi_t^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \ ,$$

где су O_i посматране вредности података, E_i очекиване вредности расподеле и n је број догађаја које посматрамо (у нашем случају број мечева).
Очекиване вредности добијене су коришћењем функције goodfit() у статистичком програму R, параметар наведене функције биће Пуасонова расподела као тип расподеле на коју тестирамо податке и method="ML" односно метод максималне веродостојности. На наведени начин добијене су следеће вредности:

> gf< −goodfit(tabela_golovi,type="poisson",method="ML") #method ML-maximum Likelihood
> gf

Observed and fitted values for poisson distribution
with parameters estimated by 'ML'

count	observed	fitted
0	83	68.4388587
1	100	113.4293062
2	90	93.9978234
3	51	51.9300881
4	20	21.5169933
5	10	7.1323738
6	3	1.9701775
7	2	0.4664766

> summary(gf)
Goodness-of-fit test for poisson distribution
X^2 df P(> X^2)
Likelihood Ratio 9.332982 6 0.1556992.

Као што можемо приметити у приказаној табели дате су вредности из посматраног скупа података и одговарајуће вредности Пуасонове расподеле са задатим параметрима. На овај начин добили смо

$\chi_t^2 = 9,332982$ са 6 степени слободе и p вредношћу $P(> X^2) = 0,1556992$. Нулту хипотезу да подаци задовољавају Пуасонову расподелу ћемо дакле прихватити. [2]

6.3 Примена Пуасоновог регресионог модела

Основни скуп података садржи: име домаћина, име госта, број голова остварених на мечу домаћина и аналогно за госта. Уз помоћ наведених података ћемо предвидети очекивани број голова оба тима .
$> head(podaci)$

	Domacin	Gost	br_golova_domacin	br_golova_gost
1	Malaga	Osasuna	1	0
2	Deportivo La Coruna	Eibar	2	1
3	Barcelona	Betis	6	2
4	Granada	Villarreal	1	1
5	Sevilla	Espanyol	6	4
6	Sporting Gijon	Athletic Club	2	1

Како бисмо дефинисали Пуасонов регресиони модел на основу датих података дефинисали смо још предиктора и они су:

- број поена домаћина (колона добија вредност 3 уколико је тим домаћина био победник на мечу, односно 1 уколико је резултат меча нерешен и 0 и уколико је тим домаћина изгубио на посматраном мечу)

- број поена госта (колона добија вредност 3 уколико је тим госта био победник на мечу, односно 1 уколико је резултат меча нерешен и 0 уколико је тим госта изгубио на посматраном мечу)

- просечан број голова као домаћин (просечан бр. голова домаћина остварен до посматраног меча)

- просечан број голова као гост (просечан бр. голова госта остварен до посматраног меча)

- просечан број бодова као домаћин (до посматраног меча просечан број бодова по мечу домаћина којег посматрамо)

- просечан број бодова као гост (до посматраног меча просечан број бодова по мечу госта којег посматрамо)

- класа тима домаћина (уместо броја поена користићемо којој групи односно класи је тим домаћина припадао у претходној сезони (1-најбољи тимови (тимови који су у претходној сезони имали више или једнако од 88 поена), 2-средњи (тимови који су у претходној сезони имали између 48 и 64 поена) и 3-најлошији (тимови који су у претходној сезони забележили мање од 48 поена)))

- класа тима госта (уместо броја поена користићемо којој групи односно класи је тим госта припадао у претходној сезони (1-најбољи тимови (тимови који су у претходној сезони имали више или једнако од 88 поена),2-средњи (тимови који су у претходној сезони имали између 48 и 64 поена) и 3-најлошији (тимови који су у претходној сезони забележили мање од 48 поена))).

- домаћи терен (предиктор који узима вредност 1 уколико је тим домаћин односно 0 уколико је тим гост).

Најпре ћемо дефинисати модел за предвиђање резултата утакмице Еспањол - Валенсија. Из почетног скупа података извучене су све утакмице за протеклих 36 недеља у којима је тим Еспањол био како домаћин тако и гост и податке смо сместили у dataset_Espanyol. Прво је дефинисан општи Пуасонов регресиони модел за тим Еспањол који садржи све предикторе:
$> summary(Espanyol\_puason\_model1)$

$Call:$
$glm(formula = br\_golova\_na\_mecu\_tima \sim br\_golova\_protivnika + prosecan\_broj\_bodova\_tima + prosecan\_broj\_bodova\_protivnika + prosecan\_br\_golova\_tima + prosecan\_br\_golova\_protivnika + factor(tip\_tima\_na\_mecu) + factor(klasa\_protivnika), family =$

$poisson, data = dataset\_Espanyol$)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7294	-0.6903	-0.1650	0.3503	1.8732

Coefficients:

