UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea in Scienze Statistiche ed Economiche

STATISTICA APPLICATA AL BASEBALL: COME COSTRUIRE UNA SQUADRA VINCENTE

Relatore: Prof. Aldo SOLARI

Tesi di Laurea di:

Alex CECCOTTI

Matricola 790497

Anno Accademico 2016-2017

«When the numbers acquire
the significance of language,
they acquire the power to do
all of the things which
language can do: to become
fiction and drama and poetry.»
Bill James

Indice

Introduzione

L'obiettivo di questo elaborato è cercare le evidenze empiriche a supporto delle teorie esposte in "Moneyball: The Art of Winning an Unfair Game" di Michael Lewis. Nel suo libro l'autore racconta come il General Manager Billy Beane ha gestito gli Oakland Athletics, una squadra della Major League Baseball. Avendo a disposizione un budget nettamente inferiore rispetto alla maggior parte delle altre squadre della MLB, Beane decide di non seguire i metodi convenzionali per la gestione della squadra, ma di utilizzare le teorie di Bill James, statistico fondatore della SABR (Society for American Baseball Research). Questa società produce analisi scientifiche riguardanti il baseball attraverso l'utilizzo di metodi statistici, cioè si occupa dello studio della sabermetrica, termine che deriva dal nome della società stessa.

> «Bill James ha definito la sabermetrica come "la ricerca per la conoscenza oggettiva del baseball". Quindi, la sabermetrica prova a rispondere alle domande oggettive riguardanti il baseball, come "quale giocatore dei Red Sox ha contribuito maggiormente all'attacco della sua squadra?" o "quanti fuoricampo farà Ken Griffey il prossimo anno?". Non potrà trattare i giudizi soggettivi che sono comunque importanti per il gioco, come "chi è il tuo giocatore preferito?" o "questa è stata una bella partita".»[**S1**]

Prima degli Oakland A's nessuna squadra della MLB aveva mai preso in considerazione una visione più completa e corretta delle statistiche di gioco disponibili. Per portare a proprio vantaggio le inefficienze presenti nel mondo del baseball, Beane decide di affidarsi ad uno staff di laureati in statistica e matematica, tra cui spicca la presenza di Paul DePodesta. L'idea dunque è chiara: un approccio diverso alla campagna acquisti è l'unico modo per colmare l'enorme differenza di budget tra gli Athletics e il resto delle squadre. Alcuni interrogativi sorgono quindi spontanei: a che strategie di mercato ha portato un'attenta analisi delle statistiche disponibili? Quali sono le statistiche più importanti per valutare l'utilità di un giocatore? Fino a che punto l'utilizzo di queste teorie può colmare il gap di payrolls? Nei seguenti capitoli si cercherà di rispondere a queste domande riportando alcune teorie esposte nel libro sopracitato (e nel "Manifesto della Sabermetrica") e analizzando dal punto di vista statistico la loro validità. Per fare ciò verranno utilizzati il database relazionale "Lahman's Baseball Database 2016 - Microsoft Access version"[1] e il software open source R. Lo script realizzato al fine di produrre i grafici ed i modelli successivamente riportati è consultabile al link `https://archive.org/details/howtobuildawinningteam_script`.

[1]Disponile all'indirizzo web `http://www.seanlahman.com/baseball-archive/statistics` (copyright 1996-2017 by Sean Lahman).

Capitolo 1

Analisi esplorativa

Una partita di baseball è composta da 9 inning (riprese) ognuno dei quali suddiviso in due parti; in ciascuna di queste due parti una squadra sarà in fase di attacco e l'altra in fase di difesa alternandosi all'interno di un inning. Durante la fase di attacco è possibile realizzare punti, mentre in fase di difesa l'obiettivo è quello di subire meno punti possibili. Per segnare un punto un corridore deve fare il giro di tutte le quattro basi. Una parte di un inning si conclude quando la squadra in difesa realizza tre out.

1.1 Attacco

Esistono diverse statistiche utilizzate per misurare la prestazione di un giocatore in fase offensiva (o di battuta):

- Valida (indicata con H), quando il battitore riesce ad arrivare in base battendo la pallina in campo, senza errori della difesa
- Base su ball (BB), quando il battitore riceve quattro ball durante il suo turno di battuta e può quindi arrivare in prima base "gratuitamente"
- Strike out (SO), quando il battitore subisce tre strike durante il suo turno di battuta e viene di conseguenza eliminato
- Fuoricampo (HR), cioè una battuta valida che permette a tutti i corridori sulle basi e al battitore stesso di arrivare a punto.

Sono presenti ovviamente molti altri indicatori che verranno spiegati successivamente. Nel baseball viene data molta importanza agli indicatori relativi, ovvero una somma di alcune statistiche rapportata alle "opportunità" che il giocatore ha avuto. Le più famose sono:

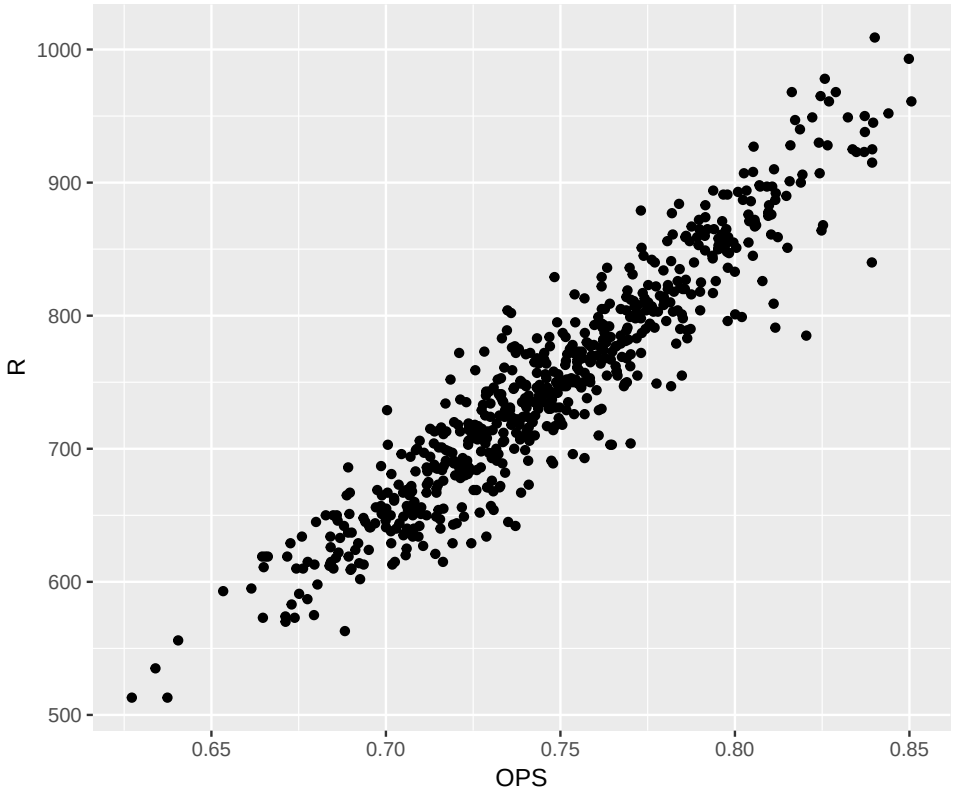
- $AVG = \frac{H}{AB}$ (media battuta) dove AB (at bat) sono le "possibilità" che il battitore ha avuto per battere una valida
- $OBP = \frac{H+BB+HBP}{H+BB+HBP+SF}$ (on-base percentage) dove HBP rappresenta il numero di volte che il battitore è stato colpito da un lanciatore (genera lo stesso effetto della base su ball) e SF è una battuta che il giocatore effettua per farsi eliminare intenzionalmente in modo da creare una situazione di gioco più vantaggiosa o per portare a casa un punto
- $SLG = \frac{1B+2*2B+3*3B+4*HR}{AB}$ (slugging percentage) dove 1B rappresenta le battute valide in cui si è arrivati in prima base, 2B in seconda base e 3B in terza base
- $OPS = OBP + SLG$

