Játékosok teljesítményének statisztikai értékelése az amerikai baseball ligában

Diplomamunka

Írta: Rácsai Mátyás Bence

Alkalmazott matematikus szak

Témavezető:

Varga László, egyetemi tanársegéd
Valószínűségelméleti és Statisztika Tanszék
Eötvös Loránd Tudományegyetem, Természettudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2015

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Varga Lászlónak a kitartó munkájáért, a rengeteg befektetett idejéért és energiájáért, valamint azért, hogy segített összegyűjteni a hasznos segédirodalmat a dolgozathoz.

Emellett szeretném megköszönni a családomnak és barátaimnak támogatását és biztatását az elmúlt három év folyamán.

Tartalomjegyzék

1. Bevezetés

Ebben a dolgozatban az amerikai baseball liga (MLB[1]) játékosait fogjuk megvizsgálni teljesítményük alapján, majd megpróbálunk következtetéseket levonni, összefüggéseket találni a fizetésükkel, illetve a csapatuk eredményével kapcsolatban. Ide értjük például azt, hogy mennyire tükrözi a kapott összeg az általuk nyújtott játék képét, vagy hogy egy-egy csapatban a fizetések kiosztása mennyire igazodik a teljesítményhez. Várhatóan szorosabb összefüggést vélünk majd felfedezni az általunk számolt teljesítményt mérő módszer segítségével az egyes játékosok csapatuk eredményéhez való hozzájárulásában és egyéni produkciójukban. A dolgozat második felében kitérünk arra is, hogy az életkor hogyan játszik szerepet a sportolók karrierjében. Megvizsgáljuk, hogy a tapasztalataink mennyire egyeznek meg az elvárásainkkal ezen a téren. Ezekhez a következtetésekhez használni fogunk különböző statisztikai becsléseket, főkomponens-analízist, lineáris regressziót. Hogy az Olvasó jobban átláthassa a leírtakat, indokolt egy rövid összefoglaló, melyben bemutatjuk a játék menetét és szabályait.

[1]Major League Baseball

2. Szükséges ismeretek a baseballról

2.1. Rövid összefoglaló a játék menetéről

A baseballban két, kilenc főből álló csapat versenyez egymás ellen kilenc meneten – inningen – keresztül. A pálya fő része – infield – négyzet alakú, minden csúcsában egy bázissal. A játék mindig a hazai bázisról – home plate – indul. A két csapat felváltva támad és védekezik, minden inning felénél cserélve, tehát egy-egy menetet a hazai csapat kezd védekezéssel, majd a felénél váltanak, és ők fognak ezután támadni. A védekező csapat áll fenn a pályán, teljes létszámban, köztük érdemes kiemelni a dobót – pitcher –, és az elkapót – catcher –, akik az ütés folyamatával közvetlen kapcsolatban vannak. Ezzel szemben a támadó játékosok állnak majd a hazai bázishoz, hogy a pályára üssék a dobó által szolgáltatott labdát, majd lehetőség szerint (akár több részletben is) körbefussanak a bázisokon, és visszaérjenek a hazai bázisra. Ez jelent 1 pontot a csapatnak. Az ütőjátékosnak 3 esélye van eltalálni a jó dobásokat – strike –, ezután kiesik – strikeout. Ha pedig a dobójátékos nem dob elég pontosan (az ütőzónán kívülre érkezik a labda, vagyis az ütőjátékos térdénél alacsonyabban, vagy könyökénél magasabban, illetve túl közel hozzá, vagy túl messze tőle), akkor az ütőjátékos 4 ilyen dobás után automatikusan, ütés nélkül sétálhat az első bázisra, potenciális pontszerző futójátékossá válva. Az utána következő ütőjátékosok célja, hogy őt, mint futót előrejuttassák a bázisokon ütésükkel, egészen a hazai bázisig. A két csapat akkor vált pozíciót, ha a pályán védekezők kiejtettek 3 támadójátékost. Ezt megtehetik a dobó segítségével (3 strike), vagy akár a bázisokon, a futókat kiejtve. Erre több mód is van, de nem lesz lényeges a következőkben, így erre nem is térünk ki. Leggyakrabban a játékosokat az ütési eredményeik alapján értékelik, mivel ez járul hozzá a legjobban a pontok, és így a győzelmek megszerzéséhez is (mint a nevéből is adódik, a védekező csapat nem tud pontot szerezni).

Játékosainkat tehát a következő szempontokból fogjuk megvizsgálni és értékelni:

G	lejátszott mérkőzések száma
AB	hányszor állt ütéshez
R	megszerzett pontok száma
H	ütések száma
1B	1 bázist érő ütések száma
2B	2 bázist érő ütések száma
3B	3 bázist érő ütések száma
HR	Home Run – hazafutások száma
RBI	a játékos ütésének következtében megszerzett pontok
SB	ellopott bázisok száma
CS	lopás közben kiejtve
BB	séták száma
IBB	szándékos séták száma
SO	strikeoutok száma
HBP	dobás által eltalálva
SH	olyan ütés, melynél az ütőjátékos önszántából kiesik, hogy a pályán lévő futók továbbhaladhassanak – nem részletezzük
GIDP	ütések, amelyek következtében két támadó is kiesett

2.2. A baseballban használatos teljesítménymutatók

A későbbiekben megnézzük, hogy egyes játékosok teljesítménye mennyire magyarázza jól a fizetésüket. Nyilván szeretnénk azt tapasztalni, hogy – ha nem is túl erős, de – legalább egy nem elvethető kapcsolat áll fent a kettő között. Ehhez szükségünk lesz a teljesítmény valamilyen mérőeszközére.

Természetesen ebben a próbálkozásban nem mi vagyunk az elsők. A baseball történetében már régóta helye van a hasonló vizsgálatoknak és számításoknak. A csapatok vezérigazgatójának (General Manager) az egyik legfontosabb feladata az új játékosok megvétele, és a nem megfelelően teljesítő játékosok eladása. Ehhez pedig nagy segítséget jelent egy jól működő módszer a sportolók aktuális értékének meghatározására.

Az erre a célra használt rendszert már sok évvel ezelőtt létrehozták és használták. Ferdinand Cole Lane mérte először az egyes támadó események hatását. Később George Lindsey rögzített több mint 1000 mérkőzést lépésről lépésre, majd megvizsgálta, hogy adott szituációkban várhatóan hány pontot szerez egy csapat. Ezek segítségével vezette be Pete Palmer a Linear Weights Systemet, ami lineárisan súlyozza az egyes ütési adatokat. Palmer annyiban tért el társaitól, hogy negatív súlyokat is tartalmazott

a formulája, ezzel "büntette" a csapatot hátravető megmozdulásokat (strikeout, CS, GIDP, ...).

Komoly nehézséget okoz, hogy sok statisztika nem tudja a játékosok értékét a csapatétól függetlenül mérni. Gyakran használatosak a következő mennyiségeket (hagyatkoztunk Michael Ross Smith [2006] disszertációjára)

BA – Batting Average, azaz ütőátlag.

$$BA = \frac{1B + 2B + 3B + HR}{AB}$$

Ez összeadja az "ütött" bázisok számát, majd elosztja az ütéshez kerülések számával. Jelentős mennyiségű információt hagy figyelmen kívül, kizárólag az ütési tevékenységgel foglalkozik, nem számolja össze a már pályán, futóként megszerzett bázisokat és pontokat.

OBP – On-Base Percentage (bázisra kijutás aránya)

$$OBP = \frac{1B + 2B + 3B + HR + BB + HBP}{AB + BB + HBP + SF}$$

Itt már figyelembe vesszük a sétákat, vagyis, amint a név is sugallja, azt mérjük, hogy emberünk milyen mértékben került ki bázisra. Mivel a séta is érhet pontot, sőt, sok esetben mérkőzések múlhatnak egy-egy sétán, ez a mennyiség már közelebb áll a pontszerzési készség valósághű tükrözéséhez.

SLG – Slugging Percentage (ütési arány)

$$SLG = \frac{1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR}{AB}$$

Szintén csak az ütési képességet méri, de immáron súlyozva. Ez egy jó gondolat, hiszen a csapat valóban előrébb van egy Home-Runnal, mint egy 1B-vel. Ezen kívül elég sok faktorral nem számol, így nem teljesen megbízható, és a mi szempontunkból szintén elhanyagolható.

A későbbiekben használni fogjuk a PA (Plate Appearance – hazai bázisnál történő megjelenés) értéket, ami azt mutatja, hányszor fejez be egy játékos egy ütéshez állást. Ez azt jelenti, hogy egy ütéshez állás csak akkor számít AB (At Bat)-nek, ha az ütő

sikeres ütést hajt végre, vagy strikeoutot kap. Ezen felül bármilyen lehetséges kimenetel PA-nak számít. Éppen ezért számítjuk a következőképpen:

$$\text{PA} = \text{AB} + \text{BB} + \text{HBP} + \text{SH} + \text{SF}$$

Végül, de nem utolsó sorban megemlítjük az ALW (Average Linear Weights) mutatót.

$$ALW = \frac{LW}{PA}$$

A Linear Weights (LW) súlyozásról nemsoká részletesebben írunk.

Természetesen fontos, hogy egy játékos hol helyezkedik el a csapat ütősorrendjében, mivel ehhez mérten például az RBI (Runs Batted In - az ő ütése eredményeképpen szerzett pontok) értéke lehet kicsi vagy nagy. A sorrendben első ütő (leadoff hitter) feladata az, hogy egy biztos ütéssel kijusson az első, vagy második bázisra (1B vagy 2B), így ha az imént említett RBI alapján próbálnánk értékelni, az nem lenne szerencsés, hiszen legtöbbször nincsenek futók a pályán, akiket az ő ütése be tud hozni. Az egy és két bázist érő ütések, vagy a bázisra kijutás (On-Base Percentage) szerint viszont már a legértékesebbek közé tartoznak ezek az ütők. Szükségszerű tehát egy olyan megközelítés, aminek segítségével egy közös eszközzel tudjuk vizsgálni az összes (ütő)játékost.

A célunk az lenne, hogy találjunk egy olyan súlyozást, ami minden egyes szempontot annyira tart fontosnak, amennyire az közelebb juttatja a játékos csapatát a pontszerzéshez. Ehhez természetesen jó lenne, ha egy-egy eseményről el tudnánk dönteni, hogy mekkora arányban vezet pontszerzéshez, és így meg tudnánk határozni a súlyát.

A Linear Weights System azon alapul, hogy vesszük az ütőjátékos cselekedetét, és megnézzük az összes szituációra (24 db: 1, 2, 3 out, illetve a bázisokon lévő futók elhelyezkedése: mindhárom bázison vagy van futó, vagy nincs, azaz $2^3 = 8$ lehetőség), hogy mennyi pontot ér a csapatnak. Ezt úgy számoljuk, hogy az esemény következtében szerzett pontokhoz hozzáadjuk a keletkezett szituáció várható pontértékét, majd levonjuk a kezdeti várható pontszerzés értékét. Így például egy Home Run értéke azzal a feltétellel, hogy nincs out és nincs futó a pályán, 1 lesz, nem meglepő módon. Egy HR következtében az ütőjátékos kiüti a labdát a játéktérről, és az összes pályán lévő futó, őt is beleérve befuthat a hazai bázisra, egy-egy pontot szerezve. Mivel nem volt futó, így az ütés következtében 1 pont jött be. Ehhez hozzáadjuk az ütés utáni szituáció (0 out és nincs futó) várható pont értékét (0.482), majd levonjuk az ütés előtti alaphelyzet (szintén 0 out és nincs futó) várható pont értékét (0.482), így a Home Run értéke: $1 + 0.482 - 0.482 = 1$.