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-0.81335	0.80711	-1.008	0.3136
br_golova_protivnika	0.33622	0.13138	2.559	0.0105 *
prosecan_broj_bodova_tima	0.21983	0.46316	0.475	0.6351
prosecan_broj_bodova_protivnika	0.16326	0.50154	0.326	0.7448
prosecan_br_golova_tima	-0.01657	0.27404	-0.060	0.9518
prosecan_br_golova_protivnika	-0.42995	0.59794	-0.719	0.4721
factor(tip_tima_na_mecu)1	-0.22300	0.16593	-1.344	0.1790
factor(klasa_protivnika)1	-1.41732	0.73924	-1.917	0.0552 .
factor(klasa_protivnika)2	0.36038	0.42409	0.850	0.3954

—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 49.223 on 35 degrees of freedom
Residual deviance: 24.605 on 27 degrees of freedom
AIC: 101.78

Number of Fisher Scoring iterations: 5 .

Класа тима противника и тип тима на мечу су у моделу дефинисани као фактори јер узимају константне вредности раније дефинисане. Видимо да се број голова тима Еспањол може предвидети уз помоћ предиктора број голова противника и класе противника.
Тестираћемо нулту хипотезу $H_0 : \beta_1 = \beta_2 = \ldots = \beta_8$, уз помоћ теста количника веродостојности.
$> lrtest(Espanyol\_puason\_model1)$

Likelihood ratio test

$Model1 : br\_golova\_na\_mecu\_tima \sim br\_golova\_protivnika + prosecan\_broj\_bodova\_tima + prosecan\_broj\_bodova\_protivnika + prosecan\_br\_golova\_tima + prosecan\_br\_golova\_protivnika + factor(tip\_tima\_na\_mecu) + factor(klasa\_protivnika)$

$Model2 : br\_golova\_na\_mecu\_tima \sim 1$
#Df LogLik Df Chisq Pr($> Chisq$)
1 9 -41.889
2 1 -54.198 -8 24.618 0.001804 **
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1 .

Као што можемо приметити, тест статистика има вредност 24,618, односно p вредност теста једнака је 0.001804 што је мање од 0.05, дакле одбацујемо нулту хипотезу да су сви предиктори једнаки нула. Ово значи да постоји барем један β_i који није једнак 0. Вредност коефицијента детерминације R^2 можемо добити из следеће једнакости:

$$1 - Residual \quad deviance/Null \quad deviance = 0.5001$$

што значи да се промене зависне променљиве са 50,01% објашњавају променама независних променљивих. Информације о коефицијентима испред предиктора се могу наћи помоћу функције $confint()$. Наведена функција даје 95% интервале поверења за коефицијенте испред предиктора и тиме добијамо :

$> confint(Espanyol\_puason\_model1)$

	2.5 %	97.5 %
(Intercept)	-2.49240937	0.70402512
br_golova_protivnika	0.07881353	0.59717645
prosecan_broj_bodova_tima	-0.65301297	1.18338727
prosecan_broj_bodova_protivnika	-0.84468214	1.13225530
prosecan_br_golova_tima	-0.63495153	0.47268549
prosecan_br_golova_protivnika	-1.59603676	0.76096657
factor(tip_tima_na_mecu)1	-0.55712099	0.09730438
factor(klasa_protivnika)1	-3.39876619	-0.20996506
factor(klasa_protivnika)2	-0.40467556	1.40694870 .

Из приложеног видимо да само коефицијенти уз предикторе број голова

противника и класе противника не садрже нулу. Затим смо проверили да ли постоји корелација између променљивих које представљају предикторе уз помоћ функције

$$> summary(Espanyol\_puason\_model1, corr = TRUE)\$cor \ .$$

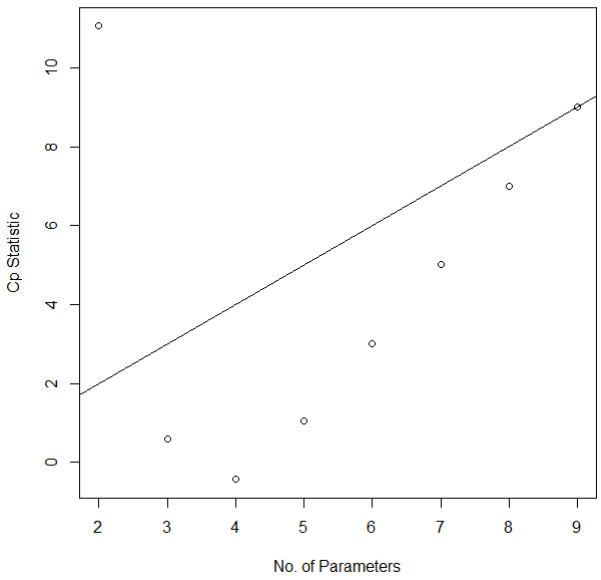
Добијена је велика корелација између просечног броја бодова противника и просечног броја голова противника, износи -0,78 (корелација између њихових коефицијената је негативна).
Према свему до сада реченом схватили смо да није неопходно да сви предиктори уђу у модел. Модел треба да буде што једноставнији јер се тиме олакшава интерпретација. Избор предиктора може се извршити на више начина, један од њих је функција $rugsubsets()$. Наведена функција предлаже које предикторе треба уврстити у модел, у зависности од броја предиктора који улазе у модел. Резултати које смо добили су следећи:

