VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

SABERMETRICS - STATISTICKÁ ANALÝZA VÝKONŮ BASEBALLOVÝCH HRÁČŮ
SABERMETRICS - BASEBALL STATISTICS THAT MEASURE IN-GAME ACTIVITY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Martin Groman**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Josef Bednář, Ph.D.**
SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání bakalářské práce

Ústav:	Ústav matematiky
Student:	**Martin Groman**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**Ing. Josef Bednář, Ph.D.**
Akademický rok:	2016/17

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Sabermetrics – statistická analýza výkonů baseballových hráčů

Stručná charakteristika problematiky úkolu:

Jedná se o statistickou analýzu výkonů baseballových hráčů. Zabývá se různými herními statistikami, za účelem předpovědi výkonu hráče, stanovení hodnoty na hráčském trhu či jeho skutečným přínosem pro tým.
Důraz bude kladen především na matematické modely založené na Markovových řetězcích a predikci.

Cíle bakalářské práce:

1) Rešerše v oblasti Sabermetrics.
2) Popis matematických modelů výkonů baseballových hráčů.
3) Aplikace modelů na konkrétní data.

Seznam doporučené literatury:

ANDĚL, Jiří. Základy matematické statistiky. Vyd. 3. Praha: Matfyzpress, 2011, 358 s. : grafy, tab. ISBN 978-80-7378-162-0.

PRÁŠKOVÁ, Zuzana. Základy náhodných procesů I. Vydání druhé, v Matfyzpressu první vydání. Praha: Matfyzpress, 2012, 158 s. ISBN 978-80-7378-210-8.

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2016/17

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt

Tato bakalářská práce se zabývá statistickou analýzou baseballových hráčů, jejichž výkony predikujeme pomocí statistických modelů. Z matematického aparátu využijeme Markovovy řetězce, indexovou analýzu a lineární regresi. Výsledkem mé práce bude porovnání predikovaných hodnot s realitou.

Abstract

This bachelor thesis is dealing with statistical analysis of baseball players, whose performances are predicted with statistical models. We will use some mathematical terms, such as Markov chains, index analysis and linear regression. The outcome of this thesis will be comparison between predicted and real values.

klíčová slova

baseball, statistický model, predikce, Markovovy řetězce, indexová analýza, lineární regrese

keywords

baseball, statistical model, prediction, Markov chains, index analysis, linear regression

Groman, M.: *Sabermetrics - statistická analýza výkonů baseballových hráčů*, Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2017. xx s. Vedoucí bakalářské práce Ing. Josef Bednář, Ph.D.

Prohlašuji, že jsem bakalářskou práci *Sabermetrics - statistická analýza výkonů baseballových hráčů* vypracoval samostatně pod vedením Ing. Josefa Bednáře, Ph.D. s použitím materiálů uvedených v seznamu literatury.

Martin Groman

Chtěl bych poděkovat panu Ing. Josefu Bednářovi, Ph.D. za odborné vedení, konzultaci a cenné rady při psaní mé práce a za pomoc při konečných úpravách.

Martin Groman

Obsah

1 Úvod

Čísla, čísla a zase čísla. Spousta lidí si pod pojmem statistika nepředstaví nic jiného. Přesvědčení většiny populace vystihuje známá česká pohádka, ve které se zpívá: "Statistika nuda je, má však cenné údaje, neklesejte na mysli, ona nám to vyčíslí." Nicméně i autoři písně uznávají, že statistika je důležitým vědním oborem, ačkoliv mnohým lidem může připadat nezáživná.

Existují ovšem oblasti statistiky, které vzbuzují vzrušení kdykoliv na ně dojde řeč a na vrcholku pomyslného žebříčku těchto oblastí se vyjímá sportovní statistika. Nepsanou povinností každého sportovního fanouška je znát alespoň základní statistiky svého oblíbeného týmu či sportovce, ať už se jedná o počet nastřílených gólů, získaných bodů anebo počet vyhraných turnajů. Z těchto a mnoha dalších pozorovaných sportovních statistik lze pomocí matematických nástrojů, mimo jiné, určit i budoucí rozvoj či výkon daného týmu nebo jednotlivce.

Jakou taktiku zvolit na příští zápas? Vyplatí se koupit tohoto hráče? Jak velkou šanci máme na zisk titulu? Kolik bodů uhrajeme tuto sezónu? Všechny tyto otázky mohou být do jisté míry zodpovězeny pomocí statistiky. V současném sportu je vyvinuto obrovské úsilí za účelem předpovídání celkových výsledků. Takovým sportem je i baseball, ve kterém výsledkové předpovědi byly dotaženy téměř k dokonalosti.

Díky ohromnému množství zaznamenaných dat se americký baseball jeví jako ideální kandidát k vytvoření predikčního modelu. Sport s více než stoletou tradicí se v USA těší daleko větší přízně než u nás v Česku a proto není divu, že nemálo matematiků, ale i nematematiků, se snaží předpovědět výkony hráčů a týmů. Kdo vyhraje tuto sezónu? Pravděpodobně ta nejlákavější hádanka pro každého fanouška. V této práci se pokusím zmíněnou hádanku rozluštit.

2 Seznámení s problematikou

V této části práce nejprve zadefinuji pojem Sabermetrics a porovnám baseball s ostatními sporty ze statistického hlediska. Dále popíšu jedny z nejdůležitější současných sabermetrických statistik a uvedu jejich využití.

2.1 Sabermetrics

Bill James v roce 1980, na počet Amerického spolku pro baseballový výzkum (angl. SABR), definoval Sabermetrics jako "hledání objektivního baseballového poznání".[1] Nástrojem k hledání našeho poznání je statistická analýza herních statistik, ať už týmových nebo individuálních. Pomocí této analýzy jsme schopni rozebírat výkony týmů či hráčů v odehraných utkáních, stejně jako do jisté míry předpovídat jejich výkony v utkáních budoucích.

Zamysleme se nyní nad tím, proč je statistická analýza v baseballu tak mocnou zbraní ve srovnání s ostatními sporty. Přestože je baseball týmovým sportem, můžeme pomocí statistiky vyhodnotit a porovnávat výkony jednotlivců, a to díky tomu, že jejich počínání na hřišti je většinou nezávislé na chování jejich spoluhráčů. Uvedu příklad – hráč, který odpálí homerun, to dokázal sám, bez jakékoliv pomoci spoluhráčů. Vezmeme-li nyní např. situaci z fotbalu – hráč vstřelí gól. Není to čistě jen jeho zásluha, ale i zásluha spoluhráčů. První mu přihrál, druhý celou akci založil, třetí clonil brankáři a čtvrtý blokoval protihráče. Statisticky "férovější" je v tomto případě baseball. [2]

Dalším rozdílem mezi baseballem a ostatními sporty je fakt, že o baseballu je možno uvažovat jako o sérii soubojů 2 hráčů - nadhazovač a pálkař, zatímco v ostatních sportech i další hráči hrají rovněž důležitou roli ve hře. Logicky ani v baseballu nelze úplně zanedbat přínoš spoluhráčů. ale ve výsledku jsou to právě souboje na pálce, které většinou rozhodují zápas. [3]

2.1.1 Pálkařské statistiky

Hlavním cílem je najít statistiku, která by popisovala pálkaře se všemi jeho kvalitami, což, jak se posledních několik desítek let ukazuje, není tak jednoduché. Moderní Sabermetrics si nevystačí pouze s klasickými statistikami jako např. počet home-runů, doběhů, či ukradených met. Zaměřme se na komplexnější statistiky, které nám o kvalitě hráče napoví něco víc. [4]
Poznámka. Seznam použitých zkratek je uveden na konci této kapitoly.

wOBA (weighted on-base average) - popisuje přínos pálkaře na celkovém počtu doběhů svého týmu. Myšlenka za jejím vznikem byla jasná - ne všechny odpaly jsou stejně hodnotné a proto je zapotřebí statistiky, která přiřadí jednotlivým odpalům váhy, abychom mohli pálkaře objektivně hodnotit. Váhy se přepočítávají každý rok, podle celkového počtu doběhů za daný rok. Definujeme tuto statistiku

$$wOBA = \frac{0,69 \cdot uBB + 0,72 \cdot HBP + 0,89 \cdot 1B + 1,27 \cdot 2B + 1,62 \cdot 3B + 2,10 \cdot HR}{AB + BB - IBB + SF + HBP}.[5]$$

BABIP (batting average on balls in play) - popisuje jak často odpálený míček skončí ve hře, čili dojde k odpalu, ale zároveň se nejedná o home-run nebo nedojde k outu. Tuto statistiku ovlivňují 3 faktory: talent pálkaře, kvalita obrany soupeře a štěstí. Lze tedy

pozorovat, že pokud má hráč stále vysokou hodnotu BABIP, jedná se o kvalitního hráče. Pokud dojde k výkyvům této statistiky, můžeme říci, že pálkař měl buď smůlu nebo hrál proti kvalitním soupeřům a tedy

$$BABIP = \frac{H - HR}{AB - K - HR + SF}. \quad [6]$$

WAR (wins above replacement) - porovnává přínos hráče k celkovému počtu výher za sezónu vzhledem k hráči z nižší ligy. V současné době se jedná o nejuniverzálnější dostupnou statistiku. Jednoduše se dá vysvětlit takto - nejlepší hráč týmu s hodnotou WAR=+6 se zraní a na jeho místo hypoteticky nasadím hráče z nižší ligy. Pak mohu očekávat, že s novým hráčem mužstvo za sezónu vyhraje v průměru o 6 zápasů méně než by vyhrálo se zraněným hráčem. Je definována

$$WAR = \frac{BR + BRR + FR + PAd + LgAd + RR}{RPW}. \quad [7]$$

RE24 (Run expectancy based on the 24 base-out states) - uvádí průměrný počet doběhů, který skóruje průměrný tým ve zbytku směny v závislosti na počtu outů a rozmístění běžců na metách. Pro výpočet RE24 se používá tabulka očekávaných doběhů, která zahrnuje všech 24 kombinací počtu outů a rozmístění běžců na metách. Tato matice se mění každý rok, podle počtu doběhů za celou sezónu. Mějmě například tabulku

Obsazené mety	1 out	2 outy	3 outy
_ _ _	0.461	0.243	0.095
1 _ _	0.831	0.489	0.214
_ 2 _	1.068	0.644	0.305
1 2 _	1.373	0.908	0.343
_ _ 3	1.426	0.865	0.413
1 _ 3	1.798	1.140	0.471
_ 2 3	1.920	1.352	0.570
1 2 3	2.282	1.520	0.736

Tabulka 1 – *Příklad RE24*

RE24 pak vypočteme podle vztahu

$$RE24 = RE\,(2.stav) - RE(1.stav) + R$$

Dejme tomu, že nás zajímá hodnota RE24 po prvním odpalu zápasu, kdy pálkař odpálí double. Hodnota počáteního stavu RE=0,461 a hodnota konečného stavu RE=1,068, počet doběhů R=0. Po dosazení

$$RE24 = 1,068 - 0,461 + 0 = 0,607$$

Můžeme tedy očekávat, že tým bude mít o 0,607 více doběhů než je průměrný počet doběhů v dané situaci. [8]

2.1.2 Nadhazovačské statistiky

Sabermetrics se ovšem nezabývá čistě jen ofenzivním počináním hráčů, ale rovněž i defenzivním. Nás budou zajímat takové statistiky, které souvisí s počtem doběhů, který jeden tým povolí skórovat druhému.
Poznámka. Seznam použitých zkratek je uveden na konci této kapitoly.