A rendszer azért is hasznos, mert pusztán az egyéni teljesítménytől függően értékel. Nem játszik szerepet, hogy mennyire jó a csapat mögöttük, vagy mennyire teljesít gyengén egy-egy csapattársuk adott szituációkban.

3. A teljesítmény hatása a fizetésre

Bár, mint látni fogjuk, a 4. fejezet elején említést teszünk korábbi eredményekről és vizsgálatokról, ebben a témában még kevés próbálkozásról számolhatunk be. Ugyan akad egy-egy elmélkedés, vagy sejtés, de a sportnak ezen területe még nem teljesen feltérképezett, így csak a saját elgondolásainkra támaszkodunk, a saját elveinket követjük.

Kiindulási alapként szolgáltak a következő cikkek, ezért érdemes őket megemlítünk: Stephen Hall egy értekezése (Hall et al. [2002]), mely részletezi a hasonlóságokat és különbségeket a baseball és az európai futball között, legalábbis a pénzügyeket tekintve. Másik forrásunk Gerald W. Scully-tól származik, ő csupán az MLB-re fókuszál (Scully [1974])

3.1. Alkalmazott matematikai eljárások

3.1.1. Lineáris modell

A statisztikában a regressziószámítást két vagy több változó közötti kapcsolat modellezésére használjuk. A mi esetünkben az egyes játékosok fizetését (illetve szerzett pontjaikat) próbáljuk magyarázni a különböző ütési adataik alapján (több változó).

A lineáris regresszió feltételezi a magyarázó – és magyarázott változók közötti lineáris kapcsolatot. Az adataink ponthalmazára egyenest fogunk illeszteni, ennek egyik lehetséges módja a legkisebb négyzetek módszere, vagyis arra törekszünk majd, hogy az egyenesünk és a pontok közötti (függőleges) távolságok négyzetösszege minimális legyen.

Feltesszük, hogy a keresett változónk (y) a magyarázó változók (x_i) lineáris kombinációja, vagyis felírható a következő egyenlet:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_n x_{i,n} + \varepsilon_i$$

ahol a β_i együtthatókat nem ismerjük, ezeket szeretnénk megbecsülni a rendelkezésre álló adatok segítségével úgy, hogy az

$$\varepsilon_i = y_i - (\beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_n x_{i,n})$$

értékek négyzetösszege minimális legyen. Itt az ε_i az i. hibát jelöli, és feltesszük róluk, hogy 0 várható értékűek, σ szórásúak és korrelálatlanok.

Vezessük be a következő jelöléseket:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

így az egyenletrendszer az alábbi mátrixos alakot ölti:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon.$$

Azt a $\hat{\boldsymbol{\beta}}$ becslést keressük, amire $\|\hat{\varepsilon}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$ minimális lesz.

A Gauss-féle normálegyenlet tétel szerint azon $\hat{\boldsymbol{\beta}}$ becslésre lesz minimális az $\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$ érték, amely megoldása az

$$\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

lineáris egyenletrendszernek. Ha $\boldsymbol{X}$ teljes rangú mátrix, akkor $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$.

3.1.2. Főkomponens-analízis (PCA)

Ez az eljárás egy többváltozós statisztikai módszer adatok tömörítésére. Működésének lényege, hogy az adott adathalmaz dimenzióját lecsökkenti, miközben megtartja a változók közt fennálló lehetséges varianciát úgy, hogy ezeket a változókat lineárisan korrelálatlan változókká - – főkomponensekké – alakítja át.

Mi ezt a módszert a kiválasztott játékosok támadó adatainak sokaságára fogjuk alkalmazni, hogy megtaláljuk, milyen súlyozással érdemes rájuk tekinteni, melyik lineáris kombináció magyarázza lehető legjobban a fizetést.

Az elméleti aszimptotikus eredmények belátásához és egyszerűség kedvéért fel szokták tenni, hogy az adatok normális eloszlásúak, de számos esetben erre nincsen szükség, és elegendő az adatok eloszlásának szimmetrikus voltát megkövetelni. Legyen X n dimenziós normális eloszlású valószínűségi vektorváltozó 0 várható értékkel, mivel ha nem lenne zérus a várható érték, azt levonva, 0 várható értékű eloszlást kapnánk. Tehát a továbbiakban

$$X \sim N_n(0, \Sigma), \qquad \text{ahol } \Sigma \text{ pozitív definit és szimmetrikus mátrix.}$$

Jelölje a kovarianciamátrix spektrálfelbontását $\Sigma = C\Delta C^T$, ahol Δ diagonális mátrix, a főátlóban a λ_i sajátértékekkel, C pedig az oszlopaiban a megfelelő sajátértékekhez tartozó c_i ortonormált sajátvektorokat tartalmazó mátrix. Mivel Σ pozitív definit, így feltehetjük, hogy $\lambda_1 \geq \ldots \geq \lambda_k \geq \lambda_{k+1} \geq \ldots \geq \lambda_n > 0$.

Célunk, hogy n-nél kevesebb, k dimenziós valószínűségi vektorváltozóba sűrítsük a rendelkezésünkre álló információkat. A feladatot tehát felírhatjuk olyan formában, hogy

$$E||X - AY||^2 \longrightarrow \min_{A,Y},$$

ahol A egy $(n \times k)$-as mátrix, Y pedig a keresett $k < n$ dimenziós főkomponensvektor.
Legyen C_k a C mátrix első k db legnagyobb sajátértékéhez tartozó ortonormált oszlopaiból álló mátrix. Ennek segítségével felírható a főkomponens-elemzési feladat megoldása:

$$A = C_k \quad Y = C_k^T X.$$

A megoldáshoz vezető matematikai levezetések megtalálhatók a Bolla and Krámli [2005], Móri [1999][2] és a Móri and Székely [1986] forrásokban.

Problémát jelenthet, ha az eloszlás ferde (itt az egyes magyarázóváltozókra gondolunk), ezért érdemes lehet ferdeségvizsgálatot végezni. Megjegyezzük még, hogy az eljárás nem skálainvariáns, azaz a mértékegység megváltoztatásával változhatnak a főkomponensek. Nem mindegy tehát, hogy a kovarianciamátrixszal, vagy a korrelációs mátrixszal dolgozunk.

A magyarázóváltozóink között találhatóak olyanok, melyek nagy értéke a játékosok teljesítményéhez negatívan járulnak hozzá. Ez alatt például a Strikeoutokat értjük, amiből természetesen, ha sok van, az azt vonja maga után, hogy az egyénünk gyengén szerepelt. Annak érdekében tehát, hogy a főkomponens-analízis jól működjön, érdemes nem az egyes "negatív" változók értékeit venni, hanem a megfigyelt értékek közül a maximálisat venni, majd abból kivonni az adott értékeket a megfelelő játékosoknál. E szerint ha például Manny Ramireznek (Tampa Bay Rays) a 2011-es szezonban mindössze 25 strikeoutja volt, míg Mark Reynolds (Cleveland Indians) 198-at tudhatott magáénak, akkor az általunk használt értékek: Manny Ramirez – $198 - 25 = 173$,
Mark Reynolds – $198 - 198 = 0$.

3.2. Szimulációs eredmények

A felhasznált adatokat Sean Lahman amerikai újságírónak köszönhetjük, aki rögzítette az összes 1871 és 2014 közötti játékos és csapat támadó és védekező teljesítményét, fizetéseiket, születési és visszavonulási dátumaikat, és még sok hasznos információt
`http://www.seanlahman.com/baseball-archive/statistics/`,
valamint segítségünkre szolgált Jim Albert egy későbbi cikke is (Albert [2010]).

Először szeretnénk választani egy olyan becslési módszert, ami vélhetően jól méri a teljesítményt. Kipróbáltunk néhányat a már fentebb részletezett mennyiségek közül,

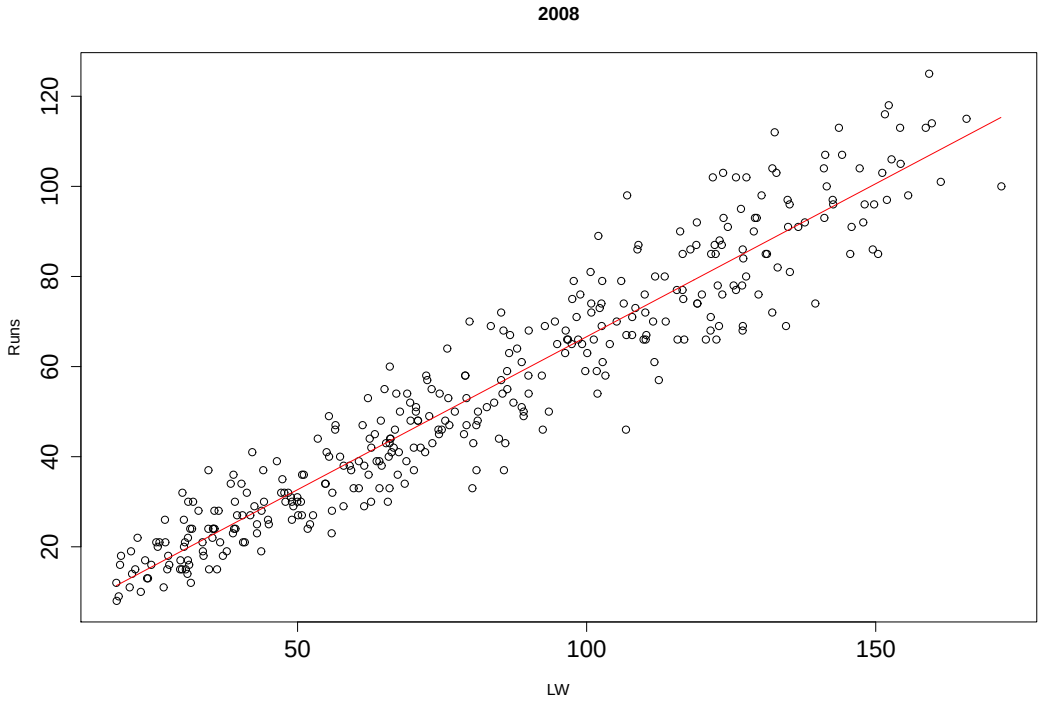
[2]Elérhetősége: `http://www.cs.elte.hu/~mori/faktor.pdf`

és jó választásnak tűnik a sima Linear Weights (LW) értéke. Bár a súlyok szezonról szezonra változnak, hiszen egy-egy adott évre lettek kiszámítva, az akkori adatokkal végzett szimuláció segítségével, azt tapasztaljuk, hogy egyes évek eltérései elhanyagolhatóan kicsik (1994 és 1996 között például az 1B érték súlya a következőképp változott: 0.489, 0.496, 0.492).

$$LW = 0.46 \cdot 1B + 0.8 \cdot 2B + 1.02 \cdot 3B + 1.4 \cdot HR + 0.33 \cdot (BB + HBP)$$

Thorn and Palmer [1985] által készített súlyok.