$$w1 < regsubsets(br\_golova\_na\_mecu\_tima \sim br\_golova\_protivnika + prosecan\_broj\_bodova\_tima + prosecan\_broj\_bodova\_protivnika + prosecan\_br\_golova\_tima + prosecan\_br\_golova\_protivnika + factor(tip\_tima\_na\_mecu) + factor(klasa\_protivnika), data = dataset\_Espanyol)$$

$$p < -summary(w1)$$

$$plot(2:9, p\$cp, xlab = "No.of Parameters", ylab = "Cp Statistic")$$

$$abline(0, 1)$$



На основу приказаног графика најбољи број предиктора је 8 и 9. Бира се онај модел где је тачка најближа правој. Пошто је модел са свих 9 предиктора раније дефинисан уопштени модел, сада ћемо дефинисати модел са 8 предиктора на следећи начин:

$> summary(Espanyol\_puason\_model2)$

Call:
$glm(formula = br\_golova\_na\_mecu\_tima \sim br\_golova\_protivnika + prosecan\_broj\_bodova\_tima + prosecan\_broj\_bodova\_protivnika + prosecan\_br\_golova\_tima + prosecan\_br\_golova\_protivnika + factor(tip\_tima\_na\_mecu), family = poisson, data = dataset\_Espanyol)$

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0562	-0.8590	-0.1037	0.5190	1.9585

Coefficients:

(Dispersion parameter for poisson family taken to be 1)

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	0.5567	0.5126	1.086	0.277
br_golova_protivnika	0.1634	0.1123	1.455	0.146
prosecan_broj_bodova_tima	0.5452	0.4358	1.251	0.211
prosecan_broj_bodova_protivnika	-0.2087	0.4692	-0.445	0.656
prosecan_br_golova_tima	-0.1555	0.2490	-0.624	0.532
prosecan_br_golova_protivnika	-0.6462	0.5829	-1.109	0.268
factor(tip_tima_na_mecu)1	-0.1754	0.1633	-1.074	0.283

Null deviance: 49.223 on 35 degrees of freedom
Residual deviance: 33.209 on 29 degrees of freedom
AIC: 106.38

Number of Fisher Scoring iterations: 6 .

Пошто не добијамо значајне параметре овим моделом дефинисали смо следећи модел 4 који садржи предикторе: број голова противника и класу противника.

$> summary(Espanyol\_puason\_model4)$

Call:
$glm(formula = br\_golova\_na\_mecu\_tima \sim br\_golova\_protivnika + factor(klasa\_protivnika), family = poisson, data = dataset\_Espanyol)$

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5862	-0.6228	-0.2387	0.2500	1.9031

Coefficients:

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 49.223 on 35 degrees of freedom
Residual deviance: 27.292 on 32 degrees of freedom
AIC: 94.465

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-0.96744	0.40970	-2.361	0.018209 *
br_golova_protivnika	0.33477	0.09812	3.412	0.000645 ***
factor(klasa_protivnika)1	-1.58590	0.67624	-2.345	0.019019 *
factor(klasa_protivnika)2	0.38889	0.38835	1.001	0.316644

—

Number of Fisher Scoring iterations: 5

Као што можемо видети број голова тима Еспањол може се предвидети уз помоћ предиктора: број голова противника и класа противника. Проверили смо да ли постоји прекорачење дисперзије овако дефинсианог модела на основу следеће функције:

$> dispersiontest(Espanyol\_puason\_model4)$

Overdispersion test

data: Espanyol_puason_model4
z = -1.5968, p-value = 0.9448
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
0.747821

Пошто је p вредност теста већа од 0.05 и износи 0.94 , прихватамо нулту хипотезу да не постоји прекорачење дисперзије. С обзиром да прекорачење дисперзије не постоји нема потребе да се Пуасонов модел мења како ни Квазипуасоновим моделом тако ни негативан Биномним моделом. На основу свега реченог модел 4 најбоље описује очекивани број голова тима Еспањол. Пуасонов регресиони модел за тим Еспањол задовољава следећу једнакост:

$$\log(\mu_i)\, 0,9674 + 0.335X_1 - 1,586X_2 + 0,389X_3,$$

где X_1 представља број голова противника, X_2 фактор класе противника 1 и X_3 фактор класе противника 2 .
Моделом 4 добијамо да је очекивани број голова тима Еспањол μ_i једнак 0.531. Вероватноћа да тим Еспањол оствари 0 голова једнака је 0.588, односно 1 гол 0.312. Модел 4 предвиђа да ће тим Еспањол на

посматраном мечу против тима Валенсије постићи 0 голова.
Аналогним поступком дефинисан је Пуасонов регресиони модел за тим Валенсије. Број голова тима Валенсије предвиђали смо уз помоћ предиктора просечан број бодова тима и просечан број бодова противника. На тај начин дефинисали смо $model\_4$ који је најбоље описивао посматране податке.

