UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Porovnání pěti významných evropských fotbalových soutěží pohledem statistika

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **doc. Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval(a): **Jakub Vodochodský**
Studijní program: B1103 Aplikovaná matematika
Studijní obor: Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Jakub Vodochodský

Název práce: Porovnání pěti významných evropských fotbalových soutěží pohledem statistika

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2024

Abstrakt: Cílem této práce je na základě týmových dat provést datovou analýzu nejlepších evropských fotbalových soutěží. Skupinu těchto lig nazýváme „velkou pětkou“. Největší pozornost je v práci věnována porovnání počtu vstřelených branek za sezónu v soutěžích „velké pětky“ modelem Poissonovy regrese. Dále pro vybrané kvantitativní veličiny provedeme testy závislosti na soutěžích buď analýzou rozptylu (ANOVA), nebo Kruskalovým-Wallisovým testem.

Klíčová slova: fotbal, nejlepší evropské soutěže, góly, Poissonova regrese

Počet stran: 83

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Jakub Vodochodský

Title: A comparison of five major European football leagues from the statistician's point of view

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2024

Abstract: The aim of this thesis is to perform a data analysis of the top European football competitions based on team data. We call the group of these leagues the Big Five. The main focus of the paper is on comparing the number of goals scored per season in the Big Five competitions using a Poisson regression model. Furthermore, for the selected quantitative variables, we perform tests of the dependence on the competitions by either analysis of variance (ANOVA) or Kruskal-Wallis test.

Key words: football, best European competitions, goals, Poisson regression

Number of pages: 83

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením pana doc. Mgr. Ondřeje Vencálka, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam tabulek

Seznam obrázků

Poděkování

Rád bych poděkoval vedoucímu práce doc. Mgr. Ondřeji Vencálkovi, Ph.D. za trpělivost a cenné rady, které mi při psaní této práce poskytnul. Také bych chtěl poděkovat své rodinně, která mi byla při psaní této práce velkou oporou.

Úvod

Fotbal je jedním z celosvětově nejpopulárnějších sportů, ne-li ten nejpopulárnější. V Evropě má tento sport ještě silnější postavení, jelikož zde vznikl a vydobyl si místo nejsledovanějšího kolektivního sportu. Coby dlouholetý fotbalový nadšenec jsem dostal nápad – porovnat tzv. „velkou pětku“ evropských fotbalových soutěží s využitím statistických metod. Nejdůležitější statistikou ve fotbale jsou góly, a proto nás může napadnout otázka. Jsou počty vstřelených branek za sezónu v soutěžích „velké pětky“ rozdílné? Na tuto a další otázky zkusíme najít odpovědi v této práci. K analýze využijeme týmová data z posledních dvou odehraných ročníků 21/22 a 22/23.

V první kapitole ve zkratce popíšeme samotné soutěže, poháry UEFA a z nich vycházející koeficienty sloužící k porovnání kvalit ligových soutěží. Ve druhé kapitole představíme datové sady a vysvětlíme, co jednotlivé proměnné znamenají. Navíc pro vybrané veličiny vykreslíme grafické zobrazení v podobě krabicových grafů a popíšeme je. V teoretické části představíme analytické metody, které využijeme k analýze dat. V poslední kapitole provedeme samotnou analýzu a nejzajímavější zjištění uvedeme v závěru práce.

Vzhledem k velkému počtu internetových zdrojů byly tyto zdroje pro větší přehlednost odděleny od seznamu literatury.

Kapitola 1

Nejlepší evropské fotbalové ligy

1.1 Pětice nejlepších evropských soutěží

Pětice nejvěhlasnějších a nejkvalitnějších evropských fotbalových soutěží tvoří tzv. „velkou pětku“. Do této prestižní společnosti patří anglická Premier League (**PL**), německá Bundesliga (**BL**), španělská LaLiga (**LL**), francouzská Ligue 1 (**LO**) a italská Serie A (**SA**). Pro soutěže budeme dále v textu používat pouze zkratky pro jednotlivé ligy.

Z těchto soutěží pochází nejkvalitnější evropské fotbalové týmy, které každoročně bojují o mistrovský titul v Champions League (Lize mistrů), nejprestižnějším evropském klubovém poháru. Fakta, která číselně potvrzují výsadní postavení těchto soutěží, uvedeme později v této kapitole. Z Obrázku 1.1 můžeme vidět, že za posledních 10 let nevyhrál Ligu mistrů žádný tým z méně kvalitní soutěže.

Nejdříve popíšeme samotnou strukturu těchto soutěží. Všechny zmíněné ligy se nachází na samotném výkonnostním vrcholu pyramidy ve svých zemích a soutěže jsou hrány ve formátu každý s každým na dva zápasy, kdy jeden zápas je hrán na domácím hřišti a druhý na stadiónu soupeře. Sezóna

Obrázek 1.1: Vítězové Ligy mistrů za posledních 10 let, zdroj: [16, Sceenshot]

v soutěžích „velké pětky“ trvá od srpna do května příštího roku. Další společnou vlastností je, že soutěže jsou otevřené – tzn. že po konci každé sezóny určitý počet nejhůře bodujících týmů sestupuje do druhé ligy (nebo hraje baráž o udržení) a stejný počet postupuje z 2. ligy výše, aby se zachoval stálý počet týmů. Sestupový klíč se může lišit v závislosti na soutěži, ale princip zůstává stejný.

Nejzásadnější odlišností, kterou v dalších kapitolách budeme muset zohlednit, je počet týmů v jednotlivých soutěžích, kdy **PL**, **LL** a **SA** hraje 20 týmů. Zatímco německou **BL** hraje pouze 18 týmů. Francouzskou **LO** dlouho hrálo 20 týmů, avšak nově od sezóny 23/24 byl počet družstev snížen na 18.

1.2 Poháry a koeficienty UEFA

Celý evropský fotbal zastřešuje organizace UEFA, která má na starosti evropské reprezentační a klubové soutěže, rozděluje odměny, hlídá dodržování pravidel a prodává vysílací práva ke svým soutěžím.

„UEFA je největší ze šesti kontinentálních konfederací pod FIFA. Je to nejsilnější konfederace z hlediska bohatství a vlivu na klubové úrovni. Většina z nejlepších hráčů světa hraje v Anglii, Španělsku, Německu, Itálii a Francii.“[7]

Pro porovnání jednotlivých klubů a soutěží vytvořila UEFA koeficienty

zakládající se na výkonech v evropských pohárech. Dělíme je na:

- klubový
- ligový

Nejdříve ve zkratce uvedeme poháry UEFA, poté se vrátíme k výpočtu samotných koeficientů. Poháry dělíme do tří výkonnostních úrovní:

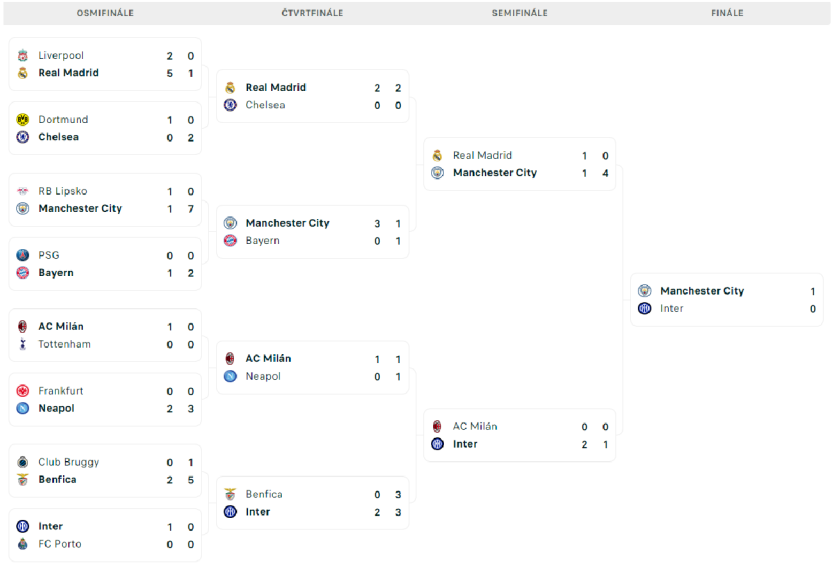
1. Liga Mistrů (CL) – nejprestižnější
2. Evropská liga (EL)
3. Evropská konferenční liga (ECL) – hrána až od sezóny 2021/22

Všechny tři soutěže mají velmi podobnou strukturu, a proto bude jednodušší vypsat základní podobnosti v bodech:

- rozdělení na část hlavní (skupina, play-off) a kvalifikační
- na podzim se hrají skupinové fáze a na jaře play-off
- skupinové fáze – rozlosování 8 skupin po 4 týmech
- play-off se do semifinále hraje na dva zápasy – jeden domácí, druhý na hřišti soupeře
- stejné bodové ohodnocení zápasů (za výhru, remízu, prohru) do koeficientů UEFA

Ze skupiny do play-off postupují vždy první dva celky z každé skupiny. U dvou kvalitativně horších soutěží EL a ECL je mezi skupinovou fázi a osmifinále vloženo předkolo play-off, které hrají celky z druhých míst ve skupinách. Vítězové skupin EL a ECL přímo postupují do osmifinále. Podrobné tabulky můžeme najít na webové stránce Livesport [16].

Obrázek 1.2: Play-off Chapions League sezóny 22/23, zdroj: [16, Sceenshot]



Pro úplnost musíme dodat, že ze skupiny CL a EL postupují do play-off i týmy ze třetího místa. Ovšem putují o pohár níže, tedy třetí tým ze skupiny CL putuje do předkola EL a třetí tým ze skupiny EL padá do předkola ECL. Pro představu, takhle vypadá play-off Champions League za sezónu 22/23 – viz Obrázek 1.2. Vidíme, že play-off CL je hráno na tři vyřazovací kola a finále. U EL a ECL by v pavouku bylo i vložené kolo celků ze druhých míst a sestupujících z kvalitativně vyšší soutěže.

Pro nový ročník 24/25 se chystají v pohárech UEFA velké změny, kdy ve skupinové fázi nahradí menší 4 členné skupiny jedna velká skupina. V tomto formátu dojde k navýšení počtu účastníků z původních 32 na 36. Také dojde ke zvýšení počtu zápasů ve skupině, a to z dosavadních 6 na 8. Poslední důležitou změnou je, že už týmy nebudou moci při nepostoupení do play-off sestoupit do kvalitativně nižší pohárové soutěže.

Nyní se můžeme vrátit k samotným koeficientům. Nejdříve představíme koeficient klubový, který dělíme do dvou kategorií, a to na 5letý a 10letý.

Tabulka 1.1: Tabulka maximálního počtu bodů do ligového koeficientu UEFA za sezónu

Bonusové body (červeně) a max. počet bodů ze zápasů	CL	EL	ECL
Skupinová fáze	4+12	12	12
Vítěz skupinové fáze (postoupivší)	-	4(2)	2(1)
Osmifinále	5+4	1+4	4
Čtvrtfinále	1+4	1+4	4
Semifinále	1+4	1+4	1+4
Finále	1+2	1+2	1+2
Maximální počet bodů	38	34	30

Důležitou roli hraje 5letý koeficient v nasazování týmů při losování klubových soutěží, který počítá s výsledky z pěti posledních sezón v evropských pohárech, kdy se jednotlivé výsledky ze sezóny sčítají. Kupříkladu klubový koeficient Bayernu Mnichov na konci sezóny 22/23 počítáme jako součet za sezóny 18/19 – 22/23.

Bodové ohodnocení za zápasy je stejné jak ve skupině, tak v play-off. Tedy 2 body za výhru, 1 bod za remízu a 0 bodů za prohru. Každý z pohárů nabízí různé maximální bodové ohodnocení za sezónu, které se odvíjí od kvality soutěže, viz Tabulka 1.1[1]. Vidíme, že největším bodovým rozdílem CL a dvou nižších pohárů jsou bonusové body za účast ve skupině a postup do osmifinále. U EL a ECL je tento deficit lehce kompenzován vítězstvím ve skupinové fázi. Také je důležité zmínit, že nejsou udělovány žádné body za vložené kolo v EL a ECL.

Pro lepší ilustraci můžeme v Tabulce 1.2[2] vidět deset nejlépe hodnocených evropských týmů za posledních pět let a všech deset týmů patří do „velké pětky“.

Jako poslední zbývá dovysvětlit ligový koeficient UEFA a z něj vyplývající žebříček. Tento koeficient a na jeho základě vytvořený žebříček slouží k ur-

[1] Tabulka podle https://kassiesa.net/uefa/calc.html

[2] Část tabulky převzata z https://www.uefa.com/nationalassociations/uefarankings/club/seasons/

Tabulka 1.2: Žebříček deseti nejlepších týmů podle koeficientu UEFA na konci sezóny 22/23

Pozice	Klub	Země	18/19	19/20	20/21	21/22	22/23	Body
1	Man City	Anglie	25	25	35	27	33	145
2	Bayern	Německo	20	36	27	26	27	136
3	Chelsea	Anglie	30	17	33	25	21	126
4	Liverpool	Anglie	29	18	24	33	19	123
5	Real Madrid	Španělsko	19	17	26	30	29	121
6	Paris	Francie	19	31	24	19	19	112
7	Man United	Anglie	19	22	26	18	19	104
8	Juventus	Itálie	21	22	21	20	17	101
9	Barcelona	Španělsko	30	24	20	15	9	98
10	Roma	Itálie	17	11	24	23	22	97

čení počtu nasazovaných týmů z evropských ligových soutěží do kvalifikace nebo skupin pohárů UEFA. Podobně jako klubový koeficient vychází z pěti posledních sezón v evropských pohárech. Výpočet se zakládá na klubových koeficientech týmů s dvěma výjimkami v uznávání bodů, viz 1.2.1.

Tedy body všech klubů z jedné soutěže (např. **PL**) sečteme, poté tento součet vydělíme počtem nasazených týmů (počítají se i týmy z kvalifikací), a tím získáme hodnotu ligového koeficientu za jeden ročník. K tomuto výsledku musíme ještě přičíst hodnoty ligových koeficientu z předchozích čtyř sezón. Tedy např. hodnota koeficientu pro kvalifikaci do pohárů UEFA pro ročník 23/24 je součtem ze sezón 18/19 až 22/23. V Tabulce 1.3[3] vidíme počet bodů do ligového koeficientu **PL** za pohárovou sezónu 22/23. Významy sloupců jsou následující:

- pořadí – postavení podle zisku bodů za ročník 22/23 týmů i soutěží (nerovná se klubovému koeficientu, viz níže 1.2.1)
- qW, qD a qL značí počet vítězství, remíz a proher v kvalifikaci
- #W, #D a #L znamená počet výher, remíz a proher v hlavní části

[3]Část tabulky převzata z https://kassiesa.net/uefa/data/method5/ccoef2023.html

Tabulka 1.3: Body do ligového koeficientu PL za ročník 22/23

Pořadí	PL – 7 týmů	Cup	qW	qD	qL	#W	#D	#L	Bonus	Body	Průměr
1	Týmy		2	0	0	46	14	12	53	161,0	23,000
1	Manchester City	CL	0	0	0	8	5	0	12	33,0	
4	West Ham United	ECL	2	0	0	12	1	0	4	31,0	
13	Manchester United	EL	0	0	0	8	2	2	4	22,0	
16	Chelsea	CL	0	0	0	5	1	4	10	21,0	
21	Liverpool	CL	0	0	0	5	0	3	9	19,0	
23	Tottenham Hotspur	CL	0	0	0	3	3	2	9	18,0	
28	Arsenal	EL	0	0	0	5	2	1	5	17,0	

Tabulka 1.4: Nejlépe hodnocené ligové soutěže UEFA na konci sezóny 22/23

Pořadí	Liga	18/19	19/20	20/21	21/22	22/23	Body
1	Anglická	22,642	18,571	24,357	21,000	23,000	109,570
2	Španělská	19,571	18,928	19,500	18,428	16,571	92,998
3	Německá	15,214	18,714	15,214	16,214	17,125	82,481
4	Italská	12,642	14,928	16,285	15,714	22,357	81,926
5	Francouzská	10,583	11,666	7,916	18,416	12,583	61,164
6	Nizozemská	8,600	9,400	9,200	19,200	13,500	59,900
7	Portugalská	10,900	10,300	9,600	12,916	12,500	56,216
8	Belgická	7,800	7,600	6,000	6,600	14,200	42,200

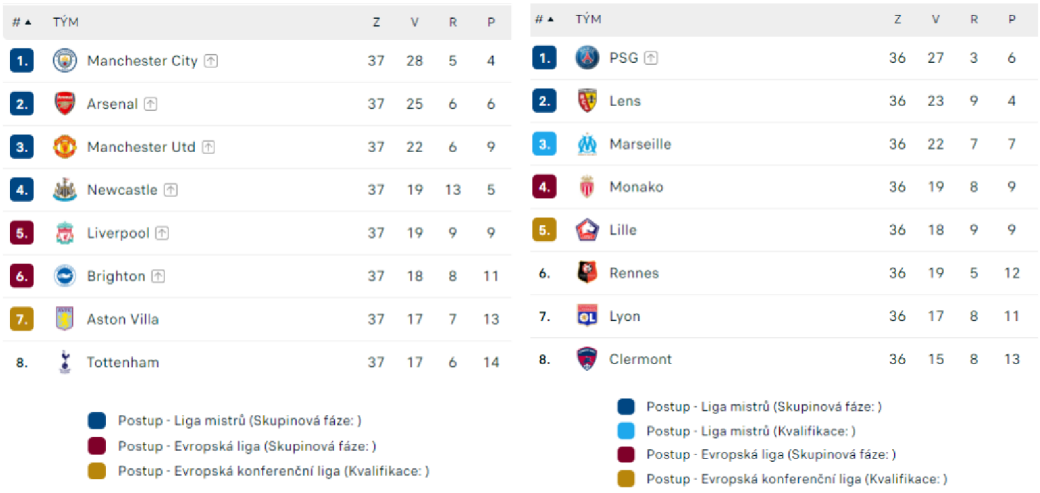
Nyní můžeme přistoupit k žebříčku soutěží, ze kterého vychází nasazování týmu do pohárů UEFA. Pořadí ke konci sezóny 22/23 je znázorněno v Tabulce 1.4[4]. Můžeme vidět, že na prvních 5 místech se umístily nám dobře známe ligy.

1.2.1 Kvalifikace a výjimky v uznávání bodů do ligového koeficientu

Již víme, že nasazování nejlepších týmů ze všech evropských soutěží do hlavní i kvalifikační části pohárů UEFA závisí na pořadí v žebříčku ligového koeficientu. Ovšem systém nasazování klubů je velmi komplexní a v této práci ho podrobněji popisovat nebudeme. Pouze pro ilustraci uvedeme roz-

[4]Část tabulky převzata z https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2023

Obrázek 1.3: Kvalifikační pozice z PL a LO do evropských soutěží pro sezónu 23/24, zdroj: [16, Screenshot]



#	TÝM	Z	V	R	P
1.	Manchester City	37	28	5	4
2.	Arsenal	37	25	6	6
3.	Manchester Utd	37	22	6	9
4.	Newcastle	37	19	13	5
5.	Liverpool	37	19	9	9
6.	Brighton	37	18	8	11
7.	Aston Villa	37	17	7	13
8.	Tottenham	37	17	6	14

Postup - Liga mistrů (Skupinová fáze:)
Postup - Evropská liga (Skupinová fáze:)
Postup - Evropská konferenční liga (Kvalifikace:)

#	TÝM	Z	V	R	P
1.	PSG	36	27	3	6
2.	Lens	36	23	9	4
3.	Marseille	36	22	7	7
4.	Monako	36	19	8	9
5.	Lille	36	18	9	9
6.	Rennes	36	19	5	12
7.	Lyon	36	17	8	11
8.	Clermont	36	15	8	13

Postup - Liga mistrů (Skupinová fáze:)
Postup - Liga mistrů (Kvalifikace:)
Postup - Evropská liga (Skupinová fáze:)
Postup - Evropská konferenční liga (Kvalifikace:)

dílné kvalifikační pozice **PL** a **LO**. Z Obrázku 1.3 můžeme kupříkladu vidět nasazení nejlepších klubů z **PL** a **LO** do evropských pohárů 23/24. Vidíme, že celkový počet nasazených týmů je odlišný a s tím souvisí i kvalita pohárových soutěží.