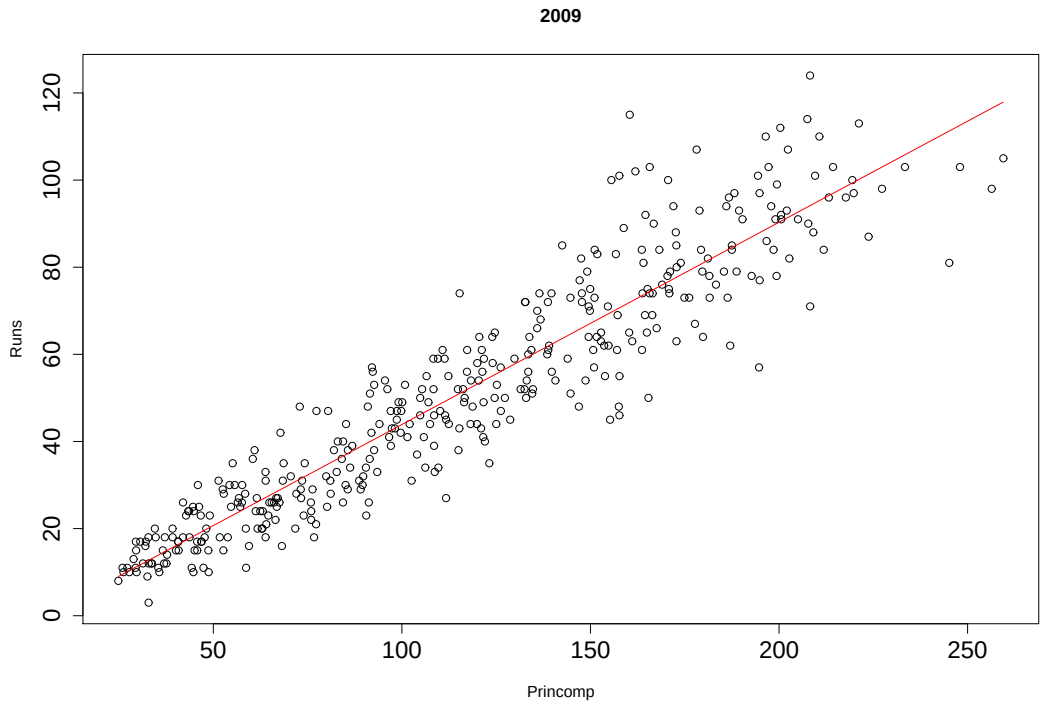
Tapasztalataink azt mutatják, hogy ezen súlyozással számolva az ütési adatok kifejezettem jól magyarázzák a pontok számát.



1. ábra. Pontok számának magyarázása LW segítségével a 2008-as idényben

Az 1. ábra az egyes játékosok által a 2008-as szezonban elért Runs-ok számát és az LW értékét mutatja meg. Minden pont egy-egy játékosnak felel meg. A piros vonal az illesztett regressziós egyenes. A lineáris modell determinációs együtthatója 0.91, ami valóban erős kapcsolatot bizonyít.

A következő, 2. ábrán is kitűnően látszik az előbb tapasztalt erős kapcsolat, de itt a saját együtthatóinkat használtuk fel, amiket a főkomponens-analízis szolgáltatott, immáron a 2009-es szezonban. Ebben ez esetben 0.84-as értéket kaptunk a determinációs együtthatóra, így ezzel a modell magyarázóereje 7%-ban elmarad az előzőétől. Ez annak fényében nem mondható jelentősnek, hogy az eljárás igencsak egyszerű, mégis



2. ábra. Pontok számának magyarázása főkomponens-analízis segítségével

majdnem olyan jó, mint a baseball-szakirodalomban évtizedeken keresztül kifejlesztett, bonyolult szimulációs technikákon alapuló LW-s módszer.

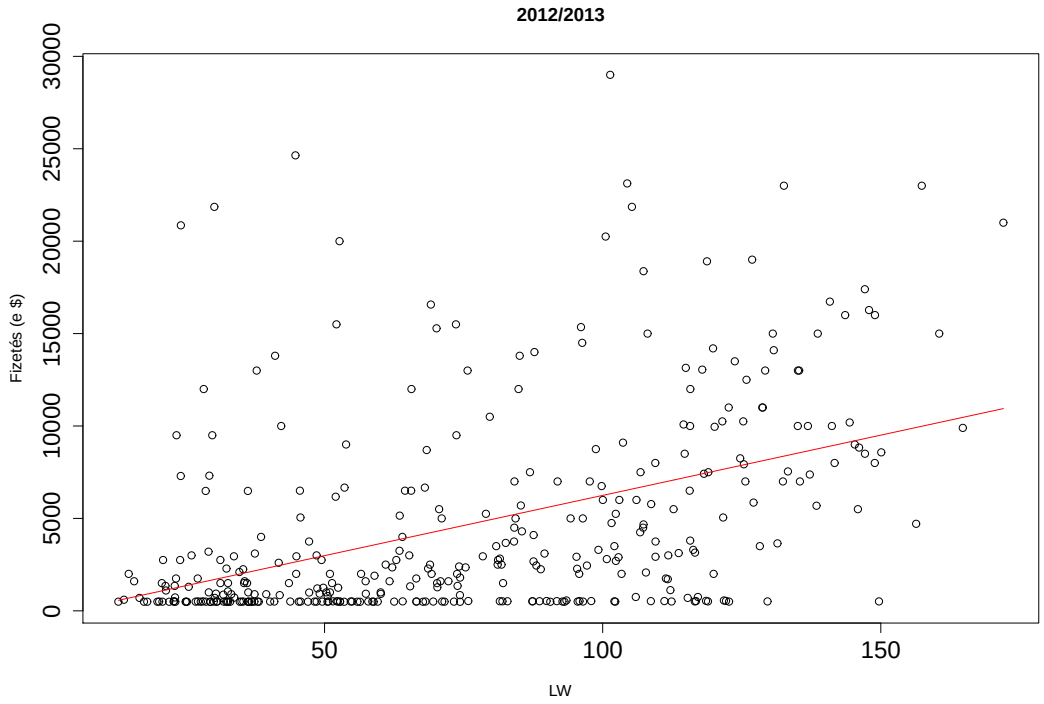
A 1. táblázat tartalmazza az adatbázisból rendelkezésünkre álló öt legfrissebb szezonra az LW és a PCA módszerrel előálló értékek mint magyarázóváltozók magyarázóerejét a pontokra vonatkozóan (lineáris kapcsolatot feltételezve).

	LW	PCA
2008	0.910	0.841
2009	0.924	0.871
2010	0.916	0.861
2011	0.908	0.834
2012	0.919	0.838

1. táblázat. A lineáris modellek magyarázóereje különböző évekre (Runs)

A tapasztalt értékek mind elég magasak, ennek fényében bátran használhatjuk mind az LW mutatót, mind a főkomponensek által készített súlyozást, ezúttal már a fizetések magyarázására. Nyilván azt feltételezhetjük, hogy a sportolók fizetésüket a teljesítményük alapján kapják, legalábbis főrészt. Azt szeretnénk látni, hogy ez valóban így van-e. Mivel a játékosok minden év elején, a bajnokság kezdetekor kapják kézhez a pénzüket, ezért figyelnünk kellett, hogy ha mondjuk a 2013-as fizetéseket magyarázzuk, akkor

fontos, hogy a 2012-es teljesítményeket használjuk fel.



3. ábra. Fizetés magyarázása LW segítségével (2012/2013)

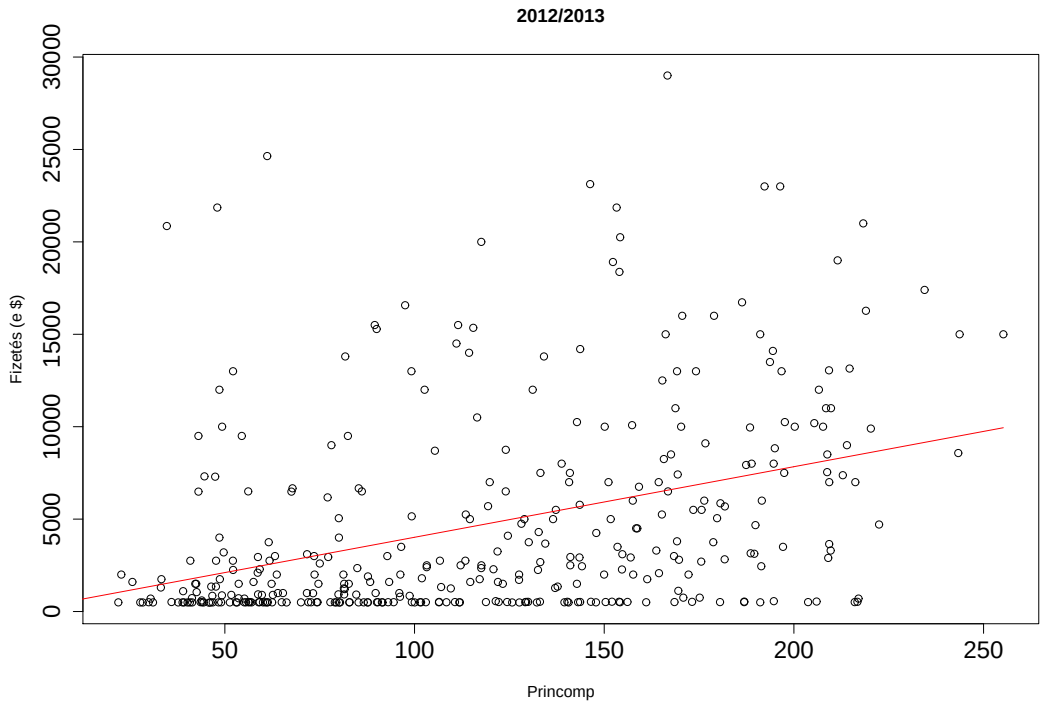
A 3. és 4. ábrák a 2012/2013-as szezonra vonatkoznak. A függőleges tengelyen a fizetést látjuk, a vízszintesen pedig sorban az LW, majd a PCA súlyozást. Jól látszik, hogy itt már nem található meg az az erős kapcsolat a magyarázó és magyarázott változók között.

A fizetésre vonatkozó eredményeket a 2. táblázat tartalmazza, amely megmutatja,

	LW	PCA
2008/2009	0.187	0.145
2009/2010	0.208	0.180
2010/2011	0.279	0.222
2011/2012	0.265	0.221
2012/2013	0.202	0.145

2. táblázat. A lineáris modellek magyarázóereje különböző évekre (fizetés)

hogy a rendelkezésünkre álló öt legfrissebb szezonra az LW és a PCA módszerrel előálló értékek mint magyarázóváltozók magyarázóerejét a fizetésekre vonatkozóan (lineáris kapcsolatot feltételezve). Nem túl meglepő módon jelentősen gyengébb összefüggést tapasztalhatunk, mint a pontok magyarázásánál.