$> summary(Valencia\_puason\_model4)$

Call:
$glm(formula = br\_golova\_na\_mecu\_tima \sim prosecan\_broj\_bodova\_tima + prosecan\_broj\_bodova\_protivnika, family = "poisson", data = dataset\_Valencia)$
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0830	-1.1153	-0.2542	0.6595	1.4509

Coefficients:

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	0.3444	0.3139	1.097	0.27264
prosecan_broj_bodova_tima	0.8061	0.3917	2.058	0.03961 *
prosecan_broj_bodova_protivnika	-0.6772	0.2446	-2.769	0.00563 **

—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 46.084 on 35 degrees of freedom
Residual deviance: 37.431 on 33 degrees of freedom
AIC: 104.12

Number of Fisher Scoring iterations: 5 .

Да ли постоји прекорачење дисперзије модела проверили смо на основу следеће функције:

$> dispersiontest(Valencia\_puason\_model4)$

Overdispersion test

data: $Valencia\_puason\_model4$
$z = -1.1691, p - value = 0.8788$
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
0.8091217

Како је p вредност је већа од 0.05 и износи 0,87, прихватамо нулту хипотезу да не постоји прекорачење дисперзије. Овако дефинисан модел 4 најбоље описује податке на основу најмање вредности AIC-а као и једноставности. Пуасонов регресиони модел за тим Валенсије задовољава следећу једнакост:

$$\log(\mu_i) = 0,344 + 0,806X_1 - 0,677X_2 \ ,$$

где X_1 представља просечан број бодова тима и X_2 просечан број бодова противника.
Очекивани број голова тима Валенсије μ_i једнак је 1.45. Вероватноћа да ће тим Валенсије освојити 0 голова на мечу једнака је 0.57, 1 гол 0.24 односно више од једног гола 0.19. Дакле наведени модел највише шанси даје да ће тим $Valencia$ постићи нула голова на мечу.

Аналогно претходно описаном поступку за меч Еспањол - Валенсија, дефинисан је Пуасонов регресиони модел за предвиђање резултата следећих утакмица: Бетис - Атлетико Мадрид и Лас Палмас - Барселона. За тим Бетис који је домаћин у мечу Бетис - Атлетико Мадрид добијена је следећа једнакост: :

$$\log(\mu_i) = 0.5562 - 0.3929X_1 \ ,$$

где X_1 представља просечан број бодова противника. Овако дефинисаним моделом добили смо да је очекивани број голова тима Бетис једнак 0.865. Највећа вероватноћа 0.42 је да посматрани тим неће имати голова на мечу, затим 0.36 да ће дати један гол.
За тим Атлетико Мадрид који је био гост у посматраном мечу добијена је следећа једнакост:

$$\log(\mu_i) = 0.3903 + 1.2109X_1 - 0.6262X_2 - 0.6875X_3 \ ,$$

где X_1 представља просечан број бодова тима, X_2 просечан број бодова противника и X_3 просечан број голова тима. На основу наведеног добили смо да је очекивани број голова тима Атлетико Мадрид 1.96. Дакле

наведени модел даје највећу вероватноћу 0.275 да ће тим дати један гол, док је вероватноћа да ће дати два гола једнака 0.271.
Овако дефинисан Пуасонов регресиони модел даје највеће шансе да ће резултат утакмице Бетис - Атлетико Мадрид бити 0-1.
За тим Лас Палмас који је био домаћин на мечу против Барселоне најбољи Пуасонов регресиони модел има следећу једнакост:

$$\log(\mu_i) = 0.9357 - 0.4727X_1 - 0.5670X_2 \ ,$$

где X_1 представља просечан број бодова тима и X_2 тип тима на мечу. Дакле очекивани број голова тима Лас Палмас је 0.758, односно највећа вероватноћа 0.4687 је да ће тим дати нула голова.
Пуасонов регресиони модел за тим Барселона, који је био гост на мечу против Лас Палмаса је:

$$\log(\mu_i) = 1.4651 - 0.3048X_1 \ ,$$

где X_1 представља просечан број голова противника. Дакле очекивани број голова тима Барселона је 2.5699, односно највећа вероватноћа 0.253 да ће тим дати два гола на посматраном мечу.
На основу наведеног Пуасонов регресиони модел предвиђа 0-2 као резултат меча Лас Палмас-Барселона.

