UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Determinanty úspěchu v play-off NHL – statistická analýza

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval(a): **Tomáš Chupáň**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2021

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Tomáš Chupáň

Název práce: Determinanty úspěchu v play-off NHL – statistická analýza

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2021

Abstrakt: Cílem práce je zjistit, do jaké míry souvisí výsledky týmů NHL v základní části s eventuálním úspěchem či neúspěchem v play-off. Na základě statistik ze základní části chceme také odhadovat pravděpodobnosti toho, kam se tým ve vyřazovací fázi soutěže dostane. Byl zjištěn významný vztah umístění týmu v tabulce a počtu vyhraných sérií, dále byl odhadnut model proporcionálních šancí, který blíže popisuje vztah některých hokejových ukazatelů a výsledků v play-off.

Klíčová slova: lední hokej, NHL, model proporcionálních šancí, metody pro analýzu ordinálních dat, chí-kvadrát test nezávislosti

Počet stran: 95

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Tomáš Chupáň

Title: Determinants of success in the NHL playoffs – statistical analysis

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2021

Abstract: The aim of the thesis is to find out how the results of NHL teams in the regular season are related to possible success or failure in the playoffs. Based on the statistics from the regular season, we also want to estimate the probabilities of where the team will go in the elimination phase of the season. A significant relationship was found between the team's position in the table and the number of series won, and a proportional odds model was estimated, which describes in more detail the relationship between some hockey statistics and the results in the playoffs.

Key words: ice hockey, NHL, proportional odds model, methods for ordinal data analysis, chi-square test of independence

Number of pages: 95

Number of appendices: 0

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Seznam tabulek

Poděkování

Rád bych poděkoval vedoucímu práce Mgr. Ondřeji Vencálkovi, Ph.D. za všechen čas a cenné rady, které mi při psaní této práce a celém studiu věnoval. Také bych chtěl poděkovat své rodině za vytvoření perfektního zázemí, které mi při psaní práce výrazně pomohlo.

Úvod

„Sezóna začíná až v play-off," nebo také „Hokej ve vyřazovací fázi je úplně jiný sport!" – tyto a podobné výroky můžeme mnohdy slyšet v rozhovorech s hokejisty i trenéry na konci základní části, kdy se už blíží očekávaný vrchol ročníku.

Je tomu však skutečně tak? Mažou se výkony mužstva ze základní části, když jim začíná boj o pohár? Napoví nám statistiky z průběhu sezóny to, kam se pravděpodobně daný celek v play-off dostane? Nebo se s poslední sirénou bojů o body do tabulky stávají tyto statistiky bezcennými údaji?

Na tyto a další otázky se snaží přinést odpovědi tato bakalářská práce, která se zaměřuje na nejkvalitnější hokejovou ligu světa – zámořskou NHL. Podíváme se na to, zda existuje vztah mezi počtem vyhraných sérií v bojích o Stanley Cup a některou z týmových či individuálních statistik počítaných za základní část, případně jejich vhodnou kombinací.

Uvažovaná data budou z ročníků 2013/14 – 2018/19, tedy z posledních šesti kompletních sezón. V roce 2020 se sezóna na jaře přerušila kvůli pandemii nemoci COVID-19 a play-off se hrálo až v létě po dlouhé pauze – případný vliv výkonů ze základní části by tak mohl být značně zkreslen, neboť zde chyběla přímá návaznost mezi jednotlivými fázemi soutěže. Proto není sezóna 2019/20 uvažována.

První kapitola se věnuje NHL, jejímu organizačnímu uspořádání a formátu vyřazovací části. Ve druhé kapitole si představíme analyzované datové sady, neboť bude třeba vysvětlit, co jednotlivé ukazatele znamenají a jaké otázky by nás v souvislosti s nimi mohly zajímat. V další kapitole si uvedeme vhodné statistické metody, které budeme v rámci analýzy dat používat. Poslední kapitola pak obsa-

huje samotnou analýzu dat, jejíž výsledky jsou zdůrazněny na konci této kapitoly a v závěru práce.

Kvůli velkému množství internetových zdrojů jsou tyto zdroje odděleny od standardního seznamu literatury, a to zejména kvůli větší přehlednosti. Stejně tak jsou odděleně citovány použité balíčky softwaru R, který byl využíván v průběhu psaní bakalářské práce jak pro zpracování a přichystání datových sad, tak pro analýzu samotnou.

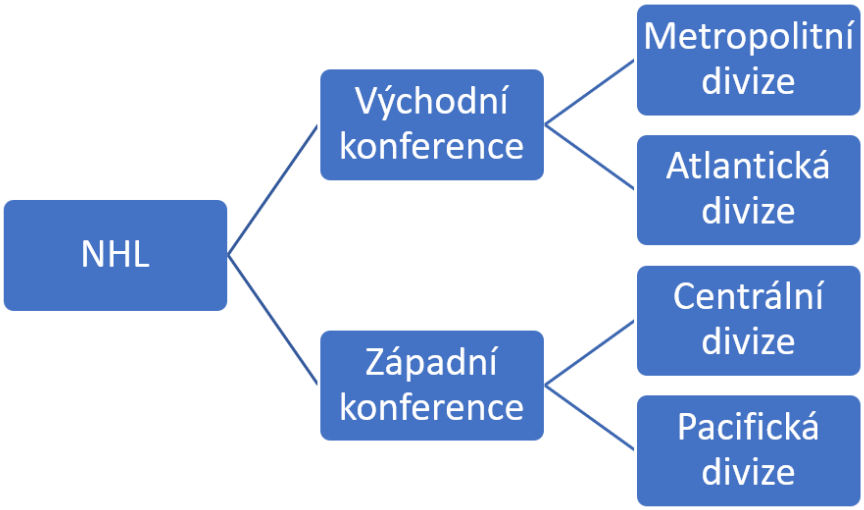
Kapitola 1

NHL a její play-off formát

1.1 Členění ligy

National Hockey League (NHL) je profesionální kanadsko-americká hokejová soutěž, jíž se v současnosti účastní 31 týmů a která je právem považovaná za nejkvalitnější hokejovou ligu světa. Každé z mužstev se rok co rok snaží dosáhnout co nejlepších výsledků v základní části (hrané na 82 utkání), postoupit do play-off a bojovat o zisk Stanley Cupu, nejcennější hokejové trofeje.

Týmy jsou v NHL rozděleny do dvou konferencí – Východní a Západní – přičemž obě se ještě dělí na dvě divize:



Obrázek 1.1: Členění ligy na konference a divize

Tabulka 1.1: Rozdělení týmů do konferencí a divizí od sezóny 2013/14

Východní konference		**Západní konference**	
Atlantická divize	**Metropolitní divize**	**Centrální divize**	**Pacifická divize**
Boston Bruins	Carolina Hurricanes	Chicago Blackhawks	Anaheim Ducks
Buffalo Sabres	Columbus Blue Jackets	Colorado Avalanche	Calgary Flames
Detroit Red Wings	New Jersey Devils	Dallas Stars	Edmonton Oilers
Florida Panthers	New York Islanders	Minnesota Wild	Los Angeles Kings
Montreal Canadiens	New York Rangers	Nashville Predators	Phoenix Coyotes
Ottawa Senators	Philadelphia Flyers	St. Louis Blues	San Jose Sharks
Tampa Bay Lightning	Pittsburgh Penguins	Winnipeg Jets	Vancouver Canucks
Toronto Maple Leafs	Washington Capitals		Vegas Golden Knights

Budeme analyzovat data ze sezón 2013/14 – 2018/19. Právě od sezóny 2013/14 se týmy rozdělily dle výše zmíněného uspořádání ligy [27], jak můžeme vidět v Tabulce 1.1.[1] Od sezóny 2017/18 se do Pacifické divize začlenil nový tým, Vegas Golden Knights, další změny nastanou až od sezóny 2021/22, kdy se Arizona Coyotes (nové jméno týmu z Phoenixu od ročníku 2014/15) přesune do Centrální divize a k lize se připojí 32. tým – Seattle Kraken – který bude zařazen do Pacifické divize, a rozložení týmů v konferencích i divizích bude vyrovnané (16 týmů v každé konferenci, 8 týmů v každé divizi).

1.2 Formát play-off

Vyřazovací fáze se účastní 16 týmů (8 z každé konference), ovšem systém postupu není tak jednoduchý, že by o Stanley Cup bojovalo 8 týmů s nejlepším bodovým ziskem z každé konference.

Jistými účastníky play-off jsou v pořadí první 3 týmy z každé divize (můžeme je značit např. A1–A3 pro Atlantickou divizi, podobně M1–M3, C1–C3, P1–P3), zbytek týmů se seřadí v konferenční tabulce „divokých karet“. Šest týmů z 1.–3. míst v divizích tak na východě i západě doplní dvojice, která získala nejvyšší počet bodů ze zbylých týmů, a to už bez ohledu na divizi. V každé konferenční větvi play-off tak přibudou 2 týmy, označené např. WC1 a WC2 (od Wild Card). Tento systém tak připouští i situaci, kdy se do play-off dostane některý z týmů

[1] Tabulka převzata z https://en.wikipedia.org/wiki/History_of_organizational_changes_in_the_NHL.

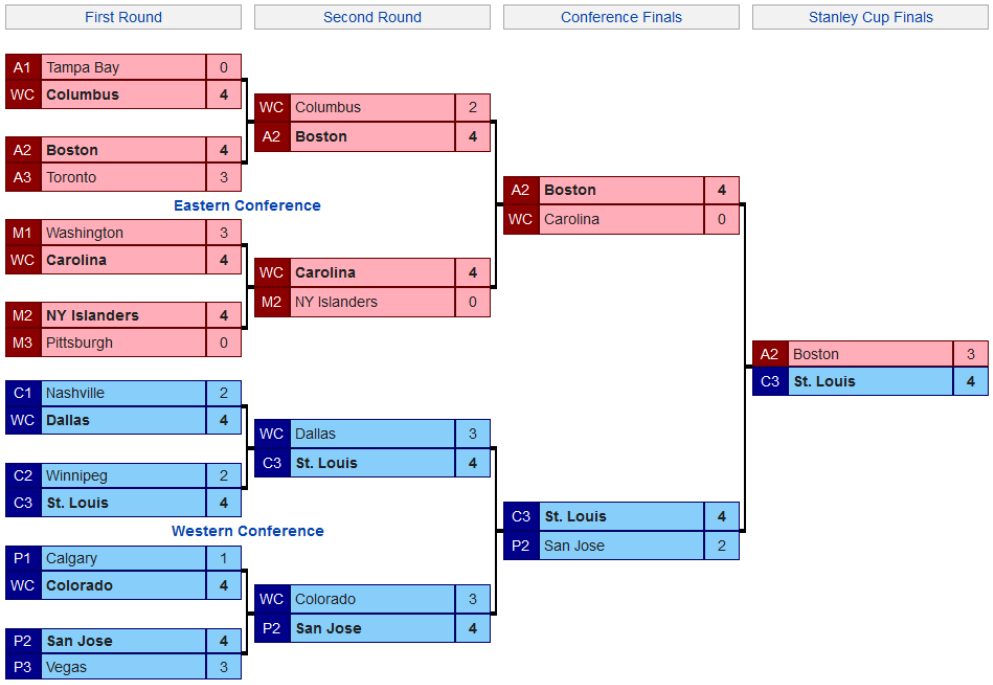
na úkor jiného, ačkoliv má méně bodů (nutno podotknout, že tento jev není zcela výjimečný). I z tohoto důvodu zastávám názor, že by se pravidla pro postup měla upravit.

Další problém, který se pokusím popsat, je nasazování jednotlivých týmů v play-off proti sobě. Play-off má celkem 4 kola, a to 1. kolo, 2. kolo, finále konferencí (Conference Final) a finále (Stanley Cup Final). V sériích hraných na 4 vítězné zápasy proti sobě nejprve nastupují týmy ze stejné konference. Pavouk se hned v 1. kole rozdělí dle divizí, úvodní dvojice vzniknou tak, že se utkají 2. a 3. týmy z každé divize (A2 vs. A3, M2 vs. M3 atd.) a vítězové divizí (A1, M1,...) nastoupí proti držitelům divokých karet. Vítěz divize s lepším bodovým ziskem ze základní části soupeří s horším týmem z tabulky divokých karet (WC2), vítěz divize s nižším počtem bodů odehraje sérii proti týmu WC1. U týmů z tabulky divokých karet se tak nepřihlíží k příslušnosti k divizi.

Vítězný tým z dvojice A2–A3 nastoupí ve 2. kole proti vítězi série A1–WC1/2[2], postupující z dvojice M2–M3 proti vítězi série M1–WC1/2, analogicky u zbylých dvou divizí. Postupující z 2. kola se pak utkají v rámci finále konferencí, z nichž vzejdou vítězové „svých“ konferencí, kteří změří síly ve finále o Stanley Cup. Finále je tak jedinou sérií, kde dojde na mezikonferenční souboj. Na Obrázku 1.2 si můžeme prohlédnout příklad takového pavouka.

Tampa Bay tuto základní část ovládla, měla nejvíce bodů ze všech týmů (tedy i více než Washington, vítěz Metropolitní divize), proto šla v prvním kole na tým Columbusu, který získal méně bodů než Carolina (resp. byl druhý v tabulce divokých karet Východní konference). Columbus sice patří do Metropolitní divize, na divizní příslušnost se však u divokých karet nebere ohled, proto může být v „atlantické“ části pavouka. V Západní konferenci Calgary bodově předčilo tým z Nashville, v 1. kole tedy soupeřilo s horším (co se zisku bodů týče) z dvojice Colorado – Dallas. Je vhodné zmínit a na tomto příkladu rovnou ukázat, jak se rozhoduje o přisouzení výhody domácího prostředí v sérii a jak se na herním plánu projeví. „Každá série se hraje ve formátu 2-2-1-1-1, což znamená, že tým

[2]WC1 či WC2 podle výše zmíněného rozhodovacího pravidla.



Obrázek 1.2: Rozlosování a průběh play-off NHL v sezóně 2018/19, zdroj: [11, Screenshot]

s výhodou domácího ledu hostí zápasy 1, 2, 5 a 7, zatímco jeho soupeř hostí zápasy 3, 4 a 6.“[28] Tým s výhodou domácího ledu se určuje dvojím způsobem. V 1. a 2. kole je to ten tým série, který je výše postavený v rámci divize, ve finále konferencí i samotném finále je to tým, který získal v základní části více bodů, bez ohledu na jeho umístění v divizi. V případě bodové shody stanovuje NHL několik dalších rozhodovacích kritérií, která lze najít např. na stránce [28]. V uvedeném pavoukovi tak měli např. ve 2. kole výhodu začínat v domácím prostředí Boston, NY Islanders, St. Louis a San Jose, ve finále začínal na domácím stadionu celek Bostonu, protože získal v základní části více bodů než St. Louis [26]. Jak jsem již uváděl výše, série se hrají na 4 vítězná utkání – je tedy zřejmé, že zápasy 5–7 se hrají pouze v případě nutnosti.

V této kapitole jsem mimo uvedené odkazy čerpal především z internetových zdrojů [16] a [24].

Kapitola 2

Data

2.1 Obecná charakteristika datových souborů

V celé práci se vyskytuje více datových souborů, které však spolu vzájemně souvisí a mají mnoho společných rysů. Například to, že jde o údaje za sezóny 2013/14 – 2018/19, tedy za 6 posledních kompletně dohraných sezón. Hlavními dvěma důvody pro tento výběr jsou tyto:

1. Od ročníku 2013/14 začala platit nová kolektivní smlouva mezi ligou a hráčskou asociací – změna některých pravidel, přeuspořádání ligy (dle Tabulky 1.1).

2. Ročník 2019/20 dle mého nemělo význam započítávat, pokud je cílem zkoumat vztahy mezi výsledky ze základní části (ZČ) a play-off. Tato sezóna byla přerušena kvůli situaci ohledně COVID-19, ZČ se nedohrála a play-off začalo po tak dlouhé pauze, že by výsledky analýz mohly být zkresleny.

Pro získání dat jsem použil dva zdroje: Hlavním zdrojem byly oficiální stránky NHL ([25] a [26]), velmi užitečné sady dat se však nacházejí také na odkazu [22] – odtud jsem čerpal především hráčské statistiky, viz později. Týmy v datech nemají plné názvy, používají se obecně užívané zkratky (tvořící v datových sadách sloupce s názvem *Te*).

Dalším společným znakem datových souborů je to, že zahrnují pro konkrétní sezónu pouze ty týmy, které se do vyřazovací části probojovaly. Výsledky ze

základní části budou totiž používány nikoliv k predikci samotného postupu do play-off, nýbrž k předpovědi úspěchu postoupivších týmů ve vyřazovací fázi sezóny.

Právě „úspěch“ v bojích o Stanley Cup bude měřen počtem vyhraných sérií, který tak vzhledem k formátu soutěže může nabývat hodnot z množiny {0,1,2,3,4}, přičemž jednotlivá čísla mají logické slovní ekvivalenty a bude tak na ně pohlíženo jako na kategorie:

- 0 ... Tým byl vyřazen v 1. kole
- 1 ... Tým byl vyřazen ve 2. kole
- 2 ... Tým byl vyřazen ve finále konferencí
- 3 ... Tým prohrál ve finále
- 4 ... Tým vyhrál Stanley Cup

2.2 Základní údaje ze ZČ a výsledek v play-off

Základní údaj o výsledku v ZČ – umístění v tabulce podle počtu získaných bodů – je k dispozici ve dvou variantách, a to Umístění v konferenci (UvK) a Umístění v lize (UvL). Preferovat přitom budeme hlavně UvK, neboť většina soubojů vyřazovací části se, jak už bylo zmíněno, odehraje v rámci konferencí. Pokud bude v některé části k porovnání dvou týmů sloužit UvL, bude to zmíněno.

S umístěním samozřejmě souvisí i počet získaných bodů, který je pro pořadí v tabulce rozhodující. Za výhru týmu v zápase základní části (ať už v základní hrací době, tj. po 60 minutách hry, nebo až v prodloužení/nájezdech) se udělují 2 body, za prohru v prodloužení/nájezdech 1 bod, po prohře v základní hrací době mužstvo nezíská žádný bod. Součet získaných bodů je také součástí datové sady.

Další, spíše pomocné sloupce datového souboru prezentovaného v Tabulce 2.1 jsou Rok play-off (RP) a jednotliví soupeři v průběhu vyřazovací části (S1, S2, SCF, SF).

Tabulka 2.1: Ukázka datového souboru – základní údaje

Te	UvK	UvL	VS	RP	S1	S2	SCF	SF	Body
BOS	1	1	1	2014	DET	MTL			117
ANA	1	2	1	2014	DAL	LAK			116
COL	2	3	0	2014	MIN				112
STL	3	4	0	2014	CHI				111
SJS	4	5	0	2014	LAK				111
PIT	2	6	1	2014	CBJ	NYR			109
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.3 Týmové statistiky

Kromě těchto základních údajů bylo potřeba získat také konkrétní statistiky, které by mohly být klíčem pro predikci úspěchu v play-off. Jelikož je takových statistik téměř nepřeberné množství, zvolil jsem jich pouze několik, a to podle toho, které mi přišly relevantní, často citované a používané při hodnocení jednotlivých týmů a jejich výkonů.

Tabulka 2.2: Ukázka datového souboru – týmové statistiky 1

Te	RP	VS	B	G.Z	InG.Z	PP	PK	ShF	ShA	FO	GDiff	SDiff
BOS	2014	1	117	3.15	2.09	21.7	83.7	31.9	29.1	51.6	1.06	2.8
ANA	2014	1	116	3.21	2.48	16.0	82.2	31.3	28.7	49.2	0.73	2.6
COL	2014	0	112	2.99	2.63	19.8	80.7	29.5	32.7	49.5	0.36	-3.2
STL	2014	0	111	2.91	2.29	19.8	85.7	29.3	26.4	51.9	0.62	2.9
SJS	2014	0	111	2.91	2.35	17.2	84.9	34.8	27.8	52.8	0.56	7.0
PIT	2014	1	109	2.95	2.49	23.4	85.0	29.9	28.8	51.0	0.46	1.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Mezi tyto základní statistiky se řadí např.:

- počet vstřelených gólů na zápas (G.Z)
- počet inkasovaných gólů na zápas (InG.Z)
- úspěšnost přesilových her (PP) – zde uváděno v %
- úspěšnost hry v oslabení (PK) – taktéž v %

- počet střel na branku soupeře na zápas (ShF)
- počet soupeřových střel na branku týmu na zápas (ShA)
- úspěšnost na vhazování (FO) – v %

Klíčem k úspěchu v hokejovém zápase je poměrně jednoduché pravidlo – vstřelit více gólů než soupeř. To, že tým vstřelí 4 góly, je sice nadprůměrný počin, avšak pokud soupeř vstřelí branek 5, bodový zisk mužstvo mine. Tato úvaha mě vedla k vytvoření sloupce GDiff, který je prostým rozdílem mezi počtem vstřelených a inkasovaných gólů na zápas. Podobně byl vytvořen i sloupec s proměnnou SDiff, který tak udává, o kolik dané mužstvo v průměru přestřílelo tým soupeře, resp. (v případě záporného výsledku) o kolik střeleckých pokusů bylo průměrně horší.

Statistiky v Tabulce 2.2 jsou však považovány za takový „základ“, řekněme rychlý přehled, jak se v hrubých obrysech týmu v základní části dařilo. Co dál by nás mohlo zajímat?

Například to, jak tvrdě tým hrál – indikátorem tvrdé hry zpravidla bývají **hity**. Schopnost „přitvrdit hru“ je v play-off velmi ceněná, má však na lepší vyřazovací část vliv množství hitů v základní části?

Také se můžeme podívat, jak moc hráči týmu puky ztráceli a kolik kotoučů naopak získávali. Tyto statistiky jsou v datech standardizovány na 60 minut hry.

Poznámka: Ztráta puku je situace, při které hráč udělá nevynucenou chybu, v důsledku čehož odevzdá puk soupeři. **Získané puky** jsou „formou změny mezi týmy v držení puku, v níž dojde k přímému odebrání puku hráčem“[23].

Modře zvýrazněné jsou ty statistiky, jejichž hodnoty jsou počítány pouze ze situací při hře 5 na 5, tedy v rovnovážném počtu hráčů na ledě. Jakmile si vysvětlíme, co ukazatele znamenají, bude zřejmé, proč tomu tak je.

Sloupec **Út.P%** označuje procento startů v útočném pásmu na začátku střídání. Čím větší podíl vhazování absolvuje tým před brankářem soupeře, tím by měl mít větší šanci gól vstřelit než dostat.

Pokud by se počítala i vhazování v přesilovkách a oslabeních, byl by tento

Tabulka 2.3: Ukázka datového souboru – týmové statistiky 2

Te	RP	VS	Hity60	Ztr60	Zisk60	SAT%	SAT%Tesne	Út.P%	Stř%	Usp%
BOS	2014	1	24.20	7.07	6.65	53.9	54.7	54.3	8.5	94.0
ANA	2014	1	24.33	7.79	5.37	50.0	49.5	50.9	9.8	92.5
COL	2014	0	24.39	5.77	7.59	46.9	47.2	49.5	8.8	92.9
STL	2014	0	22.06	3.90	6.52	53.1	53.1	52.6	8.5	92.2
SJS	2014	0	19.50	9.55	7.66	53.5	53.6	49.6	7.5	92.2
PIT	2014	1	26.01	7.01	4.64	48.6	49.5	49.2	8.3	91.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

ukazatel velmi zkreslený, neboť při své přesilovce tým vždy začíná střídání před brankou soupeře a ve většině případů se tam vyskytuje i při dalších případných vhazováních během doby, kdy má tým protivníka hráče na trestné lavici. Hodnoty této statistiky počítané z celé hrací doby by pak velmi záležely na tom, kolik času tráví tým v oslabení/přesilovce.