1.1.1 Come l'indicatore OPS incide sulla capacità di vittoria

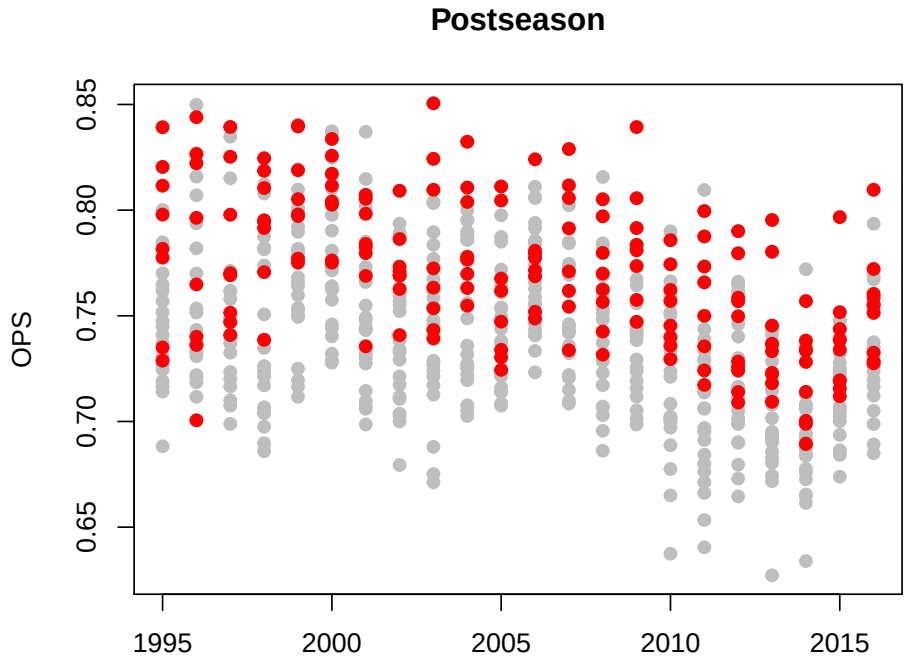
Secondo Billy Beane e i suoi analisti per arrivare alla vittoria, e quindi realizzare più punti degli avversari, è necessario cercare di creare situazioni di gioco in cui non si subiscono out. L'indicatore che meglio interpreta questa strategia è la OBP in quanto si considerano tutti gli arrivi in base rapportati praticamente a tutte le apparizioni in battuta. Questa statistica risulta essere più completa della semplice media battuta, la quale invece viene spesso vista come un'ottima sintesi dell'abilità offensiva di un giocatore. Discorso a parte per la SLG: in questo indicatore viene dato maggior peso alle valide che permettono di arrivare in basi più lontane e che quindi generano più possibilità di realizzare un punto. Dato che lo scopo finale della fase di battuta è fare più punti possibili, è evidente che la SLG può risultare determinante. Vediamo innanzitutto le correlazioni tra gli indicatori relativi, i punti messi a segno e la percentuale di vittorie in ogni stagione per ogni squadra:

	R	AVG	OBP	SLG	OPS	W%
R	1.00	0.81	0.88	0.90	0.94	0.50
AVG	0.81	1.00	0.86	0.77	0.84	0.37
OBP	0.88	0.86	1.00	0.78	0.90	0.48
SLG	0.90	0.77	0.78	1.00	0.97	0.46
OPS	0.94	0.84	0.90	0.97	1.00	0.49
W%	0.50	0.37	0.48	0.46	0.49	1.00

Notiamo che la correlazione tra punti fatti (R) e OPS è circa del 94%; inoltre R e la percentuale di vittorie in stagione (W%) sono correlate al 50%. Anche se potrebbe sembrare poco elevata rispetto alle altre correlazioni, bisogna tenere conto che alla vittoria contribuisce sia la fase offensiva sia quella difensiva e in questo caso vengono considerati solo i punti fatti e non quelli subiti. A titolo esemplativo visualizziamo graficamente le variabili OPS e R:



Le prime osservazioni del GM degli Oakland A's sembrano dunque essere corrette: per fare più punti possibili è necessario arrivare in base (e pertanto non concedere out) e fare in modo che i corridori vengano portati a punto nel minor tempo possibile, in modo da evitare di lasciare uomini sulle basi a fine inning (in tal caso gli arrivi in base andrebbero sprecati). Le statistiche che meglio approssimano questa filosofia di gioco sono OBP e SLG, che sommate danno come risultato la OPS. La OPS però incide davvero sulla probabilità che una squadra ha di raggiungere la postseason? E sulla probabilità di vincere le World Series? Rispondiamo a queste due domande prima tramite visualizzazione grafica e poi cercando di costruire un modello statistico.



In rosso le squadre che sono riuscite ad accedere ai playoff

Il grafico fa pensare a un certo tipo di legame tra accesso ai playoff e media OPS di squadra, ma lascia comunque qualche dubbio sulla sua intensità. Stimiamo ora un modello logistico avente come risposta il passaggio alla postseason (variabile dicotomica)[1]:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-23.2406	3.9078	-5.95	0.0000
HRM	99.8253	18.9675	5.26	0.0000
BBM	40.4890	7.9912	5.07	0.0000
SOM	9.7456	5.4018	1.80	0.0712
B1M	50.5406	13.0601	3.87	0.0001
B2M	26.8674	22.0888	1.22	0.2239
B3M	28.1500	62.3431	0.45	0.6516
SBratio	4.1218	1.6864	2.44	0.0145

[1]Non vengono considerati gli indicatori AVG, OBP, SLG e OPS perché sono molto correlati tra loro e potrebbero creare problemi di multicollinearità. Utilizziamo invece, oltre alla percentuale di rubate riuscite, HR, BB, SO, B1, B2, B3 rapportati ai turni di battuta effettuati (AB). Si noti che i turni di battuta in cui si riceve una base su ball non vengono conteggiati tra gli AB, quindi BBM non rappresenta una percentuale.