6.4 Примена дводимензионалног Пуасоновог регресионог модела

Основна идеја дводимензионалног Пуасоновог регресионог модела је да постоји ефекат игре на домаћем терену и да он треба бити урачунат у модел. Овај параметар се директно рачуна из пакета *bivpois*() који ћемо надаље користити. Наведене моделе поставили су први Димитрис Карлис и Јоанис Дзуфрас 2000 године. Они су дефинисали дводимензионални Пуасонов регресиони модел за који важе следеће једнакости:

$$\begin{gathered}(X_i, Y_i) \ \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \\ \log(\lambda_{1i}) = \mu + home + att_{h_i} + def_{g_i} \\ \log(\lambda_{2i}) = \mu + att_{g_i} + def_{h_i} \ ,\end{gathered}$$

где $i = 1, 2, 3, 4, , n$ представља редни број меча, h_i и g_i тим домаћина односно госта у мечу i, X_i и Y_i број голова домаћина односно госта, λ_{1i} и λ_{2i} очекивани број голова домаћина односно госта i, μ је константан параметар, $home$ представља ефекат игре на домаћем терену и att_k и def_k су нападачке односно одбрамбене способности тима k.

Основни модел који су разматрали је да у утакмици учествују два тима приближно истих снага, на неутралном терену. Затим су дефинисали параметар напада и одбране као одступања од просечне нападачке односно одбрамбене способсности тима.[1] За параметар коваријације λ_{3i}, разматрали су различите могућности линеарног предиктора који задовољава следећу једнакост:

$$\log(\lambda_{3i}) = \beta^{con} + \gamma_1 \beta_{h_i}^{home} + \gamma_2 \beta_{g_i}^{away} ,$$

где је β^{con} константан параметар, ${\beta_{hi}}^{\text{home}}$ и ${\beta_{gi}}^{\text{away}}$ параметри који зависе од тима домаћина односно госта респективно. Параметри γ_1 и γ_2 су једноставни бинарни индикатори, који узимају вредност 0 или 1 у зависности од модела који дефинишемо. Нпр. уколико је $\gamma_1 = \gamma_2 = 0$ тада посматрани модел има константну коваријацију, док уколико је $(\gamma_1, \gamma_2) = (1, 0)$ тада коваријација модела зависи искључиво од тима домаћина.... Параметар λ_3 може се интерпретирати као случајан ефекат који утиче на услове игре. [1]
Дводимензионални Пуасонов модел може се једноставно изменити уколико: постоје додатне информације о тимовима, нападачке способности тимова нису исте у игри код куће и гостима и уколико ефекат игре код куће варира од тима до тима.

Основни модел, раније дефинисан применили смо на подацима 36 недеља шпанске Примера лиге сезоне 2016/2017. Како бисмо дефинисали модел коришћен је статистички софтвер R, у коме је дефинисан пакет *bivpois*. Оно што је генерално проблем са Пуасоновим регресионим моделима је већи број нерешених резултата који често покваре модел.

За почетак треба дефинисати како параметар напада $c(home.team, away.team)$ тако и параметар одбране $c(away.team, home.team)$ [1].
Параметри λ_1 и λ_2 задовољавају следећу једнакост:

$$forma <- \sim c(Home.team, Away.team) + c(Away.team, Home.team).$$

За кодирање променљивих користимо следеће:

$$options(contrasts = c(\text{"}contr.sum\text{"}, \text{"}contr.poly\text{"})).$$

Дакле сума свих коефицијената је 0, наведена функција *contr.sum* даје ортогоналне контрасте када поредимо сваки ниво са укупним очекивањем. Односно коефицијент нпр. тима Еспањол се добија као супротан знак суме свих осталих коефицијената, док је интерпретација сваког коефицијента заправо одступање од способности напада (одбране)

сваког тима у односу на очекивање напада (одбране).
Како бисмо дефинисали дводимензионални Пуасонов регресиони модел користили смо функцију $lm.bp()$, која је саставни део пакета *bivpois* и њен основни облик је :

$lm.bp(l1, l2, l1l2 = NULL, l3 =\sim 1, data, common.intercept = FALSE, zeroL3 = FALSE, maxit = 300, pres = 1e-8, verbose = getOption("verbose"))$.

У претходној дефиницији дводимензионалног Пуасоновог модела параметри су:
- $l1$ представља модел облика $x \sim X_1 + X_2 + \ldots + X_p$ за параметар $log(\lambda_1)$
- $l2$ представља модел облика $y \sim X_1 + X_2 + \ldots + X_p$ за параметар $log(\lambda_2)$
-$l1l2$ представља модел облика $\sim X_1 + X_2 + \ldots + X_p$ за заједнички параметар од $log(\lambda_1)$ и $log(\lambda_2)$
-$zeroL3$ је логички аргумент који контролише да ли вредност параметра λ_3 узима вредност 0 или не.[1]

Дефинисали смо четири дводимензионална Пуасонова регресиона модела у зависности од коваријационог коефицијента λ_3. Различитост модела огледа се у чињеници да ли λ_3 зависи од: тима домаћина, тима госта, од оба тима или ниједног тима који посматрамо. Наведени модели задовољавају следеће једнакости:

$model\_2 < -lm.bp(ghome \sim 1, gaway \sim 1, l1l2 = forma, data = podaci)$

$model\_3 < -lm.bp(ghome \sim 1, gaway \sim 1, l1l2 = forma, l3 =\sim Home.team, data = podaci)$

$model\_4 < -lm.bp(ghome \sim 1, gaway \sim 1, l1l2 = forma, l3 =\sim Away.team, data = podaci)$

$model\_5 < -lm.bp(ghome \sim 1, gaway \sim 1, l1l2 = forma, l3 =\sim Home.team + Away.team, data = podaci)$.
[5]

Модел 2 има константан коефицијент коваријације, односно $\gamma_1 = \gamma_2 = 0$. У моделу 3 коефицијент коваријације зависи искључиво од тима домаћина, односно $\gamma_1 = 1$ $\gamma_2 = 0$. Коефицијент коваријације зависи искључиво од тима госта, односно $\gamma_1 = 0$ $\gamma_2 = 1$ у моделу 4, док у моделу

5 зависи како од тима домаћина тако и од тима госта.
Као код регуларног Пуасоновог регресионог модела тако и код дводимензионалног, модел описују две вредности AIC и BIC, које користимо за упоређивање посматраних модела. Наведене вредности задовољавају следеће једнакости:

$$AIC = -2\log(\text{вероватноће}) + 2p$$
$$BIC = -2\log(\text{вероватноће}) + plog(n)\ .$$

За почетак требамо упоредити који од наведена четири модела најбоље описује посматране податке шпанске лиге, а затим да одредимо који резултат предвиђа тако добијени модел.
Следећа табела приказује дефинисане параметре за посматране моделе:

Naziv modela	Osobine modela	Vrednost log(verovatnoce)	Vrednost AIC	Vrednost BIC
model_2	3=const.	-1020.699	2123.132	2310.767
model_3	3 zavisi od tima domac1ina	-1013.194	2145,771	2420,359
model_4	3 zavisi od gosta	-1013.431	2146,041	2420,629
model_5	3 zavisi kako od domac1ina tako i od gosta	-999.8093	2156,525	2518,066 .

Као што можемо приметити најбољу вредност параметра AIC има $model\_2$, а затим и $model\_3$. Ово значи да заправо једноставнији модели најбоље описују податке. Претпоставља се да је то последица великог броја нерешених мечева. У следећој табели приказаћемо параметре напада и одбране за $model\_2$ и $model\_3$ који су оцењени као најбољи. Наведене параметре добићемо из функција:

$$round(model\_2\$beta1, 3) \text{ односно } round(model\_3\$beta1, 3).$$

Које за резултат дају константу као и вредности параметра напада ($attack$) односно одбране ($defence$), у облику $Home.team..Away.team1$ као и $Away.team..Home.team1$ где је 1 редни број тима Алавес. Можемо приметити такође да у резултату наведених функција нема параметара за редно 20. тим односно тим Виљареал, за који ћемо наведене параметре добити на следећи начин:

-$sum(model\_2\$coef[2:20])$(параметар напада у моделу 2)
-$sum(model\_2\$coef[21:39])$(параметар одбране у моделу 2)

односно

-$sum(model\_3\$coef[2:20])$(параметар напада у моделу 3)
-$sum(model\_3\$coef[21:39])$(параметар одбране у моделу 3).

Константе које имамо у дефиницији у једнакостима $log(\lambda_1)$, $log(\lambda_2)$ и $log(\lambda_3)$ односно μ, μ_i β^{con} добијене су редом као резултати следећих функција како за $model\_2$ тако и за $model\_3$:

$round(model\_2\$beta1, 3)$(вредност $intercept$)
$round(model\_2\$beta2, 3)$(вредност $intercept$)
$round(model\_2\$beta3, 3)$
односно
$round(model\_3\$beta1, 3)$(вредност $intercept$)
$round(model\_3\$beta2, 3)$(вредност $intercept$)
$round(model\_3\$beta3, 3)$(вредност $intercept$).

Вредност коваријационог коефицијента λ_3 добијена је као резултат функције:

$exp(model\_2\$beta3)$

односно

$exp(model\_3\$beta3)(vrednostintercept)$.

Константе p и θ добијене су као резултат следећих функција:

$model\_2\$p, model3\$p$

односно

$model\_2\$theta, model\_3\$theta$.

Ефекат домаћег терена добили смо као резултат следеће функције:

$model\_2\$beta1[1] - model\_2\$beta2[1]$

односно

$model\_3\$beta1[1] - model\_3\$beta2[1]$.

На основу свих дефинисаних параметара добијена је следећа табела за меч који смо прво предвиђали, а то је Еспањол - Валенсија. Аналогним поступком можемо добити вредности и за осталих 18 тимова.