Podobně úspěšnost střelby (**Stř%** – kolik % střel skončí gólem) a úspěšnost zákroků gólmana (**Usp%** – kolik % střel gólman chytí) je lépe počítat pro situace, kdy mají oba týmy stejný počet hráčů na ledě. Při přesilové hře střílí nepotrestaný tým z mnohem lepších pozic, které znamenají statisticky větší pravděpodobnost, že z nich padne gól. Proto i úspěšnost střelby je v přesilových hrách vyšší. Procento úspěšných zákroků je ze stejného důvodu při oslabení nižší. Gólman bránícího se týmu čelí střeleckým pokusům z těžších pozic a vykazuje tak „horší čísla“, než kterých dosahuje při vyrovnané hře.

Obě tyto statistiky se počítají pouze ze střel na branku, střely mimo branku (kam patří např. i střela do tyče) a zblokované střely se do těchto ukazatelů nezahrnují.

Co se týče hodnot a významu SAT, považuji je za natolik zajímavé (rozumějte často citované, označované za moderní, relevantní statistiku s velkou vypovídací hodnotou), že se jeho charakteristice budu věnovat blíže v části 2.3.1. Tento ukazatel totiž mnozí znají pod jiným názvem – **Corsi**.

Vzhledem k tomu, co bylo psáno v předchozích odstavcích o přesilové hře, je zřejmé, v jaké nevýhodě je tým, který často čelí početní převaze. Naopak se cení, pokud je hráč schopen pro svůj tým přesilovku získat – čímž se nemyslí nafilmování pádu známé z fotbalu pro získání penalty, spíše jde o bojovnost, rychlé bruslení a důsledné napadání, kdy soupeři často nezbude nic jiného, než svého protivníka faulovat, aby ho zastavil. I proto je vhodné zahrnout do zkoumaných statistik počet obdržených **trestných minut** (jakožto nějakou „míru neukázněnosti“), případně i **rozdíl mezi tresty soupeře a obdrženými tresty** (v Tabulce 2.4 opět standardizováno na 60 minut hry). U tohoto rozdílu tedy platí, že kladný výsledek je pro tým dobrý, neboť jejich soupeři jsou trestáni více a tým tak hraje víc přesilovky než oslabení. U záporných hodnot tohoto ukazatele je tomu naopak, tým dostává více trestů než jeho protivník a více se tak brání.

Tabulka 2.4: Ukázka datového souboru – týmové statistiky 3

Te	RP	VS	TrMin	TrestProProti60	BulyPP	BulyPK
BOS	2014	1	886	-0.45	57.25	49.52
ANA	2014	1	894	0.14	52.96	41.83
COL	2014	0	891	0.06	57.49	44.42
STL	2014	0	1162	-0.14	54.83	48.37
SJS	2014	0	737	0.86	57.70	47.70
PIT	2014	1	832	0.31	59.34	44.91
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Dále je u situací s nevyrovnaným počtem hráčů na ledě důležité, kdo vyhraje buly neboli vhazování (myšleno důležitější než při hře 5 na 5). Bránící se tým má při vyhrané buly šanci vyhodit puk ze svého obranného pásma až na druhou stranu hrací plochy (oslabenému týmu se nepíská zakázané uvolnění), což ho stojí nejméně sil právě po vhazování. Odebrat kotouč již rozestavenému a kombinujícímu mužstvu v přesile je mnohem náročnější. Ze stejného důvodu chce buly vyhrát tým, který má početní výhodu – má možnost se rozestavět do přesilovkové formace, kombinovat s pukem, hrát natrénované signály atd. Proto je v Tabulce 2.4 i údaj o **úspěšnosti na buly při přesilovce a oslabení** (opět v %).

2.3.1 Corsi%

V této části jsem čerpal z internetových zdrojů [12] a [23].

V Tabulce 2.2 jsou data o střelách na branku. Pokud se však bavíme o *střeleckých pokusech*, do této kategorie kromě střel na branku spadají také střely mimo a zblokované střely (zkrátka situace, kdy se hráč pokusí o střelu, která může jít jak na bránu, tak mimo ni, případně může trefit soupeře). Pomocí střeleckých pokusů (z anglického Shot Attempts, proto SAT) můžeme komplexněji sledovat střeleckou převahu týmu. Místo SAT se většinou používá právě název Corsi, podle trenéra brankářů, který začal tyto údaje sledovat (příběh o pojmenování tohoto ukazatele je však mnohem zábavnější, jak se lze dočíst například v článku [20]). Samotné Corsi je vypočítáno jako rozdíl Corsi for (CF = střelecké pokusy na branku soupeře) a Corsi against (CA = soupeřovy střelecké pokusy na branku daného týmu), to vše je v absolutních číslech. Znovu upozorňuji – počítá se pouze ze situací 5 na 5, do této statistiky se nezapočítávají střelecké pokusy z přesilovek/oslabení, neboť tým s početní výhodou má zpravidla mnoho střeleckých pokusů, naopak bránící se tým by nasbíral mnoho zápisů do statistiky CA.

Sledované Corsi% (tzv. relativní Corsi) udává relativní počet střeleckých pokusů na branku soupeře vzhledem k celkovému počtu všech střeleckých pokusů (na libovolné straně), tedy

$$\text{Corsi\%} = \frac{CF}{CF + CA} \cdot 100\%. \tag{2.1}$$

Vyjadřuje tak, na kolika % střelecké produkce v zápasu se tým podílí. Hodnoty nad 50 % jsou brány jako nadprůměrné, tým je dominantnější co se střelby týče. Hodnoty pod 50 % jsou zase vnímány jako podprůměrné, tým s takovým výsledkem je soupeři „přestřílen“ a čelí více střeleckým pokusům, než jich vyprodukuje. Zpravidla se také Corsi% používá jako ukazatel toho, jak který tým drží puk. V některých případech se statistika označuje jako Corsi For %, jde však o stejné číslo, které jsme si ukázali v předchozím vzorci.

Statistika Corsi se počítá také na individuální rovině. Když je hráč na ledě při střeleckém pokusu svého týmu, započítá se mu bod do CF. Pokud se naopak při jeho účasti uskuteční střelecký pokus na branku jeho týmu, přibude mu zápis do CA. Individuální Corsi, dané rozdílem individuálních CF a CA, je tak ukazatelem toho, jaký vliv má hráč na schopnost týmu produkovat střelecké pokusy a v porovnání s týmovým Corsi lze vidět, zda je tento vliv pozitivní či negativní. Opět se častěji používá relativní verze této statistiky (počítaná dle individuálních CF a CA dosazených do rovnice (2.1)), která pak říká, jak se při hráčově účasti na ledě tým podílel na celkové střelecké produkci, zda spíše útočil, nebo se musel střelám bránit.

Pokud má například na konci sezóny hráč v kolonce Corsi% číslo 62, znamená to, že když byl hráč na ledě, 62 % všech sledovaných střeleckých pokusů mířilo na branku soupeře, tedy že hráčův tým protivníka přehrával. Když je při stejné situaci týmové Corsi% například „jen“ 53, pak měl hráč velmi pozitivní efekt na střeleckou produkci svého týmu, díváme se na něj jako na nadprůměrně dobrého.

Často se pro porovnání místo týmového Corsi% bere tzv. „off-ice Corsi%“, počítané ze střeleckých pokusů, které se uskutečnily, když byl daný hráč na střídačce. Jeho individuální relativní Corsi vypočítané výše popsaným způsobem se pak nazývá „on-ice Corsi“. Porovnání těchto dvou hodnot nám napoví, zda tým – zjednodušeně řečeno – víc kontroluje hru při situaci, kdy je hráč na ledě, nebo hraje lépe, pokud hráč sedí na střídačce.

V Tabulce 2.3 pak vidíme ještě sloupec **SAT%Tesne** – je spočítán stejně jako právě popsané týmové relativní Corsi, avšak s tím rozdílem, že se do něj zahrnují pouze střelecké pokusy z těch zápasů, které se označují jako těsné. Dle Slovníku pojmů na stránkách NHL [23] jsou to ty zápasy, v nichž je v průběhu první či druhé třetiny gólový rozdíl maximálně jednobrankový, nebo je ve třetí třetině stav utkání vyrovnaný, nebo zápas dojde do prodloužení.[1] Tento ukazatel je tak

[1] Těsné zápasy jsou pro play-off velmi typické.

očištěn od těch zápasů, ve kterých má tým výrazně navrch či naopak výrazně zaostává za soupeřem.

2.4 Hráčské statistiky

Hokej je bezesporu týmová hra. Přesto by byla škoda nevyužít rozsáhlé pole hráčských statistik, které se mnohdy v těch týmových přímo neobjeví, avšak mohly by hrát významnou roli. Pár takových jsem vybral z velmi obsáhlých datových sad dostupných na stránce [22], jak je již zmíněno v části 2.1. Uvažovány jsou pouze statistiky hráčů v poli, tedy útočníků a obránců. Z nich jsem vypočítal číselné charakteristiky, které by mohly být z hlediska úspěšnosti týmu zajímavé. Nejedná se jen o charakteristiky polohy, ale také charakteristiky variability, sumy či pouze počty hráčů, kteří dosáhli na určitou metu v dané statistice. Struktura dat je tak ve výsledku stejná jako v předchozí části zabývající se týmovými statistikami, ovšem je jí docíleno pomocí transformace jiných datových sad.

Samozřejmě, podobným způsobem se na oficiálních zdrojích dávají dohromady i některé týmové statistiky – počet střel jako součet dílčích střel hráčů týmu, vstřelené góly na zápas jako součet gólových zápisů hráčů vydělený počtem zápasů apod. Při transformacích hráčských statistik do druhého typu datových sad je tak kladen důraz na to, aby se ukazatele zbytečně nezdvojovaly a abychom získali unikátní informaci, kterou se nám z týmových statistik nepodařilo zachytit. Nicméně, může se stát, že výsledný ukazatel by svou povahou mohl klidně být zařazen mezi statistiky týmové – viz např. celkový počet dorážek v Tabulce 2.7. Je to však dáno tenkou hranicí mezi těmito dvěma skupinami údajů, ač některé jsou ryze týmové (získané body, umístění) či ryze individuální, hráčské (ice-time/zápas, viz další odstavec) – některé jsou však skutečně těžko zařaditelné. V této práci jsou rozlišeny zejména na základě zdroje, z kterého byly získány a který je za danou kategorii ukazatelů považoval.

Abychom nepohlíželi na výsledky hráčů, kteří téměř nic neodehráli, ale do statistik se zapsali, ponechal jsem v datech pouze ty hráče, kteří odehráli alespoň 60 minut. To je sice ekvivalent jednoho zápasu, avšak při nízkém ice-time

(čase stráveném na ledě) některých hráčů s menší rolí v mužstvu to pro ně může znamenat klidně 8–10 zápasů (přičemž běžný zápasový ice-time tahounů týmu se pohybuje kolem 20 minut u útočníků a 25 minut u obránců). Tímto způsobem jsme vyřadili statistiky 335 hráčů, což představuje přibližně 12 % celkového počtu hráčů za uvažovaných 6 sezón. Pracujeme tedy s údaji o 2425 hokejistech.

Také je potřeba upozornit na jednu malou komplikaci při práci s těmito daty. Je běžné, že během sezóny změní před uzávěrkou přestupů (konec února, přičemž sezóna startuje v první půli října) několik hráčů tým – jsou tzv. „vytrejdovaní". Z jejich statistik tak část zapsali v jiném týmu, než ve kterém ročník končili. V použité datové sadě však toto rozlišeno není a příslušný hráč je vždy s výsledky za celý ročník (ať už začal hrát kdekoliv) přiřazen k týmu, ve kterém sezónu končil, a tedy s kterým se účastnil play-off.

Musíme tedy myslet na to, že u pár jedinců jsou statistiky sice obrazem jejich hry za celou základní část, avšak nemusí plně korespondovat s tím, jak hráli v rámci daného týmu. Nemělo by to však příliš vadit vzhledem k počtu těchto případů a tomu, že s přiřazeným týmem hráli play-off. My budeme zkoumat právě vliv hry v základní části na část vyřazovací, sumarizované schopnosti všech hráčů (i trejdovaných) tak k výsledkům v play-off přispěly.

Pro ilustraci budou v následujících tabulkách statistiky počítané z čísel všech hráčů bez rozdílu pozice, v analýze (viz kapitola 4) však budu používat i totožné ukazatele počítané zvlášť pro obránce a zvlášť pro útočníky, neboť jejich role v týmu jsou dost odlišné. Bude jistě rozumnější počítat např. medián kanadských bodů (góly + asistence[2] [3], někdy uváděné pouze jako *body* – neplést ovšem s body, které tým získává do tabulky, viz 2.3) odděleně pro tyto pozice – útočníci sbírají obecně mnohem více bodů než obránci a vlivem případného velkého počtu obránců, kteří do dané sezóny zasáhli, by se tak tato hodnota mohla výrazně posunout směrem dolů, ač by třeba útočníci byli produktivní.

[2]Nahrávky nejvýše dvou spoluhráčů, které předcházely gólu.
[3]Dle pořadí – retrospektivně od gólu – se někdy dělí na primární a sekundární.

Také jsou mnohdy použity statistiky standardizované na 60 minut hry – zvlášť u produktivity nám to pomůže identifikovat týmy, kde hráči nejefektivněji využívají čas, který na ledě stráví.

Tabulka 2.5: Ukázka datového souboru – hráčské statistiky 1

Te	RP	VS	pocet_hr	icetimeZ_sd	KB_sd	KB60_med	KB_NAD30	goly_sd
BOS	2014	1	27	237.51	21.88	1.29	10	9.69
ANA	2014	1	26	239.71	21.66	1.66	9	10.28
COL	2014	0	23	281.61	22.22	1.39	10	9.21
STL	2014	0	24	275.25	20.55	1.21	11	9.35
SJS	2014	0	25	263.00	23.30	1.20	8	10.62
PIT	2014	1	32	255.46	25.26	1.21	7	9.95
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

U každé z hráčských statistik nás bude zajímat, zda nám pomůže odpovědět na určitou otázku ohledně úspěchu ve vyřazovací části.

Samotný **počet hráčů**, kteří za tým odehráli více než 60 minut, může být zajímavý (ač jde o tak obyčejnou statistiku). Je lepší, když má tým široký kádr a střídá sestavu, nebo když hraje stále stejných 18 bruslařů?[4] A může například velký počet hráčů, kteří odehráli více než 60 minut, znamenat kromě šíře kádru také větší náchylnost ke zraněním a z toho vyplývající nutnost použít v sezóně více hokejistů?

Může se tým spoléhat na hlavní tahouny s obrovskou porcí minut, přičemž někteří hráči naskočí na pár střídání? Nebo je pro úspěch týmu lepší, když posílá trenér na led formace vyváženě, s podobným ice-timem? To by nám mohla prozradit **směrodatná odchylka ice-time/zápas**, která bude tím větší, čím větší bude nevyrovnanost mezi herním vytížením hráčů.

Ohledně **kanadských bodů** i **gólů** se můžeme prostřednictvím jejich směrodatné odchylky zaměřit na podobnou otázku – nahrává dobrým výsledkům ve vyřazovací části, když má tým v základní části pár odskočených, často bodujících/skórujících hráčů (řekněme „superhvězd“, jak jsou nejproduktivnější hráči často nazýváni)? Nebo je lepší mít produkci rozloženou napříč týmem s tím, že každá

[4]Maximální počet hráčů v sestavě pro každý zápas je 20, z toho jsou 2 gólmani [18].

lajna může znamenat velké nebezpečí pro soupeřovu branku, přičemž nikdo z týmu výrazněji nevyčnívá?[5]

Ukazatel toho, jak efektivně hráči využívají časový prostor na ledě z hlediska produkce, by mohl být **medián bodů/60 minut hry**, resp. **medián gólů/60 minut hry**, jejichž hodnoty oproti klasickým kanadským bodům a gólům vyzdvihují i hráče, kteří se sice v tabulkách produktivity nepohybují na předních příčkách, vzhledem k malému ice-time jsou však týmu velmi platní. Naopak nám tato standardizace odfiltruje často bodující hráče, kteří by však vzhledem k svému hernímu vytížení měli získávat bodů mnohem víc.

Tabulka 2.6: Ukázka datového souboru – hráčské statistiky 2

Te	RP	VS	goly60_med	golyNAD10	gamescore_med	gamescore_IQR
BOS	2014	1	0.54	10	21.20	47.48
ANA	2014	1	0.55	9	27.04	24.45
COL	2014	0	0.59	10	21.90	29.70
STL	2014	0	0.36	8	23.12	47.19
SJS	2014	0	0.30	10	28.07	34.02
PIT	2014	1	0.43	10	9.48	27.28
⋮	⋮	⋮	⋮	⋮	⋮	⋮

V Tabulkách 2.5 a 2.6 jsou také proměnné udávající počet hráčů, kteří pokořili určitou metu získaných bodů, resp. vstřelených gólů. Zde jsou ukázány obecné hranice 30 bodů, resp. 10 gólů, bez rozlišení pozice.

Jak už však bylo zmíněno, statistiky máme v rámci této práce k dispozici také zvlášť pro útočníky a zvlášť pro obránce, kde jsou hranice zvoleny takto:

- útočníci – 25 bodů a 10 gólů
- obránci – 15 bodů a 6 gólů

Tyto údaje, alespoň co se týká veřejně dostupných informací a postupů, nejsou nijak oficiálně sledovaným ukazatelem, jedná se spíš o informace, které se často zmiňují v tisku, rozhovorech, různých analýzách [21][6] a podobně – proto bylo nutné hranici nějakým způsobem určit.

[5] Kromě sm. odchylek jsou ve všech případech zahrnuty i IQR – eliminace vlivu outlierů.

[6] Analýza týmu St. Louis Blues před play-off (které vyhráli) – v ZČ 13 10gólových střelců.

Statistice **gamescore**, u níž uvažujeme medián, směrodatnou odchylku, mezikvartilové rozpětí a tytéž charakteristiky ve standardizované verzi na 60 minut hry, se věnuji v samostatné části 2.4.1.

V části 2.3 již bylo zmíněno, že vstřelení gólu má různou pravděpodobnost z různých pozic. Nezáleží ovšem jen na vzdálenosti od branky či z jakého úhlu hráč střílí. Důležitým faktorem je také čas od poslední střely – čím kratší doba uplynula od předchozího střeleckého pokusu, tím větší je šance na vstřelení gólu. Je to poměrně logické, gólman nemá tolik času se po předchozím zákroku vrátit zpět do optimální pozice a přichystat se na další střelu.

Brankář dokonce ani nemusí zasáhnout, nebezpečné jsou i střely vyslané brzy po zblokovaném střeleckém pokusu či střele mimo – jde o to, že gólman musí na střelu zareagovat a změnit svůj postoj, což jej na pár okamžiků činí méně připraveného čelit následujícím střeleckým pokusům. Proto by dalším zajímavým ukazatelem mohl být **počet dorážek**. Dorážka je definována jako střelecký pokus, který následuje do 3 vteřin od střeleckého pokusu předchozího, aniž by se v tomto časovém intervalu přerušila hra [17].

V datové sadě týmů účastnících se play-off v letech 2014–2019 tvořily dorážky v průměru 7,26 % všech střel, které tým v základní části vyslal na branku soupeře. Góly z dorážky přitom představovaly v průměru 16,62 % všech vstřelených gólů daného týmu. Informace, kolik gólů z dorážek tým vstřelil a jaká byla jeho **úspěšnost** v tomto typu střelby (tedy kolik % dorážek skončilo gólem) jsou taktéž součástí datové sady.

Mohlo by se zdát, že vysoký či nízký počet dorážek je spíše věcí náhody než nějaké schopnosti či strategie týmu. Je pravdou, že k mnoha dorážkám se hokejista jistě dostane s velkou porcí štěstí díky náhodnému odrazu – není tomu tak vždy. Obecně se ví, jak nebezpečné jsou střely z dorážek, hráči proto mnohdy cíleně střílejí tak, aby se puk vhodně odrazil (např. od betonů gólmana) a jejich spoluhráči měli šanci na dorážku. Cílem střely tedy v tomto případě není gól, nýbrž možnost dorazit puk do branky. I proto jsem tuto informaci do datové sady zařadil.

Tabulka 2.7: Ukázka datového souboru – hráčské statistiky 3

Te	RP	VS	Dor_sum	Dor%	ZtrObr_sum	Corsi%AS_med	Corsi%AS_IQR
BOS	2014	1	130	26.92	276	55.21	35.57
ANA	2014	1	189	23.28	313	61.64	25.17
COL	2014	0	190	22.63	245	61.54	38.74
STL	2014	0	131	25.95	200	60.22	27.64
SJS	2014	0	220	20.00	377	57.89	34.87
PIT	2014	1	178	25.28	300	52.63	37.25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Statistika Corsi již představena byla, výpočet relativního Corsi také (část 2.3.1). Stejným způsobem dojdeme k **relativnímu Corsi After Shifts** počítanému individuálně pro každého hráče. Od již definovaného relativního Corsi se liší tím, že se každému hráči do rovnice (2.1) započítávají ty střelecké pokusy, které se uskuteční v čase mezi 1 a 5 sekundami poté, co hráč opustí ledovou plochu, tedy jde střídat – ať už pro (CF) či proti (CA). Proto název *After Shifts*, tedy „po střídání“. Znovu platí, že tento ukazatel počítáme pouze ze střeleckých pokusů při hře 5 na 5.

Tento ukazatel by měl vyzdvihovat hráče, kteří jdou střídat v momentu, kdy je jejich tým v držení puku a míří na branku soupeře (vysoké Corsi%AS). V negativním smyslu by měl identifikovat ty hráče, kteří odcházejí z ledu v případě, kdy má puk na holi soupeř a míří na branku jejich týmu – tito hráči by se správně měli vracet a podílet se na obraně, místo toho si jdou odpočinout na hráčskou lavici (nízké Corsi%AS) [22]. Pro posouzení toho, který z těchto dvou případů u hráče převažuje, používáme opět hranici 50 %.

Jako číselné charakteristiky této statistiky by mohly být vhodné medián (jakožto nějaká střední míra zodpovědnosti při střídání) a mezikvartilové rozpětí (napovídající, zda hráči týmu střídají podobně (ne)ukázněně, nebo je mezi jejich přístupem k této problematice větší variabilita). Otázkou bude, zda tento pokus o kvantifikaci týmové disciplíny ve specifické situaci může napovědět něco o výsledku týmu v play-off, např. že by týmy s vysokým mediánem této statistiky dosahovaly lepších výsledků než týmy, které v základní části střídaly v méně vhodné okamžiky.

Ztráta puku (definovaná v části 2.3) je pro tým nepříjemností, samostatně se však sledují ještě **ztráty puku v obranném pásmu**, u kterých se zřejmě předpokládá, že je po nich větší riziko inkasování gólu.