ERA (Earned run average) - jedná se o základní metriku vytvořenou k zhodnocení výkonů nadhazovačů. Popisuje jejich schopnost zabraňovat doběhům soupeře

$$ERA = \frac{ER}{IP} \cdot 9.$$

ERA se těší velké popularitě, protože se může zdát, že zodpovídá důležitou otázku, a to, kolik doběhů soupeřova týmu je čistě chyba nadhazovače? Zdání však klame a ERA není tak dokonalá statistika, jak na první pohled vypadá. Je důležité uvědomit si, že zapisovatel výsledků rozhoduje, co byla chyba a co ne, tedy hraje zde velkou roli míra subjektivity. Navíc, chyby vytvořené obranou často nejsou ani správně popsány v pravidlech a poškozují statistiky nadhazovačů.

Pokud nás tedy zajímá čistě výkon nadhazovače, neměli bychom statistiku ERA používat jako vševypovídající. Doporučuje se ji používat v kombinaci s podobnými metrikami (FIP, xFIP, RA9). [9]

WHIP (Walks plus hits per innings pitched) - vyjadřuje počet běžců, kterým nadhazovač povolí se dostat na metu, tedy

$$WHIP = \frac{W + H}{IP}.$$

Jak lze vidět, jedná se opravdu o jednoduchou statistiku, která nám poskytuje prvotní náhled na kvality daného nadhazovače. Nadhazovač musí zabránit soupeři v doběžích, čili zabránit mu dostat se na metu.

Stejně jako u předchozí statistiky nesmíme zanedbat význam obrany. Je jednou z mnoha nadhazovačských statistik, které popisují, co se ve hře odehrálo, když specifický nadhazovač nadhazoval, nejedná se čistě o míru jeho jedinečných vlastností. [10]

LOB% (Left on base percentage) - vyjadřuje v procentech kolik běžců zůstane na konci směny na metách díky nadhazovači. Je definována jako

$$LOB\% = \frac{H + BB + HBP - R}{H + BB + HBP - 1,4 \cdot HR}.$$

Většina nadhazovačů se pohybuje okolo hodnoty ligového průměru 70-72% a ti, kteří se mírně odchylují od průměru mají tendenci se k němu v budoucnu vracet. Ukázalo se, že nadhazovači, kteří nesklouzávají k průměru, mají vyšší počet strikeoutů než ostatní. Nejedná se o zvlášť překvapivé zjištění, neboť je zřejmé, že takoví nadhazovači nemusí spoléhat na obranu, takže jsou schopni udržovat svou hodnotu LOB% vysokou. [11]

2.2 Používané pojmy a zkratky

Zde uvádím seznam pojmů a zkratek a jejich vysvětlení. Rozhodl jsem se používat originální anglické pojmy, které jsou používány všude ve světě. [7],[12],[13]

Zkratka	Název	Popis
1B	Single	Počet odpálených singlů. Pálkař po odpalu doběhne na první metu.
2B	Double	Počet odpálených doublů. Pálkař po odpalu doběhne na druhou metu.
3B	Triple	Počet odpálených triplů. Pálkař po odpalu doběhne na třetí metu.
AB	At bats	Počet nadhozů absolvovaných pálkařem, které neskončí volnou metou, trefením pálkaře nebo obětováním odpalu
BB (W)	Walks	Počet volných met.
uBB	Unintentional walks	Počet neúmyslně darovaných volných met.
BR	Batting runs	Počet doběhů nad nebo pod průměrem přidaných jako pálkař.
BRR	Base running runs	Počet doběhů nad nebo pod průměrem přidaných jako běžec.
ER	Earned runs	počet získaných doběhů, které nevznikly chybou obrany
FR	Fielding runs	Počet doběhů nad nebo pod průměrem přidaných jako hráč v poli.
FIP	Fielding independent pitching	Odhad nadhazovačovy ERA na základě strikeoutů, darovaných met a povolených homerunů.
xFIP	Expected fielding ind. pitching	Odhad nadhazovačovy ERA na základě strikeoutů, darovaných met a povolených vysokých odpalů.
Fielding%	Fielding percentage	Udává v procentech bezchybnost bránících hráčů
GDP(DP)	Grounded into double play	Počet tzv. dvojitých outů.
H	Hits	Počet odpalů
HBP	Hit by pitches	Počet nadhozů, které trefily pálkaře.
HR	Homerun	Počet homerunů - míčků odpálených mimo hrací plochu.
HR/9	Homeruns per 9 innings pitched	Počet odpálených homerunů za 9 směn.
IBB	Intentional walks	Počet úmyslných darovaných met.
IP	Innings pitched	Počet nadhazovaných směn.
K	Strikeout	Počet strikeoutů. Pálkař neodpálí 3 dobré nadhozy.
LgAd	League adjustment	Korekční koeficient. Určuje se na základě ligy.

PAd	Positional adjustment	Empirický odhad počtu doběhů nad nebo pod průměrem přidaných jako hráč v poli. Záleží na pozici v obraně.
RA	Runs allowed	Počet povolených doběhů.
RS	Runs scored	Počet skórovaných doběhů.
RA9	Runs allowed per 9 innings pitched	Počet povolených doběhů za 9 směn.
RR	Replacement runs	Rozdíl doběhů mezi uvažovaným hráčem a hráčem "náhradním"(uvažuje se hráč z nižší ligy).
RPG	Runs per game	Počet doběhů za zápas.
RPW	Runs per win	Počet doběhů potřebných k výhře. Každý rok se tato hodnota mění, je závislá na součtu všech doběhů za sezónu.
SF	Sacrifice flies	Počet obětovaných odpalů. Jsou zahrávany, aby umožnily doběh spoluhráči.

Tabulka 2 – *Seznam zkratek a pojmů*

3 Matematická teorie

3.1 Regresní analýza

V důsledku všeobecné známosti tohoto tématu, jej nebudu více rozepisovat. Veškeré zmíněné pojmy jsou popsány v knize Základy matematické statistiky Jiřího Anděla. [14]

3.2 Markovovy řetězce

3.2.1 Náhodný proces

Definice 3.1. Nechť *(Ω, $\mathcal{A}$, P)* je pravděpodobnostní prostor, nechť $T \subset \mathbb{R}$. Rodina reálných náhodných veličin $\{X_t,\ t \in T\}$ definovaných na *(Ω, $\mathcal{A}$, P)* se nazývá *náhodný proces*.

V případě že $T = \mathbb{Z} = \{0, \pm1, \pm2,...\}$ nebo $T = \mathbb{N}_0 = \{0, 1,...\}$, mluvíme o *procesu s diskrétním časem* nebo o *časové řadě*. Pokud $T = [a,b]$, kde $-\infty \leq a < b \leq \infty$, říkáme, že $\{X_t,\ t \in T\}$ je *proces se spojitým časem*.

Dvojice $(S, \mathcal{E})$, kde S je množina náhodných veličin X_t a $\mathcal{E}$ je σ–algebra podmnožin S, se nazývá *stavový prostor* procesu $\{X_t,\ t \in T\}$. Pokud náhodné veličiny X_t nabývájí pouze diskrétních hodnot, říkáme, že jde o *proces s diskrétními stavy*, nabývají-li hodnot z nějakého intervalu, mluvíme o *procesu se spojitými stavy*. [15]

3.2.2 Základní vlastnosti

Mějme pravděpodobnostní prostor *(Ω, $\mathcal{A}$, P)* a uvažujme na něm posloupnost náhodných veličin $\{X_n,\ n \in N_0\}$, které nabývájí pouze celočíselných hodnot. Nechť S je množina celých čísel i takových, že $i \in S$ právě tehdy, když existuje $n \in N_0$ tak, že $P(X_n = i) > 0$. Množina S může být buď konečná nebo spočetně nekonečná. Budeme ji říkat *množina stavů* náhodného procesu $\{X_n,\ n \in N_0\}$ a její prvky budeme nazývat *stavy*. Bez omezení na obecnosti můžeme předpokládat, že $S = \{0, 1..., N\}$ nebo $S = \{0, 1,...\}$.

Definice 3.2. Posloupnost celočíselných náhodných veličin $\{X_n,\ n \in N_0\}$ se nazývá *Markovův řetězec s diskrétním časem* a množinou stavů S, jestliže

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{n+1} = j | X_n = i) \tag{3.1}$$

pro všechna $n = 0, 1,...$ a všechna $i, j, i_{n-1},...,i_0 \in S$ taková, že $P(X_n = i, X_{n-1} = i_{n-1},..., X_0 = i_0) > 0$.

Vztah (3.1) vyjadřuje *markovskou vlastnost*; znamená, že pravděpodobnost výsledku v budoucím čase $n + 1$, známe-li výsledek v přítomném čase n a výsledky z minulých časů $n - 1$, $n - 2$,..., 0, je stejná, jako když známe jen výsledek v přítomném čase.