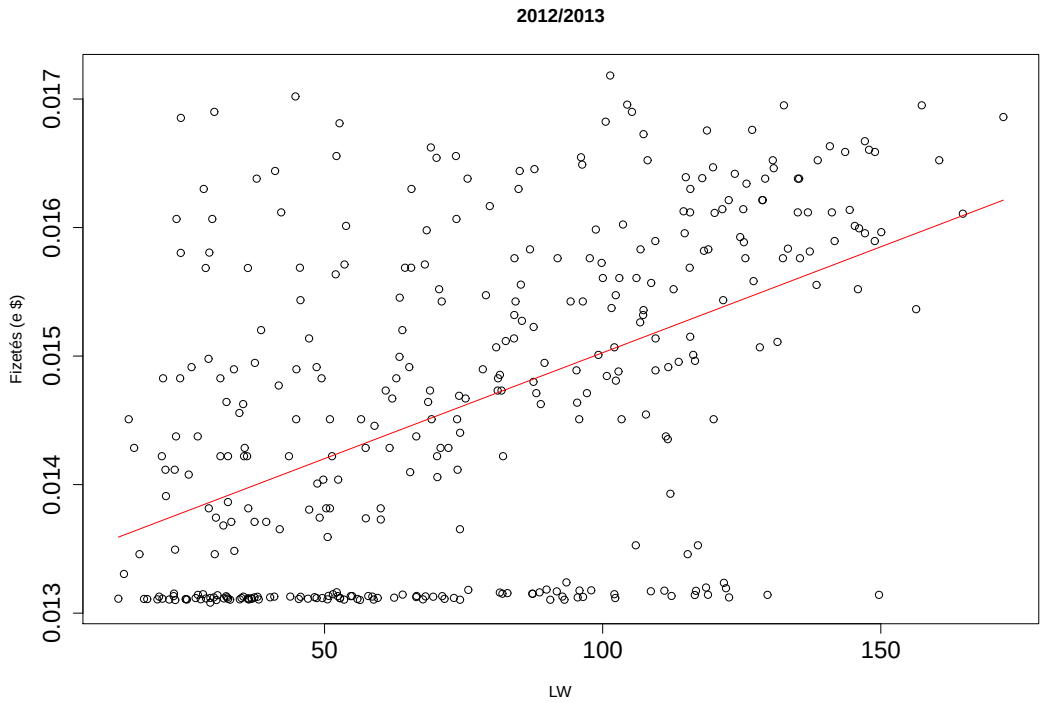
4. ábra. Fizetés magyarázása az első főkomponens segítségével (2012/2013)

A fizetésekben megjelenő nagyságrendek miatt érdemes ezen értékek logaritmusával is megnézni ugyanezt a modellt, hátha jobb eredmény kapunk. A modellek az 5. és 6. ábrákon láthatók. Ennek a próbálkozásnak eredménye, hogy az LW esetén kapott 0.202 determinációs együttható értéke 0.250-re, valamint a PCA esetében 0.145-ről 0.189-re nőtt.

Manapság vita tárgya, hogy egyes sportolók megfelelő összeget kapnak-e az alapján, amit produkálnak a pályán. Sokan állítják, hogy a fizetés és a teljesítmény nem függ össze, hiszen bizonyos játékosok már csak a "nevükért" is horribilis összegeket kapnak, miközben nem tesznek le annyit az asztalra, mint amennyi elvárható volna tőlük.

A legnépszerűbb amerikai sportok (kosárlabda, jégkorong, amerikai futball és baseball) közül a másik háromban úgynevezett fizetési sapka (salary cap) van érvényben. Ez korlátozza az egyes csapatok kiadásait a játékosok fizetésére nézve, így jobban kiegyenlítve a versenyt és az átigazolásokért vívott harcot. Ezzel szemben a baseballban ilyen szigorú megkötés nincs. Ezt a korlátot helyettesíti a fizetési adó (luxury tax), amely bünteti a sokat költekező csapatokat. Itt it meg van szabva egy korlát, ameddig büntetlenül költekezhetnek. Ezt átlépve azonban az első évben a többlet 17.5, második évben 30, harmadik évben 40, majd minden ezt követő évben 50 százalékát kell kifizetniük a ligának, amit előre meghatározott célokra fordítanak.

A baseball rajongók körében közismert az Oakland Athletics esete. A fizetési sapka



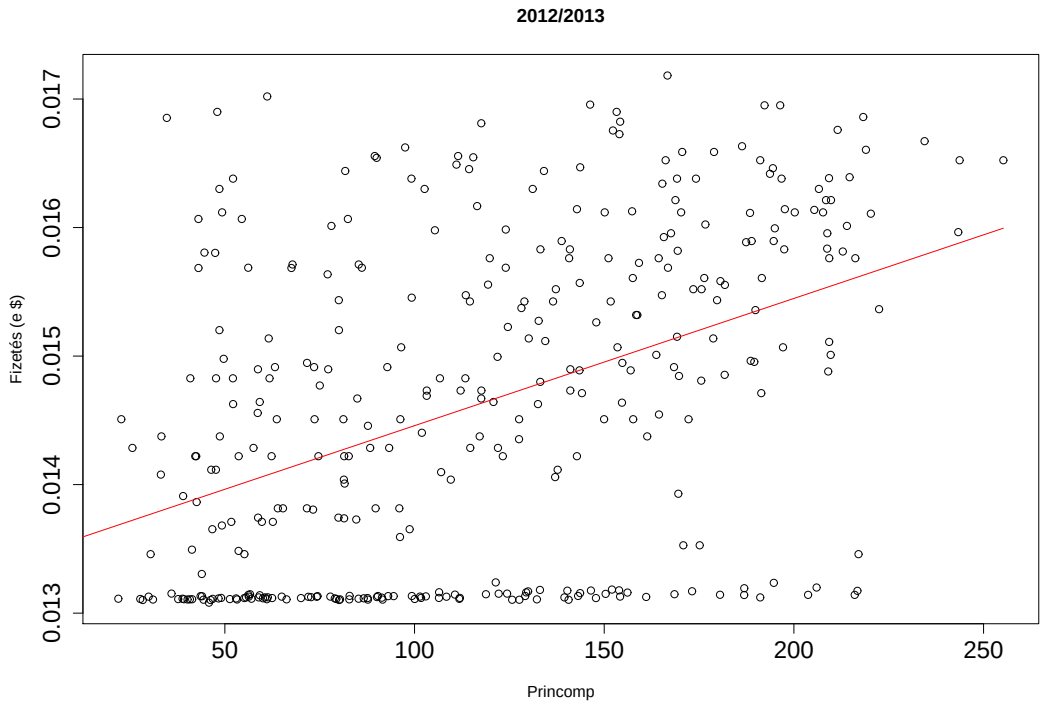
5. ábra. A fizetés logaritmusának magyarázása LW segítségével (2012/2013)

hiányában a gazdag csapatok a legtöbb tehetséges játékosra lecsaptak, így nem maradt nagy választék a liga szegényebb csapatai számára, mint például az Oakland. Illetve, ha maradt is volna, nem engedhették volna meg maguknak a kiadásokat. Egy tehetséges és jó gondolkodású general manager segítségével azonban sikerült olyan játékosokat vásárolniuk, akik megfelelően teljesítenek adott helyzetekben, és az áruk is megengedhető. A csapat 2002-ban először nyerte meg az alapszakaszt az amerikai liga nyugati régiójában, illetve híressé vált húsz egymást követő mérkőzés megnyerése miatt. Ezen is meglátszik, hogy mennyire nincsenek egyensúlyban a csapatok az amerikai baseball ligában.

Egyelőre csak a Boston Red Sox és a New York Yankees csapatai lépték át ezt a küszöböt. A bostoniak hatszor, míg a Yankees már tizenegyszer, vagyis a 2003-as érvénybe lépés óta minden évben. Ennek következtében közel 254 millió dollárt fizettek ki adósságként. Gyakran végeznek a bajnokság elején az előbbi csapatok, a rendelkezésükre álló pénzmennyiségből kifolyólag.

A fent említettek miatt érdemes lehet megvizsgálni, hogy különböző jövedelmekkel rendelkező csapatokon belül hogyan alakul a teljesítmény tükröződése az egyéni fizetésen. Ezt azért mondhatjuk, mert amikor az összes játékost vizsgáljuk, szerepet játszik az a tény, hogy egy több pénzzel rendelkező csapat arányaiban jobban meg tudja fizetni a tagjait, még ha azok nem is szerepelnek olyan kiemelkedően.

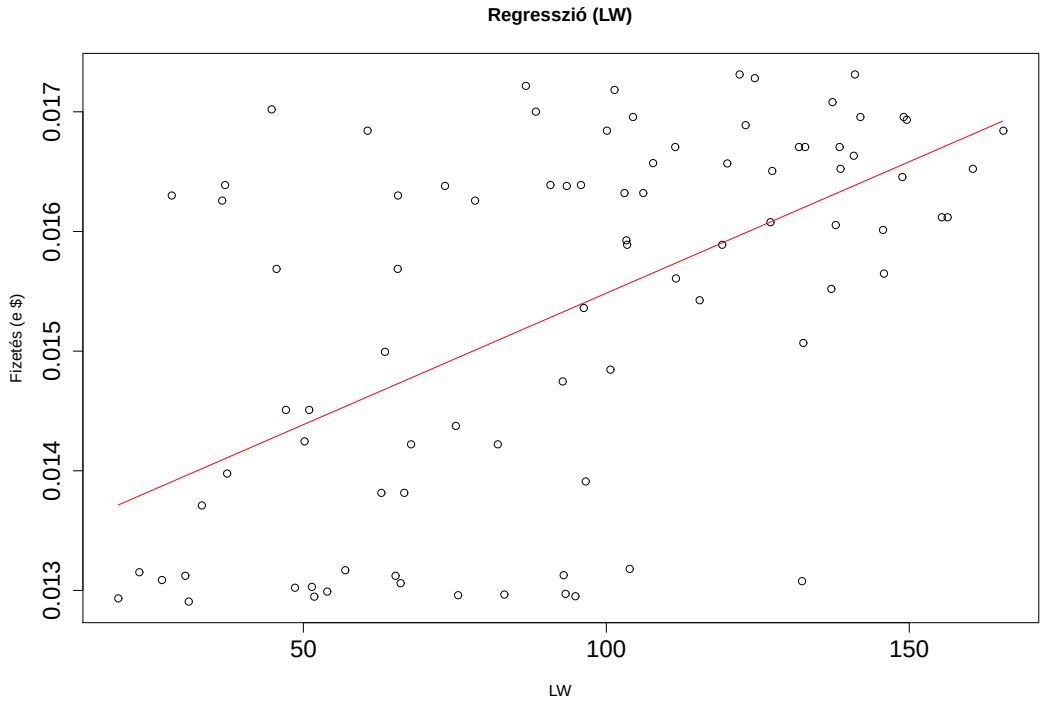
Példának hozzuk a 2013-as szezon legtöbb pénzzel rendelkező csapatát, a New York