Samozřejmě, tato statistika bude jistě pozitivně korelovaná s ukazatelem „Ztráty puku na 60 minut hry“ z části 2.3, není však záměrem mít v pozdějším modelu (viz kapitola 4) všechny v této části uváděné statistiky, cílem je spíše nasbírat rozumné množství relevantních a potenciálně přínosných ukazatelů, z kterých pak různými statistickými metodami (zmíněnými v kapitole 3) vybereme ty, jejichž zahrnutím získáme nejlepší možný model.

2.4.1 Gamescore

V této části jsem čerpal ze zdroje [29].

Gamescore je ukazatel vytvořený Domem Luszczyszynem, novinářem pracujícím pro The Atletic, který se snaží kvantifikovat celkový přínos hráče v zápase.

Uvažuje tyto individuální zápasové statistiky:

- góly (G)
- primární asistence (A1)
- sekundární asistence (A2)
- střely na branku (SOG)
- zblokované střely soupeře (BLK)
- „získané“ fauly (PD)

  = faul na daného hráče, z kterého plyne početní výhoda pro hráčův tým
- způsobené fauly (PT)
- vyhraná a prohraná vhazování (FOW a FOL)
- Corsi For a Corsi Against (CF a CA) – počítané při hře 5 na 5

- účast na ledě při vstřelené brance (GF)
- účast na ledě při obdržené brance (GA)

Všechny tyto údaje mají stanovenou váhu podle velikosti vlivu, který na konečný výsledek zápasu mají, a také příslušné znaménko váhy dle toho, zda mají na hru týmu pozitivní či negativní vliv.

Vzorec pro výpočet je následující:

$$\begin{aligned}\text{Game Score} = {} & 0.75 \cdot \text{G} + 0.7 \cdot \text{A1} + 0.55 \cdot \text{A2} + 0.075 \cdot \text{SOG} + \\ & + 0.05 \cdot \text{BLK} + 0.15 \cdot \text{PD} - 0.15 \cdot \text{PT} + 0.01 \cdot \text{FOW} - \\ & - 0.01 \cdot \text{FOL} + 0.05 \cdot \text{CF} - 0.05 \cdot \text{CA} + 0.15 \cdot \text{GF} - 0.15 \cdot \text{GA}\end{aligned} \tag{2.2}$$

V každém zápase se dle (2.2) spočítá hráčovo gamescore, za celou základní část mu sečtením těchto dílčích zápisů vyjde celkové gamescore. I tento ukazatel bere v potaz pouze některé aspekty hry, posuzovat hráčův přínos pouze na základě těchto hodnot by bylo krátkozraké – to však platí pro téměř všechny statistiky. Gamescore díky svému záběru do různých herních situací poskytuje poměrně komplexní obrázek o tom, jak dobře si hráč v zápasech vede.

V rámci datové sady je pro každý tým spočítán medián z výsledků hráčů za celou sezónu, stejně tak mezikvartilové rozpětí a směrodatná odchylka (opět jako míry variability výkonnosti v rámci týmu). Pro lepší zhodnocení toho, jak efektivně hráči využívali čas strávený na ledě, je vhodné použít i standardizovanou verzi ukazatele na 60 minut hry, u níž v datech evidujeme tytéž číselné charakteristiky.

Kapitola 3

Statistické metody

Cílem této kapitoly je představit statistické metody, které později použijeme pro zodpovězení otázek, které jsme si položili v částech předchozích. Vzhledem k tomu, že nás primárně zajímá ověření vztahu vyhraných sérií (kategoriální proměnná) a statistik ze základní části (převážně kvantitativní znaky, umístění v konferenci by se dalo považovat za kvalitativní znak), bude vhodné si ukázat pár metod pro ověření vztahu kvalitativního a kvantitativního znaku, případně dvou kvalitativních znaků, včetně způsobu, jak se dá tento vztah modelovat.

Pro zkoumání vztahu dvou kvalitativních znaků si nejprve ukážeme χ^2 test nezávislosti, u jehož popisu a konstrukce čerpám ze skript [7]. Někdy je ovšem přínosné zohlednit ordinalitu (tedy přirozené uspořádání kategorií) kvalitativních znaků – o tom pojednává část 3.1.2, která vychází z publikace [1].

Závislost kvalitativního a kvantitativního znaku můžeme za určitých předpokladů otestovat pomocí metody ANOVA (část 3.2.1), při porušení těchto předpokladů využíváme zpravidla Kruskalův–Wallisův test (část 3.2.4). U obou metod jsou doplněny i metody mnohonásobného porovnávání (používané při zamítnutí nulové hypotézy), u ANOVA je uvedena i možnost testování jednoho z předpokladů. Hlavním zdrojem pro tyto části mi byla opět skripta [7].

Pokud bychom chtěli vztah kvalitativní proměnné s jinou proměnnou či proměnnými – ať už kvalitativními či kvantitativními – nejen posoudit z hlediska závislosti, ale také modelovat, můžeme v případě uspořádatelných kategorií kvalitativní proměnné využít model popsaný v části 3.3. U této části jsem čerpal především

z monografie [1], ale také z textů [7], [9] a [10] či z internetového zdroje [15].

3.1 Vztah dvou kvalitativních znaků

3.1.1 χ^2 test nezávislosti

Tento test používáme pro ověření hypotézy nezávislosti dvou kvalitativních znaků X a Y. Budeme předpokládat, že náhodná veličina X může nabývat hodnot $a_1,\ldots,a_r$ a náhodná veličina Y hodnot $b_1,\ldots,b_s$. Tyto hodnoty jsou zástupnými hodnotami pro jednotlivé kategorie těchto znaků.

Dvourozměrný náhodný vektor $(X,\, Y)$ tak může nabývat $r \cdot s$ hodnot, které tvoří uspořádané dvojice $(a_i,\ b_j)$. To, kolikrát se v náhodném výběru o rozsahu n příslušnému náhodnému vektoru $(X,\, Y)$ realizovala uspořádaná dvojice $(a_i,\, b_j)$, můžeme přehledně zobrazit do kontingenční tabulky (někdy jako *tabulka četností*). V Tabulce 3.1 jsou tyto počty značeny jako n_{ij} (prvky kontingenční tabulky na pozici $[i,j]$). Dále $n_{i\cdot}$, resp. $n_{\cdot j}$ jsou marginální četnosti v i-tém řádku ($i \in \{1,\ldots,r\}$), resp. j-tém sloupci ($j \in \{1,\ldots,s\}$), řídící se tímto předpisem:

Tabulka 3.1: Kontingenční tabulka znaků X a Y

$X\backslash Y$	b_1	$\ldots$	b_j	$\ldots$	b_s	$\sum$
a_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_r	n_{r1}	$\ldots$	n_{rj}	$\ldots$	n_{rs}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot s}$	n

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij} \tag{3.1}$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \tag{3.2}$$

Zmíněný rozsah náhodného výběru n pak pro nás v praxi představuje celkový počet pozorování v datech. Spočítá se jednoduše jako

$$n = \sum_{j=1}^{s} n_{\cdot j} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} \tag{3.3}$$

Test

$$H_0 : \text{Náhodné veličiny } X \text{ a } Y \text{ jsou nezávislé}$$

proti alternativě

$$H_A : \text{Náhodné veličiny } X \text{ a } Y \text{ nejsou nezávislé}$$

provedeme pomocí testové statistiky

$$Z = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}, \tag{3.4}$$

která se za platnosti H_0 asymptoticky řídí χ^2 rozdělením s $(r-1)(s-1)$ stupni volnosti. To, že se tímto rozdělením řídí pouze asymptoticky, znamená, že budeme potřebovat dostatečný počet pozorování.

Obecně se tento požadavek klade na tzv. očekávané četnosti ve tvaru $\frac{n_{i\cdot}n_{\cdot j}}{n}$. Za platnosti hypotézy nezávislosti bychom v kontingenční tabulce na pozici $[i, j]$ očekávali právě tuto hodnotu. Za dostatečně velké pro provedení χ^2 testu se považují očekávané četnosti splňující $\frac{n_{i\cdot}n_{\cdot j}}{n} \geq 5 \; \forall i, j$.

Pokud testujeme na hladině významnosti α, nulovou hypotézu zamítneme, pokud $z > \chi^2_{(r-1)(s-1)}(1-\alpha)$, kde z je realizace testové statistiky (3.4) a $\chi^2_{(r-1)(s-1)}(1-\alpha)$ je příslušný kvantil.

Kategorie náhodných veličin X a Y však mohou být dvojího typu, a to

a) nominální – bez přirozeného uspořádání

   – např. barva očí (zelená, modrá, hnědá, ...) či značka používané hokejky (Bauer, CCM, Fischer, ...)

b) ordinální – mají přirozené uspořádání

   – např. velikost trička (XS, S, M, L, ...) či již zmiňovaný výsledek v play-off (tým vypadnul v 1. kole, ve 2. kole, ..., vyhrál Stanley Cup)

Popsaný χ^2 test nezávislosti však při posuzování vztahu dvou kvalitativních znaků neumí využít informaci o případné existenci uspořádání jejich kategorií. Proto bude vhodné si popsat i metody, které toto dovedou.

3.1.2 Metody pro ordinální data

Tyto metody slouží právě pro případ, kdy zkoumáme vztah dvou kvalitativních proměnných s kategoriemi, které mají přirozené uspořádání.

Můžeme chtít např. shrnout vztah mezi řádkovou a sloupcovou proměnnou (označme je znovu X a Y) pomocí jednoho čísla. Několik charakteristik, které se o to pokouší, je založeno na počtu konkordantních a diskordantních dvojic pozorování v kontingenční tabulce. Vysvětleme nyní tyto pojmy:

Mějme ordinální proměnné X a Y. Nechť může proměnná X nabývat hodnot (resp. variant) $i \in \{1, \ldots, r\}$, proměnná Y hodnot (resp. variant) $j \in \{1, \ldots, s\}$ a hodnoty v tabulce n_{ij} označují počet objektů, u kterých se proměnná X realizovala hodnotou i a proměnná Y hodnotou j. Nechť $(x_k,\ y_k)$ jsou realizace náhodného výběru příslušného náhodnému vektoru (X,Y).

Dvojice pozorování (x_1,y_1), (x_2,y_2) je

a) **konkordantní**, pokud platí: $(x_1 > x_2 \wedge y_1 > y_2) \vee (x_1 < x_2 \wedge y_1 < y_2)$,

b) **diskordantní**, pokud platí: $(x_1 > x_2 \wedge y_1 < y_2) \vee (x_1 < x_2 \wedge y_1 > y_2)$.

Př.: V datech představených v kapitole 2 by při zkoumání vztahu proměnných *Umístění v konferenci* a *Vyhrané série* konkordantní dvojicí pozorování byly takovéto výsledky týmů:

$$(\mathrm{UvK}_{team1},\mathrm{VS}_{team1})=(2,3)$$
$$(\mathrm{UvK}_{team2},\mathrm{VS}_{team2})=(4,4)$$

Naopak diskordantní dvojicí pozorování by byly např. tyto výsledky dvou týmů:

$$(\text{UvK}_{team1},\text{VS}_{team1})=(2,3)$$
$$(\text{UvK}_{team2},\text{VS}_{team2})=(8,0)$$

Poznámka: Zde je dobré upozornit na to, že v kapitole 4 budeme chtít pro tyto dvě proměnné ukázat převahu diskordantních párů, což by znamenalo, že tým s lepším umístěním (nižší hodnota první proměnné) vyhraje více sérií play-off (vyšší hodnota druhé proměnné). Dvojici týmů z výše uvedeného příkladu diskordantního páru bychom mohli popsat tak, že Tým 1 skončil 2. v konferenci a v play-off postoupil až do finále, které prohrál, zatímco Tým 2 skončil po základní části osmý v konferenci a vypadnul hned v 1. kole vyřazovací fáze.

Nechť je C počet konkordantních a D počet diskordantních párů pozorování v kontingenční tabulce. Výše uvedená definice se dá přepsat takto:

$$C = \sum_{i<k}\sum_{j<l}\sum\sum n_{ij} \cdot n_{kl}$$

$$D = \sum_{i<k}\sum_{j>l}\sum\sum n_{ij} \cdot n_{kl}$$

kde v obou případech sčítáme v první dvojité sumě přes všechny dvojice řádků $i < k$ a v druhé dvojité sumě přes všechny dvojice sloupců (pro jejichž indexy platí příslušná nerovnost dle toho, zda počítáme C, nebo D).

Charakteristiky asociace mezi proměnnými popsané níže jsou založeny právě na rozdílu $C - D$ a zobrazení tohoto rozdílu na interval $< -1, 1 >$. Říkáme, že vztah mezi proměnnými je pozitivní, pokud $C - D > 0$, a naopak negativní, pokud $C - D < 0$.

Zdůrazněme ještě, že následující charakteristiky jsou **výběrové**, určené přímo pro práci s daty.

3.1.2.1 Gamma koeficient

Tato charakteristika je dána jako rozdíl v proporci konkordantních a diskordantních dvojic:

$$\hat{\gamma} = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D} \tag{3.5}$$

Je zřejmé, že hodnota $\hat{\gamma}$ se pohybuje opravdu pouze v intervalu $< -1, 1 >$, přičemž:

- $\hat{\gamma} = 1 \Leftrightarrow D = 0$
- $\hat{\gamma} = -1 \Leftrightarrow C = 0$

Platí, že čím je větší $|\hat{\gamma}|$, tím je silnější vztah mezi proměnnými X a Y. U nezávislých náhodných veličin bychom očekávali hodnotu $\hat{\gamma}$ pohybující se kolem 0.

Mohlo by nám vadit, že pro výpočet tohoto koeficientu nevyužíváme všechna pozorování, nýbrž pouze takové dvojice, u nichž lze rozhodnout, zda jsou konkordantní či diskordantní. Když se podíváme výše na zavedení konkordance a diskordance páru, můžeme si všimnout, že nerovnosti uvedené v jejich definicích jsou pouze ostré. Některé dvojice tak budou vzhledem k těmto pojmům nezařaditelné. To platí pro ty páry, které se shodují v řádkové či sloupcové proměnné, nebo dokonce v obou zároveň. V kontingenční tabulce jsou to ty, které patří do stejného řádku či sloupce či přímo stejné buňky.

Abychom tedy využili informaci i z takových pozorování, můžeme použít následující míru asociace.

3.1.2.2 Kendallovo Tau-b

Nechť T_X je počet dvojic shodujících se v hodnotách X a T_Y počet dvojic shodujících se v hodnotách Y:

$$T_X = \sum_{i=1}^{r} \frac{n_{i\cdot}(n_{i\cdot} - 1)}{2}$$

$$T_Y = \sum_{j=1}^{s} \frac{n_{\cdot j}(n_{\cdot j} - 1)}{2}$$

Dále nechť T_{XY} je počet dvojic shodujících se v obou proměnných, tj. počet párů v rámci jedné buňky kontingenční tabulky sečtený přes všechny buňky:

$$T_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}(n_{ij} - 1)}{2}$$

Máme-li pozorování (x_k,y_k), $k = 1, \ldots, n$, pak celkový počet možných dvojic můžeme rozložit takto:

$$\binom{n}{2} = \frac{n(n-1)}{2} = C + D + T_X + T_Y - T_{XY}$$

Tuto dekompozici využijeme i při konstrukci výběrového Kendallova Tau-b, které je definováno takto:

$$\hat{\tau}_b = \frac{C - D}{\sqrt{\left(\frac{n(n-1)}{2} - T_X\right)\left(\frac{n(n-1)}{2} - T_Y\right)}} \tag{3.6}$$

Ve jmenovateli vidíme geometrický průměr dvou hodnot:

$$\frac{n(n-1)}{2} - T_X = C + D + T_Y - T_{XY},$$

tedy počet konkordantních párů, diskordantních párů a párů shodujících se pouze v proměnné Y, nikoliv v X.

Druhý výraz

$$\frac{n(n-1)}{2} - T_Y = C + D + T_X - T_{XY}$$

nám kromě počtu konkordantních a diskordantních párů přidává do součtu ty dvojice, které se shodují v proměnné X, nikoliv však v Y.

Lze vidět, že $C + D$ nemůže být větší než kterýkoliv z těchto dvou výrazů, z čehož plyne, že nepřesáhne ani jejich geometrický průměr. Kendallovo Tau-b

nám tedy v případě výskytu dvojic pozorování shodujících se v některé z proměnných vrátí menší hodnotu oproti předchozímu Gamma koeficientu, neboť z právě uvedeného plyne

$$|\hat{\tau}_b| \leq |\hat{\gamma}|.$$

Ohledně výsledné hodnoty znovu platí, že čím větší $|\hat{\tau}_b|$, tím silnější vztah mezi proměnnými X a Y (ať už v kladném či záporném smyslu, to rozhoduje znaménko $\hat{\tau}_b$). Hodnoty kolem 0 značí, že na sobě proměnné X a Y příliš nezávisí.

Přesný interval hodnot, kterých Kendallovo Tau-b může nabývat, se nedá obecně (bez známých hodnot v datech) přesně určit, neboť v případě přítomnosti dvojic shodujících se v některé z proměnných nám už C či D (při nulové hodnotě jednoho z nich) nemůže tvořit 100% proporci dvojic zahrnutých ve jmenovateli, nýbrž pouze jejich část. Obecně můžeme pouze psát, že $|\hat{\tau}_b| \leq 1$.

3.1.2.3 Somersovo d

Somersovo d je definováno jako rozdíl konkordantních a diskordantních párů ku počtu dvojic, které se neshodují v hodnotě proměnné X:

$$d = \frac{C - D}{\frac{n(n-1)}{2} - T_X} \tag{3.7}$$

Konstrukce koeficientu je podobná jako u předešlých dvou, avšak Somersovo d pohlíží na Y jako na závisle proměnnou a na X jako na vysvětlující proměnnou.

Znovu máme výraz ve jmenovateli větší nebo roven $C + D$, opět tak platí:

$$|d| \leq |\hat{\gamma}|$$

O velikosti koeficientu můžeme říct obecně jen to, že $|d| \leq 1$. Konkrétní dolní a horní hranice intervalu možných hodnot závisí na datech.

3.1.2.4 Intervalové odhady

Konstrukce (1-α)100% intervalového odhadu pro libovolnou charakteristiku z částí 3.1.2.1–3.1.2.3 je duální úlohou k testu (s pravděpodobností chyby I. druhu

α) hypotézy, že je daná charakteristika nulová, z čehož ještě nutně nevyplývá nezávislost náhodných veličin X a Y. Pro vztah nezávislosti a nulové hodnoty koeficientů platí implikace

$$X \text{ a } Y \text{ jsou nezávislé} \Rightarrow \text{koeficient asociace} = 0. \tag{3.8}$$

Opačná implikace sice neplatí, obměnou výrazu (3.8) však získáme další platný výraz

$$\text{koeficient asociace} \neq 0 \Rightarrow X \text{ a } Y \text{ nejsou nezávislé.} \tag{3.9}$$

Můžeme tedy říct, že pokud příslušný (1-α)100% interval spolehlivosti nebude obsahovat nulu, náhodné veličiny X a Y s pravděpodobností 1-α nejsou nezávislé (resp. existuje mezi nimi nějaký vztah, určitá závislost). Zda je tento vztah pozitivní (převaha konkordantních párů) či negativní (převaha diskordantních párů) závisí na tom, zda se interval spolehlivosti bude nacházet nalevo či napravo od nuly, resp. zda bude obsahovat záporná či kladná čísla.

Intervalové odhady však vždy sestavujeme pro teoretické hodnoty příslušných charakteristik. My jsme si před zavedením výše popsaných koeficientů jasně zdůraznili, že se jedná o výběrové charakteristiky počítané na základě dat, v podstatě jde tedy o bodové odhady oněch teoretických protějšků, pro které bychom chtěli získat i odhady intervalové.

Agresti však ve své publikaci [1] v rámci kapitoly 7.1 uvádí ke každé z těchto charakteristik také její teoretický protějšek, pro nějž jsme schopni intervalový odhad vytvořit. Tyto intervalové odhady jsou implementovány v softwaru R, konkrétně v balíčku `DescTools` [30] v rámci jednotlivých funkcí počítajících tyto koeficienty asociace, kde se v problematice intervalových odhadů autoři balíčku odkazují na texty [6] a [8].

3.1.2.5 Linear-by-linear association test

Alternativou k představeným mírám asociace by mohl být tzv. Linear-by-linear association test, který je založen na loglineárním modelování očekávaných četností uvnitř buněk a používá se právě pro nalezení asociace (přímo pozitivního či negativního trendu) mezi faktory s uspořádanými kategoriemi.

Nulovou hypotézou je neexistující asociace mezi proměnnými. Podrobněji je model linear-by-linear asociace a test z něho vycházející rozpracován v monografii [1] v kapitole 6.2, v softwaru R jej můžeme použít pomocí funkce `lbl_test()`[1] z balíčku `coin` [32].

3.2 Vztah kvalitativního a kvantitativního znaku

3.2.1 ANOVA

Název ANOVA (zkrácení Analysis Of Variance) sice napovídá, že se tento test zabývá rozptyly, jedná se však o velmi používaný test shody středních hodnot u k skupin, kde $k \geq 2$ (v případě $k = 2$ se používají jiné, jednodušší metody, např. dvouvýběrový t-test, jehož zobecněním je právě ANOVA). Svůj název dostala tato metoda kvůli tomu, že ve skutečnosti testování středních hodnot provádí pomocí porovnání reziduálního součtu čtverců (zbytkové variability) ve 2 různých modelech (viz dále). Někdy se místo názvu ANOVA můžeme setkat s „jednofaktorovou[2] analýzou rozptylu“, jedná se o totožný test.

Uvažujme tedy kvalitativní znak X s $k \geq 2$ kategoriemi a kvantitativní veličinu Y. Naším cílem bude zjistit, zda má náhodná veličina Y stejné rozdělení v jednotlivých skupinách daných veličinou X.

Zde nás může zarazit, že v předchozím odstavci se píše o testu středních hodnot, nyní chceme rovnou ověřovat shodu rozdělení. Je to proto, že k použití této metody se vážou poměrně přísné předpoklady, při jejichž naplnění jsou *shoda rozdělení* a *shoda středních hodnot* ekvivalentní. Předpoklady modelu jsou následující:

Mějme $k \geq 2$ nezávislých náhodných výběrů s rozsahem $n_1, \ldots, n_k$. Pro každý z těchto výběrů platí:

$$Y_{i1}, \ldots, Y_{in_i} \text{ je náhodný výběr z rozdělení } N(\mu_i, \sigma^2),\ \forall i \in \{1, \ldots, k\} \tag{3.10}$$

[1] Více je jeho použití v R zpracováno na stránce [19].

[2] Jednofaktorová kvůli tomu, že skupiny jsou dány kategoriemi jediného znaku (faktoru) X.

Všech k náhodných výběrů tak pochází z **normálního rozdělení** a mají **shodný rozptyl**. Metodám ověřování předpokladu normality se v práci nevěnuji, v kapitole 4 používám Shapirův–Wilkův test v rámci softwaru R. Ověření předpokladu shody rozptylů se věnuji v části 3.2.3.