Podmíněné pravděpodobnosti

$$P(X_{n+1} = j | X_n = i) = p_{ij}(n, n+1)$$

(pokud jsou definovány) se nazývají *pravděpodobnosti přechodu* ze stavu i v čase n do stavu j v čase $n + 1$, někdy též *pravděpodobnosti přechodu 1. řádu*. [15]

Podobně podmíněné pravděpodobnosti

$$P(X_{n+m} = j|X_n = i) = p_{ij}(n,\ n+m)$$

pro přirozené $m \geq 1$ se nazývají pravděpodobnostmi přechodu ze stavu i v čase n do stavu j v čase $n + m$, jinak též *pravděpodobnosti přechodu m–tého řádu.*

Jestliže pravděpodobnosti přechodu $p_{ij}(n,\ n + m)$ nezávisí na časových okamžicích n a $n + m$, ale jen na jejich rozdílu m, říkáme, že příslušný Markovův řetězec je *homogenní.*

Uvažujme nyní homogenní Markovův řetězec $\{X_n\}$. Pravděpodobnosti přechodu prvního řádu $P(X_{n+m} = j|X_n = i)$ jsou v tomto případě nezávislé na n; budeme je značit p_{ij} a přívlastek "prvního řádu" vynecháme. Protože pro každé $i \in S$ existuje $n \in \mathbb{N}_0$ tak, že $P(X_n = i) > 0$ a tedy podmíněná pravděpodobnost $P(X_{n+m} = j|X_n = i) = p_{ij}$ je definována pro všechna $j \in S$, můžeme všechny tyto pravděpodobnosti sestavit do čtvercové matice $\boldsymbol{P} = \{p_{ij},\ i,\ j \in S\}$. Zřejmě platí pro každé $n \in \mathbb{N}_0$

$$p_{ij} \geq 0, \quad i, j \in S; \quad \sum_{j \in S} p_{ij} = 1, \quad i \in S. \tag{3.2}$$

Čtvercová matice, jejíž prvky mají vlastnost (3.2), se nazývá *stochastická matice*; matice $\boldsymbol{P}$ pravděpodobností přechodu homogenního Markovova řetězce je tedy stochastická matice.

Označme dále

$$p_i = P(X_0 = i), \quad i \in S. \tag{3.3}$$

Zřejmě platí

$$p_i \geq 0, \quad i \in S; \quad \sum_{i \in S} p_i = 1. \tag{3.4}$$

Pravděpodobnostní rozdělení $\boldsymbol{p} = \{p_i,\ i \in S\}$ se nazývá *počáteční rozdělení* Markovova řetězce.

Uvažujme opět homogenní řetězec s maticí pravděpodobností přechodu $\boldsymbol{P}$. Položme $p_{ij}^{(0)} = \delta_{ij}$, kde δ_{ij} je Kroneckerův symbol

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Dále položme $p_{ij}^{(1)} = p_{ij}$ a pro přirozené $n \geq 1$ definujme postupně

$$p_{ij}^{(n+1)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}. \tag{3.5}$$

Lze ukázat, že řady v (3.5) jsou konvergentní pro každé $n \geq 1$; je $p_{ij}^{(2)} \leq \sum_{k \in S} p_{ik} = 1$, indukcí podle n dostaneme $p_{ij}^{(n)} \leq 1$, podobně lze ukázat, že matice $\boldsymbol{P^{(n)}}$ prvků $p_{ij}^{(n)}$ jsou stochastické matice. Ze vztahu (3.5) též plyne, že

$$\mathbf{P^{(2)}} = \mathbf{P}.\mathbf{P} = \mathbf{P^2} \text{ a obecně } \mathbf{P^{(n)}} = \mathbf{P^{(n-1)}}.\mathbf{P} = \mathbf{P}.\mathbf{P^{(n-1)}} = \mathbf{P^{(n)}} \tag{3.6}$$

Nyní můžeme ukázat souvislost s pravděpodobnostmi přechodu vyšších řádů. [15]

Věta 3.3. *Nechť $\{X_n\}$ je homogenní Markovův řetězec s maticí pravděpodobností přechodu $\boldsymbol{P}$. Potom pro pravděpodobnosti přechodu n–tého řádu platí*

$$P(X_{m+n} = j|X_m = i) = p_{ij}^{(n)}, \quad i, j \in S \tag{3.7}$$

pro všechna celá $m \geq 0$, $n \geq 0$ a $P(X_m = i) > 0$.

Vztah (3.5) lze snadno zobecnit na identitu

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)} \tag{3.8}$$

pro všechna celá m, $n \geq 0$, která se nazývá *Chapmanova-Kolmogorovova rovnost*. Přechod ze stavu i do stavu j v $m + n$ krocích lze uskutečnit tak, že nejdříve se v m krocích přejde do nějakého stavu k a potom ve zbývajících n krocích ze stavu k do stavu j. Maticově lze (3.8) vyjádřit jako $\boldsymbol{P^{(m+n)} = P^{(m)} P^{(n)}}$. [15]

3.2.3 Klasifikace stavů Markovova řetězce

Nadále se budeme zabývat jen homogenními Markovovými řetězci. Dohodněme se na tomto značení: jestliže Markovův řetězec $\{X_n\}$ vychází z počátečního stavu j, budeme podmíněné pravděpodobnosti $P(.|X_0 = j)$ značit jako

$$P(.|X_0 = j) = P_j(.).$$

Položme $\tau_{\mathrm{j}}(0) = 0$ a dále definujme

$$\tau_j(1) = \inf\{n > 0 : X_n = j\} \tag{3.9}$$

s konvencí $\inf\{\emptyset\} = \infty$. Podle této definice je $\tau_{\mathrm{j}}(1)$ náhodná veličina, která nabývá hodnot 1,2,..., nebo hodnoty ∞ a značí náhodný okamžik, ve kterém Markovův řetězec, poté, co opustil počáteční stav, poprvé vstoupí do stavu j. Někdy se nazývá *čas prvního návratu (resp. vstupu) do stavu j*. Podobně můžeme definovat časy dalších návratů (resp. vstupů) do stavu j předpisem

$$\tau_j(k+1) = \inf\{n > \tau_j(k) : X_n = j\}, \quad k = 1, 2, ... \tag{3.10}$$

Definice 3.4. Stav j Markovova řetězce se nazývá *trvalý*, jestliže řetězec, který vychází z j, se do j vrátí s pravděpodobností 1 po konečně mnoha krocích, tj.

$$P_j(\tau_j(1) < \infty) = 1.$$

Stav j se nazývá *přechodný*, jestliže řetězec, který vychází z j, se s kladnou pravděpodobností do j nikdy nevrátí, tj.

$$P_j(\tau_j(1) = \infty) > 0. \quad [15]$$

3.2.4 Rozklad množiny stavů

Definice 3.5. Řekneme, že stav j je *dosažitelný* ze stavu i, jestliže existuje $n \in \mathbb{N}_0$ tak, že $p_{ij}^{(n)}>0$. Jestliže $p_{ij}^{(n)}=0$ pro všechna $n \in \mathbb{N}_0$, říkáme, že j *není dosažitelný z* i.

Poznámka. Každý stav je dosažitelný ze sebe sama, neboť $p_{jj}^{(0)}=1$.

Definice 3.6. Neprázdná množina stavů C se nazývá *uzavřená*, jestliže žádný stav vně C není dosažitelný z žádného stavu uvnitř C. Nejmenší uzavřená množina obsahující množinu C se nazývá *uzávěr* množiny C. Uzavřená množina stavů se nazývá *nerozložitelná*, jestliže neobsahuje žádnou uzavřenou vlastní podmnožinu.

Věta 3.7. *Množina stavů C je uzavřená tehdy a jen tehdy, když $p_{ij}=0$ pro všechna $i \in C$, $j \notin C$.*

Definice 3.8. Je-li jednobodová množina $\{j\}$ uzavřená, tj. je-li $p_{jj}=1$, pak stav j se nazývá *absorpční*.

Poznámka. Vynecháme-li v matici pravděpodobností přechodu $\boldsymbol{P}$ řádky a sloupce odpovídající stavům vně uzavřené množiny C, dostaneme opět stochastickou matici. Množina C je množina stavů Markovova řetězce, kterému se říká *podřetězec* původního řetězce.

Věta 3.9. *Nechť j je trvalý a nechť k je dosažitelný z j. Potom*

1. *k je trvalý,*
2. *j je dosažitelný z k,*
3. *$f_{jk}=P_j(\tau_k(1) < \infty) = P_k(\tau_j(1) < \infty) = f_{kj} = 1$.* [15]

3.2.5 Pravděpodobnosti absorpce

Věta **3.9** nám umožňuje rozložit množinu stavů S Markovova řetězce, který obsahuje trvalé stavy, rozložit na disjunktní sjednocení

$$S = T \cup C_1 \cup C_2 \cup ...,$$

kde T je množina stavů přechodných a C_1, C_2,... jsou uzavřené nerozložitelné množiny stavů trvalých.

Uvažujme řetězec $\{X_\mathrm{n}\}$ s množinou přechodných stavů T a definujme náhodnou veličinu

$$\tau = inf\{n \geq 0 : X_\mathrm{n} \notin T\},$$

která značí *čas výstupu* z množiny přechodných stavů T. Zřejmě τ je náhodná veličina nabývající hodnot 0,1,..., může však s kladnou pravděpodobností být i $\tau=\infty$.

Věta 3.10. *V řetězci s konečně mnoha stavy je*

$$P_i(\tau = \infty) = 0, \quad i \in T.$$

Nadále budeme předpokládat, že $P_\mathrm{i}(\tau < \infty) = 1$ pro všechna i, tj. řetězec v konečném čase vystoupí z množiny přechodných stavů T a vstoupí do nějaké uzavřené množiny stavů trvalých. V této množině již pak setrvá. [15]

Nechť X_τ je ten stav, do kterého řetězec vstoupí, jakmile opustí množinu přechodných stavů T. Definujme pravděpodobnosti

$$u_{ij} = P_i(X_\tau = j) \quad i \in T, j \in T^C. \tag{3.11}$$

Je-li j absorpční stav, potom u_{ij} je pravděpodobnost, že řetězec, který byl na počátku v přechodném stavu i, je absorbován stavem j. Je-li C_k nějaká uzavřená nerozložitelná množina stavů trvalých , potom pravděodobnost, že řetězec, který vychází z i, po opuštění množiny přechodných stavů setrvá v množině C_k, je

$$u_i(C_k) = P_i(X_\tau \in C_k) = \sum_{j \in C_k} u_{ij}. \tag{3.12}$$

Ukažme nyní, jak lze pravděpodobnosti u_{ij} počítat pomocí pravděpodobností přechodu p_{kl}.