Z **PL** putují rovnou do skupiny CL celkem čtyři celky, zatímco v **LO** jsou to pouze dva kluby s přímým postupem, navíc třetí tým **LO** čeká pouze kvalifikace. Toto je způsobeno rozdílným postavením v ligovém koeficientu, viz Tabulka 1.4. V této tabulce je na prvním místě s velkým náskokem **PL**, zato **LO** se nachází až na místě pátém, a tedy rozestup čtyř pozic na žebříčku tvoří velký rozdíl v nasazování klubů do pohárové Evropy.

Závěrem přistoupíme k výjimkám, ve kterých započítáváme body pouze do ligového koeficientu. Prvním případem jsou zápasy v EL a ECL ve fázi předkol play-off. Jak již bylo zmíněno, toto předkolo najdeme mezi skupinou a osmifinále – v Tabulce 1.1 tento řádek záměrně chybí. Druhou výjimkou jsou kvalifikační zápasy, které jsou ohodnoceny poloviční počtem bodů

z hlavní části, tedy jeden bod za výhru a půl bodu za remízu. Z toho plyne, že klubové koeficienty budou za sezónu z pravidla menší než bodový zisk do ligového koeficientu.

Z této kapitoly plyne, že ve všech představených metrikách jsou soutěže z „velké pětky“ vždy na samém vrcholu pomyslné pyramidy.

V této kapitole jsem mimo uvedené zdroje využíval informace z internetových stránek [6–16].

Kapitola 2

Data

2.1 Obecný popis datových souborů

K samotné komplexní analýze „velké pětky“ použijeme tři datasety, které obsahují statistiky týmů ze soutěží „velké pětky“ za ročníky 22/23 a 21/22. Zdrojem čerpání dat byla internetová stránka FBref ([17]), která nabízí velké množství dat popisující týmy z mnoha hledisek. Seznam je následující:

- obecná data
- střelba
- defenzíva

Celkový počet týmů ve zkoumaných ligách je 98 a jak již bylo zmíněno v části 1.1, tak u francouzské **LO** budeme v této práci pracovat s 20 týmy, jelikož ke snížení počtu účastníků došlo až v probíhajícím ročníku 23/24. Pro prvotní šetření postačí data ze sezóny 22/23, druhý soubor dat využijeme pouze pro případné ověření nezamítnutých hypotéz. Zbývá dodat, že jedinými společnými charakteristikami datasetů je tým (sloupec Squad) a liga (sloupec Comp).

Tabulka 2.5: Ukázka datová sady č. 1 - obecné statistiky

Squad	Comp	#Pl	Age	Poss	90s	Gls	GA	Ast	PK	PKatt	CrdY	CrdR	LgRk
Ajaccio	Ligue 1	36	29,1	43,2	38	22	74	12	6	9	86	10	18
Almería	La Liga	29	26,4	45,1	38	49	65	33	2	3	98	4	17
Angers	Ligue 1	33	25,7	46,9	38	31	81	18	4	4	69	5	20
Arsenal	Premier League	26	24,7	59,3	38	84	43	64	3	4	51	0	2
Aston Villa	Premier League	26	27	49,3	38	49	46	35	3	4	80	1	7
Atalanta	Serie A	27	27,4	49,9	38	64	48	42	6	8	83	3	5
Athletic Club	La Liga	26	27,4	51,6	38	46	43	33	5	7	81	5	8
Atlético Madrid	La Liga	27	28,5	50,6	38	68	33	51	1	1	92	8	3
Augsburg	Bundesliga	34	25,9	41,7	34	41	63	29	4	5	96	5	15
Auxerre	Ligue 1	33	27,3	43,3	38	32	63	17	7	9	54	6	17
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.2 Popis obecných dat

Veličiny obecných dat z Tabulky 2.5 můžeme rozdělit do několika skupin. Nejdříve popíšeme proměnné, které vyjadřují celkovou gólovou produkci za sezónu. Do této skupiny patří počet vstřelených gólů (**Gls**), počet asistencí (Ast), počet kopaných a proměněných penaltových kopů (PKatt a PK).

Dalším hlediskem je skladba kádru. Do této kategorie spadá počet hráčů (#Pl). Toto číslo vyjadřuje počet fotbalistů, kteří nastoupili alespoň k jednomu soutěžnímu ligovému zápasu. Dále je tu průměrný věk týmů (Age), který je vypočten jako vážený průměr, kdy váhami jsou odehrané minuty za celou sezónu. Tím pádem průměrný věk nejvíce zohledňuje nejvytíženější hráče.

Neméně důležitou složkou pro srovnání soutěží je počet udělených žlutých (CrY) a červených (CrR) karet. V souvislosti s udílením karetních trestů můžeme mluvit o disciplíně týmů a čisté hře. Rámcově uvedeme prohřešky proti pravidlům hry, které vedou ke žluté nebo červené kartě.

- žluté karty - tvrdý faul, nesportovní chování, zdržování hry nebo protesty proti rozhodčímu
- červené karty - brutální faul (ohrožující zdraví hráče) nebo zmaření zjevné brankové situace

Posledním důležitým faktem je, že pokud hráč dostane v zápase 2 žluté karty, tak podle pravidel následuje udělení karty červené a vyloučení do konce zápasu. Navíc tým vyloučeného hráče musí dohrát zápas v oslabení.

Zbývající statistiky se přímo nehodí do žádných dříve zmíněných skupin, a tak je vypíšeme zvlášť. Výčet je následující:

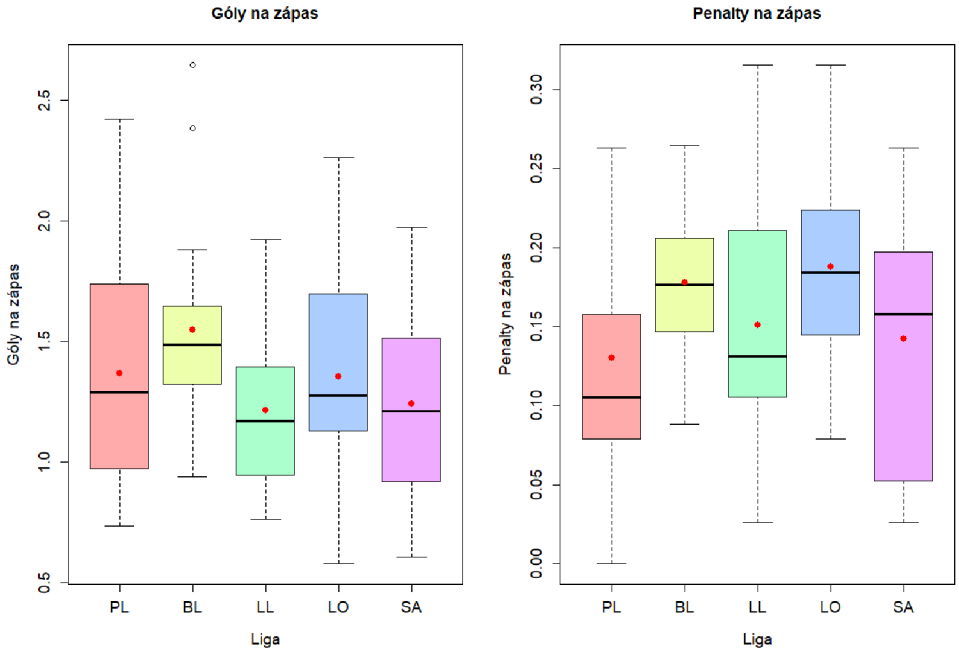
- procentuální držení míče (Poss) – uvedeno jako průměrná hodnota za celou sezónu
- počet zápasů (**90s**)
- počet inkasovaných gólů (GA)
- umístění týmů na konci soutěže (LgRk)

K veličinám GA a LgRk musíme dodat, že byly do obecných dat přidány z Overview „velké pětky“ ze stránek FBref ([17]).

2.2.1 Popisná statistika obecných dat

Pro nejzásadnější statistiky vykreslíme grafické zobrazení, abychom zjistili, zda jsou grafy odlišné v závislosti na soutěžích. U některých veličin postrádá smysl vykreslovat grafy, jelikož by neměly žádnou vypovídací hodnotu a byly by naprosto zbytečné. Mezi tyto statistiky zařadíme držení míče, počet zápasů, konečné umístění a počet obdržených branek. U prvních tří budou průměrné hodnoty pro všechny soutěže (medián, průměr) skoro totožné – ve dvou případech i stejné. U obdržených branek víme, že jsou pouze opakem vstřelených gólů. Pro nejzásadnější kvantitativní veličiny ze všech tří datových sad využijeme ke grafickému znázornění krabicové grafy (boxploty), protože obsahují nejvíce informací o vykreslených proměnných. Navíc

Obrázek 2.4: Boxploty počtu gólů a penalt na zápas

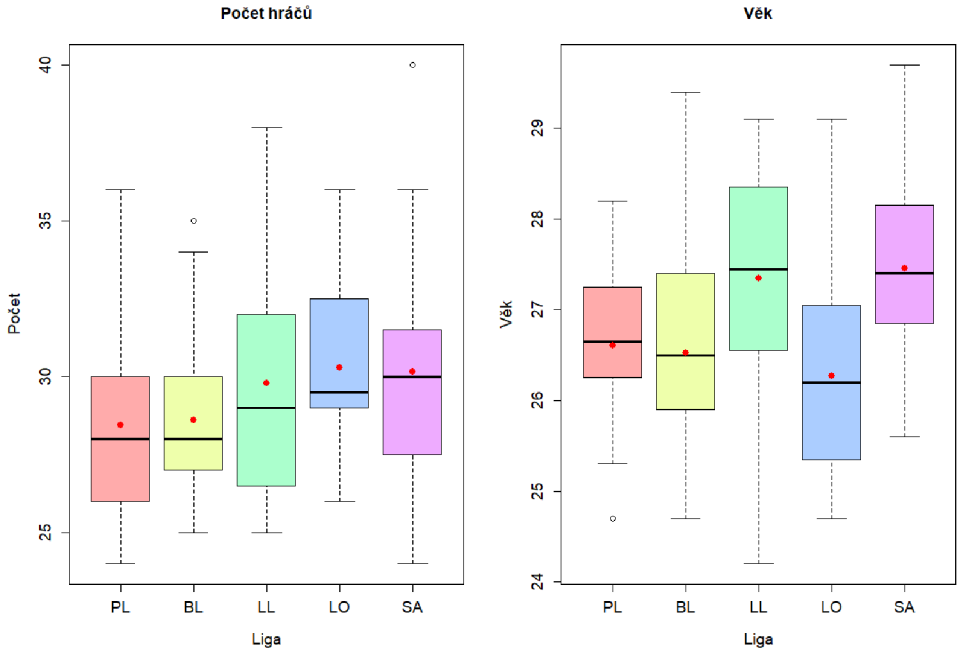


do krabicových grafů vykreslíme červené tečky, které značí průměrné hodnoty veličin pro jednotlivé soutěže.

Začneme nejvíce důležitým ukazatelem, a tím jsou góly. V samotném krabicovém grafu 2.4 je tato statistika přepočítána jako gól na zápas, jelikož má v tomto tvaru mnohem větší vypovídací hodnotu s ohledem na různý počet zápasů. Při pohledu do grafu můžeme vidět, že krabicové grafy gólů na zápas jsou celkem podobné. Jedinou výjimkou je krabice **BL**, která je posunutá do vyšších hodnot (medián je na hodnotě 1,5) a je nejméně roztažená. To znamená, že rozdíly v produktivitě týmů jsou nejmenší oproti ostatním soutěžím. Z této kompaktní skupiny vyčnívají dva kluby, které jsou výrazně vzdáleny od ostatních, což z nich dělá odlehlá pozorování. Jedná se o první dva celky konečné tabulky, jmenovitě Bayern Mnichov a Borrusia Dortmund.

Dalším v pořadí je krabicový graf 2.4 penalt na zápas (i neproměnné).

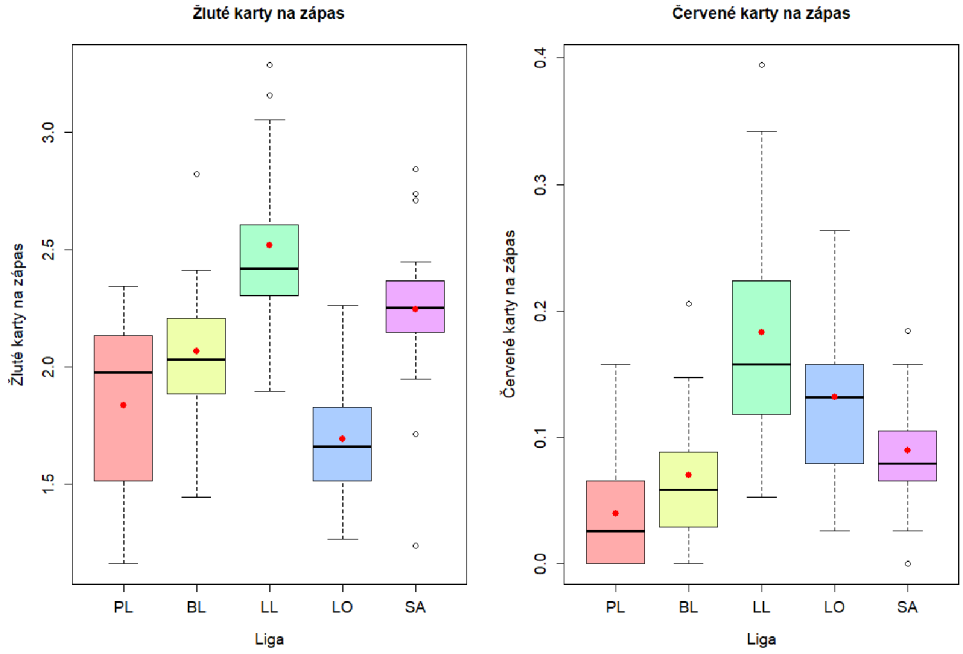
Obrázek 2.5: Boxploty počtu hráčů a věků soupisek



Krabicové grafy nejsou příliš zajímavé. Jediným vybočujícím krabicovým grafem je ten příslušný **PL**, který má nejnižší hodnotu mediánu (0,1). Ovšem průměr počtu penalt ve všech ligách osciluje kolem 0,15. Za zmínku stojí ještě říct, že graf pro **SA** je velmi roztažený k nižším hodnotám, i když je medián na podobné úrovni ostatních lig. Vysvětlením je, že největší koncentrace týmů je okolo mediánu a zbylé týmy s nižším počtem penalt na zápas patří stále do IQR.

Jako další vykreslíme krabicový graf 2.5 počtu hráčů. Znovu jsou krabice velmi podobné. Ovšem stojí za to zmínit, že **BL** týmy mají podobně početné kádry, i když odehrají za sezónu o 4 utkání méně. V podstatě 4 zápasy nehrají v kontextu celého ročníku žádnou roli. V datech se nacházejí také 2 odlehlá pozorování. Prvním je **BL** tým Schalke 04, druhým je klub z **SA** Sampdoria Janov.

Obrázek 2.6: Boxploty žlutých a červených karet na zápas



Pro krabice průměrného věku soupisek 2.5 toho není moc ke komentáři, jelikož grafy jsou opět velmi podobné. Pouze je nutné dodat, že nejvíce příležitostí dostávali mladí hráči v **LO** a suverénně nejmladším týmem **PL** byl londýnský Arsenal (odlehlé pozorování) s věkovým průměrem pod 25 let.

Zřejmě nejzajímavější jsou krabicové grafy 2.6 pro žluté a červené karty na zápas, protože obsahují velké množství odlehlých pozorování. Navíc krabice mají různé velikosti a hodnoty mediánů. Prvně začneme žlutými kartami, kde krabicové grafy soutěží můžeme rozdělit do 3 skupiny. V první je osamocená **LO**, ve které se udílelo nejméně žlutých karet – v průměru přes 1,5. Skupinu středu tvoří **PL** a **BL**. Hodnota mediánu je skoro identická a rovnající se 2. Zbylá dvojice **SA** a **LL** spadá do skupiny s nejvyššími počty žlutých karet na zápas. Z těchto dvou lig jsou více trestané týmy ze Španělska s hodnotou mediánu lehce pod 2,5 žlutých karet na zápas. Tím nejtrestanějším týmem

je Getafe, které dostávalo v sezóně 22/23 přes 3 žluté karty na zápas. Nejvíce odlehlých pozorování sledujeme u **SA**, ačkoliv je krabice této soutěže nejméně roztažená. V podstatě bychom mohli říci, že těchto 5 týmů se svojí disciplinovaností vymyká z italské soutěže.

Při pohledu do krabicového grafu 2.6 červených karet můžeme vidět, že krabice **PL** a **BL** jsou opět velmi podobné a mají nejnižší hodnoty medián̊u. U zbylých soutěží došlo ke změně, protože k nejtrestanějším týmům z krabicového grafu žlutých karet na zápas přidáváme **LO**, která měla krabici pro žluté karty na zápas posazenou nejníže, což je zajímavé. Již víme, že za 2 žluté karty v zápase následuje karta červená. To značí, že s rostoucím počtem žlutých karet by přirozeně měl vzrůst i počet červených. Tento předpoklad se projevuje u všech soutěží kromě **LO**. Vysvětlením by mohlo být, že týmy ve francouzské lize dostávaly ve větší míře rovnou červené karty. Nejvíce trestaným týmem byl Real Betis z **LL**, který za celou sezónu dostal 15 červených karet, což je obrovské číslo. V přepočtu na zápasy to znamená, že každý třetí zápas dohrával v oslabení.

2.3 Popis střeleckých dat

Pro vstřelení branky ve fotbale i obecně ve sportu platí jednoduché pravidlo – bez střely na bránu gól nelze vstřelit. Z toho plyne, že pro hlubší analýzu gólovosti soutěží využijeme podrobnější data o střelbě. Ukázku dat najdeme v Tabulce 2.5, kde jsou uvedeny proměnné – počet střel (Sh), počet střel na bránu (SoT) a jejich přepočítané hodnoty na zápas (Sh/90 a SoT/90). Dalšími ukazateli jsou úspěšnost střelby na bránu (SoT%) a průměrná vzdálenost střelby (Dist) v yardech[5]. Jelikož je jednotka průměrné vzdálenosti střelby

[5] $1 yard = 0,91 m$

Tabulka 2.6: Ukázka datové sady č. 2 - střelba

Squad	Comp	Sh	SoT	SoT%	Sh/90	SoT/90	Dist	xG	xGDiff
Ajaccio	Ligue 1	311	81	26	8,18	2,13	18,1	36,1	-14,1
Almería	La Liga	439	155	35,3	11,55	4,08	18,1	45,5	3,5
Angers	Ligue 1	367	121	33	9,66	3,18	17,5	40,9	-9,9
Arsenal	Premier League	589	194	32,9	15,5	5,11	16	71,9	12,1
Aston Villa	Premier League	427	145	34	11,24	3,82	18	50,2	-1,2
Atalanta	Serie A	506	168	33,2	13,32	4,42	16,9	57,7	6,3
Athletic Club	La Liga	541	157	29	14,24	4,13	17,1	54,2	-8,2
Atlético Madrid	La Liga	538	195	36,2	14,16	5,13	17	61,9	6,1
Augsburg	Bundesliga	354	101	28,5	10,41	2,97	18,2	34,7	6,3
Auxerre	Ligue 1	420	117	27,9	11,05	3,08	19,6	42,9	-10,9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

pro naše analýzy irelevantní, necháme veličinu v původních jednotkách.