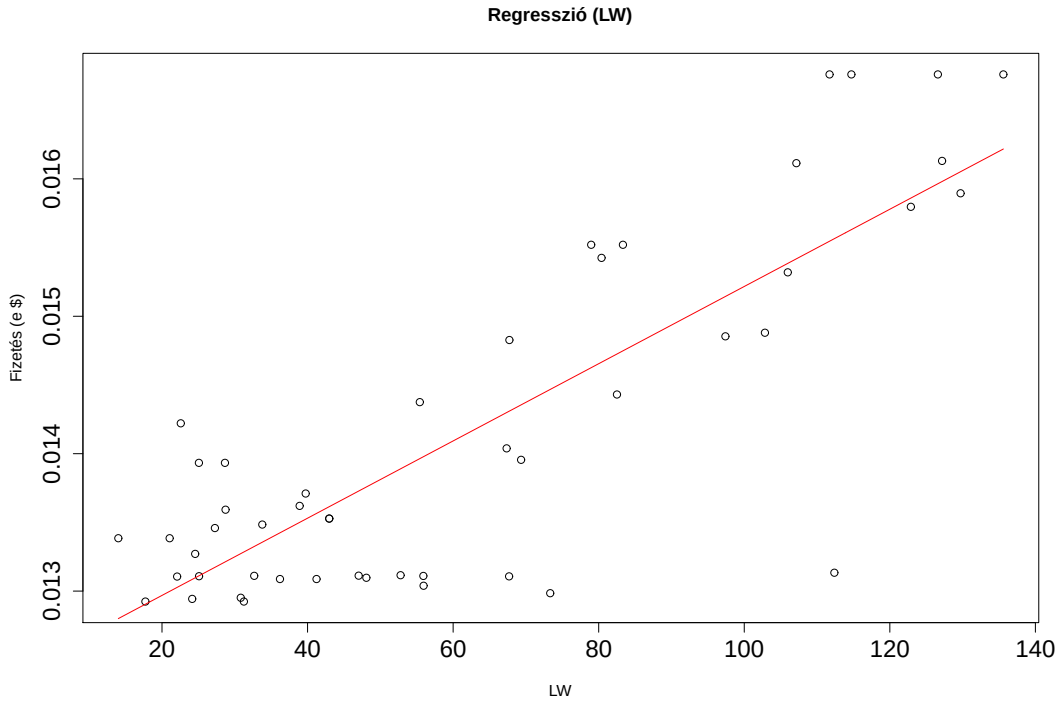
6. ábra. A fizetés logaritmusának magyarázása PCA segítségével (2012/2013)

Yankees-t (228,995,945 dollár), valamint az ugyanazon év legkevesebb bevételével bíró Houston Astros-t (24,328,538 dollár). A 7. ábra tartalmazza a New York Yankees-nél a 2008-tól 2012-ig terjedő időszak fizetéseit, valamint az LW értékét az egyes játékosokra vonatkozóan. A piros vonal az illesztett lineáris regressziós egyenes, a pontok pedig egy-egy játékos valamely évre vonatkozó teljesítményi mutatói.. A 8. ábra ugyanezt mutatja a Houston Astros csapat esetén. A lineáris regressziós modellből adódó determinációs együttható a New York esetén 31%, míg a Houstonnál 65%. Ebben az elsőre meglepő eredményben az is közrejátszik, hogy a sportemberek fizetésüket nem csak a pályán való szereplésükért, de a márkaértékükért, illetve csapatuk bevételének növeléséhez egyéb módon történő hozzájárulásukért is kapják (mezek, ütők, különböző, saját nevükkel ellátott termékek eladása által szerzett bevétel, médiaszereplések)

Phil Birnbaum amerikai kutató szerint a tapasztalt összefüggések, és relatíve gyenge korreláció nem zárja ki a sportolók játéka és a fizetésük közötti kapcsolatot. Szerinte tévesen vonják le azt a következtetést, hogy a sportolók játéka nem lesz hatással a fizetésre, és egy szezon alapján nem érdemes bírálni a sportolókat. Bemutat egy kitűnő példát is (Birnbaum [2014]):
Francisco Cervelli (New York Yankees – elkapó) +0.8-as értékelésre (WAR – Wins Above Replacement: azt mutatja, hogy hány győzelemmel többel járult hozzá valaki a csapatának győzelmeihez, mint ha egy minimális költségű játékos szerepelt volna helyette) 523.000 dollárt kapott a 2013-as szezonban, míg CC Sabathia (NYY – dobó)



7. ábra. New York Yankees, 2008-2012



8. ábra. Houston Astros, 2008-2012

+0.3-as pontszáma 24.7 millió dollárt hozott neki. Cervelli ebben a szezonban a megszokottnál (az előzőekben átlagosan +0.3 WAR) jobban teljesített valamilyen oknál kifolyólag, így elérve a +0.8-as értéket. Ezzel szemben Sabathiaról érdemes tudni, hogy 2009-ben nagyot nőtt a fizetése, miután több évig stabilan tartotta a +4-es WAR szintet, majd a 2010-2012-es időszakban meglepően magas, +15.6-os átlagot produkált. Sajnálatos módon kora a 2013-as szezonban váratlanul erősen kihatott játékára, így esett vissza a már említett 0.3-ra.

4. Az életkor hatása a teljesítményre

Ebben a fejezetben a játékosok teljesítményét fogjuk magyarázni a korukkal, felhasználva Jim Albert [2002] tapasztalatait.

Természetesnek tűnik az az elgondolás, hogy egy sportoló a profi ligába kerülve, fiatalon még nem teljesít úgy, mint a profik. Az évek haladtával azonban hozzászokik az ütemhez, belerázódik a játékba, erőnléte is javul, majd eléri a csúcspontját a pályafutásának. Ugyanakkor a lassacskán öregedő játékosok már nem olyan gyorsak, fittek, mint voltak, a reakcióidejük is romlik, és nem várható nagyobb pozitív irányú ugrás a teljesítményüket illetően.

A továbbiakban megpróbálunk olyan modellt készíteni, amely igazolja az előző bekezdésben írt intuíciónkat. Ehhez lesz szükségünk a Bayes-becslésre. Elképzelhető, hogy egy-két, különböző okokból kiugró adat teljesen elrontja az illeszkedést, ezért mindenekelőtt adattisztítást kell végrehajtani, azaz a hibás adatokat kijavítani, végső esetben pedig törölni.

4.1. Korábbi eredmények és módszerek

Jim Albert 2002-ben írt cikkében (Albert [2002]) megvizsgálja több híresebb baseball játékos karrierjét, összeveti őket különböző regressziós modellek segítségével. Mivel nyilvánvaló módon nem korrekt összehasonlítani egy 25 évest egy 40 évessel, Albert közel azonos korú játékosokat hasonlít csak össze. Mindig adott évtizeden belül születetteket vizsgál, illetve ha összeveti két sportoló teljesítményét, azt az azonos életévükben teszi, azaz az egyik 30. évét a másik 30. évével hasonlítja össze.
Albert főleg híres játékosokat véleményez, így a legtöbb olvasó ismerheti őket. Ilyenek például Mickey Mantle, Sammy Sosa, Barry Bonds, vagy Joe DiMaggio. Ők koruk kiemelkedő személyei, egy-egy velük készített diagram kimondottan látványos eredményt mutat.

Elképzelését már fentebb mi is leírtuk. Kisebb különbség van az ő bayesi megközelítése és az általunk említett között. Albert egy szinttel mélyebbre megy. Ezt úgy értjük, hogy véleménye szerint érdemes nem csak az eloszlás paraméterét valószínűségi változónak tekinteni, de a paraméter eloszlásában szereplő ismeretlen paramétert is. Mi ennyire nem merülünk el ebben a témában, ahogy azt majd a későbbiekben le is írjuk.

Másik forrásunk (Albert and Bennett [2003]) inkább pusztán a teljesítmény mérésével foglalkozik. A lehető legjobb lineáris becslést keresik regresszió segítségével, a szimulációk helyett. A szerzőpár felsorakoztat néhány kiváló megközelítést a megszerzett

pontok (Runs) magyarázására, méghozzá csapatokra nézve. Kezdetnek a TA – Total Average, azaz teljes átlag értékből indulnak ki. Ez gyakorlatilag az összes megszerzett bázis száma, leosztva a kiesések (Outs) számával. Ennek megfelelően az egyes ütések súlya egész egyszerűen az általuk nyert bázisok száma. Tehát $1B \sim 1,\ 2B \sim 2,\ 3B \sim 3,$ illetve $HR \sim 4$.

$$TA = \frac{1 \cdot 1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR + 1 \cdot (BB + HBP)}{AB - H + CS + GIDP}$$

Kezdetnek átalakították ezt a formulát, és nem a kiesések számával, osztanak le, hanem az egy szezonban lejátszott mérkőzések számával. Emögött az a logika húzódik, hogy, mivel minden mérkőzésen közel 27 darab out várható (9 inning, mindegyik 3-3 kiesésig tart) csapatonként, így közel azonos számot kaptak, viszont nem kellett az apróbb részletekkel foglalkozni, hogy honnan adódott egy-egy out. Ezután legkisebb négyzetek módszerrel keresték meg a megfelelő együtthatókat a különböző eseményekhez, így az adott változókra a lehető legjobb súlyozást kapták. Később összehasonlítottak több teljesítménymutatót, hatásosságuk alapján. Megjegyezzük, hogy mi ezt nem tettük meg, saját ízlésünk szerint választottunk egyet a sok közül a későbbi vizsgálódásokhoz.

Említést érdemel még a Koop [2002] cikk, amelyben az egyes játékosok teljesítményét koruk legjobbjának teljesítményéhez mérték, ezzel a megközelítéssel szakdolgozatomban nem foglalkozunk.

4.2. Heteroszkedasztikus négyzetes regresszió

Az intuíció alapján a játékosok teljesítményét a korukkal négyzetesen szeretnénk magyarázni, ezért modellnek a következő négyzetes regressziót fogjuk használni:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

ahol y eredményváltozó jelöli az adott játékos teljesítményét (pl. ALW), x magyarázóváltozó az életkorát, valamint x^2 a kor négyzetét , $\beta_0, \beta_1, \beta_2$ a keresett regressziós együtthatókat, ε pedig a véletlen hibát. Hagyományos regressziós modellekben feltesszik, hogy a hiba 0 várható értékű normális eloszlású, valamilyen ismeretlen σ szórással. A baseball szakértői megfigyelték, hogy a teljesítmény szórása fordítottan arányos a játékosok ütéshez állásainak (Plate Appearance) számával, ezért a továbbiakban mi is ezzel a feltétellel fogunk élni: $\varepsilon \sim N\left(0, \frac{\sigma^2}{m_i}\right)$. Vezessünk be néhány jelölést:

$$\mathbf{y} = (y_i, \ldots, y_n)^T$$
$$\mathbf{x} = (x_i, \ldots, x_n)^T$$
$$\mathbf{m} = (m_i, \ldots, m_n)^T$$

A már fentebb említett modellünk a következő alakot ölti:

$$\mathbf{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Ebben a felírásban $\mathbf{y}_i$ jelöli az i. évben a teljesítményt, mint magyarázott változót, $\beta_0, \beta_1, \beta_2$ sorra a regressziós együtthatók, x_i pedig a játékos kora az i. szezonban. Néhány alkalmas jelölés segítségével a modell könnyen láthatóan az általános lineáris modell alakjába írható:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Ebből teljes rangú esetben a β-k becslésére adódik a következő:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{y} \quad \text{amit tovább írva}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \\ x_1^2 & \dots & x_n^2 \end{pmatrix} \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} = \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum y_i & \sum x_i y_i & \sum x_i^2 y_i \end{pmatrix}^T$$

kifejezések adódnak. Az ε hibavektorról szokás feltenni, hogy normális eloszlásból származik, 0 várható értékkel és valamilyen $\frac{\sigma^2}{m_j}$ szórásnégyzettel, ahol az m_j a játékos j-edik szezonjának PA mutatója.

A mintánk tehát a következő: (x_i, y_i, m_i), ahol $i = 1, \dots, n$.