Rb.	Tim	model_2		model_3	
		attack	defence	attack	defence
9	Espanyol	-0,077	-0,129	-0,154	-0,205
19	Valencia	-0,085	0,216	-0,079	0,226
Vrednost konstante μ u jednakosti za λ_1		0,253		0,228	
Vrednost konstante μ u jednakosti za λ_2		-0,049		-0,077	
Vrednost konstante (β^{con}) u jednakosti za λ_3		-1,762		-5,180	
λ_3		0,172		0,00563	
Efekat domaceg terena		0,302		0,305	

На основу вредности добијених у претходној табели дводимензионални Пуасонов регресиони модел 2 даје следеће вредности:

$$\log(\lambda_1) = \mu + home + att_{h_i} + def_{g_i} = 0,253 + 0,302 - 0,077 + 0,216 = 0,694 \text{ односно } \lambda_1 = 2,002$$
$$\log(\lambda_2) = \mu + att_{g_i} + def_{h_i} = -0,049 - 0,085 - 0,129 = -0,263 \text{ односно } \lambda_2 = 0.7687$$
$$\log(\lambda_3) = \beta^{con} + \gamma_1\beta_{h_i}^{home} + \gamma_2\beta_{g_i}^{away} = -1,762 \text{ односно } \lambda_3 = 0,17$$

односно *model_3*

$$\log(\lambda_1) = \mu + home + att_{h_i} + def_{g_i} = 0,228 + 0,305 - 0,154 + 0,226 = 0.605 \text{ односно } \lambda_1 = 1.831$$
$$\log(\lambda_2) = \mu + att_{g_i} + def_{h_i} = -0,077 - 0,079 - 0,205 = -0,361 \text{ односно } \lambda_2 = 0.697$$
$$\log(\lambda_3) = \beta^{con} + \gamma_1\beta_{h_i}^{home} + \gamma_2\beta_{g_i}^{away} = -5,180 + 1 * 4,115 = -1,\ 065 \text{ односно } \lambda_3 = \ 0,345 \quad .$$

Модел 2 даје једнаке шансе следећим резултатима 1-0 и 2-0 , док модел 3 највеће шансе даје 1-0. Вероватноће свих резултата меча дате су следећим:
$> round(bivpois.table(6, 5, lambda = c(2.002, 0.7687, 0.17)), 3)$

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.053	0.041	0.016	0.004	0.001	0.000
[2,]	0.106	0.090	0.038	0.011	0.002	0.000
[3,]	0.106	0.099	0.046	0.014	0.003	0.001
[4,]	0.071	0.072	0.036	0.012	0.003	0.001
[5,]	0.035	0.039	0.021	0.007	0.002	0.000
[6,]	0.014	0.017	0.010	0.004	0.001	0.000
[7,]	0.005	0.006	0.004	0.002	0.000	0.000

$> round(bivpois.table(6, 5, lambda = c(1.831, 0.697, 0.345)), 3)$

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]
[1,]	0.057	0.039	0.014	0.003	0.001	0.000
[2,]	0.104	0.092	0.039	0.011	0.002	0.000
[3,]	0.095	0.102	0.051	0.016	0.004	0.001
[4,]	0.058	0.073	0.043	0.016	0.004	0.001
[5,]	0.026	0.038	0.026	0.011	0.003	0.001
[6,]	0.010	0.016	0.012	0.006	0.002	0.000
[7,]	0.003	0.005	0.005	0.002	0.001	0.000

Аналогно наведеним дводимензионалним Пуасоновим моделима за меч Еспањол-Валенсија урађени су модели и за мечеве Бетис - Атлетико Мадрид и Лас Палмас - Барселона. Добијене вредности дате су следећом табелом:

Rb.	Tim	model_2		model_3	
		attack	defence	attack	defence
5	Betis	-0.266	-0.256	-0.199	0.260
2	Atl. Madrid	0.003	-0.278	-0.083	-0.400
11	Las Palmas	-0.093	0.351	0.166	0.413
4	Barcelona	0.870	-0.397	0.894	0.367
Vrednost konstante μ u jednakosti za λ_1		0,253		0,228	
Vrednost konstante μ u jednakosti za λ_2		-0,049		-0,077	
Vrednost konstante (β^{con}) u jednakosti za λ_3		-1,762		-5,180	
λ_3		0,172		0,00563	
Efekat domaceg terena		0,302		0,305	

На основу вредности датих у табели и претходно описаном начину за меч Еспањол - Валенсија, модел 2 мечу Бетис - Атлетико Мадрид предвиђа резултат 1-0 са вероватноћом 0.148, док модел 3 предвиђа резултат 1-1 са вероватноћом 0.111. Уколико пак посматрамо меч Лас Палмас - Барселона модел 2 каже да је највероватнији резултат 1-3 са вероватноћом 0.075 као и модел 3 са истом вероватноћом. Следећом табелом дати су упоредни резултати посматрана три меча :