Věta 3.11. *Pro pravděpodobnosti u_{ij} definované v 3.11 platí*

$$u_{ij} = p_{ij} + \sum_{\nu \in T} p_{i\nu} u_{\nu j}, \quad i \in T, j \in T^C. \tag{3.13}$$

Vztah (3.13) můžeme vyjádřit i maticově. Matici pravděpodobností přechodu můžeme (po eventuálním přečíslování stavů) psát ve tvaru

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}^* & \mathbf{0} \\ \mathbf{Q} & \mathbf{R} \end{pmatrix}, \tag{3.14}$$

kde $\boldsymbol{P^*} = \{p_{ij},\ i, j \in T^C\}$, $\boldsymbol{Q} = \{p_{ij},\ i \in T,\ j \in T^C\}$, $\boldsymbol{R} = \{p_{ij},\ i, j \in T\}$.

Nechť $\boldsymbol{U} = \{u_{i,j}, \text{i} \in \text{T}, \text{j} \in T^C\}$ je matice pravděpodobností u_{ij}. Potom vztah (3.13) můžeme přepsat do tvaru

$$\mathbf{U} = \mathbf{Q} + \mathbf{RU} \tag{3.15}$$

Pro konečné matice odtud máme $\boldsymbol{(I - R)U = Q}$, kde $\boldsymbol{I}$ je jednotková matice stejného řádu jako $\boldsymbol{R}$. Jestliže k matici $\boldsymbol{I - R}$ existuje matice inverzní, potom existuje jediné řešení této sestavy

$$\mathbf{U} = (\mathbf{I} - \mathbf{R})^{-1}\mathbf{Q}. \tag{3.16}$$

Existenci inverzní matice k $\boldsymbol{I - R}$ zaručuje následující tvrzení.

Lemma 3.12. *Uvažujme řetězec s maticí pravděpodobností přechodu tvaru (3.14). Nechť T je konečná množina přechodných stavů. Potom matice $\boldsymbol{I - R}$ je regulární a platí*

$$(\mathbf{I} - \mathbf{R})^{-1} = \sum_{k=0}^{\infty} \mathbf{R}^{\mathbf{k}}$$

Poznámka. Matice $\boldsymbol{F = (I - R)^{-1}}$ se nazývá *fundamentální matice* Markovova řetězce. [15]

4 Statistické modely

V této části práce se budu věnovat statistickým modelům výkonů baseballových hráčů. Důraz bude kladen na predikční modely, které vyjadřují přínos hráčů k celkovému počtu vítezství jejich týmů.

4.1 Převod doběhů na výhry

Naším hlavním cílem je modelovat počet výher daného týmu pomocí hráčských výkonů. Nejjednodušší model může vypadat tak, že převedeme počet odhadnutých doběhů na výhry. V baseballu vyhrává tým s větším počtem doběhů na konci zápasů. Pokusme se tedy zjistit, kolik doběhů je v průměru potřeba na vítězství v zápase.

Obecně se bere odhad 10 doběhů = 1 výhra. Nicméně, tento odhad je velmi hrubý a proto ho nyní zpřesníme. K tomu nám pomůže vztah, který odvodil Bill James tzv. Pythagorejský odhad. Pomocí něj jsme schopni odhadnout počet výher v procentech

$$W\%_{Pythag} = \frac{RS^2}{RS^2 + RA^2}.$$

Bohužel ani tento vztah není dostatečně přesný, konkrétněji exponent 2 se dá dále zpřesnit.

Patriot a David Smyth upravili Pythagorejský odhad, tím, že exponent 2 nahradili exponentem závislým na tzv. skórujícím prostředí (počet doběhů za rok). Odvozený vztah můžeme psát ve tvaru

$$W\%_{Pythagenpat} = \frac{RS^X}{RS^X + RA^X}$$

kde X = $(\mathrm{RPG})^{0,287}$.

Vezměme například sezónu 2015, kdy v MLB bylo odehráno 2429 zápasů v základní části s celkovým počtem 20647 doběhů. Průměrně tedy 8,5 doběhů za zápas a pak

$$X = 8,5^{0,287} \approx 1,848.$$

Uvažujme nyní průměrný tým s průměrným počtem doběhů a povolených doběhů

$$W\%_{Pythagenpat} = \frac{688,2\overline{33}^{1,848}}{688,2\overline{33}^{1,848} + 688,2\overline{33}^{1,848}} = \frac{81}{162} = 0,500.$$

Dostáváme očekávaný výsledek - průměrný tým vyhraje přesně 50% všech svých zápasů. Nás zajímá, kolik doběhů bude znamenat rozdíl přesně jedné výhry

$$W\%_{Pythagenpat} = \frac{(688,2\overline{33} + \frac{RPW}{2})^{1,848}}{(688,2\overline{33} + \frac{RPW}{2})^{1,848} + (688,2\overline{33} - \frac{RPW}{2})^{1,848}} = \frac{82}{162} \approx 0,506.$$

Úpravami získáme

$$RPW \approx 9,196.$$

Zjistili jsme, že v roce 2015 v průměru 9,196 doběhů v zápase znamenalo vítězství. Hodnota RPW se v posledních letech mění v řádu desetin, takže nám může posloužit jako prvotní, avšak nepříliš přesný odhad.[16],[17]

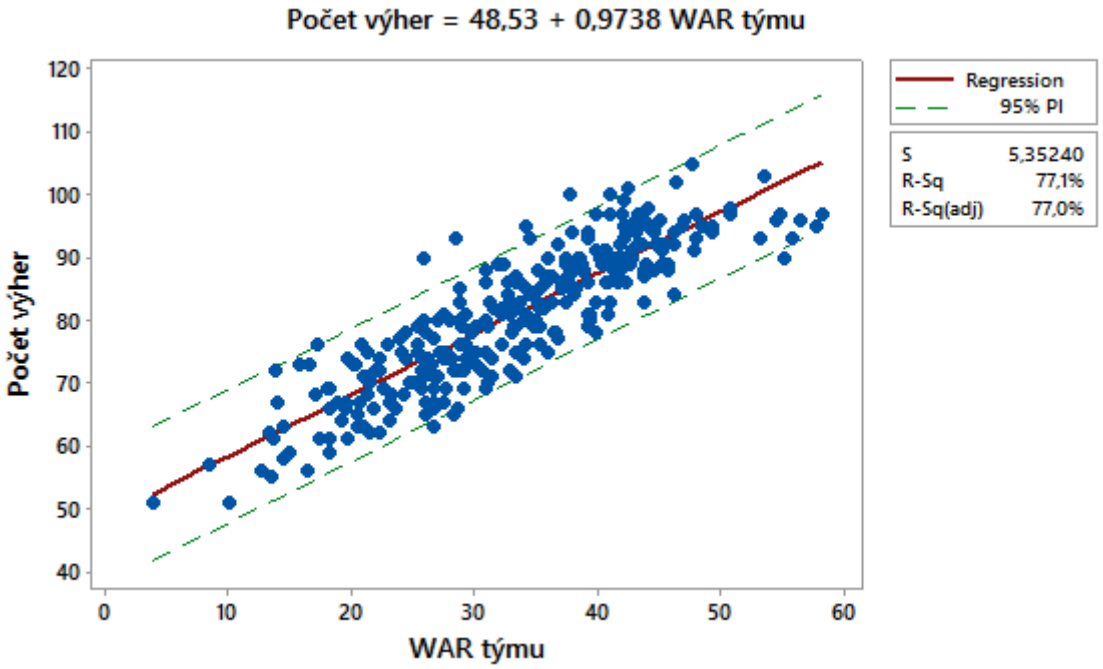
Poznámka: K výpočtu RPW se používá více metod a proto se jednotlivé hodnoty mírně odlišují dle vybrané literatury.[18]

4.2 Lineární regrese

Ve druhé části své práce jsem uvedl několik významných baseballových statistik, z nichž některé se dají využít k odhadu budoucího výkonu týmu díky regresní analýze. U obou těchto modelů vycházíme z historických dat a předpokládáme, že výkony hráčů jsou stále konstantní.

4.2.1 WAR

Pokud sečteme hodnotu WAR všech hráčů daného týmu za celou sezónu, můžeme toto číslo použít pro odhad počtu vítězství na konci sezóny. Podívejme se na hodnoty WAR týmů v rozmezí let 2003 až 2013. V grafu lze vidět závislost počtu výher jednotlivých týmů na jejich celkové hodnotě WAR.



Obrázek 1 – *Graf závislosti WAR týmu na počtu výher*

V rovnici regresní přímky lze vidět že její směrnice je velmi blízká 1 čili jednotkový přírůstek v hodnotě WAR znamená stejný přírůstek v hodnotě výher. Dále, průsečík regresní přímky s osou y v bodě [0;48,53] nám ukazuje, jak by si za těchto 10 let počínal tým složený pouze z náhradníků. Hodnota WAR takového týmu je rovna 0 a na konci sezóny by tento tým měl 49 výher. Lze tedy odhadnout počet výher týmu

$$W_{EXPWAR} = 0,974 \cdot TWAR + 48,5,$$

kde TWAR je celková hodnota WAR mužstva. Jedinou nevýhodou tohoto modelu je fakt, že jeho 95%-ní interval spolehlivosti pro individuální hodnoty (predikční interval) má rozpětí 20 výher, to však jen ukazuje, jak důležitou roli ve hře má štěstí.[19]

4.2.2 ERA

Pomocí samotné ERA počet výher zatím nedokážeme odhadnout, víme však, že ERA souvisí s počtem povolených doběhů. Můžeme tedy vytvořit model pro odhad ERA a pomocí této hodnoty odhadnout počet povolených doběhů a následně využít např. jeden z Pythagorejských odhadů k určení počtu výher.

K odhadu ERA je zapotřebí sestavit vícenásobnou lineární regresi. Využijeme pro ni některé statistiky uvedené v druhé části práce.

Pomocí dat z let 2002-2010 bylo ukázano, že nejlepším prediktorem povolených doběhů je WHIP s koeficientem determinace R^2=0,940. Do modelu byly dále zahrnuty nezávislé proměnné LOB%, HR/9, fielding%, DP.