2.3.1 Očekávané góly (xG)

V této části byl využit zdroj [18].

V posledních několika letech začaly fotbalové týmy mnohem více používat data a statistiky. Nejedná se pouze o pozápasové údaje, které můžeme kupříkladu vyhledat na webové stránce Livesport [16], ale i o komplexní scouting hráčů a vyhodnocení týmových/individuálních výkonů v zápasech. Kluby dokonce využívají datovou analýzu k přípravě na soupeře, kdy hlavním cílem této analýzy je objevit soupeřovi nejsilnější a nejslabší stránky. Očekávané góly (xG) jsou toho nejlepším příkladem. Od zavedení xG Samem Greenem ze společnosti Opta v roce 2012 se tato metrika stala jedním z nejoblíbenějších a nejrozšířenějších ukazatelů ve fotbalové analýze.

Očekávané góly vyjadřují kvalitu šance tím, že počítají pravděpodobnost vstřelení branek na základě informací o podobných kopech z minulosti. Jelikož je statistika xG pravděpodobnost, tak výsledná hodnota xG spadá do intervalu od 0 do 1, kdy 0 představuje šanci, kterou je nemožné proměnit. Na druhé straně 1 značí střelu, která by vždy měla skončit v brance. Z logiky

věci víme, že pravděpodobnost skórování z půlící čáry je oproti střele z pokutového území velmi malá. Systém xG nám umožňuje přiřadit těmto scénářům čísla. Předpokládejme například, že pro šanci ze vnitřku pokutového území je xG rovno 0,1 – tedy 10 %. To znamená, že v této situaci by měla 1 z 10 střel skončit gólem.

Problémem je, že na jeden zápas připadá v průměru 25 střel a většina soutěží má za víkend kolem 250 střel. Přiřazení pravděpodobnosti každé z těchto jedinečných situací je pro člověka nemožné, proto pro výpočet xG využívá Opta model strojového učení zvaný XGBoost (nebudeme blíže popisovat, jelikož se jedná o velmi sofistikovaný proces). Tento model vychází z historických dat 40 soutěží z let 2018-19 až 2021-22 a pracuje s přibližně jedním milionem střel. Navíc je tento výpočet modelem XGBoost velmi rychlý, protože je schopen vypočítat pravděpodobnost skórování pro všech 9 609 střel v Premier League za sezónu 2022-23 během několika vteřin. Totéž platí pro 45 764 střel v pěti nejlepších evropských ligách v minulé sezoně.

Proměnných, které jsou zohledněny v modelu, je více než 20. Nám postačí uvést pouze ty nejdůležitější. Do této skupiny patří:

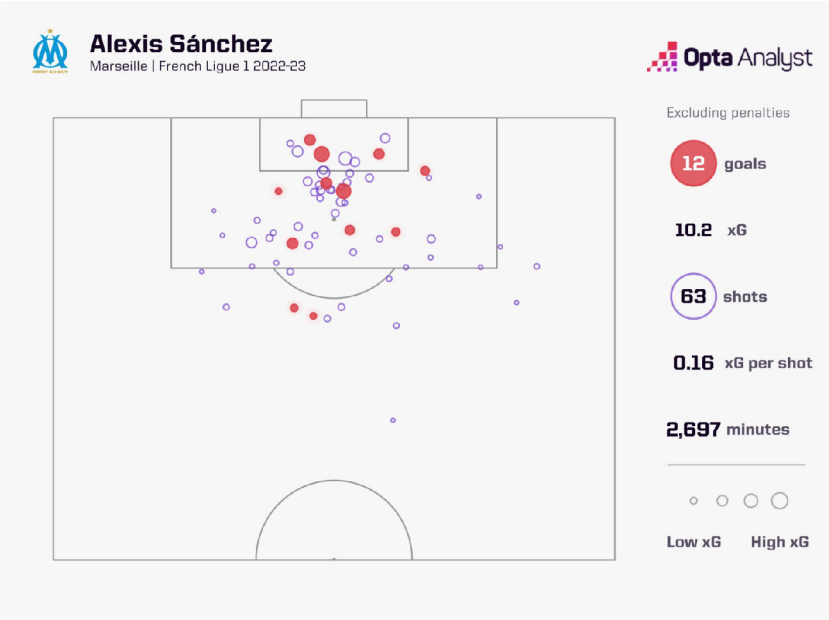
- vzdálenost střely od brány
- úhel střely
- pozice brankáře
- přehled střelce o postavení ostatních hráčů v brankovišti
- šance na zablokování střely obránci
- druh střely (nohou, hlavou, volej)
- druh situace (rohový kop, přímý kop, rychlý protiútok, atd.)

- druh přihrávky (centr, průniková přihrávka, atd.)

Speciálním případem jsou pokutové kopy, které jsou nejkonzistentnější střelou ve fotbale. Hodnota xG je konstantní a vychází z historické míry proměňování (0,79 xG).

Pro lepší představu využijeme grafické znázornění 2.7 xG u střeleckých pokusů Alexise Sáncheze, hráče francouzského Marseille, za sezónu 22/23. Můžeme vidět, že byl produktivní v proměňování svých střeleckých příležitost. Protože hodnota xG je menší než počet nastřílených gólů – rozdíl činí 1,8. Navíc můžeme vidět střelecké pozice označené kroužky, kdy velikost udává kvalitu (měřítko v obrázku) a červená barva značí gólová zakončení. Poslední bod, který stojí za zmínku a dobře ilustruje význam dříve zmíněných proměnných ovlivňujících hodnotu xG, je velikost kroužků na skoro stejné pozici. Při detailnějším pohledu do obrázku zjistíme, že jsou celkem odlišné – kupříkladu střely z malého vápna.

Obrázek 2.7: Očekávané góly Alexise Sáncheze za sezónu 22/23, zdroj: [18]



Jelikož máme v datech počet vstřelených gólů Gls a modelem odhadnutou hodnotu očekávaných gólů xG, tak se nabízí vytvořit novou veličinu, která bude rozdílem těchto statistik. V Tabulce 2.6 je označena jako **xGDiff** a definujeme ji jako:

$$xGDiff = Gls - xG.$$

Při takto formulovaném rozdílu xGDiff vyjadřuje schopnost týmů využívat své šance. A tedy záporné hodnoty značí, že se tým dostal do kvalitních střeleckých příležitostí, ale nebyl je schopen proměnit. Na druhé straně kladná čísla mohou ukazovat naprostý opak, kdy tým měl malý počet dobrých pozic pro vstřelení branky, ale byl efektivní.

Příkladem může být 10 nejlepších týmů ve statistice xGDiff za sezóny 22/23 a 21/22 v Tabulkách 2.7, 2.8. Spolu s xGDiff je také vypsáno konečné pořadí na konci soutěže LgRk, což reprezentuje výkony v dané sezóně. K těmto statistikám přidáme ještě rozdíl počtu vstřelených a obdržených branek GD, který formulujeme následovně:

$$GD = Gls - GA.$$

Z formulace GD plyne, že tato veličina poskytuje dobrý vhled do vyváženosti útočné a obranné fáze. Dále je z tabulek zřetelně vidět, že převažují týmy hrající o nejvyšší příčky ve svých soutěžích. To nás asi nepřekvapuje, protože kvalita hráčského kádru u těchto týmů je ohromná a pro nejlepší útočníky stačí málo příležitostí ke skórování. To také může přispět k vyššímu xG, jelikož tato statistika nezohledňuje hráčské schopnosti. Dodejme, že některé hodnoty xGDiff jsou velmi vysoké. Kupříkladu Lazio Řím v ročníku 21/22 nastřílelo o skoro 19 branek více, než by podle dat z modelu xG mělo skórovat. V tomto případě můžeme s trochou nadsázky říct, že v bráně skončilo vše,

Squad	LgRk	GD	xGDiff
Bayern	1	52	15,0
Man City	1	59	13,4
Arsenal	2	41	12,1
Montpellier	12	2	11,8
Union Berlin	4	11	11,1
Tottenham	8	5	10,9
Napoli	1	47	10,3
Dortmund	2	37	9,9
Salernitana	15	-16	9,3
Bremen	13	-14	8,9

Tabulka 2.7: Nejefektivnější týmy za sezónu 22/23

Squad	LgRk	GD	xGDiff
Lazio	5	16	18,7
Rennes	4	41	18,2
Dortmund	2	29	17,0
Napoli	3	43	16,2
Leicester	8	3	14,2
Leverkusen	3	29	13,4
Verona	9	4	12,6
Chelsea	3	42	11,6
Paris S-G	1	52	11,3
Lens	7	12	9,8

Tabulka 2.8: Nejefektivnější týmy za sezónu 21/22

do čeho kopli.

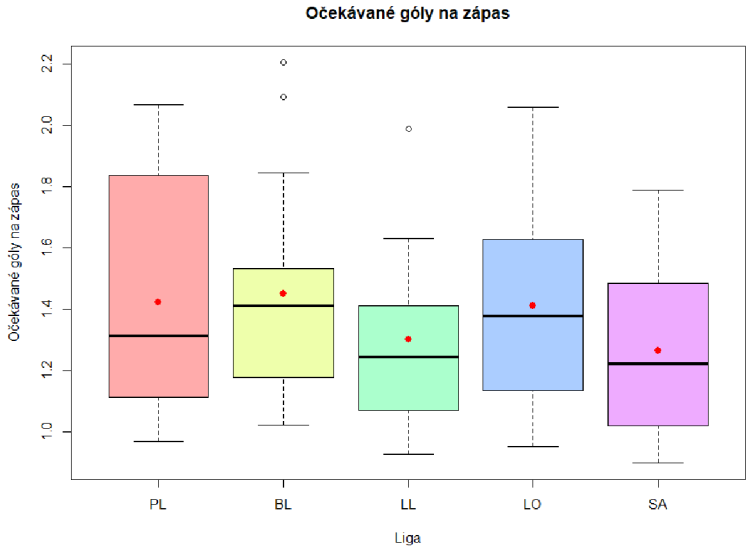
Za zmínku ještě stojí kluby s vysokým xGDiff, které skončily v nižších patrech ligových tabulek. V Tabulce 2.7 jsou tyto kluby uvedeny na posledních dvou místech. Můžeme říct, že tyto dva týmy byly sice efektivní směrem dopředu, ale pokud se zaměříme na sloupec GD, tak uvidíme záporné hodnoty. To značí velký počet inkasovaných gólů oproti gólům vstřeleným. Tedy i přes výbornou produktivitu (i s ohledem na xGDiff) byly Bremen a Salernitana velmi zranitelné do defenzívy, což je stálo lepší příčky v konečné tabulce.

Poslední zajímavostí je, že žádný tým z **LL** se neumístil mezi desítkou nejefektivnějších klubů za oba dva ročníky, což je celkem překvapivé.

2.3.2 Popisná statistika dat střelby

Pro střelecká data vykreslíme krabicové grafy pouze pro střely, střely na bránu a xG. Znovu použijeme upravené statistiky s ohledem na různý počet zápasů. Opět do krabicových grafů vykreslíme červené tečky, které značí průměrné hodnoty veličin pro jednotlivé soutěže. Zbylé proměnné přesnost střelby a průměrné vzdálenosti střelby mají větší smysl pro porovnání týmů.

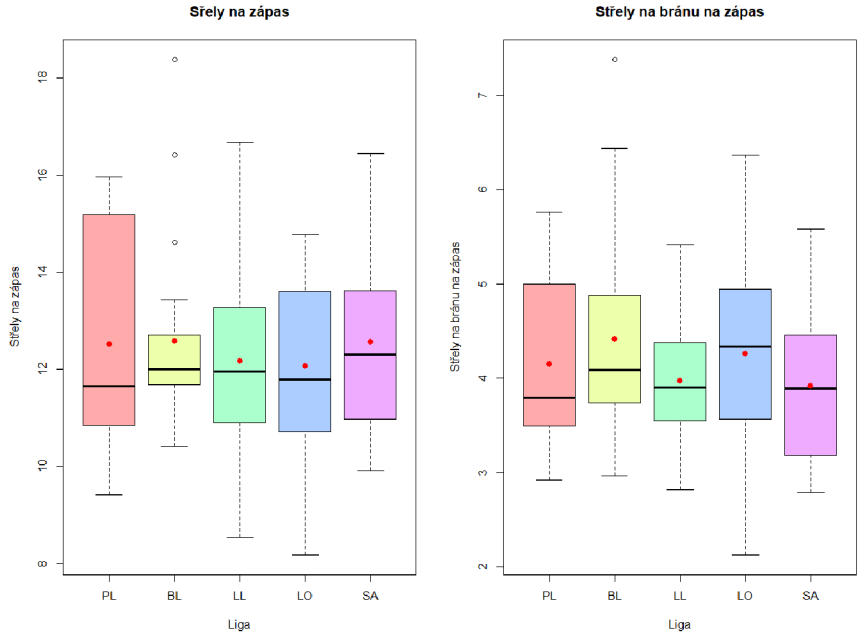
Obrázek 2.8: Boxplot očekávaných gólů (xG) na zápas



Kupříkladu vzdálenost velmi souvisí s herním stylem týmů. V podstatě záleží na taktice, kterou trenér vytvoří na základě nejsilnějších stránek svého kádru.

Prvně začneme s krabicovým grafem 2.8 očekávaných gólů na zápas. V porovnání s krabicovým grafem 2.4 vstřelených gólů na zápas je graf trochu rozdílný. Dříve než přistoupíme k detailnějšímu porovnání, musíme popsat graf xG na zápas. Všechny krabice pro xG na zápas jsou oproti vstřeleným gólům na zápas velmi podobné a žádná soutěž se nijak neodlišuje. Jedinou zajímavostí jsou odlehlá pozorování. V případě **BL** jsou odlehlými pozorováními první dva týmy z **BL** konečného pořadí v tabulce soutěže – jmenovitě Bayern a Dortmund. Pro **LL** je odlehlým pozorováním Barcelona. Ještě se tedy vraťme k porovnání krabicových grafů vstřelených branek a xG na zápas, poněvadž je mezi nimi určitý rozdíl. Můžeme říct, že týmy ze soutěží „velké pětky“ si dokázaly vytvořit podobný počet kvalitních šancí ke vstřelení branky na zápas. Ovšem samotné proměňování šancí je pro soutěže trochu odlišné – zejména pro **BL**, která má nejvyšší průměrnou hodnotu

Obrázek 2.9: Boxploty střel a střel na bránu na zápas



v počtu vstřelených branek na zápas.

Jako poslední zbývá popsat krabicové grafy 2.9 střel a střel na bránu na zápas. Krabice toho moc nenabízí podobně jako v některých dřívějších případech. Opět jsou krabice velmi podobné, pouze krabicový graf střel na zápas pro **BL** nabízí 3 odlehlá pozorování a nejméně širokou krabici, což značí malý rozptyl. V podstatě jsou na tom týmy velmi podobně, co se týče střelecké formy. Pouze Bayern, Dortmund a Leipzig mají výrazně více střel na zápas než ostatní bundesligové týmy.

2.4 Popis defenzivních dat

Nedílnou součástí hry týmů je defenzíva. Pokud se klubům nedaří střílení gólů, tak obrana je základním stavebním kamenem pro získávání bodů, proto prvním úkolem nového trenéra je vybudovat dobře fungující defenzívu.

Tabulka 2.9: Ukázka datové sady č. 3 - defenzíva

Squad	Comp	Tkl	TklDef3rd	TklMid3rd	TklAtt3rd	Blocks	BlocksSh	BlocksPass	Int
Ajaccio	Ligue 1	637	321	236	80	359	85	274	420
Almería	La Liga	557	292	202	63	435	122	313	322
Angers	Ligue 1	652	311	268	73	402	99	303	375
Arsenal	Premier League	568	238	212	118	362	86	276	237
Aston Villa	Premier League	633	305	251	77	438	118	320	324
Atalanta	Serie A	636	291	272	73	455	93	362	421
Athletic Club	La Liga	608	267	254	87	415	95	320	310
Atlético Madrid	La Liga	667	332	245	90	396	103	293	304
Augsburg	Bundesliga	522	251	202	69	429	126	303	286
Auxerre	Ligue 1	674	337	263	74	441	118	323	454
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

K porovnání defenzivní činnosti využijeme pouze základní údaje o zákrocích a blocích.

Nejdříve začneme popisem statistik zákroků (Tkl) z Tabulky 2.9, které mají vést k opětovnému zisku míče. Jinými slovy, zákrokem hráčů ve hřišti rozumíme tlak na protihráče za účelem zisku míče. Tuto statistiku máme v datech rozdělenou do tří skupin, a to podle části hřiště, ve které zákrok proběhl. Postupně jde o defenzivní zákroky v obranné třetině (TklDef3rd), ve středu (TklMid3rd) a útočné třetině (TklAtt3rd). Velmi podobnou statistikou zákroků je vypíchnutí míče (Int). Hlavním rozdílem je, že bránící hráč vyvíjí tlak na míč za účelem vypíchnutí míče buď předskočením soupeře, nebo získáním míče až po zpracování.

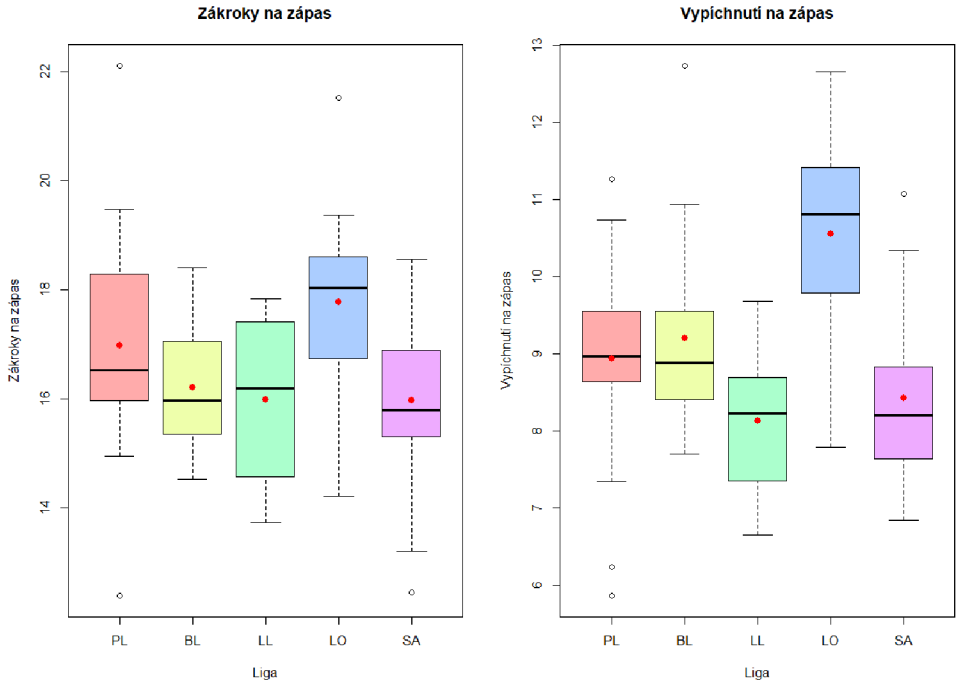
Zbývajícími veličinami jsou bloky (Blocks), zblokované střely na bránu (BlocksSh) a zmařené přihrávky (BlocksPass).