Nulovou hypotézou, kterou budeme pomocí ANOVA testovat, je shoda všech k středních hodnot $\mu_1, \ldots, \mu_k$ (3.10), alternativou pak bude tvrzení, že alespoň jedna dvojice středních hodnot se liší, tedy

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_k, \\ H_A &: \exists i \in \{1, \ldots, k\}\, \exists j \in \{1, \ldots, k\} \backslash \{i\} : \mu_i \neq \mu_j. \end{aligned} \tag{3.11}$$

Z předpokladu (3.10) a testované hypotézy je zřejmé, že v případě platnosti H_0 bude skutečně všech k náhodných výběrů pocházet ze stejného rozdělení, ta se od sebe totiž mohou lišit pouze střední hodnotou.

Samotné testování je založeno, jak už bylo zmíněno, na porovnání variability ve dvou modelech.

a) Model 1

$$\forall i \in \{1, \ldots, k\} : Y_{ij} = \mu_i + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2),\ j = 1, \ldots, n_i,\ \epsilon_{ij} \text{ nezávislé} \tag{3.12}$$

Zde uvažujeme rozdílné střední hodnoty pro každou z k skupin. Pokud bychom metodou nejmenších čtverců (MNČ) odhadovali parametry μ_i z modelu 3.12, vyšly by nám odhady

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

což jsou prosté skupinové průměry. Variabilitu uvnitř skupin můžeme vyjádřit pomocí reziduálního součtu čtverců

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

Výraz $\frac{S_e}{\sigma^2}$ má za platnosti H_0 χ^2 rozdělení s $n-k$ stupni volnosti, kde $n = \sum_{i=1}^{k} n_i$.

b) Model 2

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

V tomto modelu uvažujeme společnou střední hodnotu μ, je to tedy model za platnosti nulové hypotézy (3.11).

Odhadem společné střední hodnoty získaným metodou nejmenších čtverců je znovu obyčejný průměr hodnot, tentokrát přes všechny hodnoty Y_{ij}:

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

Zbytkovou variabilitu v tomto modelu za platnosti H_0 o shodě středních hodnot spočítáme jako

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

přičemž S_T se někdy nazývá *celkový součet čtverců*. Výraz $\frac{S_T}{\sigma^2}$ má za platnosti H_0 χ^2 rozdělení s $n - 1$ stupni volnosti.

S_T tak vlastně představuje celkovou variabilitu veličiny Y, která se dá rozložit následujícím způsobem:

$$S_T = S_A + S_e, \tag{3.13}$$

kde $S_A = \sum_{i=1}^{k} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$ je tzv. *skupinový součet čtverců*, vyjadřuje tedy variabilitu mezi jednotlivými skupinami. Výraz $\frac{S_A}{\sigma^2}$ se za platnosti nulové hypotézy řídí opět χ^2 rozdělením, tentokrát s $k - 1$ stupni volnosti (což je rozdíl stupňů volnosti rozdělení výrazů $\frac{S_T}{\sigma^2}$ a $\frac{S_e}{\sigma^2}$).

Testová statistika pro testování H_0 vznikne podílem dvou odhadů variability řídícími se χ^2 rozdělením, které ještě podělíme jejich stupni volnosti, abychom získali statistiku řídící se Fisherovým rozdělením (které vyplývá z nezávislosti veličin S_A a S_e), tedy

$$F_A = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} = \frac{S_A}{S_e} \frac{n-k}{k-1} \overset{H_0}{\sim} F_{k-1,\, n-k}. \tag{3.14}$$

Kritický obor představují příliš velké hodnoty F_A, konkrétně tvoří množinu

$$W = < F_{k-1,n-k}(1-\alpha), \infty).$$

Pokud se F_A realizuje ve W, zamítáme H_0 na hladině významnosti α. Podoba kritického oboru (a důvod, proč nám vadí zrovna velké hodnoty testové statistiky) vychází z následující myšlenky.

Při shodě středních hodnot mezi skupinami by S_e a S_T byly velmi podobné, naopak variabilita mezi jednotlivými skupinami by byla blízká 0 (což při rovnosti S_e a S_T vyplývá i z rovnice (3.13)). V případě, že se střední hodnoty napříč skupinami liší, roste i variabilita mezi skupinami (S_A) a tím i hodnota testové statistiky (3.14) – proto ve prospěch H_A mluví velké hodnoty testové statistiky a při překročení zmíněného kvantilu F-rozdělení nulovou hypotézu zamítáme.

3.2.2 Scheffého metoda mnohonásobného porovnávání

V případě zamítnutí nulové hypotézy uvedené u metody ANOVA nás může zajímat, které dvojice skupin se významně liší. Chceme tak při stejných předpokladech, jako tomu bylo u ANOVA, testovat hypotézy

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

proti alternativám

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

a to pro všechna $i, j = 1, \ldots, k$, která jsou od sebe různá.

Jedním ze způsobů, jak tyto hypotézy testovat, je použít *Scheffého metodu*, podle které zamítáme H_0^* na hladině α, pokud se od sebe příslušné skupinové průměry, velmi jednoduše řečeno, liší už příliš, přesněji pokud

$$|\overline{y}_{i.} - \overline{y}_{j.}| > \sqrt{(k-1)\frac{S_e}{n-k}(n_i^{-1} + n_j^{-1})F_{k-1,n-k}(1-\alpha)}. \tag{3.15}$$

Může se stát, že zamítneme shodu středních hodnot ve skupinách pomocí ANOVA, ale žádný z rozdílů $|\overline{y}_{i.} - \overline{y}_{j.}|$ nepřekročí hranici stanovenou v (3.15).

To by znamenalo, že důvodem zamítnutí H_0 pomocí F-statistiky (3.14) byla rozdílnost nějaké lineární kombinace středních hodnot (např. porušení toho, že $2\mu_1 - \mu_2 - \mu_3 = 0$, což by za platnosti H_0 mělo platit).

3.2.3 Test rovnosti rozptylů u $k \geq 2$ skupin

Abychom u ANOVA ověřili předpoklad shody rozptylů v k skupinách, můžeme použít některý z testů hypotézy

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2 \ \ (= \sigma^2)$$

proti alternativě, že alespoň u jedné z dvojic kategorií je tato rovnost porušena. V balíčku `stats` [34] softwaru R je implementován *Bartlettův test* (příkaz `bartlett.test`), který vychází z testu popsaného M. S. Bartlettem v textu [2].

Ve skriptech [7], přesněji v části 7.3.10, je popsána konstrukce dnes používanějšího *Leveneova testu*. V této práci však pro ověření předpokladu shody rozptylů používám výše zmíněný Bartlettův test.

3.2.4 Kruskalův–Wallisův test

Představená metoda testu shody středních hodnot proměnné Y v k skupinách daných hodnotami proměnné X a metoda následného mnohonásobného porovnávání jsou sice užitečné, dají se však použít jen v případě splnění uvedených předpokladů (normalita a shodný rozptyl ve všech skupinách).

V praxi se však často stává, že tyto předpoklady splněny nejsou, přesto potřebujeme posoudit shodu či rozdíl v hodnotách kvantitativního znaku dle jednotlivých skupin. Pro tento případ nám slouží Kruskalův–Wallisův test, který je neparametrickou alternativou k jednofaktorové analýze rozptylu.

Zcela bez předpokladů se sice také neobejdeme, jsou však snadněji splnitelné, než tomu bylo u ANOVA.

Nyní nám stačí, aby $\forall i,\ i = 1, \ldots, k \geq 2$, platilo:

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

dále předpokládáme nezávislost těchto výběrů.

Budeme testovat nulovou hypotézu[3]

$$H_0 : F_1 \equiv \cdots \equiv F_k$$

proti alternativě, že aspoň jedna dvojice náhodných výběrů pochází z různých rozdělení.

Postup testu je následující:

1. Seřadíme všechna pozorování (sdružená do jednoho společného výběru) do neklesající posloupnosti.

2. Každému číslu přiřadíme pořadí.[4]

3. Vypočítáme hodnoty T_i, $i = 1, \ldots, k$, jako součet pořadí těch pozorování, která pocházejí z i-té skupiny.

   - Pokud jsou pořadí jednoznačně určena, což teoreticky nastane s pravděpodobností jedna (pravděpodobnost shod je nulová), tvoří pořadí aritmetickou posloupnost, proto platí: $\sum_{i=1}^{k} T_i = \frac{n(n+1)}{2}$. Stejný součet však získáme i v případě shod a průměrných pořadí.

4. Vypočítáme hodnotu testové statistiky

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1). \tag{3.16}$$

   - Pokud platí H_0, má Q asymptoticky χ^2 rozdělení s $k-1$ stupni volnosti.

5. Zamítáme H_0, pokud se Q realizuje hodnotou větší než $\chi^2_{k-1}(1-\alpha)$.

[3] Symbol $\equiv$ značí, že $F_1(x) = \cdots = F_k(x)\ \forall x \in \mathbb{R}$.

[4] V případě shod to bude průměrné pořadí shodných hodnot.

3.2.5 Porovnávání dvojic bez předpokladu normality

Podobně jako v části 3.2.2 budeme chtít po zamítnutí H_0 Kruskalova–Wallisova testu zjistit, které kategorie k tomuto zamítnutí výrazně přispěly, resp. které se od sebe statisticky významně liší. Místo skupinových průměrů zde sledujeme *průměrná pořadí* v jednotlivých skupinách. Rozdělení F_i a F_j jsou od sebe významně odlišná, pokud platí:

$$\left|\frac{T_i}{n_i} - \frac{T_j}{n_j}\right| > \sqrt{\frac{1}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) n(n+1)\chi^2_{k-1}(1-\alpha)}. \tag{3.17}$$

Další možností je např. Dunnův test [4], který je implementován v softwaru R, konkrétně v balíčku `PMCMR` [33].

3.3 Model ordinální logistické regrese

ANOVA i Kruskalův–Wallisův test jsou užitečné metody pro zjištění toho, zda se rozdělení hodnot kvantitativní veličiny statisticky významně liší v různých skupinách (resp. kategoriích) kvalitativního znaku.

Může nás však zajímat také to, s jakou pravděpodobností se kvalitativní náhodná veličina realizuje v některé ze svých kategorií, pokud známe hodnotu znaku kvantitativního (případně i pokud známe hodnoty více takových znaků, ať už kvantitativních či kvalitativních) – veličinu, která odpovídá kategoriálnímu znaku, budeme nyní považovat za závisle proměnnou a budeme odhadovat její rozdělení, resp. pravděpodobnosti jednotlivých hodnot (kategorií) při daných hodnotách jiných veličin.

Upozorním, že v řešení tohoto problému, kdy se kategoriální proměnná stává závisle proměnnou, se používá jiné značení, než tomu bylo od části 3.2.1 doposud – zpravidla označujeme závisle proměnnou jako Y (znovu nabývající hodnot $1, \ldots, k$ značících jednotlivé kategorie) a vysvětlující proměnnou jako X, resp. více vysvětlujících proměnných jako $X_1, \ldots, X_p$.

Pokud se navíc budeme bavit o vysvětlujících proměnných jako i již známých realizacích těchto náhodných veličin, budeme je značit malými písmeny, tedy x,

resp. $x_1, \ldots, x_p$.

Situaci podobnou té právě popsané dokážeme modelovat pomocí logistické regrese, u které se však závisle proměnná řídí alternativním rozdělením (viz např. kapitola 7.3.12 ve skriptech [7]), což znamená pouhé dvě kategorie, v kterých se může Y realizovat. Jsou to výstupy jako např. nemocný/zdravý, prošel/neprošel (u zkoušky), zvítězil/prohrál apod.

Pokud však Y nabývá obecně k hodnot, tento model nám nestačí. Zobecněním logistické regrese je tzv. *multinomická logistická regrese* uvažující závisle proměnnou Y s k kategoriemi, které ovšem nejsou nijak přirozeně uspořádatelné (tedy Y je nominální kategoriální proměnná).

Z cíle práce je zřejmé, že nás bude zajímat možnost vytvoření modelu pro takovou proměnnou, která má kategorie s přirozeným uspořádáním (tedy je ordinální). Právě k tomu nám poslouží tzv. *ordinální logistická regrese*. Někdy je možné se setkat s více modely s tímto označením, popisován ovšem bude ten, kterému se také říká *model proporcionálních šancí* (z anglického „proportional odds model"), který je speciálním případem *modelu pro kumulativní logit* (z anglického „cumulative logit model"). Právě kumulativní logit bude vhodné si před zavedením samotného modelu vysvětlit.

3.3.1 Kumulativní logit

Mějme diskrétní náhodnou veličinu Y nabývající hodnot z množiny $\{1, \ldots, k\}$. Označme pravděpodobnosti

$$P(Y = j) = \pi_j, \tag{3.18}$$

kde hodnoty π_j určují rozdělení pravděpodobností náhodné veličiny Y a platí

$$\sum_{j=1}^{k} \pi_j = 1.$$

V celém textu budeme používat přirozený logaritmus a budeme jej značit zkráceným zápisem log.

Kumulativní logit je definován jako

$$\begin{aligned}\text{logit}[P(Y \leq j)] &= \log \frac{P(Y \leq j)}{P(Y > j)} = \\ &= \log \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \\ &= \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_k}, \quad j = 1, \ldots, k-1. \end{aligned} \tag{3.19}$$

Z této definice vidíme, že j-tý kumulativní logit je logaritmus šance[5] jevu „Náhodná veličina Y nabude hodnoty nejvýše j“, tedy že se Y realizuje nejvýše v kategorii j. Také je zřejmé, že pro veličinu s k možnými výstupy lze uvažovat pouze $k-1$ logitů, neboť $P(Y > k)$, která by se v případě logitu pro $j = k$ vyskytovala ve jmenovateli výrazu, je nulová.

3.3.2 Model proporcionálních šancí

3.3.2.1 Zavedení modelu a základních pojmů

Nechť diskrétní náhodná veličina Y nabývající hodnot z množiny $\{1, \ldots, k\}$ je závisle proměnná. Tuto veličinu budeme pozorovat u n subjektů. Hodnoty znaku Y u těchto subjektů můžeme označit $Y_1, \ldots, Y_n$, tyto náhodné veličiny nechť jsou vzájemně nezávislé. Dále nechť náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)'$ je vektor vysvětlujících proměnných, jeho realizaci u i-tého subjektu budeme značit $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$.

Na sloupce matice, jejíž řádky tvoří vektory $\mathbf{x}_i$, klademe předpoklad lineární nezávislosti. Tyto sloupce by neměly být ani téměř lineárně závislé (tzv. multikolinearita). Další důležitý předpoklad je samostatně rozebrán v části 3.3.2.3.

Vysvětlující proměnné mohou být jak kvantitativní, tak kvalitativní, ve druhém případě používáme metodu umělých proměnných (jedna kategorie referenční, zbytek je zastoupen indikátorovými proměnnými nabývajícími hodnot 1 nebo 0).

[5] Šance jevu A se vypočítá jako $\frac{P(A)}{1-P(A)}$.

Při pevné realizaci náhodného vektoru $\mathbf{X}$ budeme u každého subjektu uvažovat pravděpodobnosti uvedené u definice kumulativního logitu jako podmíněné, tedy např.

$\pi_j(\mathbf{x}_i)$ ... Pravděpodobnost, že se Y realizuje hodnotou j, za podmínky, že se náhodný vektor $\mathbf{X}$ realizoval hodnotou $\mathbf{x}_i$,

podobně

$P(Y \leq j|\mathbf{x}_i)$ ... Pravděpodobnost, že se Y realizuje hodnotou nejvýše j, za podmínky, že se náhodný vektor $\mathbf{X}$ realizoval hodnotou $\mathbf{x}_i$.

Při známém (resp. pevně daném) $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ budeme chtít modelovat kumulativní logit.

Motivace: Předpokládejme nyní, že existuje latentní[6] spojitá proměnná (označme ji např. Y^*) taková, že náhodná veličina Y vzniká její diskretizací do k uspořádaných kategorií. Hodnoty latentní proměnné tvořící hranice těchto kategorií označme $\alpha_1, \ldots, \alpha_{k-1}$. Hodnoty těchto mezí (či cutpointů) vůbec nezávisí na realizaci $\mathbf{x}_i$.

Můžeme tedy zjednodušeně psát:

$$Y = 1 \Leftrightarrow Y^* \leq \alpha_1, \tag{3.20}$$

$$Y = j \Leftrightarrow \alpha_{j-1} < Y^* \leq \alpha_j,\ j = 2, \ldots, k-1, \tag{3.21}$$

$$Y = k \Leftrightarrow Y^* > \alpha_{k-1}. \tag{3.22}$$

Tato motivace slouží k lepšímu pochopení a vhodnější interpretaci modelu, který si nyní představíme. V praxi se obvykle žádnou latentní proměnnou nezabýváme, stačí nám splnění předpokladů modelu popsaných výše a v části 3.3.2.3.

[6] skrytá, nepozorovaná

Model proporcionálních šancí, obsahující celkem $k-1$ rovnic, je ve tvaru

$$\begin{aligned}\text{logit}[P(Y \leq j|\mathbf{x}_i)] &= \log \frac{P(Y \leq j|\mathbf{x}_i)}{1 - P(Y \leq j|\mathbf{x}_i)} = \\ &= \log \frac{\pi_1(\mathbf{x}_i) + \cdots + \pi_j(\mathbf{x}_i)}{\pi_{j+1}(\mathbf{x}_i) + \cdots + \pi_k(\mathbf{x}_i)} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i = \\ &= \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad j = 1, \ldots, k-1, \quad (3.23)\end{aligned}$$

kde vektor parametrů $\boldsymbol{\beta}'$ popisuje efekt jednotlivých regresorů na závisle proměnnou Y. Všimněme si, že parametry β_i nezávisí na tom, jaký logit chceme odhadnout – předpis pro jednotlivé kategorie se tak liší pouze konstantou α_j.

Poznámka: Dále v textu se pro zjednodušení budu držet Agrestiho značení [1], kdy za $P(Y \leq j)$ budeme považovat podmíněnou pravděpodobnost, že se proměnná Y u i-tého subjektu realizuje kategorií j či nižší při daných hodnotách $\mathbf{x}_i$. Podobně pod výrazem π_j budeme chápat dříve definovaný výraz $\pi_j(\mathbf{x}_i)$.

Z toho, jak jsme si zavedli a interpretovali (v rámci motivace) absolutní členy α_j (výrazy (3.20) – (3.22)), je zřejmé, že platí nerovnosti

$$\alpha_1 < \alpha_2 < \cdots < \alpha_{k-1}. \quad (3.24)$$

Při fixní hodnotě $\mathbf{x}_i$ totiž s rostoucím j roste i $P(Y \leq j)$, tedy se zvyšuje i logit této pravděpodobnosti v modelu (3.23), $\{\alpha_j\}$ tak tvoří rostoucí posloupnost.

Parametry bychom však mohli interpretovat také v rámci modelu, a to jak absolutní členy α_j, tak parametry u příslušných regresorů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.

Parametr β_m, popisující efekt m-tého regresoru, udává to, o kolik se změní každý logit v modelu (3.23), pokud se x_{im} zvýší o jednotku a ostatní vysvětlující proměnné zůstanou stejné. To, že se $\text{logit}[P(Y \leq j)]$ změní **o** β_m, znamená to stejné, jako že se šance $(Y \leq j)$ změní e^{β_m}**krát**.

Absolutní členy jsou pak u regresních modelů obecně obtížněji interpretovatelné. V j-té rovnici modelu (3.23) je α_j hodnota $\text{logit}[P(Y \leq j)]$ v situaci, kdy se všechny vysvětlující proměnné realizují hodnotou 0. Jinými slovy, šance, že se Y realizuje v kategorii nejvýše j při nulové hodnotě všech regresorů, je rovna e^{α_j}.

Tato interpretace však postrádá na významu ve chvíli, kdy vysvětlujeme Y pomocí takových proměnných, které nulové být nemohou. Jako příklad si uveďme výšku člověka, jeho hmotnost či mzdu (za předpokladu, že nepracuje zadarmo). Při takových regresorech nemá příliš smysl absolutní člen interpretovat a slouží zkrátka jen jako určitý posun v kategoriích (viz dále) související s uvedenou motivací.

Pro bližší seznámení s absolutními členy si uveďme příklad modelu, který obsahuje pouze absolutní členy. Chtěli bychom sestavit model proporcionálních šancí pro počet vyhraných sérií v play-off NHL. Závisle proměnná V může nabývat hodnot z množiny $\{0, \ldots, 4\}$ – model tedy bude mít 4 rovnice:

$$\operatorname{logit}[P(V \leq j)] = \log \frac{P(V \leq j)}{1 - P(V \leq j)} = \alpha_j, \;\; j = 0, \ldots, 3,$$

přičemž jedinou informací, kterou můžeme využít, je samotný systém play-off. Ten garantuje to, že v každém závěru ročníku vypadne hned v 1. kole 8 týmů ($V = 0$) z celkových 16, které se do play-off kvalifikují. Ve druhém kole vypadnou další 4 týmy ($V = 1$), 2 týmy skončí po finále konferencí ($V = 2$) a ze dvou finalistů 1 finále prohraje ($V = 3$) a 1 získá Stanley Cup ($V = 4$).

Jednotlivé koeficienty $\alpha_0, \ldots, \alpha_3$ bychom tak mohli vypočítat následovně:

$$\begin{aligned}
\log \frac{P(V \leq 0)}{1 - P(V \leq 0)} &= \log \frac{8/16}{8/16} = \log 1 = 0 = \alpha_0 \\
\log \frac{P(V \leq 1)}{1 - P(V \leq 1)} &= \log \frac{12/16}{4/16} = \log 3 \doteq 1{,}0986 = \alpha_1 \\
\log \frac{P(V \leq 2)}{1 - P(V \leq 2)} &= \log \frac{14/16}{2/16} = \log 7 \doteq 1{,}9459 = \alpha_2 \\
\log \frac{P(V \leq 3)}{1 - P(V \leq 3)} &= \log \frac{15/16}{1/16} = \log 15 \doteq 2{,}7081 = \alpha_3
\end{aligned}$$

V tomto případě je samozřejmě interpretace absolutních členů relevantní, snadno odvoditelná a pochopitelná. Také vidíme, že při řešení příkladu není třeba uvažovat žádnou spojitou latentní proměnnou, která je pouze motivací modelu.

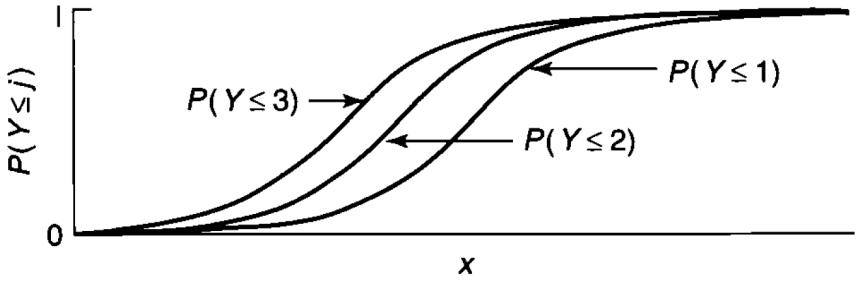
3.3.2.2 Výpočet pravděpodobností a jejich zobrazení

V úvodu této sekce byla zmíněna logistická regrese jakožto způsob modelování náhodné veličiny Y řídící se alternativním rozdělením, tedy s binárním výstupem typu 1/0. Odhad pravděpodobnosti $P(Y = 1)$ při dané hodnotě vysvětlující proměnné má při grafickém zobrazení (hodnoty regresoru na ose x, $P(Y = 1)$ na ose y) tvar tzv. *logistické křivky*.