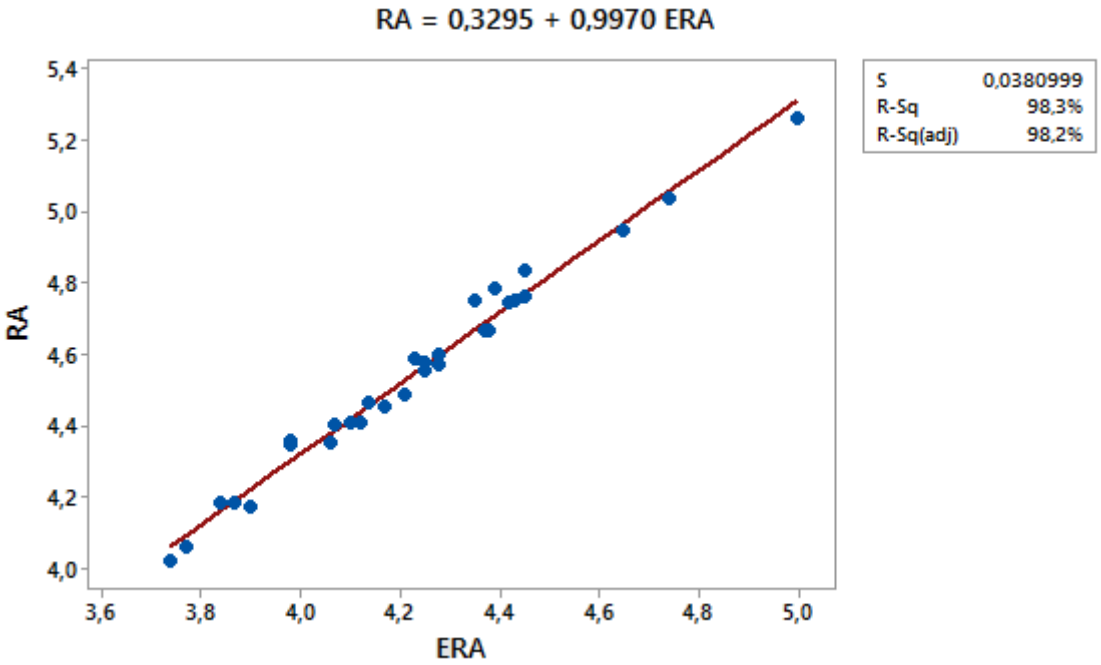
$$ERA = 2,889X_1 - 9,564X_2 + 1,006X_3 - 19,022X_4 - 0,001X_5 - 12,483$$

kde

$$X_1 = \text{WHIP},\ X_2 = \text{LOB\%},\ X_3 = \text{HR/9},\ X_4 = \text{fielding\%},\ X_5 = \text{DP}$$

Koeficient determinace tohoto modelu je R^2=0,988. Výsledek je intuitivní, hodnoty WHIP a HR/9 mají v modelu kladné znaménko, protože tým, který povolí více odpalů, volných met či homerunů bude mít logicky větší hodnotu ERA. Podobnou úvahu lze provést u hodnoty LOB% se záporným znaménkem, čím více hráčů zůstane na metách, tím méně jich udělá doběh. Hodnoty fielding% a DP jsou záporné z očividného důvodu. [20]

Podívejme se na spojitost mezi ERA a počtem povolených doběhů. Pomocí dat z let 2007-2011 byla sestavena lineární regrese. V grafu lze vidět závislost RA na ERA.



Obrázek 2 – *Graf závislosti ERA týmu na počtu povolených doběhů*

Opět jsme dosáhli vysokého koeficientu determinace R^2=0,983, což naznačuje, že doběhy vzniklé chybou obrany neovlivňují zápasy nijak významně. [21] Dále jsme otestovali hypotézu, že absolutní člen je roven nule, neboť tým s ERA=0 by měl mít 0 povolených

doběhů. Analýzou rozptylu jsme získali p hodnotu (*p=0,004*), tedy hypotézu, že konstanta je nulová zamítáme na hladině významnosti 0,05. Hodnoty nezávislé proměnné se pohybují v intervalu (3,7;5), což je dostatečně daleko od nuly a není tedy třeba řešit, zda přímka prochází počátkem. [22]

4.3 Markovovy řetězce

Markovovy řetězce slouží k popisu procesu, který se v dané chvíli nachází právě v jednom stavu v kterémkoliv čase. K popisu přechodů z jednoho stavu do druhého nám slouží matice přechodu, která obsahuje pravděpodobnosti těchto přechodů. Zdůrazněme, že každý následující stav v takovém řetězci je závislý pouze na posledním stavu a nezávislý na stavech předchozích, hovoříme o Markovově vlastnosti těchto řetězců.

Baseball můžeme považovat za Markovův řetězec, neboť v každé půlsměně se nachází právě v jednom z 25 možných stavů. 24 stavů odpovídá kombinaci žádného, jednoho, dvou nebo tří běžců na metách a žádného, jednoho nebo dvou outů. 25. stav odpovídá okamžiku konce půlsměny, když nastane 3. out.

Těchto 25 stavů můžeme znázornit pomocí tabulky jako uspořádané dvojice *(i,j)*, kde i znázorňuje obsazenost met a j znázorňuje počet outů.

	(0,j)	(1,j)	(2,j)	(3,j)	(12,j)	(13,j)	(23,j)	(123,j)
(i,0)	(0,0)	(1,0)	(2,0)	(3,0)	(12,0)	(13,0)	(23,0)	(123,0)
(i,1)	(0,1)	(1,1)	(2,1)	(3,1)	(12,1)	(13,1)	(23,1)	(123,1)
(i,2)	(0,2)	(1,2)	(2,2)	(3,2)	(12,2)	(13,2)	(23,2)	(123,2)

Tabulka 3 – *Přehled možných stavů v baseballu*

25. stav lze zapsat jako *(i,3)*, neboť nás již nezajímá rozmístění běžců na metách. [23]

Matici přechodu $\mathbf{P}$ sestavíme následujícím způsobem

$$\mathbf{P} = \begin{pmatrix} \mathbf{A_0} & \mathbf{B_0} & \mathbf{C_0} & \mathbf{D_0} \\ \mathbf{0} & \mathbf{A_1} & \mathbf{B_1} & \mathbf{E_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{A_2} & \mathbf{F_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix},$$

kde submatice $\mathbf{A}$, $\mathbf{B}$ a $\mathbf{C}$ jsou velikosti 8×8 a submatice $\mathbf{D_0}$, $\mathbf{E_1}$ a $\mathbf{F_2}$ jsou velikosti 8×1. Submatice $\mathbf{0}$ v prostředních dvou řádcích jsou velikosti 8×8 a konečně submatice $\mathbf{0}$ v posledním řádku jsou velikosti 1×8 a submatice $\mathbf{1}$ je jednotková matice 1×1. Matice $\mathbf{P}$ je tedy velikosti 25×25.

Submatice $\mathbf{A}$ představují události, které nezvyšují počet outů. Submatice $\mathbf{B}$ představují události, které zvyší počet outů o 1, ale zároveň nekončí 3 outy, čili z 0 na 1 out a z 1 na 2 outy. Submatice $\mathbf{C}$ reprezentuje události, které zvýší počet outů z 0 na 2. Submatice $\mathbf{D}$, $\mathbf{E}$, $\mathbf{F}$ reprezentují události, které vyústí ve 3 outy, popořadě z žádného, 1 a 2 outů. Nulové submatice představují události, které sníží počet outů a proto jejich pravděpodobnost je nulová.

Abychom mohli matici $\mathbf{P}$ zaplnit musíme se zamyslet nad tím, jak přistoupíme k pohybu běžců po metách. Příkladem konzervativního přístupu by byl předpoklad, že při odpálení singlu běžci postoupí jen o jednu metu. Naopak agresivní přístup předpokládá, že při odpalu singlu běžec z první mety doběhne na druhou a zbytek běžců udělá doběh.

Ukázalo se, že agresivní přístup je realističtější, obzvláště pokud se potýkáme s profesionálními ligami.[24]

Navíc bylo dokázáno, že výzkum, který využíval konzervativní přístup vykazoval 7% chybu při výpočtu doběhů, zatímco agresivní přístup snížil tuto chybu zhruba na 2%. [25]

Využijeme agresivní přístup, tedy při odpalu singlu, běžec z první mety postoupí na druhou a všichni ostatní udělají doběh. Při odpalu doublu, běžec z první mety postoupí na třetí a ostatní udělají doběh. Při odpalu triplu, všichni běžci na metách doběhnou. Při odpalu homerunu, doběhnou všichni běžci na metách i pálkař. Pokud se pálkář obětuje vysokým odpalem, běžec na třetí metě doběhne. Při chybě obrany postupují o jednu metu, pokud jsou k tomu donuceni.

Stačí už jen vypočítat jednotlivé pravděpodobnosti k vyplnění matice přechodu **P**

$$\begin{aligned}
\text{P(HR)} &= \tfrac{HR}{PA}, & \text{P(HBP)} &= \tfrac{HBP}{PA}\ , \\
\text{P(3B)} &= \tfrac{3B}{PA}, & \text{P(out)} &= \tfrac{PA-H-BB-HBP}{PA}, \\
\text{P(2B)} &= \tfrac{2B}{PA}, & \text{P(SF)} &= \tfrac{SF}{PA}, \\
\text{P(1B)} &= \tfrac{1B}{PA}, & \text{P(GDP)} &= \tfrac{GDP}{PA}, \\
\text{P(BB)} &= \tfrac{BB}{PA}, & \text{P(E)} &= \text{1- [fielding\% soupeře].}
\end{aligned}$$

Konečná podoba matice **P** je uvedena v příloze **7.1**. [23]

Markovovy řetězce poskytují mnoho různých aplikací při analýze baseballu, nás však zajímá především modelování počtu skórovaných doběhů. Tomuto problému se budeme věnovat v následující kapitole.

5 Aplikace statistického modelu

V této části se pokusím předpovědět počet výher pro všech 30 týmu americké MLB. Tedy spíše zpětně předpovědět, neboť vycházím z dat po první polovině sezóny 2015. Pomocí těchto dat se pokusím předpovědět stav na konci sezóny. Postupně využiji 2 modely.

Model **4.3** využijeme k odhadu skórovaných doběhů. U modelu **4.2.2** využijeme pouze druhou část, samotnou ERA predikovat nebudeme, použijeme její hodnotu po první polovině sezóny, pomocí které odhadneme počet povolených doběhů soupeři. Následně díky Pythagorejskému odhadu, zmíněném v **4.1**, odhadnu počet výher jednotlivých týmů na konci sezóny 2015.