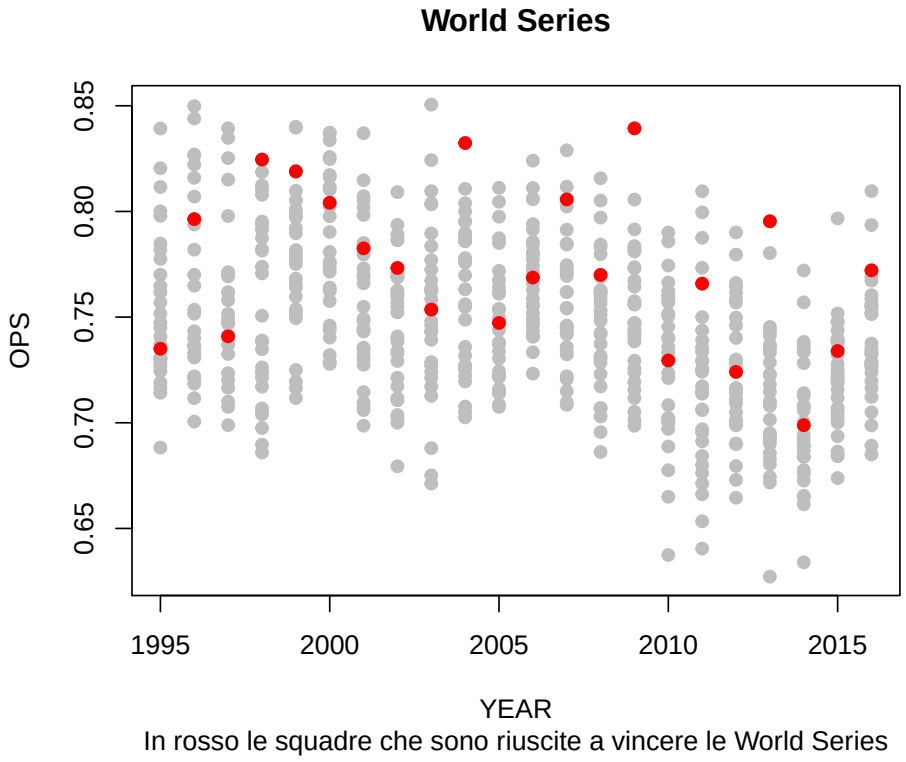
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	780.96	(Modello nullo)		
2	646	671.36	7	109.60	0.0000

Si nota come le variabili che influenzano maggiormente la capacità di una squadra di raggiungere i playoff sono HRM, BBM e B1M. Di conseguenza risultano importanti i singoli (valida in cui viene raggiunta solo la prima base), i fuoricampo e le basi su ball ricevute. Risulta essere significativa al 5% anche la percentuale di rubate[2] riuscite. Stimiamo inoltre un modello avente come esplicativa soltanto OPS per valutare statisticamente il legame supposto a partire dal grafico precedente:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-17.0731	1.9470	-8.77	0.0000
OPS	21.4389	2.5639	8.36	0.0000

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	780.96	(Modello nullo)		
2	652	698.47	1	82.50	0.0000

Viene confermata l'intuizione grafica: OPS ha un effetto significativo sull'accesso ai playoff.



In rosso le squadre che sono riuscite a vincere le World Series

[2]Una rubata è una situazione di gioco nella quale un corridore cerca di conquistare la base successiva senza che un suo compagno di squadra abbia battuto la pallina in gioco. Questo comporta dei rischi in quanto la difesa ha in mano la pallina e quindi è molto più probabile essere eliminati se si cerca di arrivare alla base successiva.

Passando alla seconda domanda, dal grafico riportato nella pagina precedente vediamo che continua ad esserci un effetto di OPS sulla capacità di vittoria della squadra, ma se consideriamo come variabile risposta la vittoria alle World Series questo legame non sembra essere più così evidente. Stimiamo subito il rispettivo modello logistico avente come variabile esplicativa solo OPS:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.5189	4.3195	-3.82	0.0001
OPS	17.3258	5.5877	3.10	0.0019

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	192.50	(Modello nullo)		
2	652	182.51	1	10.00	0.0016

Anche qui l'intuizione grafica è corretta: OPS incide significativamente sulla capacità di vittoria di una squadra, ma con minore intensità. Provando invece a considerare come esplicative le precedenti variabili rapportate ai turni di battuta, si ottengono i seguenti risultati:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.8510	8.4559	-1.76	0.0790
HRM	34.0757	42.6289	0.80	0.4241
BBM	32.3710	17.7003	1.83	0.0674
SOM	-6.7877	13.4480	-0.50	0.6137
B1M	32.3613	29.1220	1.11	0.2665
B2M	51.0020	51.7123	0.99	0.3240
B3M	169.0588	138.8274	1.22	0.2233
SBratio	-1.1706	3.8669	-0.30	0.7621

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	192.50	(Modello nullo)		
2	646	178.87	7	13.63	0.0582

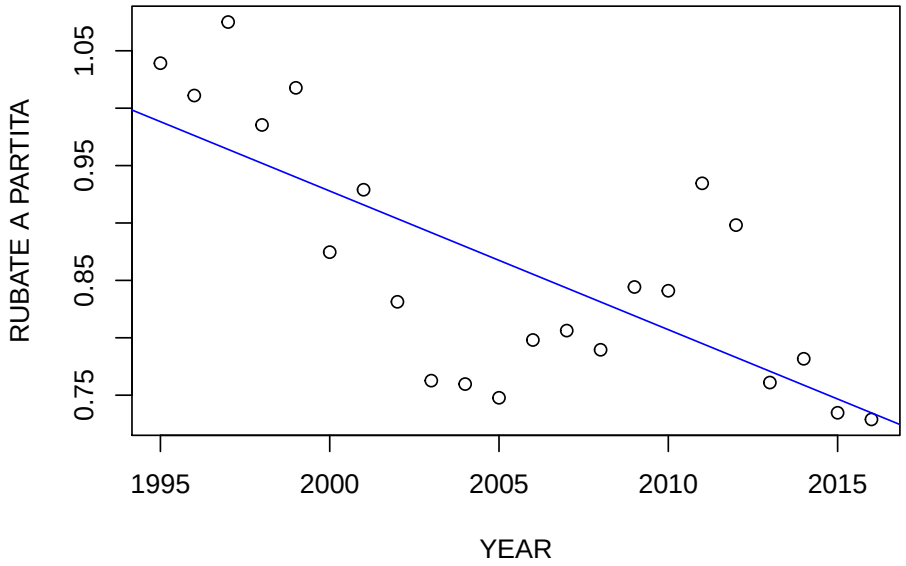
Tutte le variabili perdono significatività, tranne BBM che resta significativa al 10%. Questo potrebbe essere causato dal fatto che il passaggio di un turno durante i playoff è deciso in un numero di partite che varia da 1 a 7, mentre nella Regular Season tutte le squadre giocano 162 partite. Ovviamente valutare una vittoria delle World Series lascia molta più incertezza rispetto al passaggio dalla regular season alla postseason.