U vícekategoriální proměnné Y není realizace v jedné z kategorií binárním výstupem. Pokud si však zvolíme kategorii j z množiny možných hodnot v modelu proporcionálních šancí (libovolně, ale pevně), výstupy $Y \leq j$ a $Y > j$ již binární jsou. Pokud bychom použili jedinou vysvětlující proměnnou x, která by byla spojitá, měla by křivka vyjadřující pravděpodobnost $P(Y \leq j)$ také tvar logistické křivky.

Při různém j by měly tyto křivky díky neměnnému parametru $\boldsymbol{\beta}$ (vzhledem ke kategoriím) stejný tvar, zaznamenali bychom pouze jejich posun po ose x kvůli rozdílným „cutpointům“ $\alpha_1, \ldots, \alpha_{k-1}$.

Příklad grafického znázornění popsané situace pro $k = 4$ (tedy pro 3 křivky[7]) je zobrazen na Obrázku 3.1.



Obrázek 3.1: Křivky pro kumulativní pravděpodobnosti při dané hodnotě x získané modelem proporcionálních šancí, zdroj: [1], str. 47

Z Obrázku 3.1 také pěkně vidíme, že při fixní hodnotě x s rostoucím j roste zobrazená pravděpodobnost. Opět to souvisí s vlastností absolutních členů – viz

[7] V modelu je $k-1$ rovnic, získáme $k-1$ křivek. Navíc nemá smysl zobrazovat $P(Y \leq k) = 1$.

nerovnosti (3.24) – a jejím popisem pod těmito nerovnostmi.

Hodnoty pravděpodobností $P(Y \leq j)$ při daném x (či v případě více vysvětlujících proměnných obecně při daném $\mathbf{x}_i$) lze vyjádřit z rovnice (3.23).

Pokud z množiny $\{1, \ldots, k-1\}$ zvolíme j libovolné pevné, můžeme postupně odvodit

$$\log \frac{P(Y \leq j)}{P(Y > j)} = \alpha_j + \boldsymbol{\beta}' \mathbf{x}_i$$

$$\frac{P(Y \leq j)}{P(Y > j)} = \exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)$$

$$P(Y \leq j) = \frac{\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}{\frac{1}{P(Y>j)}} = \frac{\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}{\frac{P(Y>j)+P(Y\leq j)}{P(Y>j)}} =$$

$$= \frac{\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}, \tag{3.25}$$

kde jsme využili toho, že $1 = P(Y > j) + P(Y \leq j) \ \forall j \in \{1, \ldots, k-1\}$.

Díky tomuto odvození můžeme stanovit dílčí podmíněné pravděpodobnosti, že se Y realizuje v dané skupině při známé hodnotě $\mathbf{x}_i$. Začněme u první skupiny.

Při zavedených předpokladech platí, že $P(Y \leq 1) = P(Y = 1) = \pi_1$, pro $j = 1$ tedy můžeme rovnou s využitím (3.25) psát

$$\pi_1 = \frac{\exp(\alpha_1 + \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\alpha_1 + \boldsymbol{\beta}' \mathbf{x}_i)}.$$

Dále v tomto modelu platí, že $P(Y = j) = P(Y \leq j) - P(Y \leq j-1)$. Když navíc uvážíme, že $P(Y \leq j) = \pi_1 + \cdots + \pi_j$, můžeme z rovnice (3.25) vyjádřit také pravděpodobnosti pro $j = 2, \ldots, k-1$, a to dvěma způsoby:

$$\pi_j = \frac{\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}' \mathbf{x}_i)} = \tag{3.26}$$

$$= \frac{\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)} - \sum_{l=1}^{j-1} \pi_l.$$

Pro $j = k$ už se pak příslušná pravděpodobnost spočítá snadno jako

$$\pi_k = 1 - \sum_{l=1}^{k-1} \pi_l.$$

Při dodefinování výrazů $\alpha_k = \infty$ a $\alpha_0 = -\infty$ bychom pomocí výrazu (3.26) spočítali π_j pro libovolné j z množiny $\{1, \ldots, k\}$ – záměrně zde raději používám alternativní způsoby výpočtu, ve kterých se neobjevuje $\pm\infty$.

Toto dodatečné zavedení konstant α_k a α_0 je v souladu s výrazem (3.21), jehož platnost bychom tímto mohli rozšířit pro libovolné $j \in \{1, \ldots, k\}$, přičemž by se neporušila platnost výrazů (3.20) a (3.22), resp. by se tímto zobecněním nezměnilo původní rozhodovací pravidlo pro přiřazení kategorií proměnné Y dle hodnot spojité latentní proměnné Y^*.

3.3.2.3 Předpoklad proporcionality šancí

S pojmem *proporcionalita šancí* jsme se zatím setkali pouze u názvu modelu. Je to však klíčový předpoklad, při jehož porušení nelze model použít, resp. bychom museli pracovat s jeho zobecněnou verzí, kdy nám nepostačí jeden parametr $\boldsymbol{\beta}$, avšak museli bychom odhadovat parametry $\boldsymbol{\beta}_j$ pro každou rovnici modelu (3.23) zvlášť. Křivky pro $P(Y \leq j)$ by se tak nelišily pouze posunem o konstantu, ale také tvarem.

Zavedení tohoto předpokladu až po představení modelu má pouze ten důvod, že vysvětlit proporcionalitu šancí půjde názorněji nyní, kdy už víme, jak model vypadá. Formálně bychom předpoklad zavedli následovně:

Nechť $\mathbf{x}_1 = (x_{11}, \ldots, x_{1p})'$ jsou realizace vysvětlujících proměnných u libovolného subjektu a $\mathbf{x}_2 = (x_{21}, \ldots, x_{2p})'$ realizace vysvětlujících proměnných u libovolného jiného subjektu. Potom řekneme, že model (3.23) splňuje předpoklad proporcionality šancí, jestliže $\forall j \in \{1, \ldots, k-1\}$ platí

$$\begin{aligned} &\text{logit}[P(Y \leq j|\mathbf{x}_1)] - \text{logit}[P(Y \leq j|\mathbf{x}_1)] = \\ &= \log \frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2). \end{aligned} \tag{3.27}$$

Tento předpoklad tedy znamená, že logaritmus *kumulativního poměru šancí* (jak se někdy označuje výraz $\frac{P(Y\leq j|\mathbf{x}_1)/P(Y>j|\mathbf{x}_1)}{P(Y\leq j|\mathbf{x}_2)/P(Y>j|\mathbf{x}_2)}$) je proporcionální vzhledem ke vzdálenosti mezi realizacemi $\mathbf{x}_1$ a $\mathbf{x}_2$.

Pokud rovnici odlogaritmujeme, dostaneme se k tomu, že při hodnotách regresorů $\mathbf{x}_1$ je šance na realizaci Y v kategorii nejvýše j $\exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]$krát větší než při hodnotách vysvětlujících proměnných $\mathbf{x}_2$, a to bez ohledu na to, která kategorie je pro nás v danou chvíli „cutpointem“, tedy s jakým j aktuálně pracujeme.

Uveďme si zjednodušený příklad pro lepší pochopení. Uvažujme skupinu lidí, kteří v rámci jedné firmy vyplňují dotazník spokojenosti se svým zaměstnáním.

V odpovědi na otázku „Jak jste spokojen se svým zaměstnáním?“ jsou tři možnosti:

- spokojený (1)
- ani spokojený, ani nespokojený (2)
- nespokojený (3)

Chtěli bychom modelovat spokojenost ($Y \in \{1, 2, 3\}$) pomocí modelu proporcionálních šancí, vysvětlující proměnnou by nám byla hodinová mzda (x, v desítkách Kč). Zvolme nyní 2 hodnoty hodinové mzdy x_1 a x_2. Dejme tomu, že by platilo následující.

Ve skupině zaměstnanců s hodinovou mzdou 120 Kč (tedy $x_1 = 12$) se na základě jejich odpovědí odhadlo, že odds($Y \leq 1$) (tedy šance[8] toho, že zaměstnanec je spokojený) je rovno 1 a odds($Y \leq 2$) (tedy šance toho, že zaměstnanec je spokojený či zaujímá neutrální postoj) je rovno $\frac{7}{3}$.

V druhé skupině zaměstnanců s hodinovou mzdou 110 Kč (tedy $x_2 = 11$) se zase odhadlo, že odds($Y \leq 1$) = 0,5 a odds($Y \leq 2$) = $\frac{7}{6}$.

[8]Výraz odds je anglicky šance, používá se jako značení.

Podívejme se, jak vypadají kumulativní poměry šancí:

$$\frac{\text{odds}(Y \leq 1|x_1)}{\text{odds}(Y \leq 1|x_2)} = \frac{1}{0{,}5} = 2$$

$$\frac{\text{odds}(Y \leq 2|x_1)}{\text{odds}(Y \leq 2|x_2)} = \frac{\frac{7}{3}}{\frac{7}{6}} = 2$$

Pro každé j z uvažovaného modelu je tedy kumulativní poměr šancí stejný, konkrétně platí, že zaměstnanci s hodinovou mzdou 120 Kč mají dvakrát větší šanci být spokojení než zaměstnanci s hodinovou mzdou 110 Kč, ale také mají dvakrát větší šanci nebýt nespokojeni, než mají zaměstnanci s platem 110 Kč/h. To samotné však nestačí. Zatím vidíme, že platí

$$\frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)} = 2,\ j = 1, 2,$$

a tedy

$$\log \frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)} = \log 2 \doteq 0{,}693,\ j = 1, 2.$$

Při srovnání s rovnicí (3.27) vidíme, že jednorozměrný parametr β by musel mít (teoreticky[9]) hodnotu $\log 2 \doteq 0{,}693$, neboť $x_1 - x_2 = 12 - 11 = 1$. Ověření platnosti předpokladů se nám nyní redukuje na ověření toho, zda pro libovolné x_{k1} a x_{k2} platí:

$$\frac{P(Y \leq j|x_{k1})/P(Y > j|x_{k1})}{P(Y \leq j|x_{k2})/P(Y > j|x_{k2})} = e^{\log 2(x_{k1} - x_{k2})},\ j = 1, 2.$$

Šance, že zaměstnanec inkasující 150 Kč/hodinu bude v práci spokojený, musí být $e^{\log 2(15-11)} = 16$krát větší než stejná šance zaměstnance s platem 110 Kč/hodinu, dále šance, že zaměstnanec pobírající 110 Kč/hodinu bude v práci nejvýše s neutrální spokojeností (tedy $Y \leq 2$), musí být $e^{\log 2(11-9)} = 4$krát větší než stejná šance zaměstnance s platem 90 Kč/hodinu atd.

V této interpretaci nám poměr šancí roste exponenciálně – platový rozdíl se zvětší 4krát, poměr šancí se zvětší 2^4krát (neboť $e^{\log 2} = 2$). Pokud bychom

[9]Při práci s daty máme samozřejmě pouze odhad tohoto parametru.

pracovali s **logaritmem** kumulativního poměru šancí, zvětšil by se v případě čtyřnásobného vzrůstu platového rozdílu skutečně čtyřikrát, v případě dvojnásobného nárůstu dvakrát, mění se tedy **proporcionálně** vzhledem ke vzdálenosti hodnot, pro které výraz porovnáváme, přičemž pořád platí, že hodnoty logaritmu kumulativního poměru šancí jsou stejné bez ohledu na j. Proto má model tento název a proto musí být tento předpoklad dodržen, jinak by nám jeden společný parametr β nestačil.

Z tohoto příkladu pak vyplynula **alternativní interpretace parametru** β, která je pouze jiným vyjádřením interpretace původní. Mohli jsme si všimnout, že β udává logaritmus kumulativního poměru šancí dvou subjektů, které se v hodnotě vysvětlující proměnné liší o jednotku. Pokud bychom měli více vysvětlujících proměnných a vektorový parametr $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, zmíněnou interpretaci bychom upravili takto:

Parametr β_m popisující efekt m-tého regresoru udává logaritmus kumulativního poměru šancí dvou subjektů, které se v hodnotě tohoto regresoru liší o jednotku a v ostatních vysvětlujících proměnných nabývají subjekty stejné hodnoty.

3.3.2.4 Test předpokladu proporcionálních šancí

Příklad popisující význam proporcionality šancí byl představen pouze za účelem jejího vysvětlení, v praxi je však nutné pomocí vhodného testu platnost předpokladu otestovat.

Je známo několik možností, jak tento předpoklad testovat. Agresti se jimi zabývá v kapitole 3.5.5 [1], v této práci se jim podrobně nevěnuji. Zmíním ovšem alespoň Brantův test, který cituje i Agresti a který je rozpracován v textu [3].

Tento test využívá testovou statistiku založenou na χ^2 rozdělení, není však úplně triviální. Zmiňuji jej zejména proto, že je implementován v softwaru R v rámci balíčku `brant` [31] pod stejnojmennou funkcí `brant()`.

Nulovou hypotézou je platnost testovaného předpokladu, výstupem jsou

p-value testu jak pro celkový model (`Omnibus`), tak pro jednotlivé vysvětlující proměnné (což je užitečné pro identifikaci proměnných, které jsou „problémové“ ve smyslu dodržení předpokladu modelu). Pokud je p-value pro celý model nižší než zvolená hladina testu, zamítáme platnost předpokladu proporcionálních šancí.

3.3.2.5 Odhad parametrů v modelu

U odhadu parametrů se vychází z věrohodnostní funkce $\mathrm{L}(\{\alpha_j\}, \boldsymbol{\beta})$, kterou pro n nezávislých pozorování sestavíme (tentokrát opět při původním, nezjednodušeném značení) jako

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{k}\pi_j(\mathbf{x}_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left\{\prod_{j=1}^{k}\left[P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i)\right]^{y_{ij}}\right\} =$$

$$= \prod_{i=1}^{n}\left\{\prod_{j=1}^{k}\left[\frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}\right]^{y_{ij}}\right\},$$

kde Y_i je zařazení do jedné z kategorií $\{1, \ldots, k\}$ u i-tého subjektu a y_{ij} je indikátor zařazení i-tého subjektu do j-té kategorie, který nabývá hodnoty 1 pro $Y_i = j$ a 0 pro $Y_i \neq j$.

Obvyklým postupem je najít takové hodnoty parametrů $(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})$, které maximalizují věrohodnostní funkci (která je funkcí těchto parametrů), a tyto hodnoty pak budou odhady parametrů v modelu.

Parciálním derivováním podle jednotlivých odhadovaných parametrů získáme tolik rovnic, kolik parametrů odhadujeme, tyto parciální derivace položíme rovny 0 (jak je zvykem při hledání maxima nějaké funkce). Získáme ovšem soustavu nelineárních rovnic, kterou je potřeba řešit iteračně. Agresti v kapitole 3.4.1 [1] zmiňuje Fisherův skórový algoritmus pro vyřešení této soustavy, odkazuji se tedy na něj.

Totožný algoritmus pro odhad parametrů používá funkce `vglm()` z knihovny `VGAM` [35]. Kromě příslušné formule modelu a používané datové sady je potřeba do ní zadat parametr `family=cumulative(parallel=T)` – tím softwaru „řekneme“, že chceme použít model proporcionálních šancí.

Stejné výsledky v trochu jiné podobě (co se uspořádání parametrů ve výstupu či rozsahu dodatečných informací o modelu týká) vrací funkce `polr()` z balíčku `MASS` [36]. Zmiňuji ji kvůli tomu, že dříve popisovaná funkce `brant()` bere jako jediný možný vstup objekt vytvořený právě funkcí `polr()`. Je tedy dobré vědět o obou možnostech, jak model v softwaru R odhadnout.

Kapitola 4

Analýza datového souboru

V této části bude hlavním cílem zjistit, zda výsledky ze základní části (reprezentované různými ukazateli) souvisí s následným výsledkem týmu v play-off.

4.1 Vztah mezi umístěním týmu v ZČ a počtem vyhraných sérií v play-off

Umístění po základní části je přirozeným údajem pro určení favoritů do části vyřazovací. Reprezentuje schopnost týmu vyhrávat a zároveň ji srovnává s ostatními. Na umístění je z velké části založeno rozlosování pavouka play-off (viz část 1.2). Má však skutečně takový vliv na úspěch v play-off?

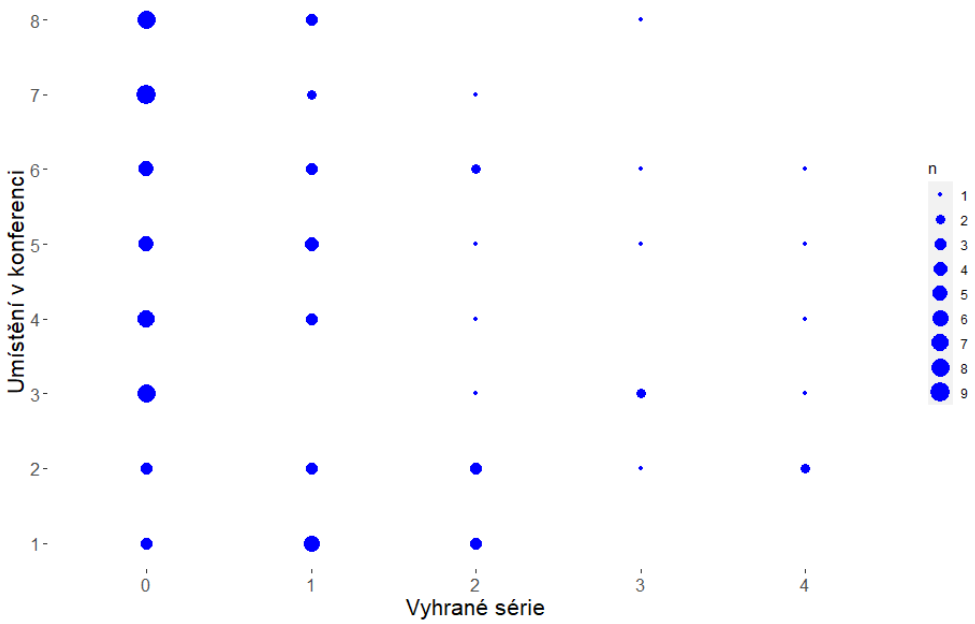
Ohledně vyhraných sérií bylo již zmíněno, že na ně v této práci bude pohlíženo jako na kategoriální proměnnou. Pokud bychom i umístění v konferenci brali pro tento případ jako zařazení do jedné z „kategorií“ {1,...,8} jakožto výstup několika proměnných – především počtu bodů, ale v případě shody rozhoduje několik dalších kritérií [14] – mohli bychom realizace obou proměnných umístit do kontingenční tabulky.

Příklad grafického zobrazení kontingenční tabulky je na Obrázku 4.1.

Na první pohled se zdá, že umístění nějaký vliv na počet vyhraných sérií má – týmy s horším umístěním se zřídkakdy dostanou v pavoukovi daleko, u lépe umístěných je tento jev častější, ovšem například vítězové konferencí se za 6 sezón

Tabulka 4.1: Vyhrané série a umístění v konferenci – kontingenční tabulka

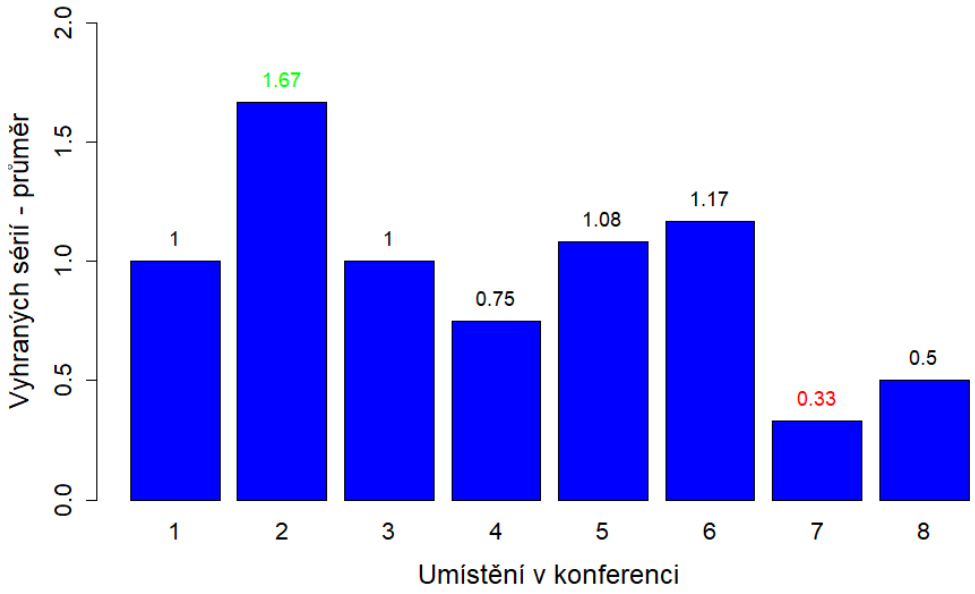
	Vyhrané série					
UvK	0	1	2	3	4	Suma
1	3	6	3	0	0	12
2	3	3	3	1	2	12
3	8	0	1	2	1	12
4	7	3	1	0	1	12
5	5	4	1	1	1	12
6	5	3	2	1	1	12
7	9	2	1	0	0	12
8	8	3	0	1	0	12
Suma	48	24	12	6	6	96



Obrázek 4.1: Počet vyhraných sérií play-off dle umístění v konferenci

nedostali ani jednou do finále (ve třech případech skončili ve finále konferencí), což však může být výrazně ovlivněno náhodou. Také posuzujeme výsledky pouhých šesti sezón.

Můžeme si ještě zobrazit průměrné počty vyhraných sérií dle pozic v tabulce jednotlivých konferencích po základní části (viz Obrázek 4.2).



Obrázek 4.2: Průměrný počet vyhraných sérií play-off pro jednotlivá umístění v konferenci

Výrazně menší průměr mají týmy, které se umístily na 7. a 8. místě, týmy na pozicích 5 a 6 však mají větší průměrný počet vyhraných sérií než týmy na 1. místě. Velký počet vyhraných sérií mají v průměru týmy na 2. místě – musíme si však uvědomit, že pro každou pozici průměrujeme pouze 12 hodnot. Právě u týmů na druhé pozici zvýšilo hodnotu průměru dvojnásobné vítězství týmu Pittsburgh Penguins v letech 2016 a 2017, kdy postupovali do play-off z druhého místa v konferenci a v obou případech si došli až pro Stanley Cup. Proto ani informace získaná z průměrných hodnot neřekne příliš mnoho o vztahu umístění a vyhraných sériích, alespoň trochu však dokreslí to, co jsme pozorovali v kontingenční tabulce.

Užitečný bývá také pohled do tabulky s četnostmi relativními. V Tabulce 4.2 jsou počítány vzhledem k řádkovým součtům.[1]

Když pomineme snahu přímo odhadnout efekt typu „čím lepší umístění, tím dál se tým ve vyřazovací fázi dostane“, můžeme alespoň říct, že mezi umístěním a počtem vyhraných sérií nějaká závislost nejspíš je – relativní četnosti se v jed-

[1]Hodnoty uvnitř tabulky tedy představují odhady pravděpodobností $P(\mathrm{VS} = j|\mathrm{UvK} = i)$, $1 \leq i \leq 8$ a $0 \leq j \leq 4$.