Každý z použitých modelů budu navíc porovnávat s primitivním modelem, který všechny hodnoty po první polovině sezóny vynásobí dvěma. Tento model představuje odhad obyčejného fanouška, který disponuje průměrnými matematickými znalostmi.

5.1 Počet skórovaných doběhů

Celý výpočet si ilustrujme na týmu New York Yankees. Nejdříve je potřeba zaplnit matici přechodu **7.1** pomocí pravděpodobností uvedených v kapitole **4.3**. První řádek matice přechodu vypadá takto

	(0,0)	(1,0)	(2,0)	(3,0)	(12,0)	(13,0)	(23,0)	(123,0)
(0,0)	0,034	0,256	0,045	0,003	0	0	0	0

	(0,1)	(1,1)	(2,1)	(3,1)	(12,1)	(13,1)	(23,1)	(123,1)
(0,0)	0,679	0	0	0	0	0	0	0

	(0,2)	(1,2)	(2,2)	(3,2)	(12,2)	(13,2)	(23,2)	(123,2)
(0,0)	0	0	0	0	0	0	0	0

Záměrně zde neuvádím poslední sloupec matice, neboť jak ukážu později, není jej potřeba. Uvedený řádek popisuje pravděpodobnosti přechodu z počátečního stavu *(0,0)* do koncového stavu *(i,j)*. Všimněme si, že spousta přechodů má nulovou pravděpodobnost. Pokud začínáme s jedním palkářem ve hře, tak se logicky nemůžeme dostat do stavu s více hráči na metách. Stejně tak jediný hráč nemůže zapříčinit více než jeden out.

Dále je potřeba sestrojit submatici matice **P**. Mějme tedy matici **R**, která je submaticí matice **P**, kterou sestrojíme odstraněním 25. řádku a 25. sloupce odpovídající absorpčnímu stavu.

Sestrojme nyní fundamentální matici Markovova řetězce

$$\mathbf{F} = (\mathbf{I} - \mathbf{R})^{-1},$$

kde **I** je jednotková matice 24×24. Z této matice snadno vyčteme očekávaný počet projití danými stavy z počátečního stavu $(i,j)_1$ do konečného stavu $(i,j)_2$ před dosažením 3-outového absorpčního stavu. [23]

Podívejme se na první řádek fundamentální matice **F** týmu New York Yankees

	(0,0)	(1,0)	(2,0)	(3,0)	(12,0)	(13,0)	(23,0)	(123,0)
(0,0)	1,051	0,280	0,050	0,004	0,093	0	0,017	0,012

	(0,1)	(1,1)	(2,1)	(3,1)	(12,1)	(13,1)	(23,1)	(123,1)
(0,0)	0,766	0,406	0,074	0,008	0,214	0,001	0,041	0,036

	(0,2)	(1,2)	(2,2)	(3,2)	(12,2)	(13,2)	(23,2)	(123,2)
(0,0)	0,580	0,450	0,085	0,013	0,334	0,002	0,068	0,070

Tento řádek vyjadřuje odhadovaný počet projití jednotlivými stavy z počátečního stavu *(0,0)* do koncového stavu *(i,j)*.

Nyní jsme schopni určit kolik pálkařů se dostane na řadu. Stačí sečíst jednotlivé počty projití stavy v každém řádku fundamentální matice **F**, abychom získali odhad počtu projití stavy před absorpcí. Vidíme, že součet prvků prvního řádku fundamentální matice pro tým New York Yankees je 4,656, což znamená, že v průměru se na pálce objeví 4,656 pálkařů ve zbytku směny, začínající ve stavu *(0,0)*.

Zjistili jsme tedy, že pomocí fundamentální matice **F** můžeme odhadnout počet pálkařů, kteří se objeví na pálce ve zbytku směny, která začíná kterýmkoliv ze 24 stavů. Tato data jsem sepsal do tabulky, kde v záhlaví sloupců je uvedeno, které mety jsou obsazeny a v záhlaví řádků je uveden počet outů.

	0	1	2	3	12	13	23	123
0	4,656	4,692	4,722	4,727	4,727	4,781	4,696	4,728
1	3,056	3,059	3,070	3,070	3,072	3,072	3,070	3,072
2	1,509	1,509	1,509	1,509	1,509	1,509	1,509	1,509

Tabulka 4 – *Odhad počtu zbývajících pálkařů*

Vraťme se však nyní ještě k fundamentální matici **F**. Podobným způsobem, jakým jsme spočítali odhad počtu pálkařů ve zbytku směny, můžeme spočítat odhad počtu skórovaných doběhů ve zbytku směny, začínající z jakéhokoliv stavu. Zaveďme sloupcový vektor $\mathbf{r_s}$, který obsahuje očekávaný počet doběhů, který tým skóruje z jednoho odpalu ve kterémkoliv ze 24 stavů. Prvky vektoru $\mathbf{r_s}$ určíme na základě maximálního potencionálního počtu skórovaných doběhů a pravděpodobností popisující tyto doběhy.

Vezměme například směnu, která začíná ve stavu *(0,0)*. Je zřejmé, že může být skórován maximálně 1 doběh anebo žádný doběh. Pravděpodobnost 1 doběhu se rovná pravděpodobnosti odpalu homerunu, čili můžeme spočítat odhad počtu doběhů v této směně po jednom odpalu

$$1 \cdot 0,034 + 0 \cdot (1 - 0,034) = 0,034.$$

Konkrétně jsme spočítali první složku vektoru $\mathbf{r_s}$. [23]

Podobným způsobem dopočítáme zbylých 23 složek a dostaneme konečnou podobu tohoto vektoru

$$\mathbf{r_s} = \begin{pmatrix} r_{s1} \\ . \\ . \\ . \\ r_{s24} \end{pmatrix},$$

kde

r_{s1}=0,034,	r_{s9}=0,034,	r_{s17}=0,034,
r_{s2}=0,071,	r_{s10}=0,071,	r_{s18}=0,071,
r_{s3}=0,260,	r_{s11}=0,260,	r_{s19}=0,260,
r_{s4}=0,269,	r_{s12}=0,269,	r_{s20}=0,260,
r_{s5}=0,297,	r_{s13}=0,297,	r_{s21}=0,297,
r_{s6}=0,306,	r_{s14}=0,306,	r_{s22}=0,297,
r_{s7}=0,495,	r_{s15}=0,495,	r_{s23}=0,486,
r_{s8}=0,633,	r_{s16}=0,633,	r_{s24}=0,624.

Abychom nyní odhadli počet skórovaných doběhů ve zbytku směny z jakéhokoliv počátečního stavu, stačí vynásobit matici $\mathbf{F}$ s vektorem $\mathbf{r_s}$. Nechť $\mathbf{r}$ značí vektor, jehož prvky jsou odhadované počty doběhů. Potom

$$\mathbf{r} = \mathbf{F} \cdot \mathbf{r_s}$$

A tedy vektor $\mathbf{r}$ pro tým New York Yankees vypadá takto

$$\mathbf{r} = \begin{pmatrix} r_1 \\ . \\ . \\ . \\ r_{24} \end{pmatrix},$$

kde

r_1=0,550,	r_9=0,297,	r_{17}=0,113,
r_2=0,946,	r_{10}=0,549,	r_{18}=0,228,
r_3=1,144,	r_{11}=0,740,	r_{19}=0,364,
r_4=1,156,	r_{12}=0,747,	r_{20}=0,364,
r_5=1,562,	r_{13}=1,008,	r_{21}=0,485,
r_6=1,579,	r_{14}=1,014,	r_{22}=0,485,
r_7=1,762,	r_{15}=1,203,	r_{23}=0,621,
r_8=2,278,	r_{16}=1,564,	r_{24}=0,812.

Vektor $\mathbf{r}$ obsahující odhadované počty doběhů ve zbytku směny je mocný nástroj v analýze baseballu. Nás bude především zajímat jeho první složka, protože udává, kolik doběhů můžeme očekávat ve zbytku směny začínající stavem *(0,0)*. Tímto stavem začíná každá směna a když tedy vynásobíme první složku vektoru $\mathbf{r}$ devíti (zápas se hraje na 9 směn) dostaneme odhad celkového počtu skórovaných doběhů daným týmem za zápas. [23]

Opět se podívejme na New York Yankees. Očekávaný počet doběhů za zápas je roven

$$RPG = 0,550 \cdot 9 = 4,953.$$

Můžeme tedy říct, že Yankees průměrně skórují 4,953 doběhů za zápas. Avšak známe-li průměrný počet doběhů za zápas, můžeme rovnou vypočítat počet doběhů za celou sezónu. Nebudeme však násobit průměrný počet doběhů za zápas 162, neboť vycházíme z předpokladu, že známe data po první polovině sezóny. Tedy víme, že Yankees v 88 zápasech skórovali 409 doběhů. Počet doběhů na konci sezóny je roven

$$RS = 4,953 \cdot (162 - 88) + 409 \approx 776.$$

Ve srovnání se skutečnou hodnotou na konci sezóny, což bylo 764, zjišťujeme, že náš první odhad se od reality příliš neliší. Do tabulky na následující straně jsem zpracoval porovnání modelu s realitou pro všechny týmy MLB v sezóně 2015. Druhý sloupec tabulky představuje primitvní model, který vynásobil počet doběhů po první polovině sezóny dvěma. [23]

Tým	Mark. řetězce	Prim. model	Realita
Angels	695	736	661
Astros	719	790	729
Athletics	702	780	694
Blue Jays	860	972	891
Braves	645	694	573
Brewers	656	720	655
Cardinals	680	710	647
Cubs	650	670	689
Diamondbacks	735	784	720
Dodgers	741	752	667
Giants	721	754	696
Indians	668	694	669
Mariners	601	624	656
Marlins	620	660	613
Mets	584	620	683
Nationals	690	702	703
Orioles	732	774	713
Padres	617	704	650
Phillies	566	616	626
Pirates	664	712	697
Rangers	695	740	751
Rays	617	664	644
Red Sox	710	752	748
Reds	663	662	640
Rockies	755	778	737
Royals	734	760	724
Tigers	793	796	689
Twins	693	766	696
White Sox	562	584	622
Yankees	776	818	764
Průměrná absolutní chyba	33,76	49,63	
Průměrná absolutní procent. chyba	5,01%	7,29%	

Tabulka 5 – *Počet skórovaných doběhů*

5.2 Počet povolených doběhů

Použití modelu budeme opět ilustrovat na týmu New York Yankees. Rozhodl jsem se použít pouze model k odhadu RA a samotnou ERA predikovat nebudu. Opět vycházíme z předpokladu, že máme k dispozici data po první polovině sezóny.