1.1.2 Scelte strategiche riguardanti rubate e bunt di sacrificio

Come già detto in precedenza, un punto fondamentale della strategia illustrata in "Moneyball: The Art of Winning an Unfair Game" è non subire out. Di conseguenza rubate e bunt di sacrificio[3] sono da considerarsi dannosi e quindi da evitare. Staccandoci però

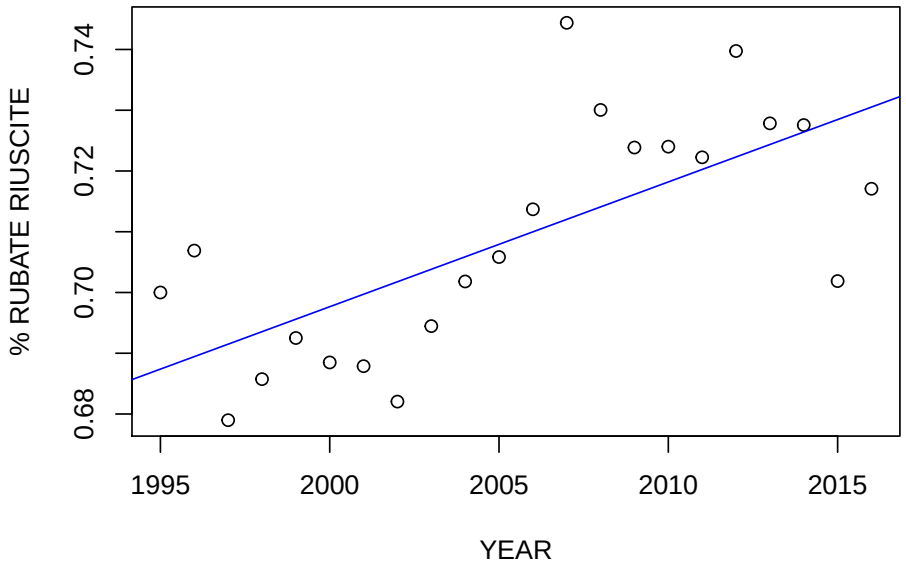
[3]Un bunt di sacrificio è un'azione che compie il battitore; consiste nello smorzare la pallina e cercare di metterla in gioco vicina. Questo permetterà (se eseguito bene) ai corridori di avanzare di una base, ma spesso comporta l'eliminazione del battitore.

da questa applicazione estrema, dai dati disponibili si nota che c'è stata comunque una responsabilizzazione a tal riguardo. Per esempio, facendo le medie annue delle rubate effettuate a partita, vediamo che sono calate significativamente nel corso del tempo:



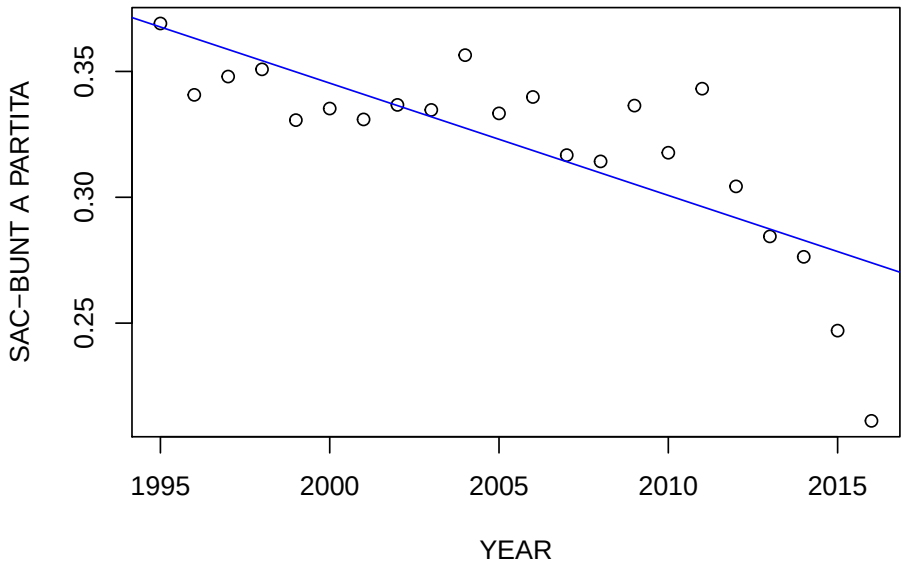
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.0861	5.1896	4.83	0.0001
Year	-0.0121	0.0026	-4.67	0.0001
	R^2=0.521		Pr(>F)<0.001	

Considerando invece la percentuale di rubate riuscite rispetto ai tentativi di rubata, osserviamo come questa sia aumentata negli anni:



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.4099	0.9567	-3.56	0.0019
yearID	0.0021	0.0005	4.31	0.0003
	R^2=0.481		Pr(>F)<0.001	

Unendo queste due informazioni, si può ipotizzare che i manager si siano accorti che non ha senso far rubare un corridore in molte occasioni solo perché è veloce, ma che questa strategia risulta utile ed efficace solo in determinate situazioni di gioco. Hanno perciò ridotto il numero di rubate totali incrementando la probabilità di riuscita. Per quanto riguarda i bunt di sacrificio, anche se è difficile ipotizzare un motivo specifico, si constata una diminuzione esponenziale del loro utilizzo:



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2671	1.5974	5.80	0.0000
yearID	-0.0045	0.0008	-5.60	0.0000
	R^2=0.611		Pr(>F)<0.001	

1.1.3 Stipendi

Visto che la ragione per cui Beane ha iniziato a basare le sue scelte sulla statistica è lo scarso budget a disposizione, è interessante vedere come determinati indicatori influenzino lo stipendio dei giocatori di baseball. Bisogna però fare una considerazione: durante una singola stagione un ottimo battitore potrebbe avere delle pessime statistiche e viceversa, di conseguenza lo stipendio (che per i buoni giocatori si presume sia maggiore) nel singolo anno potrebbe non essere legato particolarmente alle capacità oggettive. Per questo motivo la scelta è stata quella di aggregare per giocatore sia gli stipendi (facendone la media) sia le statistiche a disposizione (sommandole e creando poi gli indicatori relativi[4]). Sempre

[4]R, RBI, B1, B2, B3, HR, SO e BB si è deciso di rapportarli con AB; SB invece viene rapportato con i tentativi di rubata, cioè SB+CS. Vengono anche generati AVG, OBP, SLG e OPS.

per dare più senso e meno casualità ai risultati, verranno esclusi dall'analisi coloro che durante la carriera hanno totalizzato meno di 1'000 AB, obiettivo raggiungibile giocando almeno due stagioni complete; in questo passaggio vengono scartati implicitamente anche i lanciatori. Come prima operazione deflazioniamo[5] i singoli stipendi in base all'anno, aggreghiamo i nostri dati come illustrato in precedenza e stimiamo un modello lineare avente come variabile risposta il logaritmo dello stipendio medio[6]:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.0989	0.4238	21.47	0.0000
HRM	40.9482	5.3907	7.60	0.0000
RM	3.0846	1.9964	1.55	0.1227
RBIM	0.1543	2.1924	0.07	0.9439
BBM	3.1982	0.9495	3.37	0.0007
B1M	13.9051	1.8808	7.39	0.0000
B2M	14.7793	3.4433	4.29	0.0000
B3M	9.8451	8.4854	1.16	0.2462
SOM	-3.3891	0.5838	-5.81	0.0000
SBM	0.6995	0.1617	4.33	0.0000
	R^2=0.476		Pr(>F)<0.001	

Le uniche variabili non significative sono i punti fatti dal singolo giocatore per ogni turno in battuta, gli RBI[7] e i tripli. Anche secondo le teorie di Bill James i primi due indicatori non rispecchiano bene le abilità di un battitore in quanto dipendono molto dai compagni di squadra. Per dare un'idea più definita delle stime dei parametri, vediamo media e varianza delle variabili esplicative:

	Media	Varianza
HRM	0.02962	0.00024
RM	0.13472	0.00048
RBIM	0.12765	0.00102
BBM	0.09595	0.00117
B1M	0.17705	0.00057
B2M	0.05233	0.00007
B3M	0.00568	0.00001
SOM	0.19496	0.00335
SBM	0.63528	0.02500

[5]I deflatori sono calcolati partendo dagli indici dei prezzi per gli USA disponibili all'indirizzo web `fred.stlouisfed.org/series/FPCPITOTLZGUSA`.