Tabulka 4.2: Vyhrané série a umístění v konferenci – relativní četnosti

	Vyhrané série				
UvK	0	1	2	3	4
1	0.25	0.50	0.25	0.00	0.00
2	0.25	0.25	0.25	0.08	0.17
3	0.67	0.00	0.08	0.17	0.08
4	0.58	0.25	0.08	0.00	0.08
5	0.42	0.33	0.08	0.08	0.08
6	0.42	0.25	0.17	0.08	0.08
7	0.75	0.17	0.08	0.00	0.00
8	0.67	0.25	0.00	0.08	0.00

notlivých řádcích liší, zdá se, že řádkový a sloupcový znak nejsou nezávislé. Potřebovali bychom to však otestovat.

V tomto případě by se nabízel χ^2 test nezávislosti, nejprve však musíme ověřit, zda máme dostatečně velké očekávané četnosti, jejichž výpočet je uveden v části 3.1.1. Můžeme si je prohlédnout v Tabulce 4.3,

Tabulka 4.3: Očekávané četnosti v případě nezávislosti

	Vyhrané série				
UvK	0	1	2	3	4
1	6	3	1.5	0.75	0.75
2	6	3	1.5	0.75	0.75
3	6	3	1.5	0.75	0.75
4	6	3	1.5	0.75	0.75
5	6	3	1.5	0.75	0.75
6	6	3	1.5	0.75	0.75
7	6	3	1.5	0.75	0.75
8	6	3	1.5	0.75	0.75

V každé buňce Tabulky 4.3 (s výjimkou prvního sloupce) jsou očekávané četnosti menší než 5, musíme tak vztah těchto dvou proměnných zkoumat pomocí jiných metod.

V případě malých očekávaných četností se můžeme obrátit na *Fisherův exaktní*

test[2] (též Fisherův faktoriálový test, rozpracován je např. ve skriptu [7]), který také testuje hypotézu nezávislosti. Na rozdíl od χ^2 testu pohlíží na marginální četnosti jako na pevně dané, což v tomto případě vůbec nevadí – řádkové součty souvisí s počtem konferencí (dvě) a počtem ročníků (šest), u každé pozice tak máme vždy součet 12, sloupcové součty v Tabulce 4.1 jsou zase pevně dány systémem play-off.

V softwaru R tento test provedeme pomocí funkce `fisher.test()`. Vzhledem k tomu, že tento test počítá p-value pomocí výpočtu pravděpodobnosti všech možných kombinací četností uvnitř tabulky při daných marginálních, bude výpočetní náročnost u takto rozsáhlé tabulky veliká. Do funkce tak musíme přidat argument `simulate.p.value = T`, díky kterému funkce spočítá p-hodnotu testu na základě Monte Carlo simulací [34].

Pro Tabulku 4.1 vyjde takto vypočítaná p-value přibližně **0,30** (použito 200 000 simulací), na hladině významnosti 0,05 **nelze hypotézu nezávislosti zamítnout** – získaná data tedy neposkytují dostatečnou evidenci pro to, abychom mohli mluvit o statisticky významném vztahu umístění v konferenci po základní části a vyhraných sériích v následném play-off.

Ani jeden z těchto dvou testů však nepohlíží na znaky v kontingenční tabulce jako na uspořádané. Na kategorie obou testovaných proměnných se dívá jako na nominální, což může být důvod, proč jsme nebyli schopni zamítnout nulovou hypotézu – nezahrnutím ordinality kategorií totiž přicházíme o jistý druh informace v datech. Zkusíme se tedy na vztah umístění a vyhraných sérií podívat s využitím metod pro ordinální data.

Připomeňme „podezření“ na to, že s lepším umístěním (nižší hodnota UvK) se týmy dostanou v play-off dál (vyšší hodnota VS). Toto podezření je založeno na všeobecném povědomí – jako favorit v sérii play-off se většinou označuje tým s lepším umístěním po základní části, týmy s perfektní základní částí se pasují

[2]Tento test se sice většinou používá pro čtyřpolní tabulku (2x2), je však možné jej použít i pro test v kontingenční tabulce s obecně r řádky a s sloupci, jak se Agresti zmiňuje v úvodu kapitoly 7.6 v publikaci [1].

do role favoritů na Stanley Cup a podobně. Četnosti v kontingenční tabulce napovídají, že by toto podezření mohlo být oprávněné.

Snaha o potvrzení tohoto podezření pomocí metod pro ordinální data spočívá v tom (jak už bylo zmíněno v části 3.1.2), že pro kontingenční tabulku budeme chtít ukázat převahu diskordantních párů pozorování.

Následující tabulka obsahuje hodnoty koeficientů definovaných v části 3.1.2 pro Tabulku 4.1, a to včetně 95% konfidenčního intervalu pro teoretický protějšek těchto koeficientů (sloupce Dolní a Horní hranice).

Tabulka 4.4: Hodnoty koeficientů asociace – UvK a VS

Typ koeficientu	**Hodnota koeficientu**	**Dolní hranice**	**Horní hranice**
Gamma koeficient	-0.245	-0.423	-0.068
Kendallovo Tau-b	-0.190	-0.327	-0.052
Somersovo d	-0.218	-0.375	-0.060

Z výsledků můžeme usuzovat, že mezi umístěním týmu v konferenci a počtem vyhraných sérií je skutečně negativní vztah. **Lépe umístěné týmy tak mají tendenci vyhrávat více sérií play-off** než ty hůře umístěné.

Důležité pro tento závěr je, že žádný z 95% intervalů spolehlivosti jednotlivých koeficientů neobsahuje 0, jde tedy o zamítnutí H_0: „Koeficient asociace = 0“ proti oboustranné alternativě na hladině testu 0,05. Tím tedy vzhledem k implikaci (3.9) **zamítáme nezávislost obou proměnných**, pozorovaný vztah je tedy statisticky významný.

Podobný závěr získáme i použitím Linear-by-linear association testu – funkce `lbl_test()` spočítala pro oboustrannou alternativu (k nulové hypotéze „Mezi umístěním v konferenci a vyhranými sériiemi neexistuje žádná asociace.“) p-value rovnu **0,038**, při alternativě zahrnující negativní vztah proměnných (argument `alternative="less"`) byla hodnota p-value dokonce **0,019**. Na hladině významnosti 0,05 v obou případech **zamítáme nulovou hypotézu**.

Zopakujme tedy závěr pro vztah umístění po základní části a počtu vyhraných sérií v play-off: Lépe umístěné týmy mají tendenci dostat se v play-off dál než týmy vstupující do vyřazovací fáze z nižších příček tabulky.

4.2 Vztah mezi týmovými či hráčskými statistikami v ZČ a počtem vyhraných sérií v play-off

K posouzení vztahu statistik, na které můžeme pohlížet jako na kvantitativní znaky, a vyhraných sérií v play-off využijeme metody z části 3.2. Bude nás zajímat, zda se hodnoty těchto ukazatelů liší dle toho, kam se tým v play-off dostal, případně jde-li tam vypozorovat nějaký trend ve smyslu „čím větší/menší hodnota ukazatele X, tím více vyhraných sérií“. Na první otázku si můžeme odpovědět vhodným testem, druhou zatím posoudíme graficky a více tuto tendenci ověříme v části 4.3.

Pro grafické posouzení využijeme boxplot či mean plot (viz dále), k testování zvolíme metodu podle toho, jak budou u dané statistiky splněny předpoklady. V případě splnění předpokladů ANOVA (normalita, stejné rozptyly) je možné využít tuto metodu, při jejich porušení použijeme Kruskalův–Wallisův test.

U většiny statistik se při testování budeme dopouštět jistého zjednodušení – ANOVA i Kruskalův-Wallisův test jsou dle svých předpokladů určeny k ověřování shody středních hodnot spojité proměnné (ANOVA), resp. shody distribučních funkcí spojité proměnné (Kruskalův–Wallisův test) v jednotlivých kategoriích daných hodnotami kvalitativního znaku. Počet bodů v tabulce, dorážek či trestných minut rozhodně nemůžeme označit za spojité proměnné, když jejich hodnoty mohou být pouze celá nezáporná čísla.

Musíme si však uvědomit, že i u proměnných spojitých ze své podstaty pracujeme vždy jen s nějakou přesností, např. zaokrouhlíme hodnoty na určitý počet desetinných míst – také tedy v rámci analýzy „nemohou“ nabývat všech hodnot. Prakticky vždy se tak dopouštíme určitého zjednodušení a „obcházení“ předpokladů, které však nemusí být nutně špatné.

Pokud předpoklady ověříme a nebudeme příslušnou statistickou metodou schopni zamítnout jejich platnost, nemusí nám vadit to, že o počtu dorážek víme, že se normálním rozdělením řídit ze své podstaty nemůže – když Shapi-

rovým–Wilkovým testem nezamítneme normalitu počtu dorážek u všech skupin a Bartlettovým testem nezamítneme shodu rozptylů, můžeme ANOVA použít. U Kruskalova–Wallisova testu nám pro jednoduchost bude stačit, když realizovaných hodnot proměnných nebude příliš málo (jako je tomu např. u umístění v konferenci, kde bylo pouze 8 možných hodnot a jehož vztah s vyhranými sériemi zkoumáme pomocí jiných metod).

V případě, že by u některých proměnných nešlo použít ani jeden z uvedených testů[3], můžeme se znovu podívat na jejich hodnoty při počtu vyhraných sérií 0–4 v kontingenční tabulce a testovat nezávislost pomocí Fisherova exaktního testu či lépe Linear-by-linear association testu, pokud bude možné jejich realizace uspořádat (což v případě různých počtů jistě bude možné).

Týmových ukazatelů a číselných charakteristik hráčských ukazatelů máme obrovské množství (viz části 2.3 a 2.4), ukážu tedy jen příklady toho, jak jsem při analýze postupoval.

Jako první se můžeme podívat na vztah počtu vyhraných sérií a mediánu vstřelených gólů na 60 minut hry u útočníků.

Podívejme se nejprve na některé ze základních číselných charakteristik této kvantitativní proměnné v jednotlivých kategoriích VS.[4]

Tabulka 4.5: Základní číselné charakteristiky – *Medián gólů/60 minut – útočníci*

Počet VS	Minimum	Dolní kvartil	Medián	Průměr	Horní kvartil	Maximum
0	0.511	0.615	0.706	0.717	0.785	0.955
1	0.532	0.648	0.715	0.724	0.784	0.976
2	0.611	0.716	0.756	0.767	0.819	0.915
3	0.605	0.663	0.716	0.711	0.777	0.788
4	0.642	0.690	0.762	0.740	0.788	0.811

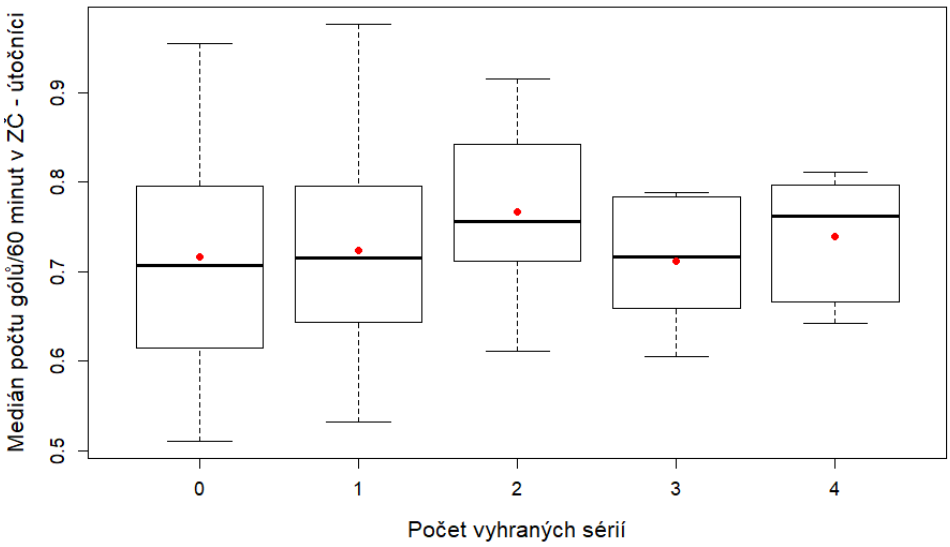
Minimální hodnota mediánu gólů/60 minut se zvyšuje spolu s VS (ovšem zároveň klesá maximum, v podstatě se spolu se snižujícím počtem týmů snižuje variabilita), u poražených finalistů jsou průměrné a mediánové hodnoty proměnné

[3]Právě kvůli malému počtu unikátních realizací.

[4]Připomeňme však, že pro 3 a 4 vyhrané série máme k dispozici pouze 6 datových bodů, neboť za 6 sezón známe stejný počet vítězů Stanley Cupu i stejný počet poražených finalistů.

na podobné úrovni jako u týmů, které vypadly hned v 1. kole. Nic výrazně podezřelého, co by poukazovalo na významný rozdíl ve skupinách či tendenci mít s vyšší hodnotou statistiky lepší výsledek v play-off, z tabulky vyčíst nelze.

Lépe se nám tyto číselné charakteristiky budou posuzovat v boxplotu (Obrázek 4.3). Na tom můžeme vidět všechny hodnoty z Tabulky 4.5, průměr je znázorněn červeným bodem.

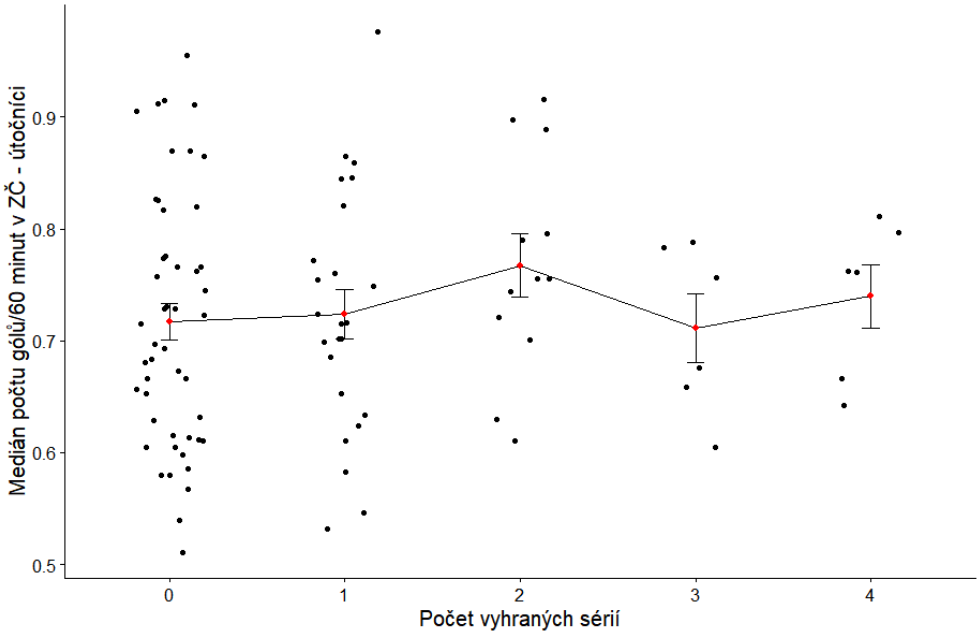


Obrázek 4.3: Medián gólů/60 minut (útočníci) dle počtu vyhraných sérií

Ani při pohledu na boxplot se nezdá, že by např. útočníci vítězů Stanley Cupu měli výrazně větší medián gólů/60 minut hry než útočníci z týmů, které vypadly v prvním či druhém kole.

Obrázek 4.4 se nazývá *mean plot* a kromě zobrazení jednotlivých datových bodů vytváří spojnici mezi odhady střední hodnoty (tedy mezi průměry) v jednotlivých skupinách, konkrétně spojí vždy dva sousední průměry úsečkou. Tato spojnice se zdá být téměř „rovnoběžná“ s osou x, nedá se říct, že by směrem k vyšším hodnotám VS rostla. Zdá se, že by střední hodnoty v jednotlivých skupinách mohly být shodné.

Shodu středních hodnot bychom mohli otestovat pomocí ANOVA, nejprve však musíme ověřit předpoklady.



Obrázek 4.4: Medián gólů/60 minut (útočníci) dle počtu vyhraných sérií

Normalitu můžeme otestovat Shapirovým–Wilkovým testem, a to v každé skupině zvlášť.

Tabulka 4.6: Výsledky Shapirova–Wilkova testu – *Medián gólů/60 minut – útočníci*

Počet VS	p-value
0	0.153
1	0.871
2	0.550
3	0.356
4	0.233

Na hladině 0,05 normalitu nezamítáme ani v jedné skupině, tento předpoklad tak můžeme považovat za neporušený.

Dále potřebujeme ověřit shodu rozptylů zkoumané proměnné v jednotlivých kategoriích. Použijeme k tomu funkci `bartlett.test()`, která při nulové hypotéze o shodě rozptylů vrátí p-value = 0,596 – ani shodu rozptylů na zvolené hladině významnosti zamítnout nemůžeme.

Předpoklady pro ANOVA jsou tak splněny, pomocí funkce `aov()` můžeme tento test provést v R, hodnota testové statistiky F_A v tomto případě vyšla 0,589, což odpovídá p-value = **0,672** $\Rightarrow$ shodu středních hodnot nelze na hladině 0,05 zamítnout.

V hodnotách mediánu vstřelených gólů útočníky za 60 minut hry není významný rozdíl napříč kategoriemi určenými počtem vyhraných sérií v play-off.

Dále bychom mohli porovnat vztah vyhraných sérií a průměrného rozdílu střel týmu (SDiff), který se vypočítá jako rozdíl střel na branku soupeře a soupeřových střel na branku daného týmu vydělený počtem zápasů.

Začněme opět pohledem na základní číselné charakteristiky proměnné SDiff.

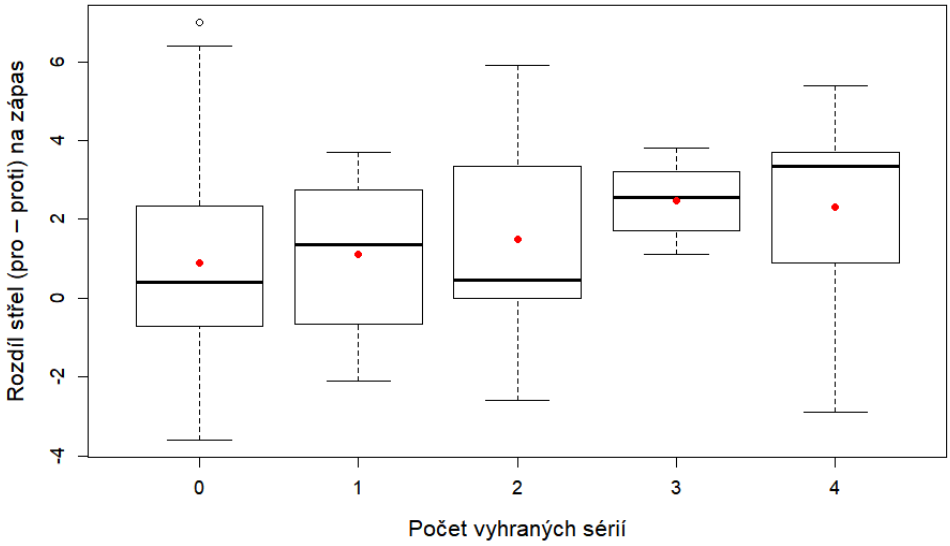
Tabulka 4.7: Základní číselné charakteristiky proměnné *Rozdíl střel pro/proti na zápas*

Počet VS	Minimum	Dolní kvartil	Medián	Průměr	Horní kvartil	Maximum
0	-3.600	-0.700	0.400	0.890	2.330	7.000
1	-2.100	-0.525	1.350	1.110	2.720	3.700
2	-2.600	0.000	0.450	1.500	2.680	5.900
3	1.100	1.800	2.550	2.480	3.150	3.800
4	-2.900	1.470	3.350	2.300	3.650	5.400

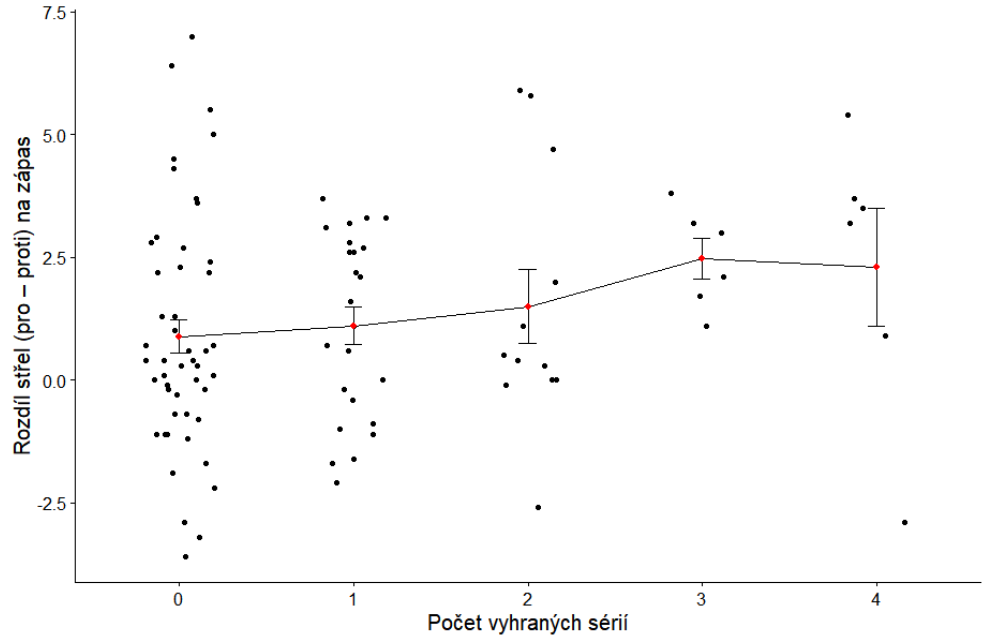
Dle hodnot v Tabulce 4.7 se zdá, že finalisti své soupeře v základní části v průměru přestříleli více než týmy, které se dostaly nejdál do finále konferencí.

I z boxplotu (Obrázek 4.5) vypadá vztah proměnných tak, že týmy, které vyhrály víc sérií v play-off, měly za základní část větší hodnotu střeleckého rozdílu na zápas. Tuto „tendenci“ bychom mohli pozorovat jak z mediánů (čáry uprostřed jednotlivých „krabic“), tak průměrů (červené body).

Na Obrázku 4.6 pak můžeme vidět i spojnici průměrů. Kdybychom tyto úsečky (resp. jejich krajní body) proložili přímkou, byla by to nejspíš přímka rostoucí – to by poukazovalo na to, že mezi těmito dvěma proměnnými je výše popsaná závislost, tedy čím víc tým v průměru soupeře v základní části přestřílí, tím dál v play-off dojde.