2.4.1 Popisná statistika defenzivních dat

Grafické znázornění v podobě krabicových grafů vykreslíme pouze pro zákroky (Tkl) a vypíchnutí (Int) na zápas. Do krabic znovu dokreslíme červené tečky, které značí průměrné hodnoty veličin pro jednotlivé soutěže.

Při pohledu do grafu 2.10 zákroků na zápas můžeme vidět, že krabicové

Obrázek 2.10: Boxplot zákroků a vypíchnutí na zápas



grafy jsou celkem rozdílné. Nejvíce vyčnívá graf **LO**, který má nejvýše posazenou krabici s hodnotou mediánu kolem 18 zákroků na zápas. Navíc můžeme vypozorovat, že krabice **LO** se jen velmi lehce překrývá s krabicemi **BL** a **SA**. Další zajímavostí jsou 2 odlehlá pozorování pro **PL**, kdy jeden tým má oproti zbývajícím celkům v soutěži velmi malý počet zákroků na zápas a druhý klub má nejvyšší počet zákroků na zápas ze všech lig.

Graf vypíchnutí na zápas je ještě více zajímavý, protože krabice **LO** je posazená výrazně výše než grafy ostatních soutěží, což značí i s přihlédnutím ke grafu zákroků, že týmy v **LO** mají taktiku postavenou na rychlém znovuzískání míče i za cenu faulů. Navíc krabicové grafy vypíchnutí na zápas můžeme rozdělit do 3 skupin. V první najdeme **LO** s nejvyšším počtem vypíchnutí na zápas, druhou skupinu (skupina středu) tvoří **PL** a **BL** a do poslední spadá **LL** a **SA**, které mají nejmenší počet vypíchnutí na zápas.

Kapitola 3
Statistické metody

Záměrem této kapitoly je představit statistické metody, které budou v další kapitole použity k zodpovězení otázek položených v předchozích částech. Vzhledem k tomu, že nás zajímá ověření některých zajímavých kvantitativních veličin v závislosti na kvalitativní veličině Comp (liga), bude potřeba uvést metody, které dokážou tento vztah modelovat.

Pro analýzu závislosti kvantitativního a kvalitativního vztahu využijeme metodu jednofaktorové ANOVY, kterou ovšem aplikujeme pouze za určitých předpokladů, proto při porušení těchto předpokladů použijeme neparametrickou obdobu ANOVY (část 3.1.1), a tou je Kruskalův-Wallisův (část 3.1.4) test. Při zamítnutí nulové hypotézy je potřeba určit, které skupiny jsou významně rozdílné, a proto jsou k oběma metodám přidány metody mnohonásobného porovnávání. Hlavním zdrojem byla skripta [1].

Dále budeme potřebovat modelovat data s Poissonovým rozdělením, a proto si nejdříve v části 3.2.1 zadefinujeme diskrétní Poissonovo rozdělení. Následující segment 3.3.1 budeme věnovat zobecněným lineárním modelům, do kterých spadá i Poissonova regrese. Zdrojem části o Poissonově rozdělení byla skripta [1], pro zobecněné lineární modely a Poissonovu regresi jsem využil zdroj [2] a část o maximálně věrohodném odhadu parametrů v Poissonově

regresi byla čerpána z diplomové práce [3].

3.1 Vztah kvalitativního a kvantitativního znaku

3.1.1 ANOVA

Tento test je jedním z nejpoužívanějších parametrických testu shody středních hodnot pro k skupin, kde $k \geq 2$ (pro $k = 2$ se využívají jednodušší metody, kupříkladu dvouvýběrový t-test). Samotný název ANOVA znamená analýza rozptylu (anglicky ANalysis Of VAriance), což je na první pohled matoucí, jelikož slouží k porovnání středních hodnot. Pojmenování vychází ze způsobu, jakým střední hodnoty testujeme, a tím je porovnání reziduálního součtu čtverců (zbytkové variability) ve 2 různých modelech. ANOVU někdy označujeme jako analýzu rozptylu jednoduchého[6] třídění.

Uvažujme náhodnou kvalitativní veličinu X s $k \geq 2$ skupinami a kvantitativní náhodnou veličinu Y. Cílem je zjistit, zda má veličina Y stejné rozdělení pro všechny skupiny X. Je zvláštní, že testujeme *shodu středních hodnot* s využitím *shody rozdělení*, protože se jedná o jiné hypotézy. Ovšem při splnění podmínek ANOVY jsou tyto hypotézy ekvivalentní.

Předpoklady modelu:

Mějme $k \geq 2$ nezávislých náhodných výběrů s rozsahem $n_1, n_2, ..., n_k$, kde celkový počet $n = n_1 + n_2 + ... + n_k$. Pro každý z těchto výběrů platí:

$$Y_{i1}, ..., Y_{in_i} \text{ je náhodný výběr z rozdělení } N(\mu_i, \sigma^2),\ \forall i \in 1, 2, ..., k \qquad (3.1)$$

Vidíme, že výše uvedené náhodné výběry pochází z normálního rozdělení se shodným rozptylem. Pro ověření předpokladu normality využijeme v kapi-

[6] Skupiny jsou určeny hodnotami jediného kategoriálního znaku X.

tole 4.1 Shapirův-Wilkův test naprogramovaný v softwaru R. Bližšímu popsaní metody se v této práci nebudeme věnovat. Porovnání rozptylů blíže představíme v části 3.1.3.

Nulová hypotéza tvrdí, že všechny střední hodnoty $\mu_1, \mu_2, ..., \mu_k$ jsou stejné. Opakem je alternativní hypotéza tvrdící, že alespoň jedna dvojice středních hodnot se liší. Píšeme:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_k \\ H_A &: \exists i \in \{1, 2, ..., k\}\ \exists j \in \{1, 2, ..., k\}, i \neq j : \mu_i \neq \mu_j \end{aligned} \tag{3.2}$$

Při platnosti předpokladů 3.1 mohou být rozdílné pouze střední hodnoty skupin, a to ověřujeme v testované hypotéze. Z toho plyne, že testování stejného rozdělení a středních hodnot je v tomto případě identické. Nyní přejdeme k porovnání variability ve dvou modelech.

První model:

$$\forall i \in \{1, ..., k\} : Y_{ij} = \mu_i + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2),\ j = 1, ..., n_i,\ \epsilon_{ij} \text{ nezávislé} \tag{3.3}$$

V tomto modelu uvažujeme rozdílné střední hodnoty pro každou z k skupin. Odhady parametrů μ_i pomocí metody nejmenších čtverců (MNČ) vypadají následovně

$$\hat{\mu}_i = \overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Jedná se o skupinové průměry. Variabilitu uvnitř skupin vyjádříme pomocí reziduálního součtu čtverců

$$S_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2.$$

Náhodná veličina $\frac{S_e}{\sigma^2}$ má za platnosti H_0 χ^2 rozdělení s $n - k$ stupni volnosti,

kde $n = \sum_{i=1}^{k} n_i$

Druhý model:

$$\forall i \in \{1, ..., k\} : Y_{ij} = \mu + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2),\ j = 1, ..., n_i,\ \epsilon_{ij} \text{ nezávislé} \quad (3.4)$$

Tento model obsahuje společnou střední hodnotu μ, a tedy splňuje nulovou hypotézu 3.2.

Odhad společné střední hodnoty metodou nejmenších čtverců je aritmetický průměr přes všechny Y_{ij}. Zápis je následující

$$\hat{\mu} = \overline{Y}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}.$$

Reziduální součet čtverců v modelu za platnosti H_0 je dán výrazem

$$S_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{..})^2.$$

S_T nazýváme *celkovým součtem čtverců* a statistika $\frac{S_T}{\sigma^2}$ má za platnosti H_0 χ^2 rozdělení s $n - 1$ stupni volnosti.

Mezi celkovou variabilitou S_T náhodné veličiny Y a S_e platí vztah:

$$S_T = S_A + S_e, \quad (3.5)$$

kde výraz $S_A = \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})$ nazýváme *skupinovým součtem čtverců*, který charakterizuje variabilitu mezi skupinami. Statistika $\frac{S_A}{\sigma^2}$ se za platnosti H_0 znovu řídí χ^2 rozdělením s $k - 1$ stupni volnosti (změna stupňů volnosti oproti statistikám $\frac{S_e}{\sigma^2}$ a $\frac{S_A}{\sigma^2}$).

Testová statistika pro testování H_0 porovnává dva různé odhady rozptylu

σ^2, tak že podělíme odhad variability mezi skupinami $\frac{S_A}{k-1}$ a odhad variability uvnitř skupin $\frac{S_e}{n-k}$. Jelikož se obě statistiky řídí χ^2 rozdělením, získáme statistiku s Fisherovým rozdělením (vyplývá z nezávislosti S_A a S_e), která vypadá následovně

$$F_A = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} = \frac{S_A}{S_e}\frac{n-k}{k-1} \overset{H_0}{\sim} F_{k-1,n-k}. \tag{3.6}$$

Pro zamítnutí H_0 svědčí velké hodnoty testové statistiky. A tak zamítáme H_0 na hladině významnosti α pro hodnoty $f_A \geq F_{k-1,n-k;1-\alpha}$, kde f_A označujeme realizaci testové statistiky a $F_{k-1,n-k;1-\alpha}$ značí kvantil Fisherova rozdělení. To tedy znamená, že při velké podobnosti skupin (mají skoro stejná rezidua), bude hodnota f_A blízká 0. S rostoucí rozdílností skupin statistika f_A narůstá, až může překročit hranici nezamítnutí.

3.1.2 Tukeyho metoda mnohonásobného porovnávání

V případě zamítnutí H_0 v analýze rozptylu potřebujeme vědět, které dvojice skupin se významně liší. Není vhodné využít porovnání hodnot za pomocí dvouvýběrových testů, jelikož bychom překročili hladinu významnosti α.

Při stejných předpokladech jako u ANOVY budeme testovat hypotézy

$$H_0^\star : \mu_i = \mu_j$$

proti alternativám

$$H_A^\star : \mu_i \neq \mu_j,$$

pro všechna $i, j = 1, ..., k;\ i \neq j$.

Pro testování výše zmíněných hypotéz využijeme tzv. *Tukeyho metodu*, podle které zamítáme $H_0^\star$ na hladině α při rozdílech odhadnutých středních

hodnot

$$|\overline{y}_{i.} - \overline{y}_{j.}| > sq_{k,n-k;1-\alpha}\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}, \tag{3.7}$$

kde čísla $q_{k,n-k;1-\alpha}$ jsou kvantily tzv. studentizovaného rozpětí (hodnoty můžeme vypočítat v softwaru R). V případě nevyváženého[7] třídění využijeme tzv. *Tukeyho HSD modifikaci*. Může nastat, že nezamítneme ani jednu shodu středních hodnot $|\overline{y}_{i.} - \overline{y}_{j.}|$ v 3.7. V tomto případě je významně rozdílná nějaká složitější kombinace středních hodnot (kupříkladu neplatí, že $\mu_1 - 3\mu_2 + 2\mu_3 = 0$)

3.1.3 Test rovnosti rozptylů pro $k \geq 2$ skupin

Jedním z předpokladů pro ANOVU je shoda rozptylů σ^2 v k skupinách. Formulace hypotézy je následující:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$

proti alternativě

$$H_A : \text{alespoň jedna dvojice rozptylů se nerovná}$$

V softwaru R použijeme pro porovnání rozptylů *Bartlettův test*. Ovšem v dnešní době je častější přibližný *Leveneův test* blíže popsaný ve skriptech [1] v části 7.3.10. K našim účelům postačí *Bartlettův test*.

3.1.4 Kruskalův–Wallisův test

V praxi se může stát, že nesplníme předpoklady (normalita a shodné rozptyly pro všechny skupiny) analýzy rozptylu, a přesto bychom chtěli střední hodnoty ve skupinách otestovat. K tomu účelu slouží Kruskalův-Wallisův

[7] hodnoty n_i pro $i = 1, ..., k$ se nerovnají

test, který je neparametrickou alternativou k analýze rozptylu jednoduchého třídění a zobecněním Wilcoxonova dvouvýběrového testu.

Kruskalův-Wallisův test řadíme do skupiny neparametrických testů, a proto předpoklady k jeho využití nejsou tak přísné jako u ANOVY. Jediným předpokladem je, aby $\forall i, i = 1, ..., k;\ k \geq 2$ platilo, že jsou náhodné výběry nezávislé. Navíc platí $n = n_1 + ... + n_k$. Zkrácený zápis předpokladů je následující:

$Y_{i1}, ..., Y_{in_i}$ je náhodný výběr z rozdělení se spojitou distribuční funkcí F_i.

Testujeme hypotézu:

$$H_0 : F_1(y) = F_2(y) = ... = F_k(y), \text{ pro všechna } y \in \mathbb{R}$$

proti alternativě

H_A : alespoň jedna dvojice náhodných výběrů pochází z různých rozdělení

Postup při testování je následující:

1. Seřadíme všechna pozorování (sdružený výběr) do neklesající posloupnosti.

2. Každé hodnotě určíme pořadí[8].

3. Vypočítáme součty pořadí T_i, $i = 1, ..., k$, kdy T_i je součtem pořadí $i-$té skupiny.

   - Celkovou hodnotu T můžeme vypočítat jako $T = T_1 + ... + T_k$. Jelikož se jedná o aritmetickou posloupnost s diferencí 1, můžeme využít vzorec pro součet této posloupnosti, tedy $T = \sum_{i=1}^{k} T_i = \frac{n}{2}(n+1)$.

[8] v případě shody - průměrné pořadí shodných hodnot

4. Vypočítáme hodnotu testové statistiky

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1). \tag{3.8}$$

5. Zamítáme H_0, pokud realizací testové statistiky je hodnota $q \geq \chi^2_{k-1,1-\alpha}$. Jde o test asymptotický. Jinými slovy, můžeme kvantil χ^2 využít pouze pro větší skupiny.

   - Pro malé počty ve skupinách využijeme tabelované hodnoty kvantilu statistiky Q.

3.1.5 Porovnání dvojic bez předpokladu normality

V případě zamítnutí H_0 Kruskalova-Wallisova testu chceme stejně jako v případě ANOVY (část 3.1.2) určit, které skupiny jsou významně rozdílné. Místo skupinových průměrů použijeme ke srovnávání *průměrná pořadí* v jednotlivých skupinách. Rozdělení F_i a F_j jsou významně odlišná, pokud platí:

$$\left| \frac{T_i}{n_i} - \frac{T_j}{n_j} \right| > \sqrt{\frac{1}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) n(n+1) \chi^2_{k-1,1-\alpha}}. \tag{3.9}$$

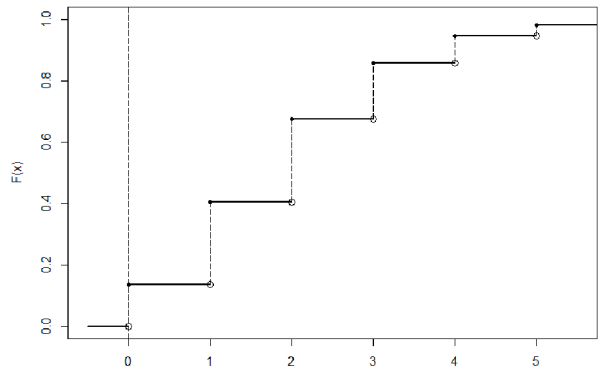
V softwaru R pro porovnávání využijeme jiný test, tzv. Dunnův test z balíčku `PMCMRplus`.

3.2 Diskrétní rozdělení pravděpodobnosti

3.2.1 Poissonovo rozdělení

Poissonovo rozdělení využíváme pro popis počtu událostí za nějaký časový úsek, v našem případě se kupříkladu jedná o vstřelené góly za sezónu.

Obrázek 3.11: Graf distribuční funkce Poissonova rozdělení ($\lambda = 2$)



Formulace pravděpodobnostní funkce Poissonova rozdělení je následující:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0, \tag{3.10}$$

kde diskrétní náhodná veličina X nabývající hodnot $k = 0, 1, 2, ...$ má Poissonovo rozdělení s parametrem λ, zkráceně zapisujeme $X \sim Po(\lambda)$. Distribuční funkce Poissonova rozdělení pro X formulujeme následovně:

$$F(x) = \begin{cases} 0, & \text{pro } \ x < 0. \\ \sum_{k \leq x} \frac{\lambda^k}{k!} e^{-\lambda}, & \text{pro } \ x \geq 0. \end{cases} \tag{3.11}$$

Střední hodnota rozdělení

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda. \tag{3.12}$$

Rozptyl rozdělení

$$var(X) = E(X^2) - [E(X)]^2 = E[X(X-1)] + E(X) - [E(X)]^2. \tag{3.13}$$

Nejdříve vypočteme hodnotu $E[X(X-1)]$

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2 e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda^2.$$

Dosazením do rovnice 3.13 získáme hodnotu rozptylu

$$var(X) = E[X(X-1)] + E(X) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Pro Poissonovo rozdělení platí, že střední hodnota a rozptyl jsou totožné a jsou rovny parametru λ.

3.3 Regresní modely

3.3.1 Zobecněné lineární modely

Dříve než přejdeme k zobecněným modelům, uvedeme model pro lineární regresi. Model zapisujeme takto:

$$Y_i = \mathbf{x_i^T}\boldsymbol{\beta} + e_i,\ e_i \sim N(0,\sigma^2),\ i = 1,2,...,n \tag{3.14}$$

kde $Y_1,...,Y_n$ jsou nezávislé náhodné veličiny vysvětlované proměnné Y. Vektor $\mathbf{x_i^T} = \{x_{i1},...,x_{ip}\}$ nazýváme vektorem regresorů (vysvětlujících proměnných), který považujeme za vektor známých hodnot. Dalším vektorem je $\boldsymbol{\beta} = (\beta_0,\beta_1,...,\beta_p)^T$, který označujeme jako vektor regresních koeficientů. Posledními nepopsanými veličinami jsou náhodné odchylky neboli náhodné chyby e_i, $i = 1,2,...,n$. Lineární model můžeme alternativně zapisovat takto:

$$E(Y_i) = \mu_i = \mathbf{x_i^T}\boldsymbol{\beta};\ Y_i \sim N(\mu_i,\sigma^2),\ i = 1,2,...,n, \tag{3.15}$$

Lineární modely můžeme také zapsat vektorově. Nejdříve označme

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Pak definujeme vektorový zápis takto:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{3.16}$$

kde $\mathbf{0}$ značí sloupcový nulový vektor a $\mathbf{I}$ jednotkovou matici. Navíc předpokládáme normalitu náhodných odchylek kvůli testování parametrů $\boldsymbol{\beta}$. Odhady $\widehat{\boldsymbol{\beta}}$ parametrů $\boldsymbol{\beta}$ získáme *metodu nejmenších čtverců*. Tedy minimalizujeme výraz

$$S(\beta_0, ..., \beta_p) = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2. \tag{3.17}$$

Pro výpočet odhadů parametrů lineární regrese je odvozen z výrazu 3.17 jednodušší vztah

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

Zobecněný lineární model, jak již jeho název napovídá, poskytuje několik zobecnění oproti lineárnímu modelu. Hlavní rozdíl spočívá v tom, že není nutné modelovat pouze střední hodnotu Y_i. Může se stát, že parametry $\boldsymbol{\beta}$, které chceme odhadnout, nejsou v lineárním tvaru, např. $\mathbf{x}_i^T e^{\boldsymbol{\beta}}$. V těchto případech nemůžeme využít model klasické lineární regrese. Proto obecně modelujeme funkci střední hodnoty veličiny Y_i, píšeme

$$g(E(Y_i)) = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}. \tag{3.18}$$

Funkci g nazýváme spojovací funkcí a platí, že je monotónní a diferencovatelná.