[6]Come nel paragrafo precedente, anche in questo caso non utilizziamo direttamente AVG, OBP, SLG e OPS per problemi di multicollinearità. Si è scelto invece di utilizzare il logaritmo dello stipendio medio per smorzare l'asimmetria e per linearizzare il legame con le covariate. Inoltre si è deciso di imputare due valori mancanti nella variabile SB ratio in quanto questi due giocatori non hanno mai effettuato una rubata in tutta la carriera. Per fare ciò, dopo aver stimato un modello avente come risposta SB ratio e come esplicative RM e BBM (le due più significative), sono stati imputati i valori previsti da questo modello.

[7]Gli RBI (runs batted-in) di un battitore rappresentano i "punti portati a casa", ovvero il numero di corridori che sono riusciti ad arrivare a punto grazie alle battute valide di quel giocatore.

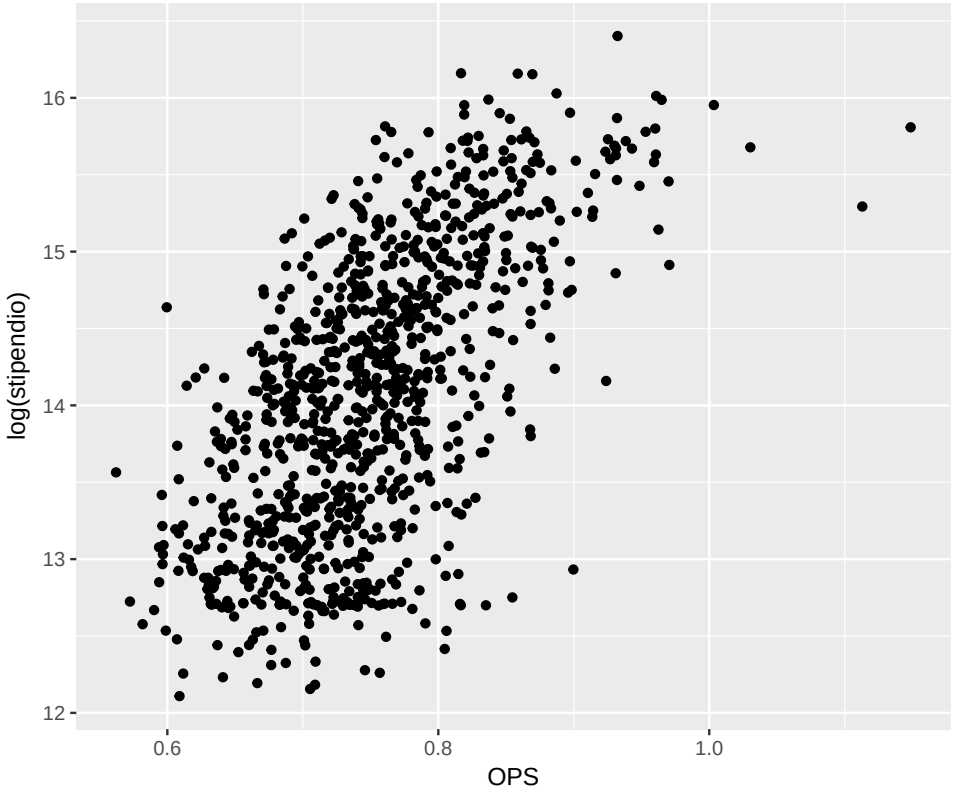
Calcoliamo poi le correlazioni tra gli indicatori classici e il logaritmo dello stipendio medio:

	AVG	OBP	SLG	OPS
Stipendio	0.508	0.551	0.613	0.651

Data l'elevata correlazione tra stipendio e OPS, stimiamo anche un modello avente come risposta il logaritmo dello stipendio medio e come esplicativa OPS:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.0218	0.2294	34.97	0.0000
OPS	8.1499	0.3046	26.76	0.0000
	R^2=0.424		Pr(>F)<0.001	

Visualizziamo infine la relazione tra le due variabili considerate:



1.2 Difesa

Durante la fase difensiva, la squadra schiera 9 giocatori, ognuno con un ruolo differente. Tra questi si distingue particolarmente la figura del lanciatore, colui che inizia ogni azione lanciando la pallina al ricevitore con lo scopo di non far battere (o far battere male) il battitore di turno. Come nel caso precedente esistono delle statistiche pensate appositamente per valutare le prestazioni dei lanciatori, tra cui:

- $ERA = 9 * \frac{ER}{IP}$ dove ER sono i punti "guadagnati" dal lanciatore[8] e IP rappresentano gli inning completati

[8]Non tutti i punti subiti da un lanciatore rientrano negli ER; infatti se un punto è stato causato da un errore della difesa questo non verrà conteggiato al fine del calcolo dell'indicatore ERA.

- BB cioè le basi su ball concesse
- SO cioè gli strike out effettuati

oltre a tutte le altre statistiche disponibili per i battitori calcolate sul singolo lanciatore. Ovviamente se per il battitore è meglio avere una determinata statistica alta, per il lanciatore è meglio averla bassa e viceversa.