Obrázek 4.5: Rozdíl střel (pro – proti) na zápas dle počtu vyhraných sérií



Obrázek 4.6: Rozdíl střel (pro – proti) na zápas dle počtu vyhraných sérií

Z mean plotu navíc vidíme, že průměr vítězům Stanley Cupu dost snižuje jedno „odlehlé“[5] pozorování, které představuje Washington v roce 2018. V základní části ho soupeři v zápase přestříleli v průměru téměř o 3 střely, přesto ovládl vyřazovací část (a ani v základní části si nevedl špatně, v konferenci se umístil na 3. místě). Podobně týmům, které prohrály ve finále konferencí (VS=2), celkový průměr v proměnné SDiff „kazí“ tým Montrealu z ročníku 2013/14, který si v zápasech základní části připsal průměrně o 2,6 střely méně než jeho soupeř – následně se dostal mezi 4 nejlepší mužstva ligy v dané sezóně.

Bude pozorovaný rozdíl v odhadech středních hodnot významný? Zkusme nejprve ověřit předpoklady pro použití ANOVA.

Ač shodu rozptylů pomocí Bartlettova testu zamítnout na hladině 0,05 nemůžeme (p-value = 0,148), s normalitou je u této proměnné problém. Výsledky Shapirova–Wilkova testu normality pro hodnoty SDiff v jednotlivých skupinách jsou zobrazeny v Tabulce 4.8.

Tabulka 4.8: Výsledky Shapirova–Wilkova testu – *Rozdíl střel (pro – proti) na zápas*

Počet VS	p-value
0	0.107
1	0.045
2	0.058
3	0.851
4	0.283

Na hladině 0,05 musíme – ač těsně – zamítnout předpoklad normality pro hodnoty průměrného střeleckého rozdílu těch týmů, které vyhrály jednu sérii ve vyřazovací části (resp. vypadly ve druhém kole play-off).

Předpoklady pro použití ANOVA nejsou splněny, využijeme tedy Kruskalův–Wallisův test. Ten v R provedeme pomocí funkce `kruskal.test()`, jako výsledek

[5] Nikoliv detekované boxplotem jako odhlehlé – tím je v datech při pohledu na boxplot jen největší pozorování pro VS= 0, tedy prázdný bod nad čarou neodlehlého maxima – zde je to myšleno jako pozorování, které je svou hodnotou opticky výrazně odtržené od ostatních v rámci skupiny, založeno je to tedy jen na subjektivním posouzení.

získáme p-value = **0,172**, na hladině 0,05 tedy nelze zamítnout shodu rozdělení proměnné SDiff napříč jednotlivými skupinami určenými počtem vyhraných sérií v play-off.

U obou výše zmíněných příkladů si musíme znovu uvědomit, že ANOVA ani Kruskalův–Wallisův test nevyužívají ordinalitu jednotlivých kategorií, na skupiny určené počtem vyhraných sérií pohlížejí oba testy jako na nominální.

Jiný postup zvolíme v případě počtu obránců, kteří vstřelili za tým alespoň 6 gólů. Pro zkoumání vztahu této veličiny a počtu vyhraných sérií bude vhodnější zvolit některé z metod pro posouzení vztahu dvou kategoriálních veličin či přímo ordinálních proměnných.

Počty obránců s minimálně 6 góly v sezóně se v datech realizují pouze hodnotami z množiny $\{1,\ldots,6\}$. Počet týmů, které dané množství alespoň 6gólových obránců v sezóně měly, je uveden v Tabulce 4.9.

Tabulka 4.9: Počet obránců s 6 a více góly v týmu

Počet obránců s 6 a více góly	1	2	3	4	5	6
Četnost	12	36	31	14	2	1

Podívejme se na kontingenční tabulku, kde si počet těchto obránců v týmu porovnáme s vyhranými sériemi play-off.

Tabulka 4.10: Vyhrané série a počet obránců s 6 a více góly v týmu – kont. tabulka

	Vyhrané série					
Obránci s 6+ G	0	1	2	3	4	Suma
1	8	2	1	1	0	12
2	16	9	4	4	3	36
3	15	9	3	1	3	31
4	7	4	3	0	0	14
5	1	0	1	0	0	2
6	1	0	0	0	0	1
Suma	48	24	12	6	6	96

Nezdá se, že by větší počet obránců s alespoň 6 góly zvyšoval šanci na větší úspěch v play-off či že by spolu obě proměnné závisely nějak jinak.

Pokud otestujeme jejich nezávislost pomocí Fisherova exaktního testu (opět s využitím Monte Carlo simulací), získáme p-value = **0**,**927**, hypotézu nezávislosti tedy na žádné rozumné hladině zamítnout nelze.

Metody pro ordinální data poskytnou stejný závěr. P-value Linear-by-linear association testu vyšla při oboustranné hypotéze **0**,**711**, hodnoty koeficientů asociace jsou v Tabulce 4.11.

Tabulka 4.11: Hodnoty koeficientů asociace – Počet obránců s 6+ góly a VS

Typ koeficientu	**Hodnota koeficientu**	**Dolní hranice**	**Horní hranice**
Gamma koeficient	-0.0009	-0.2426	0.2408
Kendall Tau-b	-0.0006	-0.1666	0.1653
Somersovo d	-0.0007	-0.1731	0.1718

Všechny intervaly spolehlivosti obsahují nulu, bodové odhady teoretických protějšků vyšly u všech koeficientů také téměř nulové – můžeme tedy vyvodit závěr, že tímto způsobem stanovený počet skórujících obránců nemá na počet vyhraných sérií statisticky významný vliv.

Podobně bychom mohli analyzovat vztah dalších ukazatelů s počtem vyhraných sérií. Výsledky porovnávání dvou znaků vyjádřené pomocí p-value a s informací, který test byl použit (přičemž rozhodování bylo stejné jako ve výše uvedených příkladech), jsou zapsané v Tabulce 4.12 a seřazené vzestupně podle hodnoty p-value.

Označení K–W je pro Kruskalův–Wallisův test, LbL zase značí použití Linear-by-linear association testu. Zobrazeno je prvních 30 testovaných proměnných, ostatní (dalších 53 ukazatelů) měly p-value ještě vyšší. Jejich názvy jsou stejné jako v části 2, případně je za podtržítkem doplněno písmeno D (počítané pouze pro obránce) či F (to samé pro útočníky).

Tabulka 4.12: Výsledky ověřování vztahu ukazatelů s počtem vyhraných sérií

Proměnná	P-value	Test
Corsi%AS_med_F	0.070	ANOVA
gamescore_med_F	0.074	K–W
goly60_med	0.108	K–W
KB_NAD25_F	0.117	LbL
gamescore60_sd	0.120	ANOVA
Corsi%AS_sd	0.122	ANOVA
Corsi%AS_IQR	0.130	K–W
gamescore_med	0.155	ANOVA
B	0.166	ANOVA
SDiff	0.172	K–W
SAT%	0.183	ANOVA
SAT%Tesne	0.185	ANOVA
gamescore60_IQR_F	0.185	ANOVA
gamescore60_med	0.188	ANOVA
GDiff	0.193	K–W
goly_IQR_D	0.209	K–W
Corsi%AS_sd_F	0.210	K–W
gamescore60_sd_F	0.218	ANOVA
gamescore60_med_F	0.248	ANOVA
Corsi%AS_med_D	0.256	K–W
ShF	0.275	ANOVA
TrestProProti60	0.288	ANOVA
Út.P%	0.342	ANOVA
KB_IQR	0.372	K–W
IcetimeZ_sd_F	0.374	K–W
IcetimeZ_sd_D	0.396	K–W
gamescore_IQR_F	0.398	ANOVA
gamescore_IQR	0.405	ANOVA
goly_sd_F	0.422	K–W
IcetimeZ_IQR_F	0.423	K–W

Na hladině významnosti 0,05 nebyl vztah vyhraných sérií s jakoukoliv proměnnou významný, tedy nebyli jsme schopni zamítnout shodu středních hodnot, resp. shodu rozdělení kvantitativní proměnné napříč skupinami danými proměnnou VS,

resp. jsme nebyli schopni zamítnout neexistující asociaci mezi vyhranými sériemi a jinou ordinální veličinou.

To, že jsme nenalezli statisticky významný vztah pro některou dvojici, ještě neznamená, že nám hodnoty některého z ukazatelů neposkytnou zajímavou informaci o následném počtu vyhraných sérií ve vyřazovací části. Pokud bychom modelovali proměnnou VS pomocí modelu proporcionálních šancí, některý z ukazatelů by přesto mohl být statisticky významný v rámci tohoto modelu, pomocí kterého bychom pak mohli stanovit odhad pravděpodobnosti toho, v jakém kole play-off tým vypadne při daných hodnotách regresorů.

Tento model bude tvořit obsah poslední sekce této analýzy.

4.3 Model proporcionálních šancí pro vyhrané série play-off

V předchozích částech analýzy jsme došli k závěru, že pouze u umístění týmu v rámci konference jsme byli schopni nalézt statisticky významný vztah k počtu vyhraných sérií v play-off.

Shodu středních hodnot, resp. rozdělení, resp. neexistující asociaci jsme sice pro vztah týmových a hráčských statistik spolu s VS nebyli schopni zamítnout, i tak nám některé z nich mohou vyjít v modelu proporcionálních šancí jako významné vysvětlující proměnné.

Pokud bychom totiž vycházeli pouze ze závěrů porovnávání dvou znaků v předchozí části, mohli bychom zjednodušeně konstatovat, že ať je hodnota daného ukazatele jakákoliv, nenapoví nám to nic moc o tom, kam se tým v play-off dostane – hodnoty statistik se přeci napříč skupinami významně neliší.

Jinými slovy, pokud by nám někdo řekl, že daný tým v základní části např. přestřílel soupeře v průměru o 6 střel na jeden zápas (SDiff = 6), což by byla naše jediná informace o počínání daného týmu v základní části, a zeptal by se nás, jaké pravděpodobnosti bychom přiřadili možným výsledkům týmu v play-off, nebyli bychom na základě závěrů části 4.2 schopni říct nic víc, než co vyplývá ze samotného systému play-off – tým s 50% pravděpodobností vypadne hned v 1. kole,

s pravděpodobností 0,25 vypadne až ve druhém kole, ..., s pravděpodobností 0,0625 (tedy 1/16) vyhraje Stanley Cup.

Tento výpočet pravděpodobností je v podstatě totožný s příkladem uvedeným na konci části 3.3.2.1, kdy jsme pro počet vyhraných sérií sestavili proportional odds model, který obsahoval pouze absolutní členy spočítané právě na základě znalosti systému play-off, v modelu nebyl žádný regresor, který by nám přinesl informaci navíc.

Když se však vrátíme k Obrázku 4.6, napadne nás, že ač se rozdíl v rozdělení hodnot SDiff neukázal na hladině 0,05 jako statisticky významný, tak při znalosti toho, že SDiff = 6, bychom nejspíš pravděpodobnosti přiřadili jednotlivým hodnotám VS trochu jinak. Jak přesně? A skutečně nám tato informace pomůže odhad výsledku týmu v play-off vylepšit? Na to nám může odpovědět právě model proporcionálních šancí.

Těchto modelů bychom mohli sestavit velké množství vzhledem k počtu potenciálních vysvětlujících proměnných, které máme k dispozici. Postupně přidávat a ubírat proměnné a sledovat, zda se model vylepšil (dle určitého kritéria, viz dále), by bylo poměrně neefektivní, software R však v sobě má implementován efektivnější způsob hledání optimálního modelu.

Pro nalezení nejlepšího modelu využijeme funkci `step4vglm()`, která jako kritérium vhodnosti modelu používá *Akaikeho informační kritérium* (AIC), které – zjednodušeně řečeno – udává to, kolik informace jsme použitím modelu ztratili. Ztrátu informace chceme minimalizovat, proto pro nás bude nejlepším modelem ten, který bude mít hodnotu AIC nejmenší.[6] Tato funkce je adaptací funkce `step()`, obě na základě zvoleného parametru `direction` postupují takto [5, kapitola 5.3]:

a) `direction = "forward"` – *vzestupný výběr*: funkce vyjde z modelu, který obsahuje pouze absolutní člen, postupně přidává regresory a vyhodnocuje, pro jaký přidaný regresor bude hodnota AIC nejnižší. Pokud by se přidáním

[6] Podrobněji je AIC popsáno např. v kapitole 5 skript [5].

jakéhokoliv regresoru AIC zvýšilo, máme již optimální model, už nic nepřidáváme a funkce proces zastaví.

b) `direction = "backward"` – *sestupný výběr*: funkce vyjde z plného modelu, tedy se všemi regresory, které máme k dispozici. Postupně odstraňuje z modelu ty vysvětlující proměnné, jejichž vynecháním docílíme nejnižší hodnoty AIC. Pokud by se vynecháním jakéhokoliv dalšího regresoru AIC zvýšilo, máme již optimální model, už nic neodstraňujeme a funkce proces zastaví.

c) `direction = "both"` – *kroková regrese*: kombinace předchozích dvou postupů, funkce postupuje podle vzestupného výběru, v každém kroku však ověří, zda by se odebráním některého z dříve přidaných regresorů model nevylepšil (tedy zda by se hodnota AIC nesnížila).

Přejděme tedy k nalezení modelu. Ze tří představených postupů zvolíme **krokovou regresi**, neboť plný model se všemi regresory by se nám nepodařilo sestavit kvůli lineární závislosti sloupců matice, která je v rámci předpokladu nezávislosti sloupců definována v části 3.3.2.1. To znemožňuje použití sestupného výběru. Když už tedy budeme vycházet z nejjednoduššího modelu, bude při postupném nabírání regresorů lepší mít možnost některý z nich vyřadit – proto volba krokové regrese.

Nejprve budeme vysvětlující proměnné uvažovat bez interakcí, abychom zúžili skupinu možných regresorů minimalizujících ztrátu informace. Tím jsme dosáhli modelu s devíti vysvětlujícími proměnnými (z původních 83 ukazatelů), AIC tohoto modelu vyšlo přibližně 245,43.

Pro těchto 9 regresorů jsme pro další krokovou regresi umožnili zařazení interakcí mezi vysvětlujícími proměnnými, čímž jsme získali druhý model a snížili jak AIC (na 239,53), tak počet statistických ukazatelů zapojených do modelu (na šest). Přidáním možnosti mít vysvětlující proměnné kromě lineárního členu také v členu kvadratickém byl pomocí opětovného použití funkce `step4vglm()` získán finální model s AIC = 238,79.

Vysvětlující proměnné, které tento model obsahuje, jsou tyto:

- umístění v konferenci (UvK),
- směrodatná odchylka ice-time/zápas u obránců (IcetimeZ_sd_D),
- rozdíl střel (pro − proti) na zápas (SDiff),
- počet obránců, kteří vstřelili alespoň 6 gólů (golyNAD6_D),
- mezikvartilové rozpětí ukazatele Gamescore (gamescore_IQR),
- počet útočníků, kteří získali alespoň 25 kanadských bodů (KB_NAD25_F).

Podle Brantova testu žádný z regresorů neporušuje předpoklad proporcionality šancí, celková p-value tohoto testu vychází **0,12**, přičemž i pro tento test uvažujeme hladinu významnosti 0,05.

Podívejme se tedy, jak vypadá výsledný model. Vztah počtu vyhraných sérií a vysvětlujících proměnných model popisuje takto:

$$\begin{aligned}\log \frac{P(\mathrm{VS} \leq j)}{1 - P(\mathrm{VS} \leq j)} = \alpha_j &+ \beta_1(\mathrm{UvK}) + \beta_2(\mathrm{IcetimeZ\_sd\_D}) + \beta_3(\mathrm{SDiff}) + \\ &+ \beta_4(\mathrm{golyNAD6\_D})^2 + \beta_5(\mathrm{gamescore\_IQR}) + \\ &+ \beta_6(\mathrm{KB\_NAD25\_F}) + \beta_7(\mathrm{UvK} \cdot \mathrm{IcetimeZ\_sd\_D}) + \\ &+ \beta_8(\mathrm{UvK} \cdot \mathrm{SDiff}) + \beta_9(\mathrm{SDiff} \cdot \mathrm{KB\_NAD25\_F}), \quad j = 0, \ldots, 3.\end{aligned}$$

Interakce regresorů jsou v modelu znázorněny symbolem „ · “, jde totiž o součin hodnot příslušných vysvětlujících proměnných. Pro lepší interpretovatelnost modelu jsou některé proměnné objevující se v interakcích jistým způsobem „centrovány“ tak, aby při jejich nulové hodnotě dávala interpretace smysl. Toto centrování se u regresorů IcetimeZ_sd_D a KB_NAD25_F provedlo odečtením mediánu od všech hodnot, u UvK (které nabývá hodnot 1–8) se tento problém vyřešil odečtením čísla 1 od všech hodnot – nyní tedy pro umístění v konferenci

máme k dispozici hodnoty 0–7, přičemž 0 značí nejlepší možné umístění a 7 nejhorší. Proměnnou SDiff nebylo třeba centrovat, neboť její nulová hodnota má přirozenou interpretaci – v případě SDiff = 0 tým v průměru vystřelil na branku soupeře stejný počet střel, jaký jeho soupeř vyslal na branku tohoto týmu.

Abychom tyto centrované regresory odlišili od původních necentrovaných, budeme je dále značit pomocí původního značení, za které však bude přidáno `centr`.

Podívejme se nyní na odhady parametrů přímo v summary modelu vytvořeného pomocí funkce `vglm()` – tyto odhady jsou ve sloupci `Estimate`:

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept):1                -4.0454261  1.4396342  -2.810 0.004954 **
(Intercept):2                -2.5930036  1.4047044  -1.846 0.064901 .
(Intercept):3                -1.5983766  1.3940166  -1.147 0.251548
(Intercept):4                -0.7419713  1.4053369  -0.528 0.597522
UvKcentr                      0.5668285  0.1491716   3.800 0.000145 ***
IcetimeZ_sd_Dcentr            0.0106406  0.0072143   1.475 0.140232
SDiff                         0.0632661  0.1965995   0.322 0.747603
I(golyNAD6_D^2)               0.0840114  0.0414302   2.028 0.042583 *
gamescore_IQR                 0.0553728  0.0352344   1.572 0.116054
KB_NAD25_Fcentr              -0.0002131  0.2288783  -0.001 0.999257
UvKcentr:IcetimeZ_sd_Dcentr  -0.0052698  0.0016855  -3.127 0.001769 **
UvKcentr:SDiff               -0.1202151  0.0531158  -2.263 0.023620 *
SDiff:KB_NAD25_Fcentr        -0.1799102  0.0887915  -2.026 0.042743 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vidíme, že máme některé z regresorů na hladině významnosti 0,05 nevýznamné (dle sloupce `Pr(>|z|)`, který obsahuje p-value testu nulovosti příslušného parametru). To je většinou proto, že jejich přínos modelu se projeví až v interakcích s jinými vysvětlujícími proměnnými – tyto interakce už pak významné jsou. Jedinou výjimkou je proměnná gamescore_IQR, která není na stanovené hladině významná a v žádné interakci se neobjevuje. Přesto ji v modelu ponecháme, p-value není příliš vysoká (poměrně blízká zvýšené hladině významnosti 0,1) a vynecháním tohoto regresoru by se AIC modelu zvýšilo.

Pojďme si tedy jednotlivé odhady parametrů β_i interpretovat, abychom vyjádřili jejich efekt na počet vyhraných sérií v play-off.

Obecně bude platit, že pokud je odhad koeficientu u daného regresoru kladný, bude se s rostoucí hodnotou tohoto regresoru (či případné interakce, v summary značená pomocí znaku „:“) zvyšovat logit$[P(\text{VS} \leq j)]$, tím pádem i samotné odds$(\text{VS} \leq j)$, tudíž větší hodnoty budou mluvit v neprospěch velkého počtu vyhraných sérií, naopak snižováním hodnot těchto vysvětlujících proměnných bude dle modelu pravděpodobnější delší setrvání v bojích o Stanley Cup.

U regresorů se záporným odhadem koeficientu to bude přesně obráceně – jejich vysoké hodnoty budou pro úspěch týmu v play-off dobré, nízké hodnoty budou napovídat dřívějšímu vypadnutí týmu.

Níže v textu budou odhady jednotlivých koeficientů kvůli větší přehlednosti zaokrouhleny na 4 desetinná místa. Hodnotu j z množiny $\{0,\ldots,3\}$ budeme považovat za libovolné, ale pevně zvolené číslo z této množiny. Začněme s interpretací pro jednoduchost u koeficientů těch regresorů, které se nevyskytují v žádné interakci.

U mezikvartilového rozpětí ukazatele Gamescore je odhad koeficientu 0,0554. S rostoucí hodnotou tohoto mezikvartilového rozpětí se bude zvyšovat šance $\text{VS} \leq j$, při jednotkovém nárůstu hodnoty regresoru se logaritmus této šance zvýší o 0,0554. Při tomto nárůstu a zachování hodnot všech ostatních regresorů se tedy zmíněná šance daného týmu (tedy že vyhraje nejvýše j sérií) přibližně 1,057krát zvětší. Čím nižší toto mezikvartilové rozpětí bude, tím pravděpodobněji se tým dostane v play-off dál.

Co to tedy znamená? Že pro více vyhraných sérií ve vyřazovací části je lepší mít v základní části menší variabilitu ukazatele Gamescore napříč hokejisty v týmu, tedy mít spíše hráče s malými rozdíly v přínosu každému zápasu, jinými slovy – vyrovnanější mužstvo. Pokud bude v týmu např. perfektní první lajna dosahující v každém zápase velkého Gamescore, ale zároveň bude mít tým skupinu hráčů s velmi špatným celkovým přínosem v zápasech, variabilita (re-

prezentovaná v tomto případě IQR) naroste a jejich šance na úspěch v play-off se sníží. To koresponduje s tím, co se tvrdí v hokejové veřejnosti – pro úspěch týmu v play-off je důležité mít vyrovnaný kádr a dobré výkony od všech 4 formací.

Druhou vysvětlující proměnnou, která se nevyskytuje v interakcích, je počet obránců s alespoň šesti góly, který je však v modelu pouze v kvadrátu. Odhad koeficientu je, pro někoho možná překvapivě, kladný, konkrétně 0,0840. Se zvyšujícím se počtem obránců, kteří v základní části pokoří hranici šesti vstřelených gólů, se zvyšuje logaritmus šance na vítězství v nejvýše j sériích, a to kvadraticky. Konkrétní nárůst šance $\mathrm{VS} \leq j$ tak závisí na tom, ke kterému číslu jednotkový nárůst počtu takových obránců vztahujeme.

Pokud bychom si například vzali poměr šancí dvou týmů, které se shodují ve všech ostatních vysvětlujících proměnných, avšak první tým měl v základní části 4 obránce s alespoň šesti brankami a druhý měl tyto obránce pouze 3, bude šance na nejvýše j vyhraných sérií prvního týmu 1,801krát vyšší než stejná šance u druhého týmu. Pokud by však první tým měl 5 alespoň šestigólových beků a druhý tým 4 (stále rozdíl jednoho obránce), byla by šance na nejvýše j vyhraných sérií prvního týmu už 2,13krát vyšší než tato šance u druhého mužstva.

Je tedy poměrně zajímavé, že větší počet beků, kteří pokoří hranici šesti branek, zvyšuje šanci vypadnout z play-off dříve. Mohlo by to souviset s tím, že pro úspěch v play-off je důležitá precizní obrana, jejíž součástí jsou i defenzivní beci, kteří sice nedají tolik gólů, jejich přínos mužstvu však spočívá v jiných, obranných činnostech. Čím více je tedy v týmu ofenzivních beků, tím horší výsledek v play-off se od mužstva dá čekat (snižuje se tím počet čistě defenzivních obránců), přičemž mezi týmy s jedním a dvěma takovými obránci není ještě efekt na vyhrané série tak velký, ale mezi týmy se čtyřmi a pěti ofenzivními beky je poměr jejich šancí na brzké vypadnutí již velký (viz příklad výše).