Ennek következtében

$$\mathbf{Y}|\mathbf{X} \sim N\left(\beta_0 + \beta_1\mathbf{X} + \beta_2\mathbf{X}^2, \frac{\sigma^2}{m_i}\right)$$

A likelihood függvény az alábbi:

$$
\begin{aligned}
& f_{Y_1,\ldots,Y_n|X_1,\ldots,X_n}\left(y_1,\ldots,y_n|x_1,\ldots,x_n\right) = \\
&= \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} \left(\frac{\sigma}{\sqrt{m_i}}\right)^{-1} \exp\left\{-\frac{1}{2}\frac{\left[y_i-(\beta_0+\beta_1 x_i+\beta_2 x_i^2)\right]^2}{\left(\frac{\sigma}{\sqrt{m_i}}\right)^2}\right\} = \\
&= (2\pi)^{-\frac{n}{2}}\sigma^{-n}\prod_{i=1}^{n} m_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} m_i\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right]^2\right\},
\end{aligned}
$$

és ezt szeretnénk maximalizálni $\beta_0, \beta_1, \beta_2, \sigma$ szerint. Ennek logaritmusaként kapjuk a log-likelihood függvényt:

$$
\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = C - n\log\sigma + \frac{1}{2}\sum_{i=1}^{n}\log m_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} m_i\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right]^2
$$

Ezt a kifejezést deriválva a négy ismeretlen paraméter szerint, majd az egyenleteket nullára rendezve keresünk lehetséges szélsőértékhelye(ke)t:

$$
\partial_{\beta_0}\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} m_i\cdot 2\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right](-1) = 0
$$

$$
\partial_{\beta_1}\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} m_i\cdot 2\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right](-x_i) = 0
$$

$$
\partial_{\beta_2}\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} m_i\cdot 2\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right](-x_i^2) = 0
$$

$$
\partial_{\sigma}\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} m_i\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right]^2 = 0
$$

Tegyük fel, hogy $\sigma > 0$. Az egyenletrendszer egyszerűbb alakban is írható, ha az egyes konstansoktól eltekintünk:

$$
\begin{aligned}
&\sum_{i=1}^{n} m_i\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right] = 0 \iff \\
&\qquad \iff \sum_{i=1}^{n} m_i y_i = \beta_0\sum_{i=1}^{n} m_i + \beta_1\sum_{i=1}^{n} m_i x_i + \beta_2\sum_{i=1}^{n} m_i x_i^2
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i=1}^{n} m_i\left[y_i-\left(\beta_0+\beta_1 x_i+\beta_2 x_i^2\right)\right]x_i = 0 \iff \\
&\qquad \iff \sum_{i=1}^{n} m_i x_i y_i = \beta_0\sum_{i=1}^{n} m_i x_i + \beta_1\sum_{i=1}^{n} m_i x_i^2 + \beta_2\sum_{i=1}^{n} m_i x_i^3
\end{aligned}
$$

$$\sum_{i=1}^{n} m_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] x_i^2 = 0 \iff$$

$$\iff \sum_{i=1}^{n} m_i x_i^2 y_i = \beta_0 \sum_{i=1}^{n} m_i x_i^2 + \beta_1 \sum_{i=1}^{n} m_i x_i^3 + \beta_2 \sum_{i=1}^{n} m_i x_i^4$$

Az egyenletrendszert az alábbi mátrixos alakra lehet írni:

$$\underbrace{\begin{pmatrix} \sum m_i & \sum x_i m_i & \sum x_i^2 m_i \\ \sum x_i m_i & \sum x_i^2 m_i & \sum x_i^3 m_i \\ \sum x_i^2 m_i & \sum x_i^3 m_i & \sum x_i^4 m_i \end{pmatrix}}_{=:\mathbf{X}'} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \underbrace{\begin{pmatrix} \sum y_i m_i \\ \sum x_i y_i m_i \\ \sum x_i^2 y_i m_i \end{pmatrix}}_{=:\mathbf{y}'}$$

Ezek után, a lineáris modellnél láthatóak mintájára a becslésünk így írható fel:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$$

Most már megvannak a β-k becslései, ezeket felhasználva kaphatjuk meg a σ becsült értékét. A negyedik egyenletben σ^3-nal beszorozva (feltettük, hogy $\sigma > 0$), majd átrendezve a következő adódik:

$$\hat{\sigma} = \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} m_i \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2\right)\right]^2}$$

Könnyen lehet igazolni, hogy a regressziós együtthatók klasszikus, legkisebb négyzetes becslése és a maximum likelihood becslés ezzel a kis módosítással is megegyezik. A két módszer ekvivalenciája abban látható, hogy a maximum likelihood-módszernél a log-likelihood függvény maximalizálása megegyezik az exponensében található kifejezés minimalizálásával, ugyanis a

$$\prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} \left(\frac{\sigma}{\sqrt{m_i}}\right)^{-1} \exp\left\{-\frac{1}{2} \frac{\left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right]^2}{\left(\frac{\sigma}{\sqrt{m_i}}\right)^2}\right\}$$

kifejezés logaritmusának β-k szerinti maximalizálása, azaz

$$C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} m_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right]^2 \to \max_{\beta_0, \beta_1, \beta_2}$$

ekvivalens azzal, hogy

$$\sum_{i=1}^{n} m_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right]^2 \to \min_{\beta_0, \beta_1, \beta_2}$$

4.3. Bayes-becslés és négyzetes regresszió

Matematikai statisztikában a minta ismeretlen paramétereinek becslésére három leggyakrabban alkalmazott paraméterbecslési eljárás a maximum likelihood, a momentum-módszer, valamint a bayes-i megközelítéssel számított ún. Bayes-becslés. A bayesi hozzáállás lényege és egyben a legfőbb eltérés az előző kettőhöz képest, hogy a minta eloszlásának paraméterét is egy valószínűségi változónak tekintjük. Tehát nem egy konkrét valós számot keresünk, hanem a véletlen is szerepet játszhat a paraméter értékében. A normális eloszlás magasabb dimenziós általánosítására és egy állításra lesz szükségünk a négyzetes regresszió paramétereinek korrigálásához.

Az $\mathbf{X}$ valószínűségi változó d dimenziós normális eloszlást követ m várható érték vektorral és $\Sigma = A^T A$ kovarianciamátrixszal, ha $\mathbf{X} = A\mathbf{Y} + \mathbf{m}$, ahol $\mathbf{Y} = (Y_1, \ldots, Y_k)^T$, $Y_1, \ldots, Y_k$ függetlenek és standard normális eloszlásúak, $A \in \mathbb{R}^{d \times k}$ valamint $m \in \mathbb{R}^d$. Jelölése: $\mathbf{X} \sim N_d(\mathbf{m}, \Sigma)$. Amennyiben $\Sigma > 0$ (pozitív definit), akkor $\mathbf{X}$ sűrűségfüggvénye a következő alakot ölti:

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} |\det(\Sigma)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \Sigma^{-1} (\mathbf{x}-\mathbf{m})\right\}.$$

Állítás. Tegyük fel, hogy rendelkezésünkre áll egy n elemű $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ független minta d dimenziós, valamilyen ϑ ismeretlen várható értékű és ismert $\Sigma > 0$ kovariancia-mátrixú normális eloszlásból. Ekkor a ϑ paraméter Bayes-becslése $\vartheta \sim N_d(\mu, V)$ apriori eloszlás esetén a következő:

$$\hat{\vartheta}_B = (\Sigma^{-1} + V^{-1})^{-1} \left(\Sigma^{-1} \sum_{i=1}^{n} \mathbf{X_i} + V^{-1}\mu\right).$$

Bizonyítás. Meg fogjuk állapítani a $\vartheta|\mathbf{X}$ feltételes eloszlást, amiből – négyzetes rizikófüggvényt feltételezve – azonnal adódni fog a Bayes-becslés.

$$\begin{aligned} f_{\mathbf{X}|\vartheta}(\mathbf{x}|\mathbf{t}) \cdot f_\vartheta(\mathbf{t}) = \prod_{i=1}^{n} (2\pi)^{-\frac{d}{2}} (\det(\Sigma))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x_i}-\mathbf{t})^T \Sigma^{-1} (\mathbf{x_i}-\mathbf{t})\right\} \cdot \\ \cdot (2\pi)^{-\frac{d}{2}} (\det(V))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{t}-\mu)^T \Sigma^{-1} (\mathbf{t}-\mu)\right\} = \\ = (2\pi)^{-\frac{d(n+1)}{2}} (\det(\Sigma)^n \cdot \det(V))^{-\frac{1}{2}} \cdot \\ \cdot \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n} (\mathbf{x_i}-\mathbf{t})^T \Sigma^{-1} (\mathbf{x_i}-\mathbf{t}) + (\mathbf{t}-\mu)^T \Sigma^{-1} (\mathbf{t}-\mu)\right]\right\} \end{aligned}$$

A hatványkitevő kapcsos zárójelében lévő kifejezést tovább alakíthatjuk:

$$\sum_{i=1}^{n} \mathbf{x_i}^T \Sigma^{-1} \mathbf{x_i} - \sum_{i=1}^{n} \mathbf{x_i}^T \Sigma^{-1} \mathbf{t} - \mathbf{t}^T \Sigma^{-1} \sum_{i=1}^{n} \mathbf{x_i} + \mathbf{t}^T \Sigma^{-1} \mathbf{t} +$$

$$+ \mathbf{t}^T V^{-1} \mathbf{t} - \mathbf{t}^T V^{-1} \mu - \mu^T V^{-1} \mathbf{t} + \mu^T V^{-1} \mu =$$

$$= \mathbf{t}^T (\Sigma^{-1} + V^{-1}) \mathbf{t} - \left(\sum_{i=1}^{n} \mathbf{x_i}^T \Sigma^{-1} + \mu^T V^{-1} \right) \mathbf{t} - \mathbf{t}^T \left(\Sigma^{-1} \sum_{i=1}^{n} \mathbf{x_i} + V^{-1} \mu \right) +$$

$$+ \sum_{i=1}^{n} \mathbf{x_i}^T \Sigma^{-1} \mathbf{x_i} + \mu^T V^{-1} \mu =$$

$$= \left[\mathbf{t} - (\Sigma^{-1} + V^{-1})^{-1} \left(\Sigma^{-1} \sum_{i=1}^{n} \mathbf{x_i} + V^{-1} \mu \right) \right]^T \left[\Sigma^{-1} + V^{-1} \right] \cdot$$

$$\cdot \left[\mathbf{t} - (\Sigma^{-1} + V^{-1})^{-1} (\Sigma^{-1} \sum_{i=1}^{n} \mathbf{x_i} + V^{-1} \mu) \right] -$$

$$- \left(\Sigma^{-1} \sum_{i=1}^{n} \mathbf{x_i} + V^{-1} \mu \right)^T (\Sigma^{-1} + V^{-1}) \left(\Sigma^{-1} \sum_{i=1}^{n} \mathbf{x_i} + V^{-1} \mu \right) +$$

$$+ \sum_{i=1}^{n} \mathbf{x_i}^T \Sigma^{-1} \mathbf{x_i} + \mu^T V^{-1} \mu$$

Ebből tehát azt látjuk, hogy az aposteriori eloszlás nem más, mint

$$\vartheta | \mathbf{X} \sim N_d \left((\Sigma^{-1} + V^{-1})^{-1} (\Sigma^{-1} \sum_{i=1}^{n} \mathbf{X_i} + V^{-1} \mu), (\Sigma^{-1} + V^{-1})^{-1} \right)$$