Po lehké úpravě závislosti pro RA v **4.2.2** jsme schopni odhadnout RA pro tým Yankees, jejichž ERA byla rovna 3,96 po 88 zápasech

$$RA = (0,997 \cdot 3,96 + 0,330) \cdot (162 - 88) + 383 \approx 700.$$

Ukazuje se, že model se od reality opět příliš neliší, neboť na konci sezóny měli Yankees na kontě 698 povolených doběhů. V následující tabulce je zpracováno porovnání modelu s realitou a s primitivním modelem pro všechny týmy MLB v sezóně 2015.

Tým	Použitý model	Prim. model	Realita
Angels	628	670	675
Astros	622	690	618
Athletics	611	690	729
Blue Jays	723	808	670
Braves	703	770	760
Brewers	734	818	737
Cardinals	485	528	525
Cubs	591	638	608
Diamondbacks	732	784	713
Dodgers	575	628	595
Giants	653	702	627
Indians	671	732	640
Mariners	662	720	726
Marlins	654	708	678
Mets	571	624	613
Nationals	627	672	635
Orioles	648	696	693
Padres	707	800	731
Phillies	833	936	809
Pirates	527	584	596
Rangers	731	790	733
Rays	616	682	642
Red Sox	766	838	753
Reds	712	758	754
Rockies	823	886	844
Royals	609	634	641
Tigers	755	826	803
Twins	661	720	700
White Sox	686	730	701
Yankees	700	766	698
Průměrná absolutní chyba	32,43	43,90	
Průměrná absolutní procent. chyba	4,75%	6,32%	

Tabulka 6 – *Počet povolených doběhů*

5.3 Počet výher

Nyní, když máme k dispozici odhady pro počet skórovaných a povolených doběhů, se můžeme zaměřit na odhadnutí počtu výher jednotlivých týmů. Pro připomenutí znovu uvádím vzorec pro Pythagorejský odhad uvedený v části **4.1**

$$W\%_{Pythagenpat} = \frac{RS^X}{RS^X + RA^X}.$$

Nejdříve musíme určit koeficient X. Za první polovinu sezóny 2015 bylo odehráno 1330 zápasů, během nichž bylo skórováno 10915 doběhů, tedy průměrně 8,207 doběhů za zápas. Můžeme tedy určit koeficient X

$$X = 8,207^{0,287} \approx 1,830.$$

Teď už se můžeme pustit do odhadu počtu výher. Nezapomeňme na předpoklad, že známe počet skórovaných a povolených doběhů, počet výher a počer odehraných zápasů z první poloviny sezóny, díky tomu můžeme s využitím Pythagorejského odhadu odhadnout počet výher na konci sezóny

$$W_{\text{EXP}} = \frac{(RS_2 - RS_1)^X}{(RS_2 - RS_1)^X + (RA_2 - RA_1)^X} \cdot (162 - GS_1) + W_1,$$

kde

RS_2 = Počet skórovaných doběhů odhadnutých v části **5.1**,
RS_1 = Počet skórovaných doběhů za 1. polovinu sezóny,
RA_2 = Počet povolených doběhů odhadnutých v části **5.2**,
RA_1 = Počet povolených doběhů za 1. polovinu sezóny,
GS_1 = Počet odehraných zápasů za 1. polovinu sezóny,
W_1 = Počet vyhraných zápasů za 1. polovinu sezóny.

Pro ilustaci provedeme výpočet pro New York Yankees

$$W_{\text{EXP}} = \frac{(776 - 409)^{1,830}}{(776 - 409)^{1,830} + (700 - 383)^{1,830}} \cdot (162 - 88) + 48 \approx 90$$

Ve skutečnosti Yankees vyhráli v sezóně 2015 celkem 87 zápasů, jedná se tedy o velice příznívý odhad. Podívejme se však na toto porovnání s realitou a primitivním modelem pro všechny týmy MLB v roce 2015 v tabulce na následující straně.

Tým	Pythagorejský odhad	Prim. odhad	Realita
Angels	89	96	85
Astros	90	98	86
Athletics	82	82	68
Blue Jays	86	90	93
Braves	76	84	67
Brewers	71	76	68
Cardinals	105	112	100
Cubs	89	94	97
Diamondbacks	80	84	79
Dodgers	98	102	92
Giants	87	92	84
Indians	81	84	81
Mariners	76	82	76
Marlins	73	76	71
Mets	85	94	90
Nationals	91	96	83
Orioles	86	88	81
Padres	72	82	74
Phillies	54	58	63
Pirates	99	106	98
Rangers	78	84	88
Rays	83	92	80
Red Sox	77	84	78
Reds	77	78	64
Rockies	75	78	68
Royals	97	104	95
Tigers	86	88	74
Twins	86	98	83
White Sox	73	82	76
Yankees	90	96	87
Průměrná absolutní chyba	5,00	8,70	
Průměrná absolutní procent. chyba	6,51%	11,08%	

Tabulka 7 – *Počet výher*

5.4 Umístění týmů

V poslední kapitole aplikační části své práce se budu věnovat porovnání umístění jednotlivých týmů na konci sezóny. Předmětem mého zájmu je pouze prvních 10 týmů, neboť právě 10 týmů postupuje do play-off a má šanci bojovat o titul.

V následující tabulce uvádím umístění spočítané pomocí Pythagorejského odhadu (POU), umístění spočítané primitivním modelem(PMU) a reálné umístění(RU). Tedy pouze seřazujeme týmy podle počtu výher, neboť konečné umístění závisí pouze na tomto počtu. Týmy, které by se podle modelů umístily mimo elitní desítku, nemají v tabulce doplněné umístění.

Tým	MŘU	PMU	RU
Cardinals	1.	1.	1.
Pirates	2.	2.	2.
Cubs	8-9.	10.	3.
Royals	4.	3.	4.
Blue Jays	-	-	5.
Dodgers	3.	4.	6.
Mets	-	-	7.
Rangers	-	-	8.
Yankees	6-7.	7-9.	9.
Astros	6-7.	5-6.	10.

Tabulka 8 – *Umístění*

Bez ohledu na přesné umístění, vidíme, že oba modely správně předpověděly umístění 7 týmů v první desítce. Tyto týmy hrály celou sezónu vyrovnaně bez větších výkonnostních výkyvů. Naopak zbylé 3 týmy, které by se podle modelů neumístili v top 10, měly pravděpodobně slabší první polovinu sezóny a silnou druhou polovinu. Při porovnání odhadu přesných umístění, lze konstatovat, že o trochu lépe si vedl Pythagorejský odhad, avšak rozdíly v přesnosti jsou téměř zanedbatelné.

Na jednu stranu nás tento výsledek může překvapit, protože v předchozí části jsme ukázali, že Pythagorejský odhad je přesnější než odhad primitivní a přirozeně bychom očekávali, že se tato skutečnost projeví i při porovnání umístění. Na druhou stranu jsme se zajímali jen o prvních 10 týmů a navíc při umístění mnohdy závisí na jediné výhře, jak lze vidět v tabulce 7. Vzhledem k průměrné absolutní chybě obou odhadů se tedy nelze divit, že jsme neobdrželi přesnější výsledky.

6 Závěr

V úvodu a ve druhé kapitole své práce jsem definoval pojem Sabermetrics a uvedl, proč právě baseball je lehce statisticky popsatelný. Rovněž jsem popsal nejdůležitější hráčské herní statistiky. Obzvláště bych připomenul statistiky WAR a ERA, které se v současné době řadí mezi ty nejvýznamnější.

Ve třetí kapitole jsem se věnoval Markovovým řetězcům a popsal jejich základní vlastnosti. Dále jsem matematicky přesně definoval související pojmy, které jsem použil později v aplikační části práce.

V následující kapitole jsem popsal některé statistické modely, založené na lineární regresi či na Markovových řetězcích. Také jsem uvedl jednoduchý model využívající tzv. Pythagorejský odhad.

Konečně, v páté kapitole své práce jsem zmíněné modely aplikoval na reálná data. Mým cílem bylo předpovědět počet výher pro všech 30 týmů v americké MLB na konci sezóny 2015, za předpokladu, že znám data po její první polovině.

Nejdříve jsem využil Markovových řetězců k odhadu počtu skórovaných doběhů. Tuto část jsem podrobně rozepsal, neboť právě Markovovy řetězce jsou hlavním matematickým aparátem mé práce. Z tabulky 5 vidíme, že model s Markovovými řetězci měl průměrnou absolutní chybu 33,76 doběhů, což můžeme považovat za příznivé zjištění, avšak když jsem tento model porovnal s modelem primitivním, zjistil jsem, že můj model je pouze o 2,28% přesnější. Nicméně tento rozdíl, necelých 16 doběhů, není zanedbatelný, neboť ve skutečnosti znamená přibližně 2 výhry (viz. **4.1**).

V další části jsem využil lineární závislosti RA na ERA a odhadnul počet povolených doběhů opět pro všech 30 týmů. V tabulce 6 jsem porovnal použitý model s modelem primitivním. Mnou zvolený model je v průměru o 11 doběhů přesnější, čili v průměru přibližně o jednu výhru přesnější.

V závěru páté kapitoly jsem pomocí spočítaných dat a Pythagorejského odhadu mohl předpovědět počet výher pro všechny týmy MLB. Zmíněný odhad se ukázal o 4,57% přesnější než primitivní odhad a jeho průměrná absolutní chyba byla 5 výher. Pokud se však bavíme o konečném umístění na konci sezóny, nelze ani jeden z odhadů považovat za obzvlášť spolehlivý.

Můžeme tedy konstatovat, že při sázení na baseballové zápasy, obyčejný fanoušek nebude výrazně znevýhodněn oproti matematikovi se znalostí Sabermetrics, neboť odhady obou jedinců budou velmi podobné.