Nyní se pustíme do složitějších interpretací, neboť u proměnných obsažených i v interakcích bude třeba zohlednit, pro jaké hodnoty regresorů, s nimiž inter-

agují, bude jejich efekt na vyhrané série roven danému číslu, resp. zda je tento efekt vůbec významný.

Podívejme se na umístění v konferenci. Odhad koeficientu pro tento regresor je roven 0,5668. Kdy však platí, že se při jednotkovém nárůstu umístění – propadu v tabulce o jednu příčku – zvýší logaritmus šance VS $\leq j$ o tuto hodnotu, resp. se zvýší tato šance 1,763krát? Je to pouze tehdy, když je hodnota SDiff nulová (již bylo zmíněno, co tato situace znamená) a směrodatná odchylka ice-time/zápas u obránců je rovna mediánu (její centrovaná hodnota je tedy také nulová). V tomto případě je horší umístění v tabulce spojeno s horším výsledkem v play-off.

Pokud je však IcetimeZ_sd_D větší než medián hodnot této vysvětlující proměnné v datech či roste SDiff (tým své soupeře v průměru přestřílí), snižuje se efekt umístění v konferenci, přičemž může být až zcela opačný – pokud by byla směrodatná odchylka ice-time/zápas u beků výrazně nad mediánem či by tým v průměru své soupeře přestřílel o velký počet střel, potom by horší umístění týmu v tabulce způsobilo naopak zvýšení šance pro počet vyhraných sérií **větší** než j, což je zajímavé.

V podstatě to říká, že čím se tým umístí v tabulce hůř, tím spíš potřebuje své soupeře v utkáních přestřílet a mít raději jednu elitní obrannou dvojici s obrovským ice-time/zápas (z čehož plyne nízký ice-time pro ostatní dvojice a velká směrodatná odchylka), aby měl nižší pravděpodobnost[7] vypadnutí v dřívějších kolech play-off (a dorovnal tak deficit horšího umístění).

Naopak velmi vyrovnané dvojice beků, kterým bude trenér čas na ledě rozdělovat rovnoměrně (s nízkou variabilitou), případně to, že bude tým soupeři v průměru přestřílen, zvyšují šanci brzkého konce v play-off tím víc, čím hůř se tým umístí. Špatné umístění tak zvýrazňuje efekt těchto dvou faktorů (ať už v pozitivním či negativním smyslu), s lepším umístěním se naopak efekt těchto faktorů snižuje, resp. je méně výrazná závislost vlivu umístění na hodnotách zmíněných dvou regresorů interagujících s UvK.

[7] Ze snížení šance jevu VS $\leq j$ plyne také snížení hodnoty P(VS $\leq j$)

Z právě uvedené interpretace by se mohlo zdát, že je pro úspěch týmu v play-off lepší, pokud má elitní obrannou dvojici s velkým časem na ledě na úkor ostatních beků. To však právě záleží na pozici týmu v tabulce. Pokud je tým v konferenci na prvním místě (UvKcentr=0), bude se s rostoucí směrodatnou odchylkou času na ledě/zápas u obránců zvyšovat šance VS $\leq j$, pro větší počet vyhraných sérií by pak pro lídra konference bylo lepší dávat obráncům prostor na ledě co nejrovnoměrněji (a mít tak směrodatnou odchylku jejich času na ledě výrazně pod mediánem).

To však platí jen pro týmy na prvním a druhém místě konference – lídrovi se s nárůstem hodnoty tohoto regresoru o jednotku zvýší zmiňovavný logit o 0,0106, druhému týmu v konferenci (UvKcentr=1) se při každém jednotkovém nárůstu této směrodatné odchylky zvýší hodnota logitu o 0,0054 – pro týmy na těchto pozicích tedy růst variability ve vytěžování obránců znamená snižování pravděpodobnost vysokého počtu vyhraných sérií.

Od třetí pozice dál se však sitauce změní – zvýší-li se IcetimeZ_sd_D o jednotku, nedojde u třetího mužstva v konferenci k téměř žádné změně v hodnotě logitu (zvýší se o 0,0001, tedy velmi nepatrně), na čtvrté a horší pozici pak bude mít variabilita zvýšená o jednotku opačný efekt a sníží pravděpodobnost nejvýše j vyhraných sérií – pro mužstva na těchto pozicích je tak lepší mít ve vytěžování obránců větší variabilitu, u první či druhé příčky v tabulce je výhodnější dát prostor na ledě bekům rovnoměrněji, u týmů na třetím místě má tato vysvětlující proměnná malý vliv (i vzhledem k významnosti samotného koeficientu pro tuto vysvětlující proměnnou v rámci modelu).

Vliv proměnné SDiff velmi závisí na umístění v konferenci a počtu útočníků s alespoň 25 kanadskými body. Pokud by byl tým v tabulce své konference na prvním místě a měl mediánový počet útočníků, kteří stanovenou hranici bodů překonali (medián byl 9 takových útočníků v týmu), prakticky nemůžeme mluvit o významném efektu průměrného střeleckého rozdílu, neboť p-value testu nulovosti

samotného koeficientu je přibližně 0,75, v dané situaci se tedy nedá zamítnout absence jakéhokoliv efektu tohoto regresoru.

Jakmile se však zhoršuje umístění mužstva, roste i efekt průměrného střeleckého rozdílu. Hůře umístěným týmům může velmi zvýšit šanci na počet vyhraných sérií větší než j v případě, že tým přestřílí své soupeře. Stejně tak ale může SDiff špatně umístěným týmům zvýšit šanci na vítězství v nejvýše j sériích, pokud by na branku soupeře vyslal takový tým méně střel, než kolik jich vyprodukuje protivník na opačné straně.

Pokud by např. tým byl v tabulce na 2. místě, bude mít jednotkový nárůst v SDiff za následek snížení logitu v uvažovaném modelu o 0,0569. Pro poslední tým v konferenci už by jednotkový nárůst hodnoty SDiff znamenal snížení logitu dokonce o 0,7782. Nutno zdůraznit, že v těchto příkladech musíme uvažovat mediánový počet útočníků s 25+ kanadskými body (a tedy vypadnutí jedné z interakcí).

Pokud bude počet takto produktivních forwardů v týmu větší než medián, bude se snižovat pravděpodobnost brzkého vypadnutí z play-off při nárůstu SDiff ještě více než v příkladech uvedených v předchozím odstavci (a dokonce bude jednotkový nárůst snižovat šanci VS $\leq j$ i u lídrů konference). To dává smysl, pokud tým s velkým počtem produktivních útočníků zároveň vyprodukuje mnoho střel a soupeře přestřílí, mělo by to o to víc zvýšit jeho šanci na více vyhraných sérií.

Pokud by však byl počet takových útočníků nižší než medián, rostoucí počet SDiff by měl při fixním umístění v konferenci opačný vliv a zvyšoval by tedy šanci na výsledný počet vyhraných sérií nejvýše j. Pokud tedy tým nedisponuje velkým množstvím produktivních útočníků, je pro výsledek ve vyřazovací fázi lepší, pokud soupeře během sezóny tolik nepřestřílí. Neznamená to však, že by takový tým měl nechat svého gólmana v každém zápase základní části zasypat střelami – při velmi záporném SDiff by mohl převážit interakční člen UvK $\cdot$ SDiff, čímž by se ve výsledku celková pravděpodobnost nejvýše j vyhraných sérií zvýšila.

Interpretovat samotný počet forwardů s alespoň 25 kanadskými body při nulovém SDiff by nemělo smysl, neboť je příslušný koeficient zcela nevýznamný (s p-value testu nulovosti 0,999). Význam tak má tato vysvětlující proměnná pouze při situacích s kladným či záporným rozdílem střel na zápas.

Celkově tedy vidíme, že u některých proměnných je směr jejich efektu zcela jasně daný, u jiných si však musíme dát velký pozor na to, že jejich efekt ovlivňují hodnoty jiných proměnných. Rostoucí hodnota určitého ukazatele tak může mít jak pozitivní, tak negativní efekt, a to právě vlivem různých hodnot s ním interagujících regresorů.

Pokud bychom měli dle známých hodnot regresorů říct pro každý tým to, kolik vyhraných sérií je pro něj nejpravděpodobnější, zvolíme tu hodnotu VS, pro niž je na základě modelu pravděpodobnost nejvyšší.

Tabulka 4.13: Confusion matrix pro predikci počtu vyhraných sérií

	Predikce				
Skutečnost	0	1	2	3	4
0	42	4	0	0	2
1	12	10	0	0	2
2	6	5	0	0	1
3	1	5	0	0	0
4	2	2	2	0	0

Vidíme, že např. vítěze Stanley Cupu bychom „neuhodli“ ani jednou, model by dokonce po základní části nikoho neurčil poraženým finalistou (VS = 3 není predikováno ani jednou). Odhadnuté pravděpodobnosti samozřejmě nereflektují zcela přesně systém play-off. Jinými slovy, model neví, že za 6 sezón musí nutně predikovat 6 vítězů Stanley Cupu, 6 poražených finalistů apod.

Tady však bude dobré si uvědomit rozdíl mezi jakousi absolutní predikcí (počet vyhraných sérií, na který bychom si dle modelu měli, dejme tomu, vsadit) a predikcí vyjádřenou pravděpodobnostmi. V tom je docela zásadní rozdíl. To, že

zvolený model dle rozhodovacího kritéria „Zvol pro daný tým nejpravděpodobnější počet VS“ neurčí danému týmu správný počet vyhraných sérií v play-off, ještě neznamená, že stanovuje odhad nutně špatně. Pouze se na základě odhadnutých vztahů a hodnot regresorů daného týmu zdál být pravděpodobnější jiný výsledek, než který zrovna nastal – ostatní výsledky však byly též pravděpodobné, jenom méně.

Model žádnému týmu nepřiřadil P(VS = 3)=0, tuto možnost tedy nijak principiálně nevylučoval. Jen zkrátka pravděpodobnost pro výsledek „Poražen ve finále“ nebyla u žádného z týmu tou nejdominantnější (u několika mužstev je sice velmi blízká nule, u některých se však pohybuje kolem 20 %).

K posouzení predikční úrovně tohoto modelu bychom tak v případě cíle odhadnout co nejlépe dané pravděpodobnosti museli použít jiné než standardní metody[8], a to takové, které slouží k posouzení kvality pravděpodobnostní predikce a z dlouhodobého hlediska i míry kalibrace, např. Brierovo skóre (kterému se tu však více věnovat nebudu, základní údaje o tomto skóre lze nalézt např. na stránce [13]).

Model navíc vůbec nebere v potaz to, co se děje v samotném play-off. Začne tým hrát jinak a bude mít jiné hodnoty ukazatelů, podle kterých jsme predikovali výsledek po základní části? Bude se tým pokládat za jednoho z favoritů, ale hned v prvním kole narazí na favorita ještě většího? Další věcí by byla úprava predikce tak, aby vše sedělo na systém play-off a pravidla pro nasazování týmů do pavouka vyřazovací části. Tyto spojitosti by musely být předmětem další, důkladnější analýzy, která však již nebude obsahem této práce.

[8] Jako by zde bylo např. procento správně predikovaných počtů vyhraných sérií.

Závěr

Cílem práce bylo zjistit, do jaké míry souvisejí výsledky týmů NHL v základní části s výkony ve vyřazovací části. Toho jsme chtěli docílit za pomoci rozsáhlých datových sad a konkrétních statistických metod.

První dvě kapitoly uvedly čtenáře do problematiky systému play-off NHL a také do světa hokejových statistik, na jejichž sledování a analýzu se klade čím dál větší důraz. Na některé ukazatele bylo zaměřeno více pozornosti, pokud byl jejich význam méně zřejmý a pokud bylo cenné si tento význam přiblížit.

Statistiky popisované v druhé kapitole souvisely s daty a jejich analýzou – cílem bylo zjistit, zda existuje vztah mezi různými ukazateli reflektujícími hru týmu v základní části a počtem vyhraných sérií v play-off, který je zase měřítkem úspěchu týmu v rámci celé sezóny. V tomto smyslu tak byla tato kapitola nutnou podmínkou k úspěšné analýze, neboť bez znalosti základních pojmů a spojitostí by názorná interpretace výsledků byla velmi složitá.

Ve třetí kapitole byly představeny statistické metody, o které jsme se mohli opřít při samotné analýze. V této teoretické části se čtenář mohl seznámit s některými metodami pro posouzení vztahu dvou kvalitativních znaků, a to jak při nominálních kategoriích (χ^2 test nezávislosti), tak v případě existence přirozeného uspořádaní kategorií těchto kvalitativních znaků (metody pro ordinální data).

Dále jsou v rámci této kapitoly popsány základní testy vztahu kvalitativního a kvantitativního znaku – ANOVA a Kruskalův–Wallisův test. Volba mezi těmito metodami spočívá v ověření předpokladů pro jejich použití. Pokud jsou splněny předpoklady pro testování pomocí ANOVA, je lepší využít tuto metodu. U obou testů jsou popsány také postupy pro mnohonásobné porovnávání, které se používa-

jí při zamítnutí nulové hypotézy daného testu. Při analýze tato situace nicméně nenastala, proto tyto metody nebyly využity.

Stěžejním tématem práce je model ordinální logistické regrese, přesněji model proporcionálních šancí, který je speciálním typem této regrese a na který klademe předpoklad proporcionality šancí. V poslední části třetí kapitoly je tak popsán tento model a základní pojmy s ním související. Práce se věnuje také interpretaci parametrů a vysvětlení zmíněného předpokladu tohoto modelu. Kromě teoretického zavedení jsou uvedeny i jednoduché příklady pro lepší pochopení problematiky. Závěr této části se věnuje i testování předpokladu a odhadu parametrů pomocí softwaru R.

Poslední kapitola je věnována analýze datových sad, která měla pomoci se zodpovězením otázek položených v úvodu a prvních dvou kapitolách. Pomocí metod pro ordinální data jsme zjistili, že mezi umístěním v konferenci a počtem vyhraných sérií je statisticky významný vztah, konkrétně mají týmy s lepším umístěním tendenci se v play-off dostat dál než týmy umístěné hůře, což svým způsobem není příliš překvapivé a potvrzuje to relevanci vnímání favoritů pro play-off na základě pozice v tabulce.

Při porovnávání hodnot týmových či hráčských ukazatelů pro jednotlivé kategorie dané počtem vyhraných sérií jsme došli k závěru, že na hladině významnosti 0,05 se hodnoty kteréhokoliv z ukazatelů neliší napříč jednotlivými kategoriemi. Konkrétní podoba porovnávání hodnot se pak lišila dle splnění předpokladů (a tedy použitého testu) – šlo vždy buď o ověření shody středních hodnot, nebo ověření rovnosti distribučních funkcí rozdělení daného ukazatele, případně jsme zkoumali neexistenci pozitivního či negativního trendu v datech ve smyslu převahy konkordantních a diskordantních dvojic pozorování. V některých případech mohlo k nezamítnutí nulové hypotézy dojít z důvodu relativně malého počtu pozorování (zejména ve skupinách poražených finalistů a vítězů Stanley Cupu), uvažování většího počtu sezón však bylo zavrhnuto vzhledem k dynamice vývoje hry v ledním hokeji, organizačním změnám a také výluce v sezóně 2012/13 a pandemii v ročníku 2019/20.

Pomocí údajů ze základní části jsme pak byli schopni vysvětlit počet vyhraných sérií v rámci modelu proporcionálních šancí. Některé statistiky se ukázaly být z hlediska modelu významné, a to buď samy o sobě, nebo v rámci vzájemných interakcí. Předpoklady modelu byly ověřeny pomocí Brantova testu, použité vysvětlující proměnné proporcionalitu šancí neporušují. Klíčová je pak interpretace modelu a popis efektu jednotlivých faktorů – ty představují hlavní výstup analýzy a přibližují vliv některých údajů ze základní části na délku doby setrvání týmu v play-off.

Pomocí získaného modelu také můžeme pro každý tým odhadnout pravděpodobnost jednotlivých výsledků ve vyřazovacích bojích. Pokud na tuto predikci pohlížíme jako na pravděpodobnostní (tedy nepožadujeme jako výstup jednu hodnotu, ale pravděpodobnostní ohodnocení jednotlivých možností), mohlo by být posouzení schopnosti predikce tohoto modelu v uvažovaných datech (a v případném dlouhodobém horizontu pro další sezóny) předmětem dalších analýz, které by využily některé z poznatků této práce. Model však nepočítá s pevně danou strukturou výstupů, resp. jasně daným počtem týmů, které se v jednotlivých kategoriích (daných počtem vyhraných sérií v play-off) musí nutně realizovat – i to by mohlo být dalším ze způsobů vylepšení výstupů tohoto modelu.

Seznam literatury

[1] AGRESTI, A.: *Analysis of Ordinal Categorical Data*. 2nd ed. Hoboken, New Jersey: A John Wiley & Sons, Inc., 2010. ISBN 978-0-470-08289-8.

[2] BARTLETT, M. S.: *Properties of sufficiency and statistical tests*. Proc. R. Soc. Lond., A 160: s. 268–282, 1937. Dostupné z: http://doi.org/10.1098/rspa.1937.0109

[3] BRANT, R.: *Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression*. Biometrics, 46(4), s. 1171–1178, 1990. Dostupné z: www.jstor.org/stable/2532457

[4] DUNN, O. J.: *Multiple Comparisons Using Rank Sums*, Technometrics, 6:3, s. 241-252, 1990. doi: 10.1080/00401706.1964.10490181

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

[6] GOODMAN, L., KRUSKAL, W.: *Measures of Association for Cross Classifications III: Approximate Sampling Theory*. Journal of the American Statistical Association, 58(302), s. 310–364, 1963. Dostupné z: www.jstor.org/stable/2283271

[7] HRON, K., KUNDEROVÁ, 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.

[8] KENDALL, M. G.: *Rank correlation methods*. 2nd ed. London: Charles Griffin, 1955.

[9] ROSOĽANKA, M.: *Analýza ordinálních dat* [online]. Olomouc, 2016 [cit. 2021-04-18]. Dostupné z: https://theses.cz/id/5ii4ah/. Diplomová práce. Univerzita Palackého v Olomouci, Přírodovědecká fakulta.

[10] ŘEHÁKOVÁ, B.: *Introducing Logistic Regression*. Sociologický časopis / Czech Sociological Review, 36(4), s. 475–492, 2000. doi: 10.13060/00380288.2000.36.4.06

Internetové zdroje

[11] *2019 Stanley Cup playoffs*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2020-09-02]. Dostupné z: https://en.wikipedia.org/wiki/2019_Stanley_Cup_playoffs

[12] *Blueshirts Breakaway: Lexicon* [online]. [cit. 2020-07-19]. Dostupné z: https://www.blueshirtsbreakaway.com/hockey-lexicon/

[13] *Brier score*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2020-04-20]. Dostupné z: https://en.wikipedia.org/wiki/Brier_score

[14] FITZPATRICK, J.: *How Does the NHL Break a Tie in the NHL Standings?*. Liveabout.com [online]. [cit. 2020-07-19]. 2017. Dostupné z: https://www.liveabout.com/breaking-tie-in-the-nhl-standings-2778920

[15] FOLEY, M.: *My Data Science Notes* [online]. [cit. 2021-04-19]. 2020. Dostupné z: https://bookdown.org/mpfoley1973/data-sci/

[16] *History of organizational changes in the NHL*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2020-09-02]. Dostupné z: https://en.wikipedia.org/wiki/History_of_organizational_changes_in_the_NHL

[17] *Hockey Reference* [online]. © 2000-2021 [cit. 2021-03-30]. Dostupné z: https://www.hockey-reference.com/about/advanced_stats.html

[18] JONES, W.: *NHL roster size: How many players are on a hockey team?*. Hockey Answered [online]. [cit. 2021-03-27]. Dostupné z: https://hockeyanswered.com/nhl-roster-size-an-easy-to-follow-guide/

[19] MANGIAFICO, S. S.: *Association Tests for Ordinal Tables*. Summary and Analysis of Extension Program Evaluation in R [online]. [cit. 2021-01-28]. Dostupné z: https://rcompanion.org/handbook/H_09.html

[20] McKENZIE, B.: *McKenzie: The real story of how Corsi got its name*. TSN.ca [online]. [cit. 2021-03-12]. 2014. Dostupné z: https://www.tsn.ca/mckenzie-the-real-story-of-how-corsi-got-its-name-1.100011

[21] MELENDEZ, A.: *Multiple St. Louis Blues Franchise Records Power Team To Playoffs*. Last Word On Sport [online]. [cit. 2021-03-29]. 2019. Dostupné z: https://lastwordonsports.com/hockey/2019/04/05/st-louis-blues-franchise-records/

[22] MoneyPuck, Data. Skaters 2013–14 až 2018–19 [dataset]. [cit. 2020-07-23]. Dostupné z: http://moneypuck.com/data.htm

[23] *National Hockey League: Slovník pojmů* [online]. © NHL 2018 [cit. 2021-03-12]. Dostupné z: http://www.nhl.com/stats/cs/glossary

[24] *National Hockey League: Stanley Cup Playoffs format, qualification system* [online]. [cit. 2020-09-02]. 2019. Dostupné z: https://www.nhl.com/news/stanley-cup-playoffs-format-qualification-system/c-711015

[25] *National Hockey League: Statistiky* [dataset]. © NHL 2018 [cit. 2021-01-30]. Dostupné z: http://www.nhl.com/stats/cs/teams

[26] *National Hockey League: Tabulky 2018–2019* [online]. © NHL 2018 [cit. 2020-09-02]. Dostupné z: https://www.nhl.com/cs/standings/2018/league

[27] ROSEN, D.: „NHL introduces new division names with schedule“. National Hockey League [online]. [cit. 2020-09-02]. 2013. Dostupné z: https://www.nhl.com/news/nhl-introduces-new-division-names-with-schedule/c-678456

[28] *Stanley Cup playoffs*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2020-09-02]. Dostupné z: https://en.wikipedia.org/wiki/Stanley_Cup_playoffs

[29] WAY, D.: „Hockey Lexicon Spotlight Series – Game Score“. Blueshirts Breakaway [online]. [cit. 2020-07-19]. 2018. Dostupné z: https://www.blueshirtsbreakaway.com/2018/03/29/hockey-lexicon-spotlight-series-game-score/

Balíčky softwaru R

[30] Andri Signorell et mult. al. (2021). DescTools: Tools for descriptive statistics. R package version 0.99.40.

[31] Benjamin Schlegel and Marco Steenbergen (2020). brant: Test for Parallel Regression Assumption. R package version 0.3-0. https://CRAN.R-project.org/package=brant

[32] Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a class of permutation tests: The coin package." Journal of Statistical Software, *28*(8), 1-23. doi: 10.18637/jss.v028.i08 (URL:https://doi.org/10.18637/jss.v028.i08).

[33] Pohlert T (2014). The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR). R package, URL:https://CRAN.R-project.org/package=PMCMR.

[34] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[35] Thomas W. Yee (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

[36] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0