1.2.1 Pregi e limiti dell'indicatore ERA

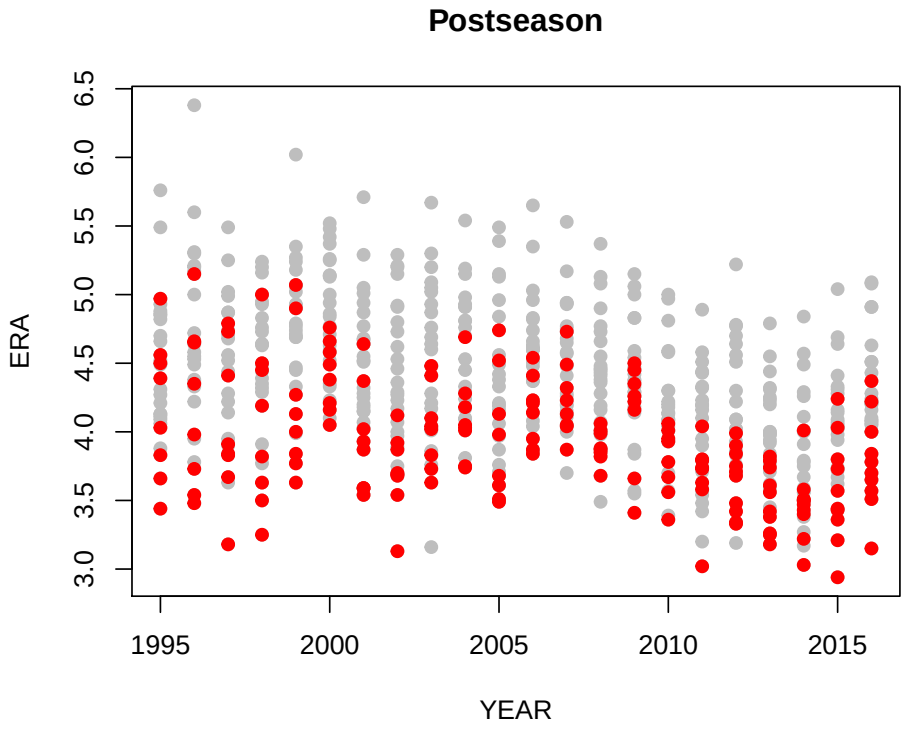
Per una squadra, oltre ai punti realizzati nella fase offensiva, è importante anche subire meno punti possibili. Analizzando la fase di battuta non ci siamo soffermati troppo sui punti fatti in quanto, se osservati sul singolo giocatore, non danno molta informazione sulle sue abilità. Infatti i punti realizzati da un battitore dipendono sì dalle sue capacità di arrivare in base, ma anche dalla bravura dei compagni che vengono dopo di lui nell'ordine di battuta. Passando alla fase difensiva invece, l'abilità del singolo lanciatore può essere ben interpretata considerando i punti subiti (rapportati agli inning lanciati). Questo perché, sebbene l'abilità difensiva degli altri 8 giocatori è importante, non incide particolarmente sui punti "guadagnati". Per questo motivo, ad esempio, in "Moneyball: The Art of Winning an Unfair Game" si sostiene che l'analisi approfondita della fase di lancio non porta a un livello di conoscenza superiore. Difatti i lanciatori vengono già valutati sufficientemente bene in base alla loro ERA. Quindi, una volta capito che non è necessario utilizzare nuovi indicatori per valutare la prestazione di un lanciatore, le domande da porsi sono le seguenti: avere una buona rotazione iniziale e un buon bullpen[9] fino a che punto incide sulla capacità di vittoria? Quale strategia converrà adottare per tenere bassa la propria ERA? Come primo approccio al problema calcoliamo le correlazioni tra alcuni indicatori:

	W%	ERA	H1AM	HRAM	BBAM	SOAM
W%	1.00	-0.60	-0.50	-0.41	-0.43	0.32
ERA	-0.60	1.00	0.80	0.74	0.64	-0.60
H1AM	-0.50	0.80	1.00	0.38	0.38	-0.66
HRAM	-0.41	0.74	0.38	1.00	0.36	-0.36
BBAM	-0.43	0.64	0.38	0.36	1.00	-0.35
SOAM	0.32	-0.60	-0.66	-0.36	-0.35	1.00

Vediamo come W% (% di partite vinte durante la regular season) sia altamente correlato con ERA e notiamo che gli indicatori singoli H1AM (valide che non siano fuoricampo subite rapportate agli out lanciati[10]), HRAM (fuoricampo subiti rapportati agli out lanciati), BBAM (basi su ball concesse rapportate agli out lanciati) e SOAM (strike out effettuati rapportati agli out lanciati) sono molto correlati con ERA. Proviamo dunque a rispondere alla prima domanda utilizzando un grafico per inquadrare meglio la questione:

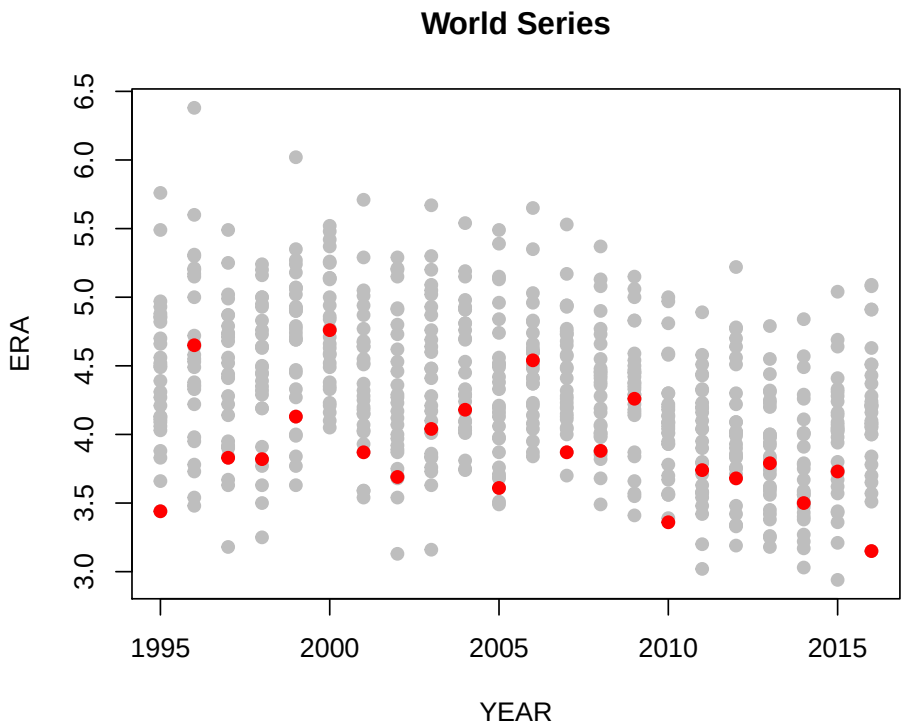
[9] La rotazione iniziale è solitamente composta da 5 lanciatori che hanno il compito di iniziare la partita. Generalmente riescono a stare sul monte di lancio per 5 inning o più. È necessario avere più giocatori che ricoprano il ruolo del "partente" dato che si gioca 6 giorni a settimana circa. Il bullpen invece è composto da circa altri 7 lanciatori che subentrano al partente durante il corso della partita.

[10] Durante l'analisi verranno considerati gli out lanciati e non gli inning perché spesso capita che un lanciatore venga sostituito a inning in corso, quando potrebbero essere già stati effettuati uno o due out. Per evitare di utilizzare dati con frazioni di inning risulta più comodo considerare gli out.