Innen pedig könnyen leolvasható a Bayes-becslés:

$$\hat{\vartheta}_B = E(\vartheta | \mathbf{X}) = (\Sigma^{-1} + V^{-1})^{-1} (\Sigma^{-1} \sum_{i=1}^{n} \mathbf{X_i} + V^{-1} \mu) \qquad \square$$

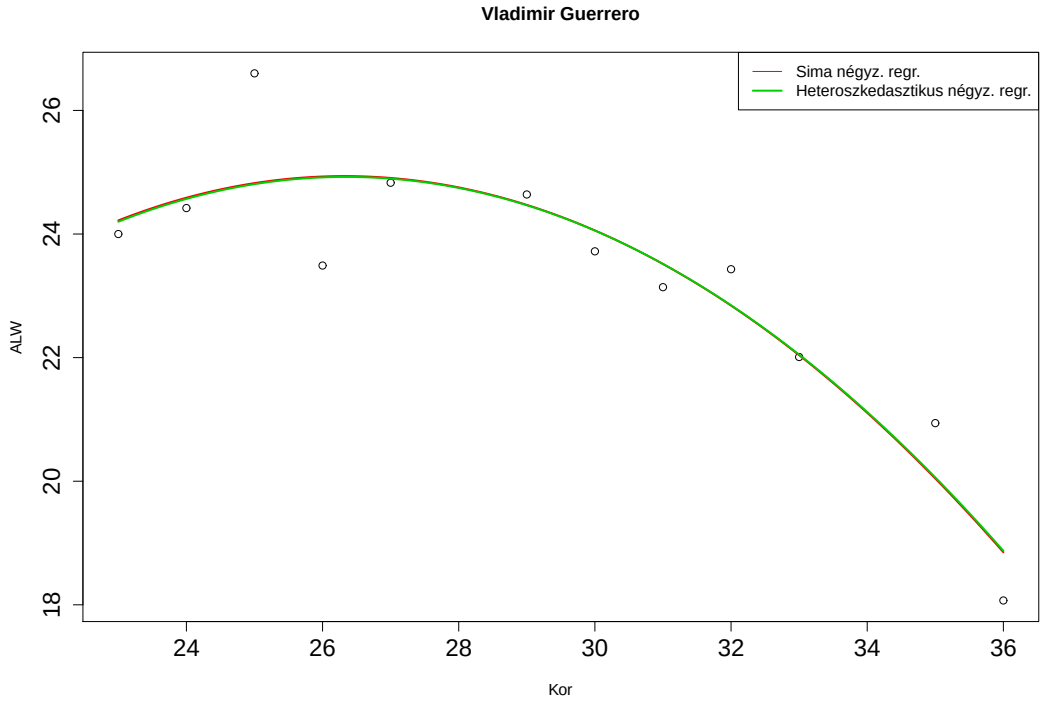
Az imént belátott állítást egyelemű mintára és 3 dimenziós normális mintára fogjuk alkalmazni (azaz $n = 1,\ d = 3$) – feltesszük, hogy az egyes játékosokra kapott becsült $\hat{\beta}_0, \hat{\beta}_1$ és $\hat{\beta}_2$ együtthatók függnek a véletlentől, mivel az emberi teljesítményt mindig sok-sok tényező befolyásolhatja, például a játékos hangulata, az időjárás, a csapatnál tapasztalható körülmények, a szakértők elvárásai. Feltesszük, hogy $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ normális eloszlást követ ismeretlen ϑ várható értékkel és ismert Σ kovarianciamátrixszal. Az ismeretlen ϑ paraméterről pedig azt tételezzük fel, hogy normális eloszlást követ μ várható értékkel és V kovarianciamátrixszal. Ebben a leírásban Σ mindig a megfelelő évtizedben kiszámított együtthatók kovarianciamátrixa, μ az összes együttható átlaga (azaz $\mu_0 = \frac{\beta_{0,1} + \ldots + \beta_{0,n}}{n}$), V pedig szintén az összes együttható kovarianciamátrixa.

4.4. Szimulációs eredmények

Vizsgálataink között szerepet kapnak a játékosok teljesítményi "görbéi" (vagyis a tényleges adatok alapján négyzetes regresszióval kapott görbék, ahol a kor függvényében tekintjük a teljesítmény változását), a becsült görbék (Bayes-becslés segítségével kapott, az elvárásainkhoz igazított görbék), heteroszkedasztikus regresszióval kapott görbék és néhány sportoló összehasonlítása.

Mint azt Jim Albert is csinálta, megnéztük, hogyan alakulnak a trajektóriák, ha a játékosoknak nem csak az LW értékeit nézzük, hanem azt az ütéshez állásaik számával leosztjuk. Mivel csak olyan játékosokat vizsgáltunk, akiknek egy-egy szezonban legalább 500 PA-juk (Plate Appearance) volt, mindenkinél közel azonos ez az érték, így nem várunk különösebb eltéréseket a heteroszkedasztikus regresszióval előállított görbék és a sima négyzetes regresszióval kapottak között.

A 9. ábra egy adott játékos esetén a sima négyzetes és a heteroszkedasztikus regressziós egyenest tartalmazza, és jól látható, hogy alig észlelhetünk eltérést a két görbe között.



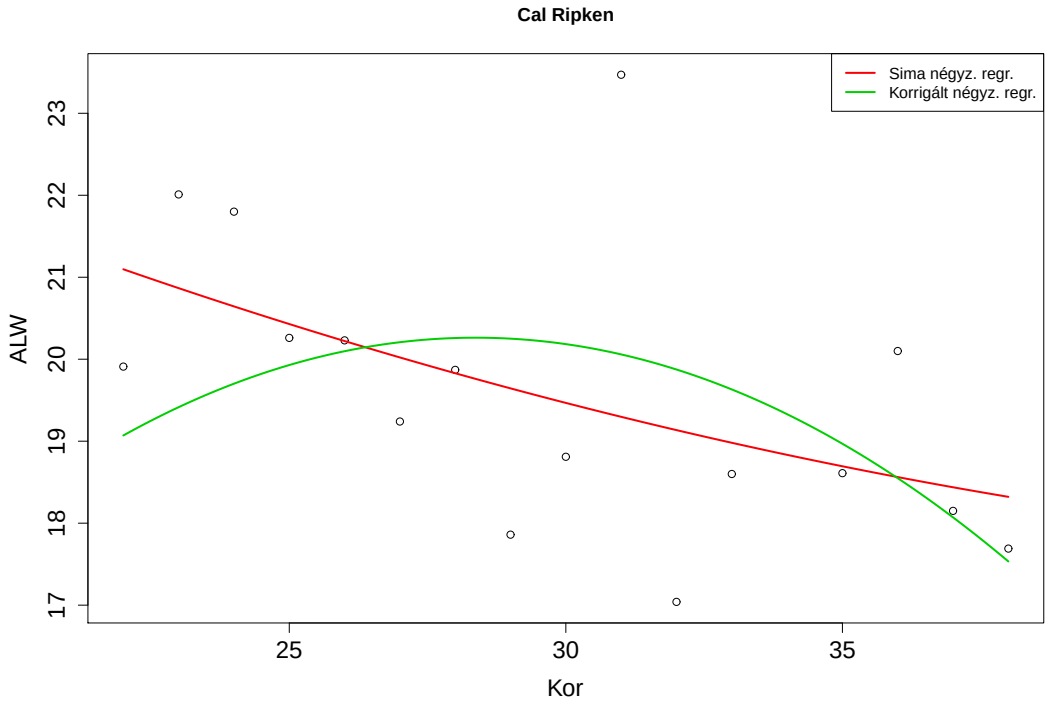
9. ábra. Vladimir Guerrero (sima regresszió - heteroszkedasztikus regresszió)

4.4.1. A Bayes-becslés hatása

Ahogy már említettük, szeretnénk elérni, hogy a kapott eredmények tükrözzék azt az elképzelést, miszerint a játékosok karrierjük elején még fejlődő tendenciát mutat-

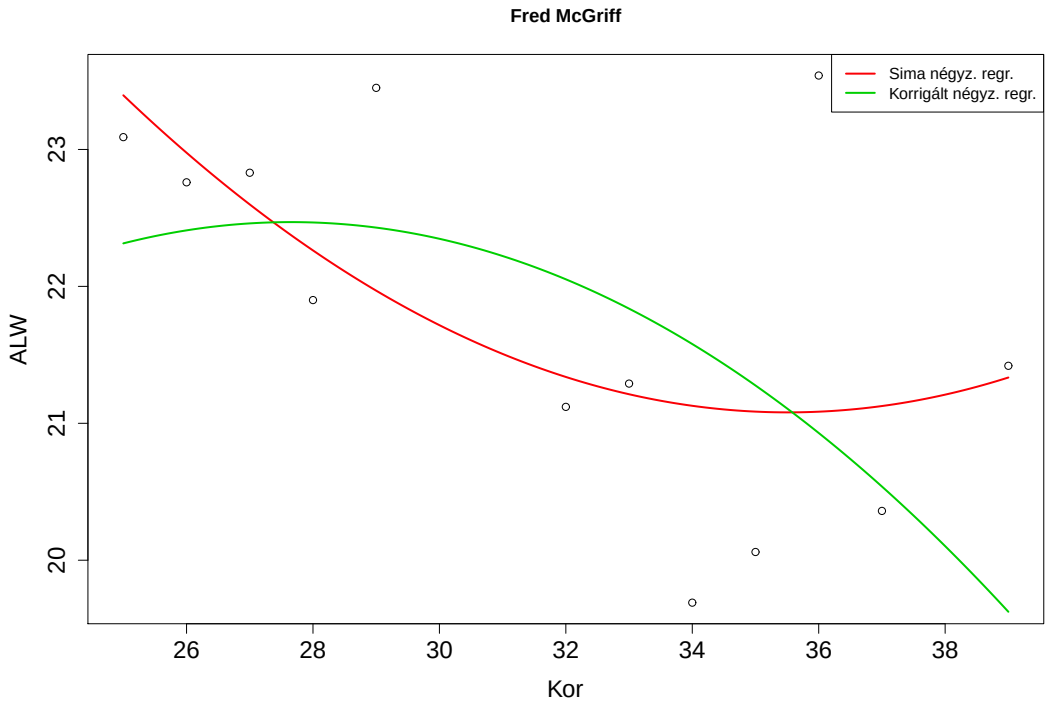
nak, pályafutásuk közepére elérik a legjobb formájukat, majd visszavonulásukat megelőzően csökkenni kezd a teljesítményük. Elvégeztük a becsléseket az összes, általunk vizsgált alanyra. Megszabtuk, hogy legalább 500 ütéshez állással, illetve legalább 8 évnyi használható adattal rendelkezzen az illető. Ezek alapján közel 150 sportolóval sikerült dolgoznunk.

Az elvégzett méréseink alapján a valóság nagyjából megfelel az elvárásainknak. Természetesen akad néhány kivétel, ahol kisebb (néhol pedig meglepően nagy) eltérést tapasztaltunk – erről a szakirodalomban is lehet olvasni. A következő ábrák szemléltetik a sikereinket, azaz hogy a Bayes-becslés mennyit javít egy játékos kor-teljesítménygörbéjén. A 10. és a 11. ábrákon észrevehetjük, hogy egyes görbék főegyüttható pozitív előjelű volt, majd a becslésünk hatására negatívra változik, talán itt látszik meg legjobban a munkánk.