Меч	Стварни резултат	Предв. једно.Пуас.	Предв. дво. Пуас. моде. 2	Предв. дво. Пуас. моде. 3
Espanjol - Valensija	0 - 1	0 - 0	1 - 0	1 - 0
Betis - Atletiko Madrid	1 - 1	0 - 1	1 - 0	1 - 1
Las Palmas - Barselona	1 - 4	0 - 2	1 - 3	1 - 3

Из наведене табеле прво што се запази је једино тачно предвиђање дводимезионалног Пуасоновог модела 3 за резултат меча Бетис - Атлетико Мадрид. Наведено такорећи погрешно предвиђање тачног броја голова тимова условљено је: малим бројем мечева које смо посматрали, великим бројем нерешених резултата у посматраној сезони као

и једнакошћу треће класе тимова по снази. Уколико погледамо матрице вероватноћа посматраних модела, стварни резултати мечева имају приближно једнаке вероватноће са највећом вероватноћом посматраног модела. Спортске кладионице поред наведених модела користе и резултате претходних сезона тимова који су играли, као и огромно практично искуство кладионичара који одговарајуће вероватноће трансформишу у квоте. Једнодимензионални Пуасонов регресиони модел је за нијансу боље предвидео победнике посматраних мечева, док је дводимензионални тачније предвидео број голова тимова.

Поглавље 7

Закључак

У раду смо дали дефиницију уопштених линеарних модела. Након упознавања са уопштеним карактеристикама модела, дефинисали смо Пуасонов регресиони модел и дводимензионални Пуасонов модел.

Наведене моделе дефинисали смо у циљу дефиниције регресионих модела за тим Еспањол и Валенсију, односно њиховог очекиваног броја голова у заједничком мечу. Затим смо аналогним поступцима рачунали очекивани број голова за учеснике следедећих мечева: Бетис - Атлетико Мардид и Лас Палмас - Барселона. Пуасонова регресија била је добар избор јер су у питању пребројиви подаци, где су догађаји ограничени у неком временском интервалу, односно у нашем случају 90 мин уколико меч нема продужетака и при чему су тимови међусобно независни.

Случај дводимензионалне Пуасонове регресије обрађен је готовим пакетом у статистичком софтверу R, који су дефинисали статистичари Димитрис Карлис и Јоанис Дзуфрас једни од најцитиранијих у области предвиђања исхода спортских мечева.

Након дефинисања модела Пуасонов регресиони модел као и дводимензионални нису са потпуном прецизношћу предвидели резултат посматрана сва три меча шпанске Примера лиге. Међутим уколико погледамо резултате матрице вероватноћа за све сценарије за сваки меч, приметићемо да Пуасонов регресиони модел прихватљиво добро предвиђа победника меча, док дводимензионални Пуасонов модел боље предвиђа број голова тимова на мечу. Наведени дводимензионални модели дали су добре резултате али је у нашем случају дводимензионални Пуасонов модел био прецизнији, што не мора бити случај и са осталим мечевима. Стручњаци у овој области сматрају да је код наведених модела проблем што или прецене или потцене нерешене резултате. На основу реченог математичари Димитрис Карлис и Јоанис Дзуфрас тврде да уколико има доста нерешених мечева у неком такмичењу дијагоналин-

флаторни модели дају најпрецизније резултате. Оно што је дефинитивно је да постоји утицај домаћег терена, односно да он утиче на крајњи исход меча као и параметри напада односно одбране тимова.

Литература

[1] Dimitris Karlis and Ioannis Ntzoufras, Analysis of sports data by using bivariate Poisson models, The Statistician (2003) 52, Part 3, pp. 381393

[2] Gavin Whitaker, The Bivariate Poisson Distribution and its Applications to Football, School of Mathematics and Statistics, May 5, 2011 Newcastle University

[3] James Adam Gardner,Modeling and Simulating Football Results, MATH5003 - Assignment in Mathematics, May 6, 2011

[4] Sanja Bojovic, Puasonova regresija i primene, Univerzitet u Novom Sadu, Prirodno-Matematichki fakultet, Departman za Matematiku i informatiku, Novi Sad, Jun 2014.

[5] Dimitris Karlis and Ioannis Ntzoufras, Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R, Journal of Statistical Software, September 2005; Volume 14, Issue 10.

[6] Dimitris Karlis and John Ntzoufras, Bayesian and Non-Bayesian Analysis of Soccer Data using Bivariate Poisson Regression Models, Department of Statistics Athens University of Economics and Dept. of Business Administration University of the Aegean, Kavala, April 2003

Биографија аутора

Рођења 05.08.1988. године у Крушевцу, где је завршила Гимназију. Након тога 2012.године завршила Математички факултет, смер статистичка, актуарска и финансијска математика, Универзита у Београду. Од 2013. године запошљена у компанији Моцарт, на пословима аналитичара. Уже области интересовања су : пословна интелигенција, складиштење података и израда модела за предвиђање резултата спортских такмичења. Удата, мајка једног детета.