Reference

[1] JAFFE, Chris. Bill James Interview. In: *The Hardball Times* [online]. 2008 [cit. 2017-04-01]. Dostupné z: http://www.hardballtimes.com/bill-james-interview/

[2] GRABINER, David. The Sabermetric Manifesto. In: *Sean Lahman* [online]. 2011 [cit. 2017-04-01]. Dostupné z: http://www.seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/

[3] BUKIET, Bruce. The Probability Of Winning A Baseball Game - And A Post-Season Prediction. In: *Science 2.0* [online]. 2013 [cit. 2017-04-01]. Dostupné z: http://www.science20.com/bruce_bukiet/probability_winning_baseball_game_and_postseason_prediction-121526

[4] WEINBERG, Neil. How to evaluate a hitter, sabermetrically. In: *Beyond the box score* [online]. 2014 [cit. 2017-04-01]. Dostupné z: http://www.beyondtheboxscore.com/2014/5/26/5743956/sabermetrics-stats-offense-learn-sabermetrics

[5] SLOWINSKI, Steve. wOBA. In: *FanGraphs* [online]. 2010 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/offense/woba/

[6] SLOWINSKI, Steve. BABIP. In: *FanGraphs* [online]. 2010 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/offense/babip/

[7] SLOWINSKI, Steve. WAR for Position Players. In: *FanGraphs* [online]. 2012 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/war/war-position-players/

[8] WEINBERG, Neil. RE24. In: *FanGraphs* [online]. 2014 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/misc/re24/

[9] SLOWINSKI, Steve. ERA. In: *FanGraphs* [online]. 2010 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/pitching/era/

[10] SLOWINSKI, Steve. WHIP. In: *FanGraphs* [online]. 2010 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/pitching/whip/

[11] SLOWINSKI, Steve. LOB%. In: *FanGraphs* [online]. 2010 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/pitching/lob/

[12] WEINBERG, Neil. Complete List (Offense). In: *FanGraphs* [online]. 2014 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/offense/offensive-statistics-list/

[13] WEINBERG, Neil. Complete List (Pitching). In: *FanGraphs* [online]. 2014 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/pitching/complete-list-pitching/

[14] ANDĚL, Jiří. Základy matematické statistiky. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-807-3781-620.

[15] PRÁŠKOVÁ, Zuzana a Petr LACHOUT: *Základy náhodných procesů*, 1. vyd. Praha: Karolinum, 1998, 146 s. ISBN 80-7184-688-0.

[16] GROSNICK, Bryan. Converting runs to wins in 2013. In: *Beyond the box score* [online]. 2014 [cit. 2017-04-01]. Dostupné z: http://www.beyondtheboxscore.com/2014/4/10/5591522/converting-runs-to-wins-in-2013-wins-above-replacement-sabermetrics-ugh-math

[17] Guts!. In: FanGraphs [online]. [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/guts.aspx

[18] SLOWINSKI, Steve. Converting Runs to Wins. In: *FanGraphs* [online]. 2010 [cit. 2017-04-01]. Dostupné z: http://www.fangraphs.com/library/misc/war/converting-runs-to-wins/

[19] DOLINAR, Sean. Predicting baseball wins with WAR. In: *Stats Sean Dolinar* [online]. 2014 [cit. 2017-04-01]. Dostupné z: http://stats.seandolinar.com/predicting-baseball-wins-with-war/

[20] BENEVENTANO, Philip, Paul D. BERGER a Bruce D. WEINBERG. Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. *International Journal of Business, Humanities and Technology* [online]. 2012, **2**(4), 67-75 [cit. 2017-04-01]. ISSN 2162-1381. Dostupné z: http://www.ijbhtnet.com/journals/Vol_2_No_4_June_2012/7.pdf

[21] Developing a System for Predicting Wins, Part II. In: *LiteSABERs* [online]. 2012 [cit. 2017-04-01]. Dostupné z: http://litesabers.blogspot.cz/2012/03/developing-system-for-predicting-wins.html

[22] Úvod do regresní analýzy. In: *StatSoft* [online]. [cit. 2017-04-01]. Dostupné z: http://www.statsoft.cz/file1/PDF/newsletter/2014_26_03_StatSoft_Uvod_do_regresni_analyzy.pdf

[23] TESAR, Naomi. *Estimating Expected Runs Using a Markov Model for Baseball* [online]. [cit. 2017-04-01]. Dostupné z: https://www.edsolio.com/media/2/265/files/Tesar_FinalDraft.pdf

[24] BUKIET, Bruce, Elliotte Rusty HAROLD a José Luis PALACIOS. A Markov Chain Approach to Baseball. *Operations Research* [online]. 1997, **45**(1), 1-22 [cit. 2017-04-01]. ISSN 1526-5463. Dostupné z: http://www.math.cornell.edu/∼levine/4740/markov-baseball.pdf

[25] SOKOL, Joel S. An Intuitive Markov Chain Lesson From Baseball. *INFORMS Transactions on Education* [online]. 2004, **5**(1), 50 [cit. 2017-04-01]. ISSN 1532-0545. Dostupné z: http://isye.umn.edu/courses/ie5112/mc/sokol.pdf

Seznam obrázků

Seznam tabulek

7 Přílohy

7.1 Matice přechodu

$$\mathbf{P} = \begin{pmatrix} \mathbf{A_0} & \mathbf{B_0} & \mathbf{C_0} & \mathbf{D_0} \\ \mathbf{0} & \mathbf{A_1} & \mathbf{B_1} & \mathbf{E_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{A_2} & \mathbf{F_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix},$$

kde

$$\mathbf{A_0} = \begin{pmatrix} P(HR*) & P(1X) & P(2B) & P(3B) & 0 & 0 & 0 & 0 \\ P(HR**) & 0 & 0 & P(3B*) & P(1X) & 0 & P(2B) & 0 \\ P(HR**) & P(1Y*) & P(2B*) & P(3B*) & P(1Z) & 0 & 0 & 0 \\ P(HR**) & P(1Y*) & P(2B*) & P(3B*) & 0 & P(1Z) & 0 & 0 \\ P(HR***) & 0 & 0 & P(3B**) & P(1Y) & 0 & P(2B*) & P(1Z) \\ P(HR***) & 0 & 0 & P(3B**) & P(1Y*) & 0 & P(2B*) & P(1Z) \\ P(HR***) & P(1Y**) & P(2B**) & P(3B**) & 0 & 0 & 0 & P(1Z) \\ P(HR****) & 0 & 0 & P(3B***) & P(1Y*) & 0 & P(2B**) & P(1Z*) \end{pmatrix},$$

$$\mathbf{A_1} = \begin{pmatrix} P(HR*) & P(1X) & P(2B) & P(3B) & 0 & 0 & 0 & 0 \\ P(HR**) & 0 & 0 & P(3B*) & P(1X) & 0 & P(2B) & 0 \\ P(HR**) & P(1Y*) & P(2B*) & P(3B*) & P(1Z) & 0 & 0 & 0 \\ P(HR**) & P(1Y*) & P(2B*) & P(3B*) & 0 & P(1Z) & 0 & 0 \\ P(HR***) & 0 & 0 & P(3B**) & P(1Y*) & 0 & P(2B*) & P(1Z) \\ P(HR***) & 0 & 0 & P(3B**) & P(1Y*) & 0 & P(2B*) & P(1Z) \\ P(HR***) & P(1Y**) & P(2B**) & P(3B**) & 0 & 0 & 0 & P(1Z) \\ P(HR****) & 0 & 0 & P(3B***) & P(1Y**) & 0 & P(2B**) & P(1Z*) \end{pmatrix},$$

$$\mathbf{A_2} = \begin{pmatrix} P(HR*) & P(1X) & P(2B) & P(3B) & 0 & 0 & 0 & 0 \\ P(HR**) & 0 & 0 & P(3B*) & P(1X) & 0 & P(2B) & 0 \\ P(HR**) & P(1Y*) & P(2B*) & P(3B*) & P(1Z) & 0 & 0 & 0 \\ P(HR**) & P(1Y*) & P(2B*) & P(3B*) & 0 & P(1Z*) & 0 & 0 \\ P(HR***) & 0 & 0 & P(3B**) & P(1Y*) & 0 & P(2B*) & P(1Z) \\ P(HR**) & 0 & 0 & P(3B**) & P(1Y*) & 0 & P(2B*) & P(1Z) \\ P(HR***) & P(1Y**) & P(2B**) & P(3B**) & 0 & 0 & 0 & P(1Z) \\ P(HR****) & 0 & 0 & P(3B***) & P(1Y**) & 0 & P(2B**) & P(1Z*) \end{pmatrix},$$

$$\mathbf{B_0} = \begin{pmatrix} P(out) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & P(out) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & P(out) & P(SF) & 0 & 0 & 0 & 0 \\ P(SF*) & 0 & 0 & P(out) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & P(out) & 0 & 0 & 0 \\ 0 & P(SF) & 0 & 0 & 0 & P(out) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & P(SF*) & P(out) & 0 \\ 0 & 0 & 0 & 0 & P(SF) & 0 & 0 & P(out) \end{pmatrix},$$

$$\mathbf{B_1} = \begin{pmatrix} P(out) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & P(out) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & P(out) & P(SF) & 0 & 0 & 0 & 0 \\ P(SF*) & 0 & 0 & P(out) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & P(out) & 0 & P(SF) & 0 \\ 0 & P(SF*) & 0 & 0 & 0 & P(out) & 0 & 0 \\ 0 & P(SF) & 0 & 0 & 0 & P(SF) & P(out) & 0 \\ 0 & 0 & 0 & 0 & P(SF*) & 0 & 0 & P(out) \end{pmatrix},$$

$$\mathbf{C_0} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ P(GDP) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ P(GDP) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ P(GDP) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & P(GDP) & P(GDP) & 0 & 0 & 0 & 0 \\ P(GDP*) & P(GDP*) & P(GDP*) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P(GDP*) & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{D_0} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{E_1} = \begin{pmatrix} 0 \\ P(GDP) \\ 0 \\ P(GDP) \\ P(GDP) \\ P(GDP) \\ 0 \\ P(GDP) \end{pmatrix}, \quad \mathbf{F_2} = \begin{pmatrix} P(out) \\ P(out) \\ P(out) \\ P(out) \\ P(out) \\ P(out) \\ P(out) \\ P(out) \end{pmatrix},$$

kde

$$\begin{aligned} &\mathrm{P(1X) = P(1B) + P(BB) + P(HBP) + P(E)}, \\ &\mathrm{P(1Y) = P(1B) + P(E)}, \\ &\mathrm{P(1Z) = P(BB) + P(HBP)}. \end{aligned}$$

Poznámka: Počet hvězdiček ukazuje, kolik doběhů bylo v daném přechodu mezi stavy skóróváno. [23]