3.3.2 Poissonova regrese

Mezi zobecněné lineární modely spadá Poissonova regrese, kdy pro veličiny $Y_1, Y_2, ..., Y_n$ řídící se Poissonovým rozdělením platí

$$Y_i \sim Po(\mu_i),\ E(Y_i) = \mu_i \ \ i = 1, 2, ..., n. \tag{3.19}$$

Z definice 3.10 pravděpodobnostní funkce Poissonova rozdělení víme, že se střední hodnota λ nachází v exponentu, proto jako spojovací funkci využijeme přirozený logaritmus. Tento druh zobecněného lineárního modelu nazýváme log-linearní model a zapisujeme ho jako

$$g(\mu_i) = \ln \mu_i, \ \ i = 1, 2, ..., n. \tag{3.20}$$

Díky této transformaci můžeme parametry modelu odhadnout metodou maximálně věrohodných odhadů – podrobněji je probereme v části 3.3.3.

Pro μ_i platí

$$\mu_i = n_i \lambda_i, \tag{3.21}$$

kde n_i značí časový úsek pro Y_i pozorování. Může tedy nastat situace, kdy je délka expozice[9] pro rozdílná pozorování Y_i odlišná. V podstatě uvažujeme časový úsek, za který se pozorování realizují, a tak rozlišujeme dva druhy modelů. V prvním obecném modelu, který využijeme pro další analýzu, musíme vzít v potaz délku expozice. Pro příklad bychom chtěli otestovat počet zákazníků za den v různých obchodech, expozicí n_i by byly různě dlouhé ote-

[9] míra vystavení nějakému jevu

vírací doby. Ve druhém modelu je pro všechny skupiny expozice konstantní. Zobecněný lineární model s různou délkou expozice n_i formulujeme následovně:

$$E(Y_i) = \mu_i = n_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}, \;\; Y_i \sim Po(\mu_i). \tag{3.22}$$

Rovnice pro model spojovací funkce ln zapisujeme takto:

$$\ln \mu_i = \ln n_i + \ln \lambda_i = \ln n_i + e^{\mathbf{x}_i^T \boldsymbol{\beta}}, \;\; i = 1, 2, ..., n \tag{3.23}$$

Při tvoření modelu Poissonovy regrese je nutné pro různé hodnoty n_i využít funkci `offset()` v softwaru R, do které vkládáme zlogaritmované hodnoty $\ln(n_i)$. Tento krok musíme udělat, protože hodnoty n_i jsou známé konstanty, které nepotřebujeme odhadovat a lze je snadno začlenit do postupu odhadu.

Interpretace parametrů v Poissonově regresi je trochu odlišná od klasické lineární regrese, protože odhadované parametry jsou v exponentu. Pro porovnání vlivu kvalitativní veličiny na podmíněnou střední hodnotu vysvětlované proměnné musíme vypočítat **rate ratio** (RR). Pro binární veličinu nabývající hodnot $x_j = 0$, $x_j = 1$ platí

$$RR = \frac{E(Y_i|x_j = 1)}{E(Y_i|x_j = 0)} = e^{\beta_j}.$$

Stejný postup využijeme i pro interpretaci kvalitativní veličiny o více kategoriích, kdy v modelu určíme referenční skupinu a vypočítáme RR pro všechny skupiny. Kupříkladu kvalitativní veličina o 4 kategoriích nabývající hodnot $x_j = 1$ (referenční skupina), $x_j = 2$, $x_j = 3$ a $x_j = 4$ má hodnoty RR

následující:

$$RR = \frac{E(Y_i|x_j = 2)}{E(Y_i|x_j = 1)} = e^{\beta_2}$$
$$\vdots$$
$$RR = \frac{E(Y_i|x_j = 4)}{E(Y_i|x_j = 1)} = e^{\beta_4}$$

Vidíme, že pro interpretaci nemusíme počítat RR jako podíl, pouze stačí vypočítat hodnoty e^{β_j}. Podobně funguje interpretace pro kvantitativní proměnné x_j. Tedy stejně jako u lineární regrese zafixujeme všechny ostatní proměnné kromě x_j a při posunutí x_j o 1 dojde k násobné změně podmíněné střední hodnoty, a to e^{β_j}krát. V tomto modelu se hodnota podmíněné střední hodnoty nezvyšuje lineárně, ale e^{β_j} má multiplikativní efekt na střední hodnotu μ.

Vyrovnané hodnoty modelu:

$$\widehat{Y}_i = \widehat{\mu}_i = n_i e^{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}}, \;\; i = 1, ..., n$$

Pro vyhodnocení kvality modelu využíváme **Pearsonova rezidua**, která vypočítáme z výrazu

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}, \tag{3.24}$$

kde o_i značí pozorovanou hodnotu (observed) Y_i a e_i vyrovnanou hodnotu (expected). Jelikož v Poissonově rozdělení platí $var(Y_i) = E(Y_i)$, pak můžeme provést zjednodušení a směrodatnou odchylku e_i vypočítat jako $\sqrt{e_i}$.

Celkový reziduální rozptyl pro model Poissonovy regrese je vypočítán pomocí reziduí ze vztahu 3.24, které souvisí s chí-kvadrát testem dobré shody [10] a platí

$$X^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}. \tag{3.25}$$

[10] Využívá se pro protestování v kontingenčních tabulkách.

Další možností jak vypočítat reziduální rozptyl je pomoci **deviance reziduí**, kdy celkovou hodnotu vypočítáme ze vztahu

$$D = 2\sum_{i=1}^{n}[o_i \ln(o_i/e_i) - (o_i - e_i)]. \tag{3.26}$$

Ovšem v některých modelech můžeme tento vztah zjednodušit, a to když se $\sum o_i = \sum e_i$. Zjednodušený vztah vypadá takto:

$$D = 2\sum_{i=1}^{n}[o_i \ln(o_i/e_i)].$$

Obě statistiky X^2 a D mají χ^2 rozdělení s $n - p$ stupni volnosti, kde n je počet pozorování a p je počet odhadnutých parametrů.

3.3.3 Maximálně věrohodný odhad parametrů v Poissonově regresi

Maximálně věrohodný odhad zakládáme na maximalizaci věrohodnostní funkce, která je definovaná jako sdružená hustota nebo pravděpodobnostní funkce pozorovaných náhodných veličin. V případě Poissonova rozdělení maximalizujeme věrohodnost jako funkci parametru λ pomocí realizací $y_1, ..., y_n$ náhodných veličin $Y_1, ..., Y_n$ a věrohodnostní funkci formulujeme následovně:

$$L(\lambda) = \prod_{i=i}^{n} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} y_i}}{y_1! \cdot y_2! \cdot ... \cdot y_n!}. \tag{3.27}$$

Věrohodnostní funkci můžeme zlogaritmovat, protože zachovává polohu maxima, neboť logaritmus je rostoucí funkcí. Logaritmická věrohodnostní funkce

vypadá následovně:

$$\ln L(\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \ln \lambda - \ln(y_1! \cdot y_2! \cdot ... \cdot y_n!) \tag{3.28}$$

Model Poissonovy regrese má pro realizace $y_1, y_2, ..., y_n$ náhodného výběru $Y_1, Y_2, ..., Y_n$ a vektor regresních parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ věrohodnostní funkci ve tvaru

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\mu_i(\boldsymbol{\beta})^{y_i}}{y_i!} exp\{-\mu_i(\boldsymbol{\beta})\}. \tag{3.29}$$

Logaritmická věrohodnostní funkce má tvar

$$l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \left[\frac{\mu_i(\boldsymbol{\beta})^{y_i}}{y_i!} exp\{-\mu_i(\boldsymbol{\beta})\} \right]. \tag{3.30}$$

Maximálně věrohodný odhad vektoru parametrů $\hat{\boldsymbol{\beta}}$ je dán předpisem

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} l_i(\boldsymbol{\beta}), \tag{3.31}$$

kde l_i odvozeno z rovnice 3.30 a rovná se

$$\begin{aligned} l_i(\boldsymbol{\beta}) &= -\ln \left[\frac{\mu_i(\boldsymbol{\beta})^{y_i}}{y_i!} exp\{-\mu_i(\boldsymbol{\beta})\} \right] \\ &= -y_i \ln(\mu_i(\boldsymbol{\beta})) + \ln(y_i!) + \mu_i(\boldsymbol{\beta}) \end{aligned}$$

Výsledné hodnoty vektoru $\widehat{\boldsymbol{\beta}}$ hledáme tak, že parciálně zderivujeme výraz $\sum_{i=1}^{n} l_i(\boldsymbol{\beta})$ podle všech parametrů β_i. Protože minimalizujeme daný výraz, tak parciální derivace položíme rovny 0 a vznikne nám soustava p rovnic, kterou nazýváme soustavou rovnic věrohodnosti. Většinou při výpočtu maximálně věrohodných odhadů využíváme numerické metody, jelikož pro tyto složité soustavy neexistuje explicitní vyjádření jejich řešení.

Kapitola 4

Analýza soutěží

V této části bude hlavním cílem ověřit, zda se průměrné počty vstřelených gólů za sezónu liší v závislosti na soutěži. Dalším cílem je otestování středních hodnot veličin v závislosti na lize (ANOVA, Kruskalův-Wallisův test). Tyto veličiny jsou graficky znázorněny krabicovými grafy v sekcích 2.2.1, 2.3.2 a 2.4.1. Pro všechna testování zvolíme hladinu významnosti $\alpha = 0,05$.

4.1 Závislost kvantitativních veličin na ligách

Pro vybrané kvantitativní proměnné z úvodu do této kapitoly provedeme testy závislosti na soutěžích. Chceme tedy zjistit, zda jsou střední hodnoty těchto veličin pro soutěže „velké pětky“ odlišné. Pro testování využijeme buď jednofaktorovou ANOVU, nebo Kruskalův-Wallisův test (K-W test).

Jelikož z části 3.1.1 víme, že pro použití ANOVY (funkce `aov()`) je nutné splnit dvě podmínky, a to že skupiny (ligy) proměnné se řídí normálním rozdělením a mají shodný rozptyl. Když není splněn předpoklad normality, musíme k otestování využít neparametrický Kruskalův-Wallisův test (funkce `kruskal.test()`). Normalitu otestujeme funkcí `shapiro.test()` ze softwaru R (H_0: data se řídí normálním rozdělením). Před otestováním normality mu-

Tabulka 4.10: Tabulka p-value pro testování středních hodnot vybraných veličin v závislosti na lize (Z - zamítáme, NZ - nezamítáme)

p-value Shapiro-Wilkova testu

Liga	PKatt/90s	Pl	Age	CrdY/90s	CrdR/90s	Sh/90s	SoT/90s	Tkl/90s	Int/90s
PL	0,147	0,218	0,342	0,059	0,003	0,015	0,023	0,345	0,156
BL	0,589	0,047	0,487	0,404	0,030	0,001	0,025	0,180	0,014
LL	0,565	0.167	0,157	0,038	0,212	0,965	0,650	0,049	0,612
LO	0,686	0.401	0,269	0,631	0,094	0,497	0,987	0,517	0,304
SA	0,077	0,267	0,408	0,036	0,094	0,211	0,500	0,287	0,114
H_0	NZ	Z (těsně)	NZ	Z	Z	Z	Z	Z (těsně)	Z

p-value ANOVY a Kruskalova-Wallisova testu

Bartlettův test	0.156	-	0,873	-	-	-	-	-	-
ANOVA	0,029	-	0,002	-	-	-	-	-	-
K-W test	-	0,152	-	< 0.001	< 0.001	0,957	0,559	0,001	< 0.001

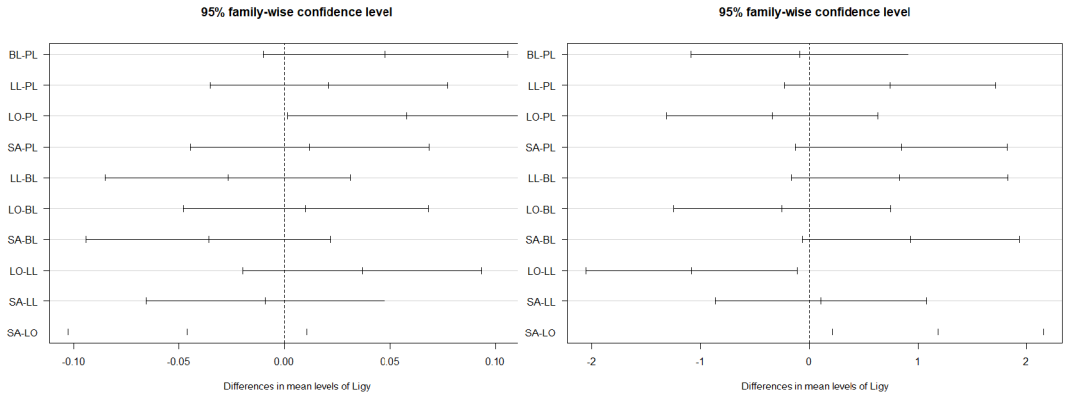
síme jednotlivé realizace proměnných rozdělit do skupin (ligy), teprve poté můžeme přistoupit k ověření normality ve skupinách. ANOVU využijeme pouze v případě nezamítnutí normality ve všech ligách. Vypočítané hodnoty p-value funkce `shapiro.test()` jsou uvedeny v Tabulce 4.10. Dále tabulka obsahuje hodnoty p-value Bartlettova testu (`bartlett.test()`) shody rozptylů lig, který používáme jen v případě nezamítnutí normality pro všechny soutěže. V posledních dvou řádcích najdeme hodnoty p-value ANOVY a K-W testu. Vždy pro ověření rovnosti středních hodnot ve skupinách využíváme test s většími předpoklady, jsou-li tyto požadované předpoklady splněny. Zbývá dodat, že významné rozdíly lig budeme pro veličiny přepočítané na zápas interpretovat za celou sezónu, jelikož předpokládáme, že střední hodnoty za zápas jsou v průběhu sezóny neměnné.

Nulovou hypotézu rovnosti středních hodnot v ligách nezamítáme pouze u tří veličin, a to u Pl (počet hráčů v kádru), Sh/90s a SoT/90s. Krabicové grafy střel a střel na bránu v části 2.3.2 jsou velmi podobné, a tak jsme neočekávali statisticky významný rozdíl, což se potvrdilo. Pro krabice Pl v části 2.2.1 jsme opět významný rozdíl nepředpokládali, a to jsme také potvrdili. To značí, že menší počet zápasů v **BL** nemá vliv na velikosti kádrů.

Jako první popíšeme veličiny testované ANOVOU, u kterých zamítáme shodu středních hodnot v závislosti na soutěžích. Těmito proměnnými jsou Pkatt/90s a Age. Pro mnohonásobné porovnávání využijeme Tukeyho metodu (funkce `TukeyHSD()`), jejíž grafickým výsledkem jsou Obrázky 4.12, které zobrazují 95% intervaly spolehlivosti (CI) rozdílů průměrných hodnot pro všechny dvojice soutěží. Významný rozdíl ve dvojicích poznáme tak, že CI neobsahuje 0. Pro lepší přehlednost budeme popisovat významné rozdíly v bodech. Výčet je následující:

- PKatt/90s (penalty na zápas) – významný rozdíl mezi **LO** a **PL**
  - CI je posunut lehce nad 0, což při zápisu LO-PL značí, že střední hodnota **LO** je větší než **PL**.
- Age – **LO** je významně rozdílný od **LL** a **SA**
  - **LO** je v zápisech na dvou různých pozicích (LO-LL, SA-LO). V prvním případě je CI posunut do záporných a ve druhém do kladných hodnot, z toho plyne, že v **LO** dávají více šancí mladším hráčům než v **LL** a **SA**.

Nyní přejdeme k veličinám, u kterých zamítáme nulovou hypotézu K-W testu. Jedná se o žluté a červené karty na zápas spolu s zákroky a vypíchnutími na zápas. K mnohonásobnému porovnávání využijeme Dunnův test, který najdeme v softwaru R pod příkazem `kwAllPairsDunnTest()`. Pro správné fungování testu musíme zapsat do argumentu příkazu (`p.adjust.method`) některou z metod, která upravuje hodnoty p-value jednotlivých dvojic, aby nebyla překročena hranice hladiny významnosti $\alpha = 0.05$. V našem případě zvolíme Bonferonniho metodu (α vydělíme počtem dvojic), a tedy píšeme `p.adjust.method = "bonferroni"`. Výsledkem jsou krabicové



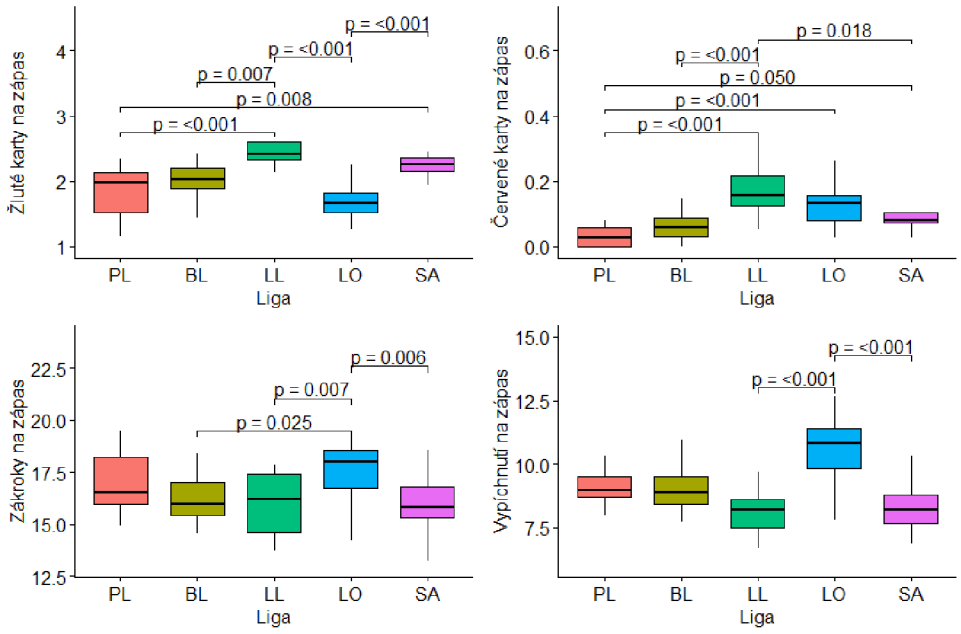
(a) Intervaly spolehlivosti pro PKatt/90s (b) Intervaly spolehlivosti pro Age

Obrázek 4.12: Intervaly spolehlivosti výsledků Tukeyho metody

grafy 4.13, které jsou doplněny o hranatou závorku, která spojuje významně rozdílné dvojice. Navíc jsou nad závorky vykresleny hodnoty p-value Dunnova testu. Opět významné rozdíly soutěží popíšeme pro veličiny v bodech. Výčet je následující:

- CrdY/90s – **LL** je významně rozdílná od všech soutěží kromě **SA**
  - **PL**, **BL** a **LO** tvoří skupinu v počtu žlutých karet na zápas. Navíc z krabicového grafu 4.13 je zřejmé, že hráči **LL** jsou nejvíce trestaní.
- CrdR/90s – **LL** je významně odlišná od všech soutěží kromě **LO**
  - V **PL** se udílí nejméně červených karet a nezamítáme shodu pouze s **BL**. Z krabicového grafu 4.13 **LL** opět vidíme, že hráči ze španělské ligy jsou nejvíce trestaní.
- Tkl/90s – **LO** je významně rozdílná od zbylých lig kromě **PL**
- Int/90s – **LO** je významně rozdílná od **LL** a **SA**

Obrázek 4.13: Boxploty lig s významně odlišnými dvojicemi Dunnova testu



Z krabicového grafu 4.13 můžeme usoudit, že týmy **LO** mají významně větší střední hodnotu zákroků a vypíchnutí na zápas, než výše zmíněné soutěže.