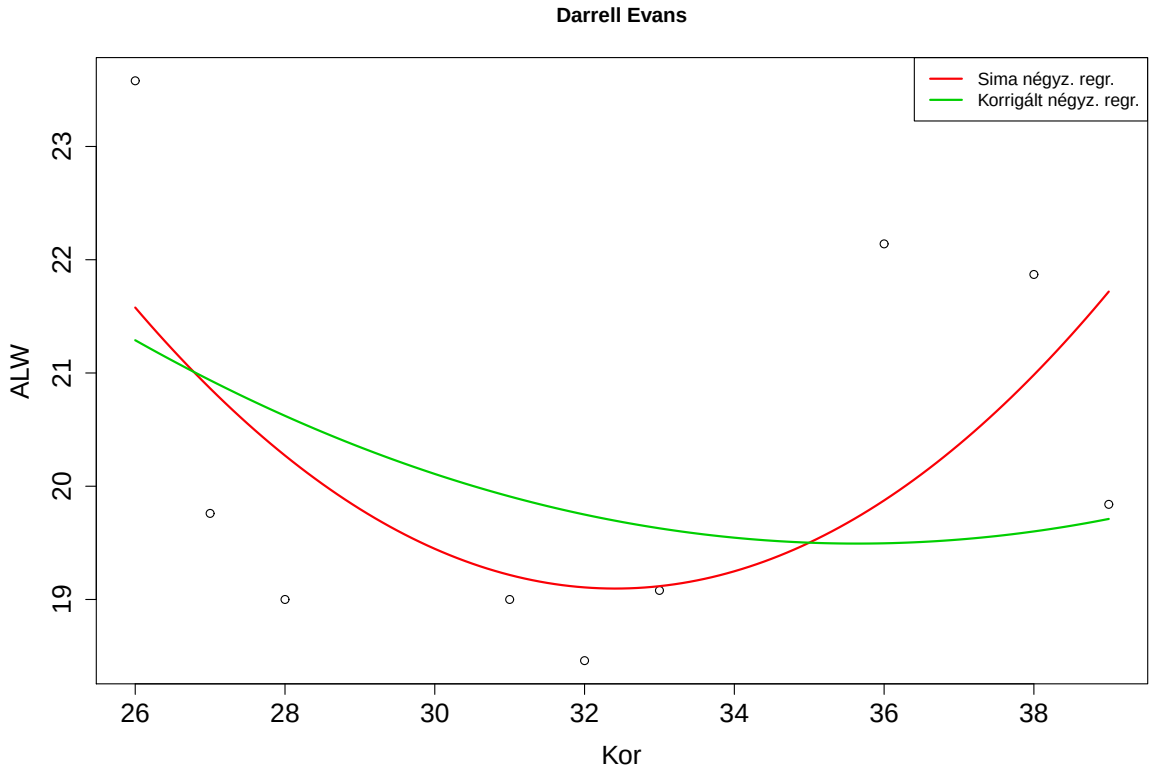


10. ábra. Cal Ripken görbéi a Bayes-becslés előtt és után

Sajnos néhány személy esetében még ez a módszer sem bizonyult elégnek, de nem is várhatjuk el, hogy minden adat, ami rendelkezésünkre áll, tökéletes legyen, valamint minden játékos karrierje elvárásainknak megfelelően alakuljon. Ezt jól szemlélteti az 12. ábra.



11. ábra. Fred McGriff görbéi a Bayes-becslés előtt és után

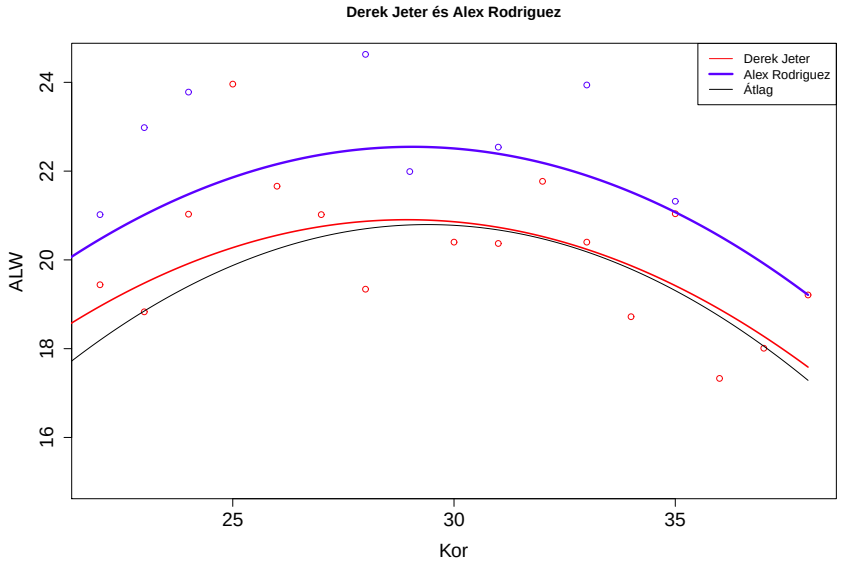


12. ábra. Darrell Evans – személetünkkel nem egyező eredmény

4.4.2. Baseball játékosok összehasonlítása teljesítményük alapján

A következőkben néhány játékost összevetünk karrierjük folyamán nyújtott játékuk alapján. Ezen vizsgálatok azt is megmutatják, hogy miként változott a baseball képe

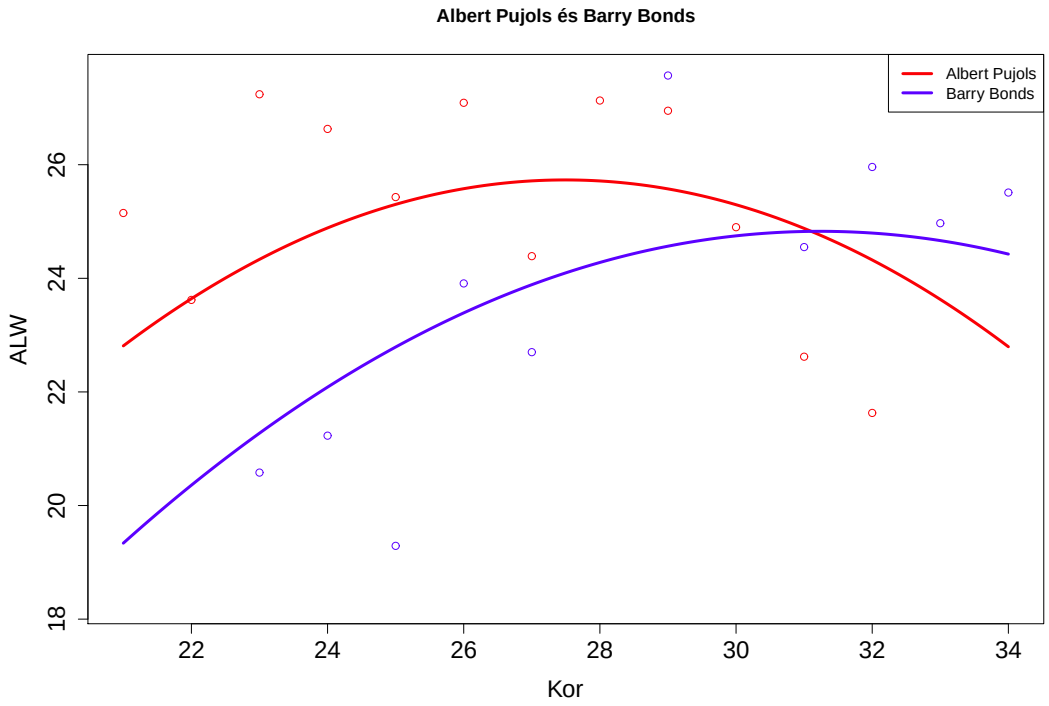
az idő múltával. A rendelkezésünkre álló játékosok mindegyike az 1940-es és 1980-as évek között született. Érdemes megnézni, hogyan teljesítettek a régi nagyágyúk az ifjú titánokhoz képest, illetve az új nemzedék felülmúlja-e az őt megelőzőt.



13. ábra. Derek Jeter és Alex Rodriguez összehasonlítása Bayes-becslés segítségével

Alex Rodriguez és Derek Jeter az elmúlt évtized kiemelkedő tagjai voltak a New York Yankees csapatában (Alex Rodriguezről már tettünk említést rendkívüli fizetése kapcsán). A 13. ábrán látható módon mindkét játékos megfelel az elvárásainknak, valamint teljesítményük kiemelkedő (mindkét ALW szint átlagon felüli). Itt már a Bayes-becslés eredménye látható, tehát a trajektóriák módosított állapotban láthatók.

Természetesen Barry Bonds (1964) volt kora egyik legkimagaslóbb baseball játékosa, de Albert Puyols (1980), aki nála 16 évvel később született, is teljesített legalább olyan erősen, mint Bonds: a 14. ábrán az látszik, ami a sportokban manapság érezhető, az új generáció felnő a régihez, és túl is teljesíti azt. Ez nem meglepő eredmény, más sportokban, mondjuk az úszásban is hasonlókat tapasztalhatunk: Darnyi Tamás olimpiai és világcsúcs időket úszott a maga korában (1988 és 1992 között) 200 és 400 méter vegyesen, ám ezek az idők a mai mezőnyben már a dobogóra se lennének elegendők.



14. ábra. Barry Bonds (1964) és Albert Pujols (1980) összehasonlítása Bayes-becslés segítségével

Függelék

Ábrák jegyzéke

Táblázatok jegyzéke

Hivatkozások

Jim Albert. Smoothing career trajectories of baseball hitters. *Unpublished manuscript, Bowling Green State University, at bayes. bgsu. edu/papers/career_ trajectory. pdf*, 2002.

Jim Albert. Baseball data at season, play-by-play, and pitch-by-pitch levels. *Journal of Statistics Education*, 18(3), 2010.

Jim Albert and Jay Bennett. *Curve ball: Baseball, statistics, and the role of chance in the game*. Springer Science & Business Media, 2003.

Phil Birnbaum. Do baseball salaries have "precious little" to do with ability? 2014. URL `http://blog.philbirnbaum.com/2014/10/do-baseball-salaries-have-precious.html`.

Marianna Bolla and András Krámli. *Statisztikai következtetések elmélete*. Typotex Kft, 2005.

Stephen Hall, Stefan Szymanski, and Andrew S Zimbalist. Testing causality between team performance and payroll the cases of major league baseball and english soccer. *Journal of Sports Economics*, 3(2), 2002.

Gary Koop. Comparing the performance of baseball players. *Journal of the American Statistical Association*, 97(459), 2002.

Tamás F. Móri. Főkomponens- és faktoranalízis, 1999. URL `http://cs.elte.hu/~mori/faktor.pdf`.

Tamás F. Móri and Gábor J. Székely. *Többváltozós statisztikai analízis*. Műszaki Könyvkiadó, 1986.

Gerald W Scully. Pay and performance in major league baseball. *The American Economic Review*, pages 915–930, 1974.

Michael Ross Smith. *Modeling the Performance of a Baseball Player's Offensive Production*. PhD thesis, Birmingham Young University, 2006. URL `http://scholarsarchive.byu.edu/etd/362/`.

John Thorn and Pete Palmer. *Hidden Game of Baseball*. Doubleday Books, 1985.