4.1.1 Závislost kvantitativních veličin na soutěžích v sezóně 21/22

V předcházející části jsme ověřili, že některé střední hodnoty proměnných jsou v závislosti na soutěžích významně rozdílné, proto provedeme další sérii testů pro předcházející ročník 21/22, abychom ověřili, jestli se statisticky významné rozdíly opakují.

V Tabulce 4.11 můžeme opět vidět hodnoty p-value testů normality a vhodného (veličina splňuje předpoklady) testu pro ověření závislosti středních hodnot veličin na soutěžích. Tedy vidíme, že jediná proměnná, u které nezamítáme shodu středních hodnot je PKatt/90s. Další změnou je nezamít-

Tabulka 4.11: Tabulka p-value pro testování středních hodnot významně odlišných veličin v závislosti na lize za ročník 21/22 (Z - zamítáme, NZ - nezamítáme)

p-value Shapiro-Wilkova testu

Liga	PKatt/90s	Age	CrdY/90s	CrdR/90s	Tkl/90s	Int/90s
PL	0.192	0.878	0.688	0.001	0.849	0.703
BL	0.722	0.326	0.014	0.034	0.417	0.575
LL	0.332	0.041	0.804	0.446	0.878	0.798
LO	0.322	0.922	0.856	0.485	0.924	0.526
SA	0.037	0.500	0.237	0.730	0.113	0.777
H_0	Z (těsně)	Z (těsně)	Z	Z	NZ	NZ

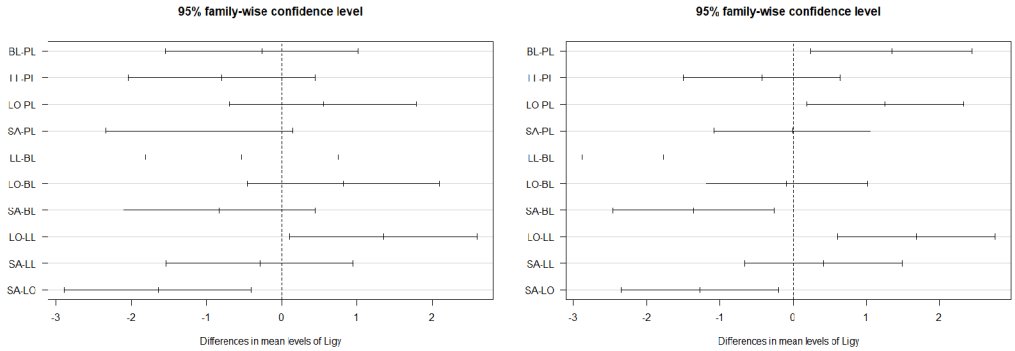
p-value ANOVY a Kruskalova-Wallisova testu

Bartlettův test	-	-	-	-	0.149	0.198
ANOVA	-	-	-	-	0.003	< 0.001
K-W test	0.185	< 0.001	< 0.001	< 0.001	-	-

nutí normality u Tkl/90s a Int/90s, tudíž můžeme na rozdíl od dat ze sezóny 22/23 použít k testování ANOVU. Opak nastává u Age, pro který musíme využít K-W test.

Pro lepší přehled vypíšeme v bodech nejzajímavější významné rozdíly veličin v závislosti na soutěžích za ročník 22/23 a následně i změny, které nastaly v ročníku 21/22. Nejdříve začneme veličinami ze sezóny 21/22 testované ANOVOU, u kterých zamítáme nulovou hypotézu. Intervaly spolehlivosti Tukeyho metody jsou pro všechny dvojice soutěží vykresleny v grafu 4.14. Výčet je následující:

- Tkl/90s – za sezónu 22/23 má **LO** významně vyšší střední hodnotu než zbylé ligy kromě **PL**
  - Pro předcházející ročník je **LO** významně odlišná od **LL** i **SA**, kdy z grafu 4.14 vidíme, že při rozdílu dvojice (LO-LL) je CI posunut do kladných hodnot a v druhém případě (SA-LO) je CI v záporných hodnotách. Z toho plyne, že **LO** má vyšší střední hodnotu než výše zmiňované soutěže.
- Int/90s – za sezónu 22/23 má **LO** významně vyšší střední hodnotu než



(a) Intervaly spolehlivosti pro Tkl/90s (b) Intervaly spolehlivosti pro Int/90s

Obrázek 4.14: Intervaly spolehlivosti výsledků Tukeyho metody za sezónu 21/22

LL a **SA**

  - Pro ročník 21/22 je více soutěží významně odlišných než v sezóně 22/23. Ovšem **LO** má znovu významně větší střední hodnotu než **LL** i **SA**, jelikož CI mají podobnou polohu jako v případě Tkl/90s pro ročník 21/22.

Jako poslední zbývá popsat veličiny s významně rozdílnými dvojicemi soutěží Dunnova testu (při zamítnutí K-W testu). Výčet je následující:

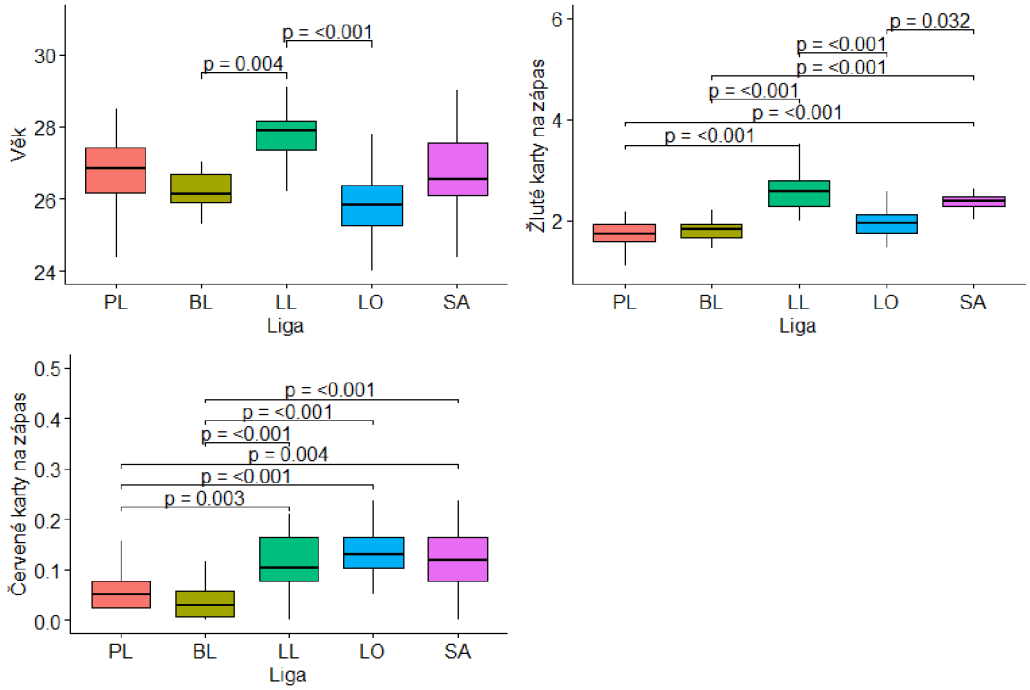
- Age – v sezóně 22/23 měla **LO** významně nižší střední hodnotu než **LL** a **SA**

  - V ročníku 21/22 má **LO** opět významně nižší střední hodnotu než **LL**, což je vidno z krabicového grafu 4.15.

- CrdY/90s – za sezónu 22/23 bylo v **LL** udíleno významně více žlutých karet než ve zbylých ligách kromě **SA**

  - Z grafu 4.15 vidíme, že pro ročník 21/22 platí obdobná situace jako pro sezónu 22/23.

Obrázek 4.15: Boxploty lig s významně odlišnými dvojicemi Dunnova testu za sezónu 21/22



- CrdR/90s – v ročníku 22/23 bylo v **LL** udíleno významně více červených karet než všech soutěží kromě **LO**
  - Pro ročník 21/22 je situace obdobná s tím, že k nevýznamně rozdílným ligám vzhledem k **LL** přidáváme **SA**. Při pohledu do krabicového grafu 4.15 vidíme, že soutěže tvoří dvě skupiny.

Shrnutím této podkapitoly je, že menší počet utkání v sezóně nemá vliv na počet hráčů v kádru. Dalším zjištěním je, že v **LO** mají mladší hráči větší herní vytížení než v **LL**. Dále jsme odhalili, že počet obranných zákroků v **LO** byl v obou sezónách významně vyšší něž v **LL** a **SA**. Pro žluté karetní tresty platí, že v **LL** jich bylo udíleno více než ve zbylých soutěží kromě **SA**, a to platí pro data z obou sezón. V poslední řadě jsme zjistili, že

počet udílených červených karet byl v **LL** pro oba ročníky významně vyšší než v **PL** a **BL**.

4.2 Porovnání gólové produkce soutěží

Vstřelené a obdržené branky jsou nejzásadnější statistikou nejen ve fotbale, ale i v dalších kolektivních sportech. Proto by nás mohlo zajímat, zda jsou počty vstřelených gólů v soutěžích „velké pětky“ za sezónu rozdílné. Z taktického pohledu se příliš nevyplatí střílet spoustu gólů, protože za velký rozdíl vstřelených a obdržených branek v zápasech nejsou udělovány žádné bonusové body. Navíc při ofenzivních výpadech jsou týmy více zranitelné v obraně. Z tohoto úhlu pohledu by gólová produkce měla být více méně stejná. Pojďme tedy tuto domněnku ověřit.

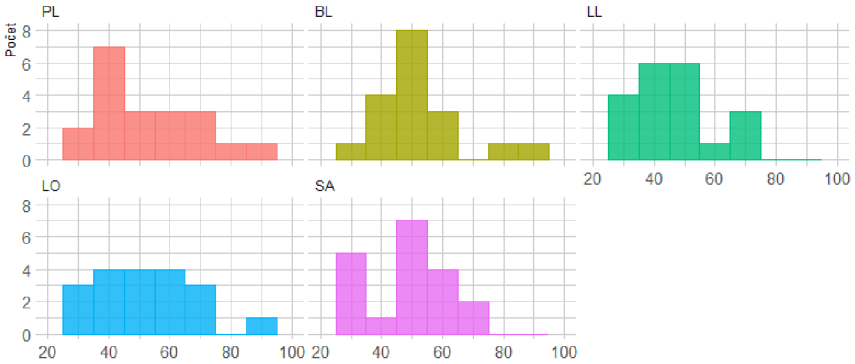
Z popisu krabicového grafu 2.4 gólů na zápas v části 2.2.1 víme, že střední hodnoty a mediány jsou velmi podobné a oscilují pod 1,5 gólů na zápas. Jedinou výjimkou je **BL**, která má medián a střední hodnotu posazenou výše (přibližně 1,5) a obsahuje 2 odlehlá pozorování s velkou gólovou produkcí. Jedná se o první dva týmy konečné tabulky – Bayern a Dortmund. Pro lepší představu do tabulky zapíšeme průměrné hodnoty a mediány, tyto výpočty můžeme vidět v Tabulce 4.12.

Tabulka 4.12: Tabulka průměrných hodnot (a mediánů) počtu gólů za sezónu 22/23

	PL	BL	LL	LO	SA
Průměr	51,95	52,67	46,20	51,45	47,15
Průměr na zápas	1,37	1,55	1,22	1,35	1,24
Medián na zápas	1,29	1,49	1,17	1,28	1,21

Poslední věcí, kterou potřebujeme prověřit, je rozložení týmů v soutěžích s ohledem na střílení branek. Pro tento účel poslouží histogram 4.16. Z něj

Obrázek 4.16: Histogramy počtu vstřelených gólů týmů (za sezónu)



můžeme vyčíst, že střední počet gólů za sezónu se pohybuje kolem 45-55. Výjimkou je **PL**, která má modus (odpovídající hodnotám 7 týmů) mezi 35-45 vstřelenými brankami. Další zajímavostí je, že **LL** a **SA** mají bimodální rozdělení. Speciálně **SA** má první modus v nízkých hodnotách (mezi 25-35), poté je jeden tým s 35-45 vstřelenými góly, dále následuje druhý modus pro skupinu mezi 45-55 góly. To značí, že oproti ostatním soutěžím jsou v **SA** dvě velké rozdílné skupiny, kdy jedna obsahuje týmy s malým počtem vstřelených branek a druhá je podobná produktivitou k ostatním soutěžím. Úplný opak platí pro **LL**, kdy lokální maximum hustoty není v nízkých hodnotách, ale nachází se ve velmi vysokých číslech (65-75). Dále musíme blíže popsat histogram **BL**, protože je to jediná soutěž s 18 účastníky. Velmi zajímavé je porovnání **BL** a **SA**, protože i přes nižší počet zápasu v německé lize je počet týmů ve skupině 25-35 minimální v porovnání s ostatními soutěžemi, speciálně s **SA**. K **BL** zbývá ještě dodat, že histogram je nejvíce koncentrován kolem lokálního maxima hustoty. Posledním nepopsaným grafem je histogram **LO**, ve kterém žádná skupina početně neodskakuje. Jediným vybočujícím týmem je PSG, které nastřílelo mezi 85-95 góly.

Po analýze krabicových grafů a histogramů jsme objevili určité odlišnosti, tou největší je velmi malý počet **BL** týmů ve skupině 25-35 vstřelených gólů za sezónu. Když navíc zohledníme, že tyto kluby odehrály o 4 zápasy méně než týmy v ostatních ligách, tak jsou tyto počty ještě zajímavější.

Víme, že při zápase může padnout pouze diskrétní počet gólů $k = 0, 1, 2, ...$ a navíc se tento počet odvíjí od časového intervalu, kterým je 90 minut herní doby – tedy fotbalový zápas. Z těchto předpokladu (viz 3.2.1) plyne, že počty vstřelených branek se řídí Poissonovým rozdělením, kde parametr λ (střední hodnota i rozptyl) značí průměrný počet gólů na zápas. Ovšem v datech máme počty gólů za celou sezónu, což by mohl být problém. Avšak platí, že pokud se počet gólů za zápas řídí Poissonovým rozdělením, pak má i počet gólů za sezónu Poissonovo rozdělení.

Matematicky formulujeme (zápas zkracujeme na z):

$$\begin{gathered} Gls_z \sim Po(\lambda_z), \text{ pro všechna } z = 1, ..., 38 \\ \text{pak platí} \\ \sum_{z=1}^{38} Gls_z \sim Po(\sum_{z=1}^{38} \lambda_z), \ pro \ Gls_1, Gls_2, ..., Gls_{38} \text{ nezávislé.} \end{gathered} \tag{4.32}$$

Zdrojem čerpání pro větu 4.32 byla přednáška [5]. Pro branky na zápas navíc můžeme uvažovat speciální případ, kde

$$\lambda_1 = ... = \lambda_{38} = \lambda.$$

A to při rozumném předpokladu, že střední hodnota počtu vstřelených gólů je po celou sezónu neměnná. Věta 4.32 platí i pro BL, pouze bychom v zápise snížili počet zápasů na $\sum_{z=1}^{34} Gls_z$. Navíc z věty vyplývá, že střední hodnota gólů za sezónu by pro BL měla být menší. Ovšem když se podíváme do

Tabulky 4.12, tak vidíme, že odhadnutá střední hodnota je nejvyšší a o trochu větší než **PL** a **LO**.

Nyní, když víme, že i góly za sezónu se řídí Poissonovým rozdělením, tak můžeme přistoupit k analýze za využití statistických metod. K porovnání středních hodnot vstřelených gólů v soutěžích „velké pětky“ za sezónu 22/23 využijeme model Poissonovy regrese. Z části 3.3.1 víme, že Poissonovu regresi řadíme do skupiny zobecněných lineárních modelů (Generalized Linear Model) – zkráceně glm. Proto pro vytvoření modelu v softwaru R využijeme naprogramovanou funkci `glm()`, kde do argumentu funkce musíme zadat, že chceme vytvořit model Poissonovy regrese (`family=poisson`). Než přejdeme k vytvoření modelu Poissonovy regrese, musíme dodat, že veličiny, které se řídí Poissonovým rozdělením, jsou v praxi často nadměrně rozptýlené (anglicky *overdispersed*). To znamená, že rozptyl dat je násobně větší než jejich průměrná hodnota. Veličinu Gls z našich dat také řadíme do této kategorie, ale v této práci se tímto problémem zabývat nebudeme, a tak k vytvoření modelu využijeme Poissonovu regresi.

Do modelu Poissonovy regrese gólů za sezónu zahrneme vybrané veličiny z datasetů obecných dat a střelby, které nejvíce souvisí se skórováním branek. Po několika úpravách modelu jsme na základě *Akaikeho informačního kritéria* (AIC[11]), které minimalizuje ztrátu informace v modelu a počet regresorů, vybrali tyto veličiny:

- Comp – liga
- LgRk – umístění na konci soutěže (zohledňuje kvality týmů)
- Dist – průměrná vzdálenost střelby
- SoT – střely na bránu

[11] Podrobnější popis můžeme najít ve skriptech [4] v kapitole 5.

- 90s – zlogaritmované hodnoty počtu zápasů

Samotný model zapisujeme následovně:

$$\ln(Gls) = \beta_0 + \beta_1 I_{[\text{Comp=BL}]} + \beta_2 I_{[\text{Comp=LL}]} + \beta_3 I_{[\text{Comp=LO}]} + \beta_4 I_{[\text{Comp=SA}]} + \beta_5 LgRk + \beta_6 Dist + \beta_7 SoT + \text{offset}(\ln(90s)), \quad (4.33)$$

kde I jsou umělé proměnné (identifikátory) nabývající hodnot 0 a 1 po nesplnění/splnění nějaké podmínky. V našem případě se jedná o podmínku, zda i-té pozorování patří do určité ligy (např. Comp=BL). Parametr β_0 značí hodnotu průsečíku s osou souřadnic. Dále můžeme vidět, že v modelu 4.33 chybí parametr pro **PL**. Ten do rovnice záměrně nepíšeme, protože jsme za referenční skupinu zvolili právě **PL**. Do regresního modelu jsme také zahrnuli funkci `offset()`, jelikož z části 3.3.2 víme, že v poissonovské regresi musíme zohlednit rozdílné délky expozice, a to je v našem případě menší počet odehraných zápasů v **BL**. Proto do funkce `offset()` vložíme zlogaritmované hodnoty počtu zápasů $\ln(90s)$, ovšem pro většinu týmů je počet zápasů stejný. Proto můžeme vytvořit novou veličinu – pojmenujme ji id_Nemecko, do které vložíme upravené počty zápasů pro německé týmy, pro zbytek se bude rovnat 0. Novou veličinu formulujeme takto

$$\text{id\_Nemecko} = \ln(34/38) = \ln(34) - \ln(38).$$

V tomto tvaru vypočítáme pro bundesligové týmy kompenzaci expozice, a můžeme tedy v modelu 4.33 ve funkci `offset()` nahradit $\ln(90s)$ veličinou id_Nemecko. Ovšem musíme id_Nemecko vložit do funkce bez logaritmu, jelikož hodnoty jsou už zlogaritmované.

Než přejdeme k vytvoření modelu v softwaru R, tak předvedeme fungo-

vání funkce `offset()` pro rozdílené expozice na příkladu, kdy vygenerujeme počty branek pro 5 smyšlených soutěží s Poissonovým rozdělením se střední hodnotou $\lambda_z = 2,5$. Pomocí funkce `rpois()` v softwaru R vygenerujeme následující vektory:

- 4 ligy – 41 týmů, 40 odehraných zápasů; označíme ligy $i = 1, 2, 3, 4$ a `rpois(41,100)`
- poslední soutěž – 31 týmů, 30 odehraných zápasů; označíme ligu $i = 5$ `rpois(31,75)`

Střední hodnoty počtu gólů za sezónu vypočítáme podle věty 4.32, kdy vynásobíme hodnotu λ_z počtem odehraných zápasů za sezónu. Následně spojíme vektory do jednoho společného, vytvoříme vektor soutěží a počtu zápasů (veličina 90s se zlogaritmovanými počty utkání 40 a 30). Pak můžeme do funkce `glm()` s argumentem (`family=poisson`) zadat model bez a s zohledněním počtu zápasů. Tyto modely zapisujeme:

$$\ln(Gls) = \beta_1 + \beta_2 I_{[\text{Comp}=2]} + \beta_3 I_{[\text{Comp}=3]} + \beta_4 I_{[\text{Comp}=4]} + \beta_5 I_{[\text{Comp}=5]}$$

$$\ln(Gls) = \beta_1 + \beta_2 I_{[\text{Comp}=2]} + \beta_3 I_{[\text{Comp}=3} + \beta_4 I_{[\text{Comp}=4]} + \beta_5 I_{[\text{Comp}=5]} + \text{offset}(90s).$$

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.6287914  0.0154340 299.909   <2e-16
sim_ligalig2 -0.0226461  0.0219516  -1.032    0.302
sim_ligalig3 -0.0233774  0.0219557  -1.065    0.287
sim_ligalig4  0.0009524  0.0218218   0.044    0.965
sim_ligalig5 -0.3048724  0.0257984 -11.817   <2e-16
```

(a) Parametry pro model bez `offset`

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.6287914  0.0154340 299.909   <2e-16
sim_ligalig2 -0.0226461  0.0219516  -1.032    0.302
sim_ligalig3 -0.0233774  0.0219557  -1.065    0.287
sim_ligalig4  0.0009524  0.0218218   0.044    0.965
sim_ligalig5 -0.0171904  0.0257984  -0.666    0.505
```

(b) Parametry pro model s `offset`

Odhady parametrů ze summary funkce `glm` pro oba modely můžeme vidět níže na této stránce. Ve sloupci `Pr(>|z|)` je uvedena p-value testu rovnosti parametrů (středních hodnot) vzhledem k referenční skupině – liga1. Pro model bez funkce `offset()` je p-value pro soutěž s menším počtem zápasů

skoro nulová. Ovšem když zohledníme počet utkání, odhadnutý parametr se stane nevýznamným, což je výsledek, který jsme očekávali, protože ve všech soutěžích padá stejný počet branek na zápas. V podstatě zlogaritmované konstanty kompenzují kratší časový interval expozice.

Přejděme zpět k Poissonově regresi počtu vstřelených gólů. Odhady parametrů ze summary funkce `glm()` jsou v tabulce níže ve sloupci `Estimate`.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.7563769  0.3820598   9.832  < 2e-16
CompBL       0.1517815  0.0465210   3.263   0.0011
CompLL      -0.0437127  0.0493142  -0.886   0.3754
CompLO      -0.0214195  0.0451464  -0.474   0.6352
CompSA      -0.0218456  0.0488889  -0.447   0.6550
LgRk        -0.0171195  0.0041671  -4.108 3.99e-05
Dist        -0.0269623  0.0202442  -1.332   0.1829
SoT          0.0050392  0.0006841   7.367 1.75e-13
```

Vidíme, že jediná soutěž, která je statisticky na hladině $0,05$ (sloupec `Pr(>|z|)`) významně odlišná od **PL** je **BL**. Navíc je hodnota p-value malá ($< 0,01$), a tedy je střední hodnota vstřelených branek za sezónu velmi odlišná vzhledem k **PL**, ale i ostatním soutěžím, protože zbylé ligy nejsou významně odlišné od **PL**. Jelikož už v modelu nemáme žádné další kvalitativní proměnné, přejdeme ke kvantitativním. Z této množiny jsou statisticky významné SoT a LgRk. Veličina střel na bránu má hodnotu p-value blížící se nule. Což nás asi nepřekvapuje, poněvadž góly jsou podmíněny střelami na bránu. V podobné situaci se nachází i LgRk s hodnotou p-value $< 0,001$.

Dříve než přejdeme k interpretaci parametrů, musíme ověřit kvalitu modelu pomocí reziduálního rozptylu. K dispozici máme dvě statistiky, podle kterých můžeme ověřit, zda je rozptyl reziduí dostatečně malý, model tedy dobře aproximuje naše pozorování. Výše zmíněnými statistikami jsou X^2 a

Tabulka 4.13: Tabulka odhadů RR a reziduálních rozptylů pro Poissonovu regresi gólů za zápas

	BL	LL	LO	SA	LgRk	Dist	SoT
$\widehat{RR}$	1,16	0,96	0,98	0,98	0,98	0,97	1,01
			Reziduální rozptyl				
statistika X^2	72,83		statistika D		72,76	R^2	0,838
kvantil $\chi^2_{90;\,0,95}$	113,15		W$=<113,15,\infty)$			S^2_Y	447,8

D, které můžeme vypočítat z rovnic 3.25 a 3.26 z teoretické části o Poissonově regresi, kde obě verze reziduálních rozptylů se řídí χ^2 rozdělením s $n-p$ stupni volnosti. V Tabulce 4.13 máme vypočítané hodnoty těchto rozptylů a hodnotu 95% kvantilu χ^2 rozdělení o 90 stupních volnosti – odečteme od 98 pozorování 8 odhadnutých parametrů. Navíc je v tabulce vypsán kritický obor W a můžeme vidět, že vypočtené statistiky do tohoto intervalu nespadají. Z toho plyne, že model dostatečně dobře aproximuje naše data. Posledními statistikami v tabulce jsou S^2_Y a R^2, kde první určuje odhad reziduálního rozptylu pro nejjednodušší model (pouze odhadujeme parametr β_0) a druhý nazýváme koeficientem determinace, který udává procentuální hodnotu vysvětlené variability v datech. Ve skriptech [1] v kapitole 7.2.12 jsou popsané vztahy a výpočty této statistiky pro lineární regresi. V případě zobecněného lineárního modelu se budou pouze lišit výpočty reziduálního rozptylu a celkové variability (nejjednodušší model), ale princip výpočtu R^2 zůstává stejný. Platí tedy, že $R^2 = 1 - (D/S^2_Y)$ (funkce `glm()` využívá statistiku D), což pro náš model vychází zhruba na 84 % vysvětlené variability.

Teď když jsem ukázali, že je tento model vhodný pro otestování shody středních hodnot počtu vstřelených branek za sezónu, můžeme přistoupit k interpretaci parametrů v modelu. Z části 3.3.2 víme, že odhady $\boldsymbol{\beta}$ jsou zlogaritmované, a tak musíme parametry vložit do exponenciální funkce, a

tím získáme odhady parametrů $e^{\boldsymbol{\beta}}$. Poté pro interpretaci musíme vypočítat odhady **rate ratio** (RR), ty jsou ovšem v případě Poissonovy regrese stejné jako hodnoty odhadnutých parametrů $e^{\boldsymbol{\beta}}$. Interpretaci RR budeme vztahovat k zápasům, protože víme, že ligové průměry vstřelených gólů za sezónu jsou odvislé od počtu utkání.

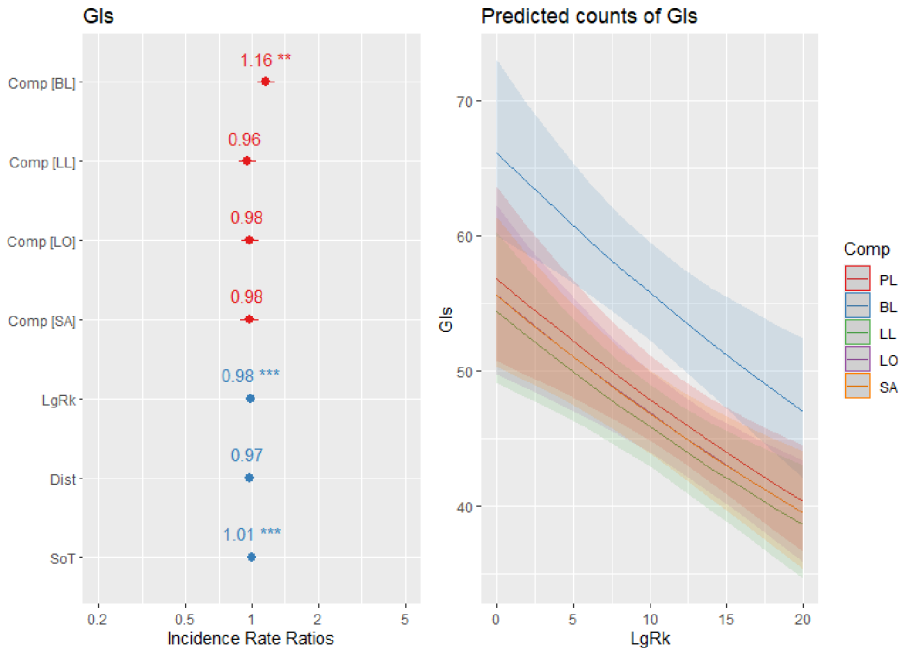
Nejdříve začněme interpretací RR pro **BL**, která se jako jediná soutěž významně odlišuje průměrným počtem vstřelených branek za zápas. Hodnoty odhadů $\widehat{RR}$ jsou v Tabulce 4.13 společně s odhady reziduálních rozptylů, kde $\widehat{RR}$ pro již zmíněnou **BL** se rovná 1,16. Což znamená, že oproti referenční lize (**PL**) padá v **BL** 1,16krát více gólů za zápas. Pro zbývající soutěže nejsou parametry β statisticky významné a hodnoty $\widehat{RR}$ jsou blízké 1 (podíl je roven 1, a to značí shodu). Z toho vyplývá, že **BL** je významně odlišná i do zbývajících soutěží.

Nyní přejdeme na interpretaci statisticky významných kvantitativních veličin. Z části 3.3.2 víme, že se tyto proměnné interpretují podobně jako v lineárním modelu s jednou změnou, kdy při posunutí veličiny o 1 se střední hodnota modelu nemění lineárně, ale má multiplikativní efekt na střední hodnotu. První veličinou je LgRk (umístění v konečné tabulce ligy), která má hodnotu $\widehat{RR}$ rovnu 0,98, to značí, že při zafixování ostatních proměnných a posunutí LgRk o 1 se střední hodnota počtu vstřelených gólů v modelu zmenší na zhruba 98 % původní hodnoty (sníží se o 2 procenta). Stejný postup použijeme pro SoT (střely na bránu), ovšem v tomto případě se střední hodnota počtu vstřelených gólů zvýší o 1 % .

V poslední řadě se podíváme na vývoj střední hodnoty Gls pro Comp v modelu. K vykreslení grafu využijeme funkci `plot_model()` ze softwaru R, která se nachází v balíčku `sjPlot` [12], kde do argumentu (`type="pred"`)

[12] Je potřeba také stáhnout balíček `strengejacke`.

Obrázek 4.18: Grafické znázornění RR a očekávaných gólů pro model Poissonovy regrese za sezónu 22/23



zadáme, že chceme zobrazit vyrovnané hodnoty.

V Obrázku 4.18 máme vykresleny dva grafy, kde graf vlevo zobrazuje odhady $\widehat{RR}$ s hvězdičkami naznačující významnost a intervaly spolehlivosti. Význam hvězdiček je následující:

- $\star$ – p-value<0.05
- $\star\star$ – p-value<0.01
- $\star\star\star$ – p-value<0.001

Zajímavější je pro nás pravý graf, ve kterém vidíme vývoj očekávaných hodnot dle modelu pro jednotlivé ligy v závislosti na LgRk. Tuto veličinu jsme zvolili záměrně, jelikož vývoj vyrovnaných hodnot modelu v závislosti na LgRk je pro nás nejvíce uchopitelný (vývoj střední hodnoty vstřelených gólů v závislosti na síle týmů). Dva zbylé regresory SoT a Dist jsou

v grafu zafixovány (zafixované hodnoty nevykreslených veličin zjistíme příkazem `print(plot_model()$data)`). Funkce `plot_model()` využívá průměrné hodnoty za sezónu – SoT je přibližně rovno 154, pro Dist je přibližná hodnota rovna 17,5. Dále jsou v grafu vykresleny 95% intervaly spolehlivosti, které kolem křivek tvoří pás spolehlivosti. Tento pás vzniká tak, že pro každý bod na regresních křivkách vypočítáme interval spolehlivosti. Bod na regresní křivce libovolné soutěže v závislosti na hodnotě LgRk můžeme zapsat takto $E(Gls|Comp, LgRk)$. Vidíme, že regresní křivka **BL** je posazená výše a pás spolehlivosti kolem této křivky se skoro neprotíná s pásy ostatních soutěží, až na konce křivek, kde máme největší intervaly spolehlivosti z důvodu větší nejistoty (málo pozorováni, odlehlá pozorování). Naopak pásy jsou nejužší kolem středu křivek, v našem případě se jedná o hodnotu $LgRk \doteq 10$. Navíc si při bližším pohledu do grafu můžeme všimnout, že nevidíme fialovou regresní křivku pro **LO**, a to z důvodu překrytí s křivkou SA. Tento jev je zapříčiněn velmi podobnou hodnotou $\widehat{RR}$, která je odhadnuta na 0,98, což je vidět z levého grafu. Ještě zbývá dodat, že slovo křivka jsme v tomto odstavci použili záměrně, jelikož změna střední hodnoty se odvíjí od $0,98^x$ (exponenciální funkce). V našem případě jsou regresní křivky velmi podobné přímkám, a to z jednoduchého důvodu – základ 0,98 je velmi blízký 1.

Nejzásadnějším zjištěním z analýzy Gls modelem Poissonovské regrese je, že BL je významně odlišná od **PL** (i od ostatních soutěží) v počtu vstřelených branek za zápas. S předpokladem neměnné střední hodnoty vstřelených branek za zápas můžeme tento významný rozdíl uvažovat i pro celou sezónu. Proto by bylo záhodno ověřit, zda se tento trend objevuje i v dřívějších ročnících. Pro tento účel využijeme data se stejnou strukturou z ročníku 21/22.

Z krabicového grafu 2.8 víme, že počty xG jsou pro soutěže velmi po-

dobné, a proto dříve než přistoupíme k modelu Poissonovy regrese Gls za sezónu 21/22, zkusíme vytvořit skoro totožný model, do kterého přidáme veličinu xGDiff. Tato veličina určuje rozdíl počtu skutečně vstřelených branek Gls a teoretické hodnoty vycházející z historických dat xG. Jinými slovy, přidáním tohoto rozdílu do modelu vykompenzujeme vstřelené/nevstřelené branky, a to by mělo zapříčinit, že střední hodnota Gls pro **BL** nebude významně odlišná. V podstatě vytvoříme ideální svět, ve kterém odstraníme vliv lidského faktoru ve smyslu hráčských/brankářských chyb i šťastných gólových střel s malou hodnotou xG.

Model s xGDiff vypadá, jak bylo řečeno výše, podobně jako model gólů 4.33. Upravený model zapisujeme takto:

$$\ln(Gls) = \beta_0 + \beta_1 I_{[\text{Comp=BL}]} + \beta_2 I_{[\text{Comp=LL}]} + \beta_3 I_{[\text{Comp=LO}]} + \beta_4 I_{[\text{Comp=SA}]} + \beta_5 LgRk + \beta_6 Dist + \beta_7 SoT + \beta_8 xGDiff + \text{offset}(\ln(90s)).$$

Jelikož jsme už více do hloubky popsali první model Poissonovy regrese, tak pro tento model budeme komentovat pouze nejzajímavější změny. Odhady parametrů můžeme opět vidět níže v tabulce ve sloupci `Estimate` a hodnoty p-value ve sloupci `Pr`$(>|z|)$.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.9153066  0.3860226  10.143  < 2e-16
CompBL       0.0719773  0.0480930   1.497  0.13449
CompLL      -0.0219444  0.0495382  -0.443  0.65778
CompLO      -0.0168924  0.0452642  -0.373  0.70900
CompSA      -0.0454755  0.0492514  -0.923  0.35583
LgRk        -0.0135367  0.0041653  -3.250  0.00115
Dist        -0.0296224  0.0204103  -1.451  0.14668
SoT          0.0042420  0.0006906   6.142 8.13e-10
xGDiff       0.0148622  0.0023942   6.208 5.38e-10
```

Hodnota p-value se pro **BL** velmi změnila a rozdíl ve střední hodnotě počtu gólů **BL** oproti **PL** je nyní nevýznamný (i v porovnání s ostatními soutěžemi). To znamená, že se naše domněnky z odstavce výše vyplnily a rozdíl je

pro soutěže statisticky nevýznamný. Pro kvantitativní proměnné SoT a xGDiff je hodnota p-value velmi nízká ($< 0,001$) a jejich efekty jsou statisticky významné.

Odhady RR a reziduálních rozptylů najdeme v Tabulce 4.14, ze které vidíme, že v **BL** znovu padá více gólů na zápas, a to zhruba 1,08krát více než u ostatních soutěží. Ovšem tento rozdíl je v tomto modelu nevýznamný. Co se týče interpretace xGDiff (efektivita), vidíme, že s rostoucím počtem gólů roste i efektivita, a to při posunu o 1 se zvýší střední hodnota o 1,5 %. Což značí, že kvalitnější týmy, které střílí více branek, jsou v zakončení efektivnější. To je způsobeno větší kvalitou kádru. Pro zbylé významné kvantitativní veličiny zůstávají odhady podobné jako v předchozím modelu.

Ve zkratce si popíšeme statistiky vypovídací o kvalitě modelu. V porovnání s předchozím neupraveným regresním modelem se odhady reziduálních rozptylů snížily přibližně o polovinu. Což znamená, že nezamítáme nulovou hypotézu o dobrém modelu. Z důvodu snížení reziduálního rozptylu došlo k navýšení koeficientu determinace $R^2 \doteq 0,92$. Navíc se změnil i kvantil χ^2 rozdělení, protože se zvýšil počet odhadovaných parametrů na 9.

Tabulka 4.14: Tabulka odhadů RR a reziduálních rozptylů pro Poissonovu regresi gólů za sezónu s přidáním xGDiff

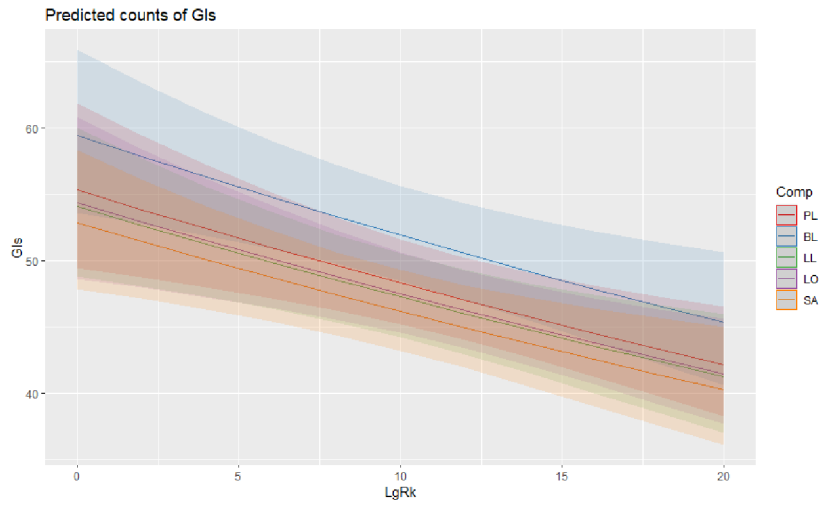
	BL	LL	LO	SA	LgRk	xGDiff	SoT
$\widehat{RR}$	1,075	0,978	0,983	0,956	0,987	1,015	1,004
			Reziduální rozptyl				
statistika X^2	33,81		statistika D		34,158	R^2	0,924
kvantil $\chi^2_{89;\,0,95}$	112,022		W=$< 112,022, \infty)$			S^2_Y	447,8

Zbývá vykreslit graf vyrovnaných hodnot modelu dle soutěží. Opět pro vykreslení využijeme funkci `plot_model()` a vykreslíme graf v závislosti na umístění v soutěži. Proměnné SoT a Dist jsou znovu pro vyrovnané hodnoty

v grafu zafixovány na průměrných hodnotách, pro nově přidanou veličinu xG-Diff je průměrná přibližně rovna -1,07. Výsledkem je Obrázek 4.19, ve kterém vidíme, že rozdíl mezi soutěžemi není statisticky významný, jelikož se pásy spolehlivosti kolem křivek velmi překrývají. Tento výsledek jsme očekávali s ohledem na hodnoty p-value pro odhady $\boldsymbol{\beta}$. Zbývá dodat, že křivky mají velmi podobný tvar jako v modelu bez veličiny xGDiff, a to z důvodu velmi podobného odhadu parametru LgRk.

Shrnutím modelu Poissonovy regrese gólů s přidáním veličiny xGDiff (efektivity) je naplnění domněnky neodlišnosti jednotlivých soutěží při zohlednění proměněných a neproměněných šancí týmů.

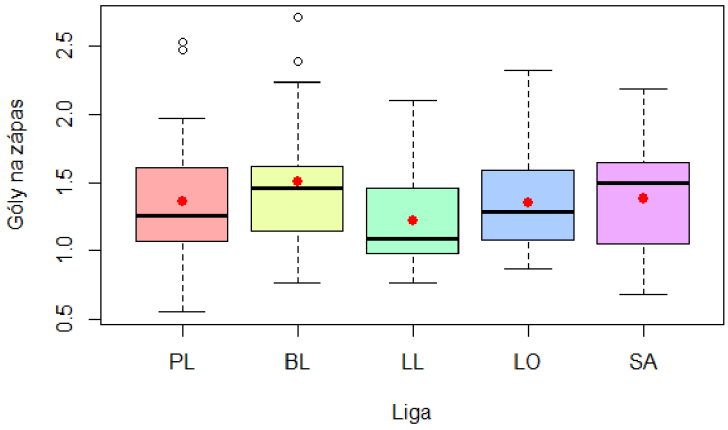
Obrázek 4.19: Očekávané góly pro model Poissonovy regrese s přidání xGDiff za sezónu 22/23



4.2.1 Poissonova regrese gólů za ročník 21/22

Jelikož jsme modelem poissonovské regrese zjistili, že v **BL** padalo za sezónu 22/23 více branek než v ostatních soutěžích, a tak zkusíme ověřit, zda se tento jev opakuje i v předchozím ročníku. K analýze využijeme datasety

Obrázek 4.20: Boxploty počtů vstřelených gólů za sezónu 21/22



obecných dat a střelby za ročník 21/22. Regresní model gólů za sezónu 22/23 zkráceně označíme jako model 22/23.

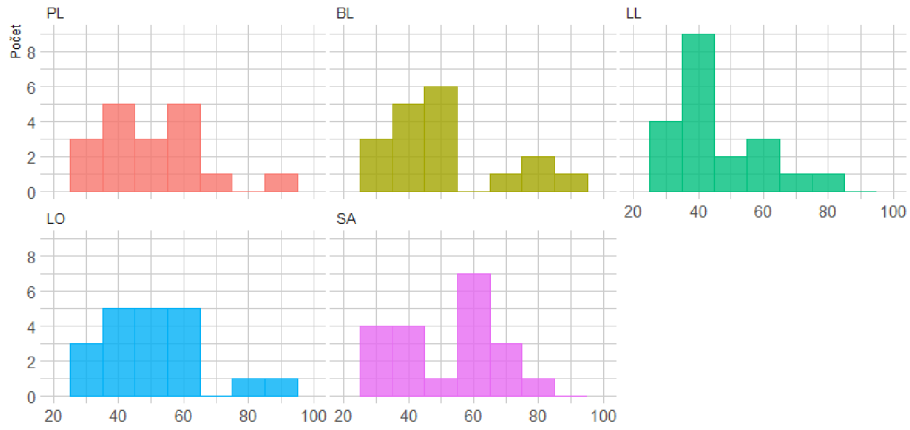
Nejdříve vykreslíme krabicový graf 4.20 počtu vstřelených gólů na zápas pro ročník 21/22. Do ligových krabic zobrazíme i průměrné hodnoty počtu vstřelených gólů na zápas. Při pohledu do grafů zjistíme, že se góly na zápas pro soutěže příliš neliší. Jedinou velkou změnou v porovnání s krabicemi 2.4 z ročníku 22/23 je velké navýšení počtu vstřelených gólů na zápas pro italskou **SA**. Toto navýšení můžeme vysledovat v Tabulce 4.15, kde v porovnání se sezónou 22/23 jde o nárůst přibližně 0,15 gólů na zápas, což je zajímavé. Abychom zjistili, které průměrné počty branek na zápas jsou pro **SA** běžné, musíme vypočítat průměrné hodnoty pro předcházející ročníky. Tyto výpočty nám poskytne webová stránka FBref ([17]), kde v záložce Squad Stats a položce Standard Stats jsou vypočítané průměrné hodnoty počtu vstřelených gólů na zápas pro předchozí ročníky. Tyto průměrné hodnoty jsou podobné ročníku 21/22, a tedy sezóna 22/23 byla pro **SA** výjimečná v malém počtu

Tabulka 4.15: Tabulka průměrných hodnot (a mediánů) počtu gólů za sezónu 21/22

	PL	BL	LL	LO	SA
Průměr	51,85	51,33	46,35	51,55	52,60
Průměr na zápas	1,36	1,51	1,21	1,36	1,38
Medián na zápas	1,26	1,46	1,1	1,29	1,5

vstřelených gólů.

Obrázek 4.21: Histogramy počtu vstřelených gólů za sezónu 21/22



Lepší vhled do rozdělení týmů podle počtu vstřelených gólů můžeme vidět v histogramu 4.21. Největší rozdíl můžeme spatřit pro italskou **SA** v porovnání s ročníkem 22/23. Hlavní změnou je velmi zvýšený počet klubů ve skupině 55-65 a snížená koncentrace ve skupinách s nízkým počtem vstřelených gólů. Pro **BL** je situace v porovnání s následující sezónou neměnná, pouze můžeme evidovat více týmů ve skupině 25-35 a větší koncentraci pro skupiny s velkým počtem vstřelených gólů.

Do modelu Poissonovy regrese gólů pro ročník 21/22 zahrneme stejné regresory jako v případě modelu 22/23, jelikož parametry kvantitativních regresorů byly velmi významné. Jako referenční skupinu opět zvolíme **PL**.

Zápis modelu je následující:

$$\ln(Gls_{21/22}) = \beta_0 + \beta_1 I_{[\text{Comp=BL}]} + \beta_2 I_{[\text{Comp=LL}]} + \beta_3 I_{[\text{Comp=LO}]} + \beta_4 I_{[\text{Comp=SA}]} + \beta_5 LgRk + \beta_6 Dist + \beta_7 SoT + \text{offset}(\ln(90s)). \quad (4.34)$$

Výstupem ze summary funkce `glm()` jsou odhadnuté parametry ve sloupci `Estimate` a hodnoty p-value ve sloupci `Pr(>|z|)`. Jediný významný odhad parametru pro soutěže náleží **BL**, což znamená, že **BL** je opět významně odlišná od **PL** (i od ostatních soutěží). Avšak hodnota p-value (< 0.05) pro parametr **BL** je nižší než v modelu 22/23. Odhady parametrů kvantitativních veličin SoT a LgRK jsou znovu velmi významné (p-value< 0.001), dokonce je i statisticky významný odhad parametru pro proměnnou Dist s hodnotou p-value< 0.01.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.1186419  0.3116550  13.215  < 2e-16
CompBL       0.1118949  0.0461734   2.423   0.0154
CompLL       0.0081367  0.0482093   0.169   0.8660
CompLO       0.0732852  0.0454812   1.611   0.1071
CompSA       0.0306882  0.0440241   0.697   0.4858
LgRk        -0.0209486  0.0043271  -4.841 1.29e-06
Dist        -0.0447365  0.0173205  -2.583   0.0098
SoT          0.0049465  0.0006344   7.797 6.32e-15
```

Odhady $\widehat{RR}$ a vypočtené statistiky reziduálních rozptylů opět najdeme v Tabulce 4.16, kde můžeme vidět, že vypočítané hodnoty statistik X^2 a D nenáleží do kritického oboru W, a tedy model dostatečně dobře aproximuje naše data.

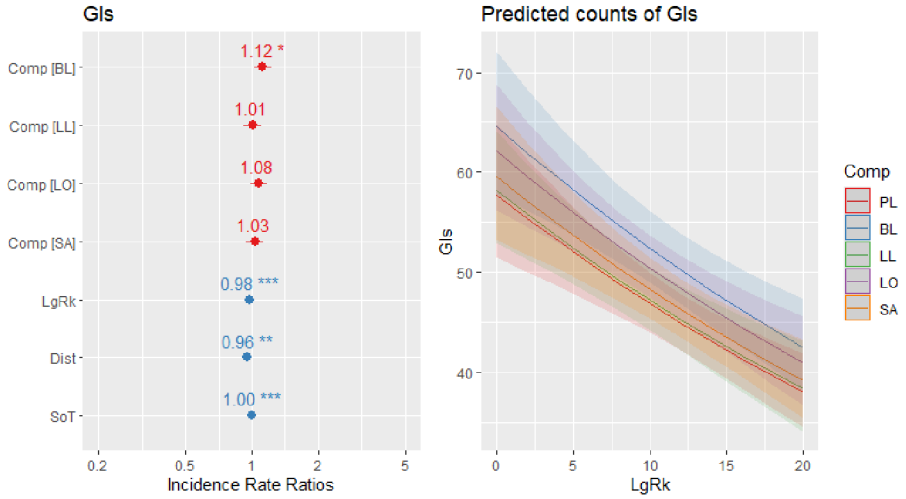
Když jsme ověřili kvalitu modelu, přistoupíme k interpretaci RR. Pro **BL** platí, že padá přibližně 1,12krát více gólů na zápas než v **PL** (i ve zbylých soutěžích). Odhady $\widehat{RR}$ jsou pro SoT a LgRk velmi podobné jako pro model 22/23, a tak je už blíže interpretovat nebudeme. Avšak oproti

Tabulka 4.16: Tabulka odhadů RR a reziduálních rozptylů pro Poissonovu regresi gólů na zápas pro ročník 21/22

	BL	LL	LO	SA	LgRk	Dist	SoT
$\widehat{RR}$	1,118	1,008	1,076	1,031	0,979	0,956	1,005
			Reziduální rozptyl				
statistika X^2	54,834		statistika D		54,315	R^2	0,89
kvantil $\chi^2_{90;\,0,95}$	112,022		W$=<112,145,\infty)$			S^2_Y	504,215

modelu 22/23 je veličina Dist v tomto modelu významná, a tak si uvedeme její vliv na podmíněnou střední hodnotu počtu vstřelených gólů. Tedy při zafixování ostatních proměnných a posunutí Dist o 1 se střední hodnota počtu vstřelených gólů v modelu sníží přibližně na 95 % původní hodnoty, což znamená, že při střelbě z větší vzdálenosti průměrný počet vstřelených gólů klesá, což je logické.

Obrázek 4.22: Grafické znázornění RR a predikovaných gólů pro model Poissonovy regrese za sezónu 21/22



Posledním bodem je graf vyrovnaných hodnot, pro jehož vykreslení znovu využijeme funkci `plot_model()`. Výsledkem je Obrázek 4.22, kde v levém grafu jsou vykresleny odhady $\widehat{RR}$ s hvězdičkami naznačující významnost a

v pravém vývoj očekávaných hodnot modelu pro jednotlivé ligy v závislosti na LgRk. Zbylé kvantitativní proměnné SoT a Dist jsou zafixovány na průměrných hodnotách za sezónu 21/22, které jsou velmi podobné průměrným hodnotám z ročníku 22/23. Vidíme, že tvar regresních křivek je skoro identický s modelem 22/23 (odhad $\widehat{RR}$ proměnné LgRk je přibližně roven 0,98). Dále je vidno, že křivka **BL** je dle očekávání posazena výše, ovšem i přes významnou odlišnost se více překrývají pásy spolehlivosti kolem křivek.

Modelem Poissonovy regrese vstřelených gólů za sezónu 21/22 jsme zjistili, že v **BL** opět padalo více branek na zápas, a tedy i za celou sezónu, než v ostatních ligách.

Závěr

Cílem této práce bylo statistickými metodami porovnat soutěže „velké pětky“. Samotnou analýzu jsme rozdělili do dvou částí, a to na otestování vybraných kvantitativních veličin z datových sad v závislosti na lize s využitím ANOVY nebo Kruskalova-Wallisova (K-W test) testu a porovnání počtu vstřelených branek v soutěžích modelem Poissonovy regrese.

S využitím modelů Poissonovy regrese jsme zjistili, že v obou sezónách padalo v **BL** významně více branek za sezónu než v ostatních soutěžích. Z regresního modelu za ročník 22/23 plyne, že v **BL** padalo 1,16krát více gólů za sezónu než ve zbylých soutěžích. Pro model předcházejícího ročníku je odhad $\widehat{RR}$ pro **BL** trochu nižší a přibližně roven 1,12. Pro modely Poissonovy regrese zbývá dodat, že při přidání veličiny xGDiff (efektivity zakončení) do modelu pro ročník 22/23 se rozdíl v počtu vstřelených gólů za sezónu pro ligy stává nevýznamným.

Pro veličiny testované ANOVOU nebo K-W testem na hladině významnosti 0,05 jsme pro ročník 22/23 zjistili, že menší počet zápasů v **BL** nemá vliv na počet hráčů v kádru. Další veličiny jsme testovali v obou sezónách, a tak uvedeme významné rozdíly, které byly pro oba ročníky identické. Jako první jsme odhalili, že v **LO** mají mladí hráči větší herní vytížení než v **LL**. Z toho plyne, že pro trenéry španělských klubů je důležitější zkušenost hráčů. Dále jsme zjistili, že je více žlutých karet na zápas udíleno v **LL** než ve všech

zbylých soutěžích kromě **SA**, což značí nedisciplinovanost hráčů z těchto soutěží. U červených karet jsme došli k podobnému závěru, pouze s tím rozdílem, že k italské a španělské lize řadíme **LO**. V poslední řadě uvedeme rozdíly pro zákroky a vypíchnutí na zápas, kdy pro tyto veličiny evidujeme, že **LO** má více zákroků a vypíchnutí na zápas než **LL** a **SA**. To značí, že hráči v **LO** chtějí dříve získat míč zpět pod kontrolu než ve výše zmíněných soutěžích.

Na závěr jsem rád, že jsem si vybral toto téma bakalářské práce. Jako fanda fotbalu jsem zjistil spoustu zajímavých zjištění o srovnání soutěží „velké pětky“.

Seznam literatury

[1] HRON, K., KUNDEROVA, P., VENCÁLEK, O.: *Základy počtu pravděpodobnosti a metod matematické statistiky*. 3. přepracované vydání. Olomouc: Univerzita Palackého v Olomouci, 2018. ISBN 978-80-244-5398-9.

[2] DOBSON, A. J., BARNETT, A. G.: *An introduction to Generalized Linear Models*. Third Edition. Boca Raton: CRC Press, 2008. ISBN 978-1-58488-950-2.

[3] NOVOTNÁ, J.: *Robustní odhady a testy v modelech poissonovské regrese* [online]. Diplomová práce. Praha: FJFI ČVUT, 2021. Dostupné z: https://dspace.cvut.cz/handle/10467/98232

[4] FIŠEROVÁ, E.: *Lineární statistické modely*. 2. dopl. vydání. Olomouc: Univerzita Palackého v Olomouci, 2015. ISBN 978-80-244-4797-1.

[5] KLAZAR, M.: *The Poisson distribution and approximation* [online]. Přednáška. Praha: MFF UK, 2021. Dostupné z https://kam.mff.cuni.cz/ klazar/l13L_PRTE_20.pdf

Internetové zdroje

[6] *UEFA.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/UEFA

[7] *UEFA.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://cz.wikipedia.org/wiki/UEFA

[8] *Bundesliga.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/Bundesliga

[9] *Serie A.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/Seria_A

[10] *La Liga.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/La_Liga

[11] *Ligue 1.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/Ligue_1

[12] *Premier League.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/Premier_League

[13] *UEFA coefficient* [online]. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2023-05-19]. Dostupné z: https://en.wikipedia.org/wiki/UEFA_coefficient

[14] *UEFA rankings.* [online]. © 1998-2024 UEFA [cit. 2024-02-03]. Dostupné z: https://www.uefa.com/nationalassociations/uefarankings/.

[15] *UEFA European Cup Football* [online]. KASSIES, Bert. [cit. 2023-06-01]. Dostupné z: https://kassiesa.net/uefa/index.html

[16] *Livesport: Fotbal* [online]. © 2006-23 Livesport.cz [cit. 2023-05-26]. Dostupné z: https://www.livesport.cz/

[17] *2022-2023 Big 5 European Leagues Stats* [online]. © 2000-2023 Sports Reference LLC [cit. 2023-06-01]. Dostupné z: https://fbref.com/en/comps/Big5/2022-2023/shooting/squads/2022-2023-Big-5-European-Leagues-Stats

[18] *Expected goals (xG) explainer* [online]. © 2024 The Analyst. All Rights Reserved. [cit. 2024-02-03]. Dostupné z: https://theanalyst.com/eu/2023/08/what-is-expected-goals-xg/.