In rosso le squadre che sono riuscite ad accedere ai playoff

È evidente che le squadre con ERA più bassa hanno maggiori possibilità di accedere alla postseason. Vediamo ora lo stesso grafico valutando però la vittoria delle World Series:



In rosso le squadre che sono riuscite a vincere le World Series

Come abbiamo potuto constatare nel primo paragrafo, capire quali fattori incidono sulla vittoria finale è più complicato in quanto ai playoff il passaggio del turno viene deciso in un numero ristretto di partite. Stimiamo dunque due modelli per provare a confermare quanto detto:

- Accesso ai playoff

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.5700	0.9306	9.21	0.0000
ERA	-2.2774	0.2272	-10.02	0.0000

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	780.96	(Modello nullo)		
2	652	640.57	1	140.39	0.0000

- Vittoria delle World Series

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.2137	1.8289	1.76	0.0789
ERA	-1.6073	0.4650	-3.46	0.0005

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	192.50	(Modello nullo)		
2	652	178.84	1	13.66	0.0002

Notiamo come, sebbene i p-value nel secondo modello siano aumentati, l'esplicativa ERA rimane comunque molto significativa. Ciò potrebbe suggerire una certa rilevanza di questo indicatore sulla capacità di vittoria di una squadra. Cerchiamo ora di capire quali fattori condizionano maggiormente la statistica in esame:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.0580	0.1680	-18.20	0.0000
HRAM	40.4011	1.0444	38.68	0.0000
H1AM	15.2332	0.3937	38.69	0.0000
BBAM	9.5690	0.3868	24.74	0.0000
SOAM	0.3184	0.2606	1.22	0.2222
	R^2=0.933		Pr(>|f|)<0.001	

Su quattro variabili considerate, solo tre risultano significative e queste spiegano il 93% della variabile risposta (ERA). L'unica non significativa è il numero di strike out effettuati per ogni out lanciato. Questo modello potrebbe indicare una strategia vincente per ridurre il numero di punti subiti: concedere poche basi su ball anche a discapito del numero di strike out. Infatti per effettuare uno strike out spesso si tende a lanciare a filo della zona dello strike concedendo talvolta delle basi su ball. Risulta più utile invece, sempre stando al modello, "far battere male" piuttosto che "non far battere" e di conseguenza concedere "basi gratis".

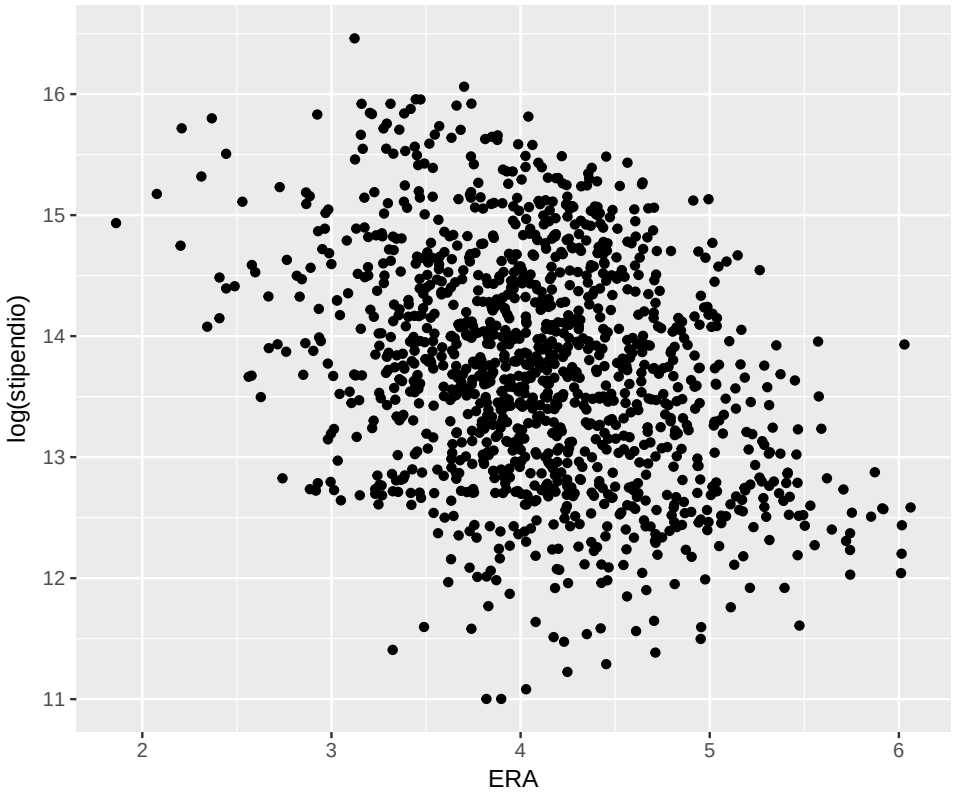
1.2.2 Stipendi

Seguendo la procedura adottata nel precedente paragrafo, analizziamo i fattori che influiscono sugli stipendi deflazionati dei lanciatori. Utilizziamo come variabile risposta il logaritmo degli stipendi medi di ogni lanciatore che abbia raggiunto in carriera 1'000 out (per un partente si raggiungono in circa due anni completi, per un rilievo in circa quattro). Considerando i singoli fattori come esplicative, il modello ottenuto è il seguente:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.3808	0.6501	26.74	0.0000
HRAM	-7.3809	1.1153	-6.62	0.0000
BBAM	-3.3106	0.3155	-10.49	0.0000
SOAM	0.4249	0.2803	1.52	0.1300
H1AM	1.7979	0.4754	3.78	0.0002
	R^2=0.219		Pr(>F)<0.001	

Vediamo che i fattori che determinano lo stipendio tra quelli considerati sono HRAM, BBAM e H1AM, mentre SOAM non risulta significativa. Questo rispecchia ciò che viene esposto in "Moneyball: The Art of Winning an Unfair Game" e cioè che i lanciatori vengono valutati correttamente, diversamente da quanto accade per i battitori. Infatti le esplicative significative in questo modello sono le stesse che influiscono sull'indicatore ERA. Volendo considerare invece come unica covariata ERA, il risultato è il seguente:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.4060	0.1705	96.25	0.0000
ERA	-0.5549	0.0398	-13.93	0.0000
	R^2=0.184		Pr(>F)<0.001	



Bisogna però fare una considerazione aggiuntiva sugli stipendi (sia dei battitori sia dei lanciatori): nella MLB, quando un giocatore firma il suo primo contratto viene vincolato alla squadra con cui ha firmato per sei anni dall'esordio nella massima serie (o per sette anni nelle leghe minori). Questo permette a squadre con basso budget di "sottopagare" per molti anni giocatori con ottime abilità, generando così un "rumore" non indifferente nei modelli stimati in questo elaborato.

1.2.3 Rilevanza statistica della Fielding Percetage

Fino ad ora, in questo paragrafo, è stato discusso esclusivamente il ruolo del lanciatore. Merita attenzione però anche il ruolo svolto dagli altri otto componenti. La statistica più utilizzata per descrivere l'abilità difensiva di un giocatore è la Fielding Percentage ($FP = \frac{E}{E+A+PO}$) dove E sono gli errori commessi dal giocatore, A sono gli assist forniti e PO gli out effettuati. Indubbiamente l'abilità difensiva di un giocatore non è una caratteristica da ignorare, ma il modo in cui questa viene valutata potrebbe essere sicuramente rivisto in chiave più moderna. Infatti, il concetto di "errore" è stato introdotto nel diciannovesimo secolo quando i giocatori non indossavano ancora i guantoni (nella forma in cui li conosciamo oggi). Questo impediva ai giocatori di quel tempo di prendere palline battute a poco più di un metro da loro; di conseguenza risultava facile determinare un errore, dato che le palline "giocabili" erano soltanto quelle che arrivavano praticamente addosso. Attualmente però è possibile per un giocatore effettuare prese una volta impensabili. Per questo motivo le palline "giocabili" sono difficili da quantificare e talvolta capita che un errore venga commesso in quanto in precedenza c'è stata una buona giocata (o un buon posizionamento) da parte dello stesso giocatore. Tenendo conto di questa critica esposta da Bill James molti anni fa, andiamo ad analizzare l'effetto della FP su ERA, capacità di vittoria e stipendi. Aggiungiamo dunque FP come esplicativa al modello precedentemente stimato avente come variabile risposta ERA:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.0528	2.2199	-2.73	0.0066
HRAM	40.3584	1.0442	38.65	0.0000
H1AM	15.3477	0.4025	38.13	0.0000
BBAM	9.6917	0.3971	24.41	0.0000
SOAM	0.3031	0.2607	1.16	0.2454
FP	3.0030	2.2196	1.35	0.1765

	Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)
1	649	13.32	(Modello senza FP)		
2	648	13.29	1	0.04	0.1761

Vediamo che gli errori difensivi non incidono sulla statistica ERA, anche perché i punti causati da un errore non sono a carico del lanciatore. Se invece includiamo FP come esplicativa nei modelli logistici stimati in questo paragrafo otteniamo i seguenti risultati:

- Accesso ai playoff

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-79.6059	37.6762	-2.11	0.0346
ERA	-2.1379	0.2335	-9.16	0.0000
FP	89.0924	38.0908	2.34	0.0193

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	652	640.57	(Modello senza FP)		
2	651	634.99	1	5.58	0.0182

- Vittoria della World Series

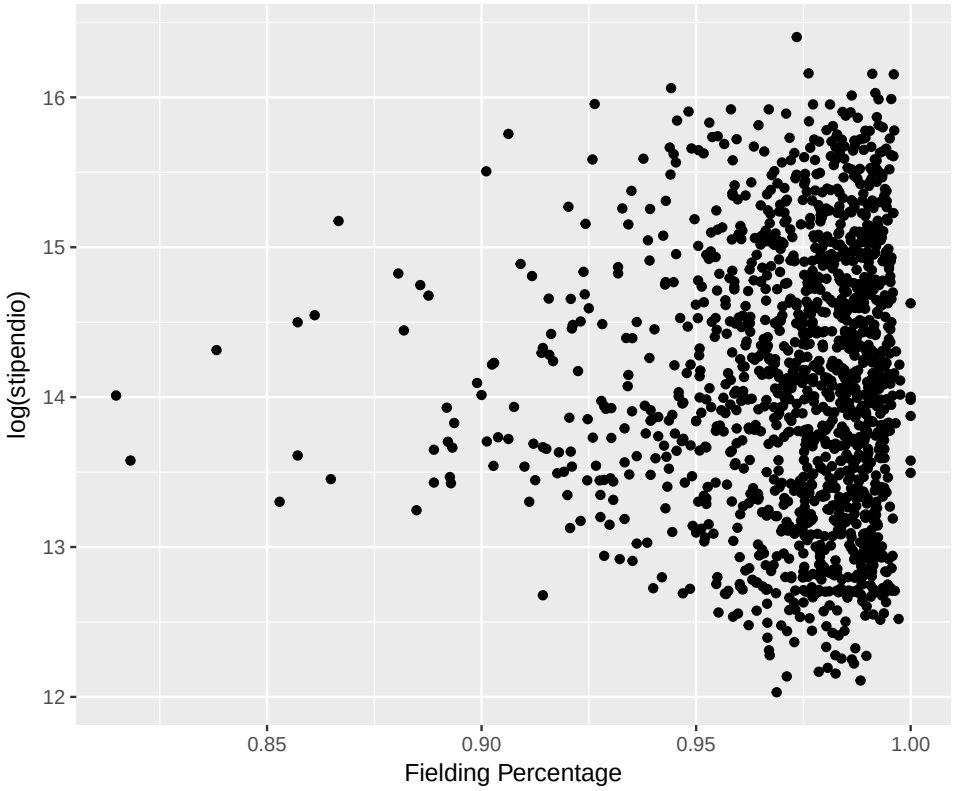
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-92.9712	87.8433	-1.06	0.2899
ERA	-1.4636	0.4869	-3.01	0.0026
FP	97.2018	88.7636	1.10	0.2735

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	652	178.84	(Modello senza FP)		
2	651	177.61	1	1.22	0.2684

Quindi FP risulta significativa al 5% per il primo modello, ma non per il secondo. Per quanto riguarda gli stipendi invece, prendiamo come variabile risposta il logaritmo dello stipendio medio deflazionato di tutti i giocatori con più di 300 partite giocate in carriera e consideriamo come esplicativa soltanto FP:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.5434	1.0394	13.03	0.0000
FP	0.6153	1.0682	0.58	0.5650
	R^2<0.001		Pr(>F)=0.565	

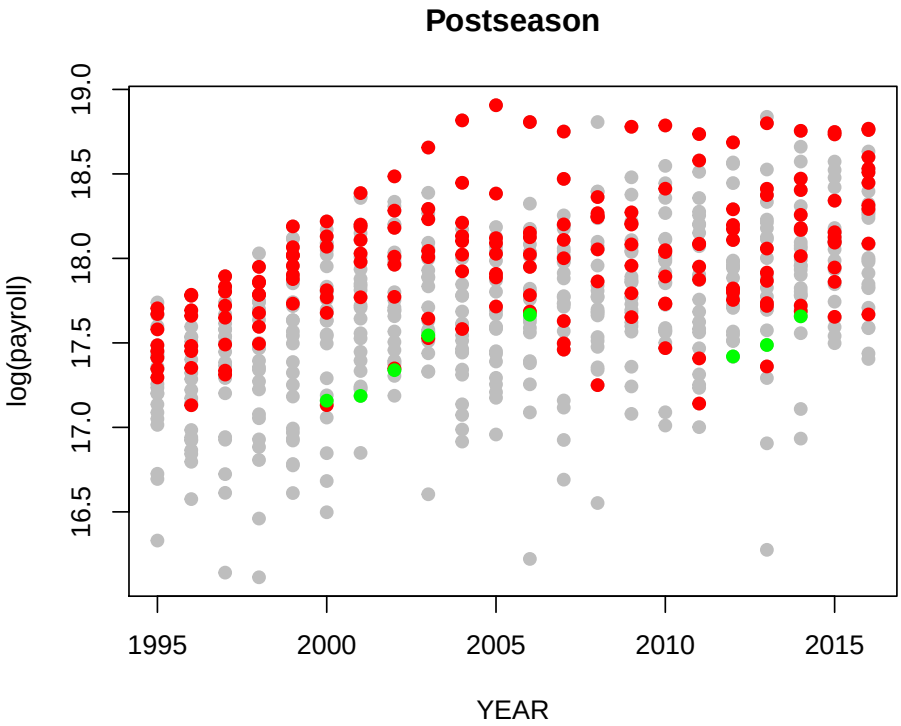
Il risultato ottenuto ci fa capire che i giocatori di baseball non vengono pagati in base alle loro abilità difensive o per lo meno non se valutate con questo indicatore. Una visualizzazione grafica del problema è la seguente:



1.3 Payrolls

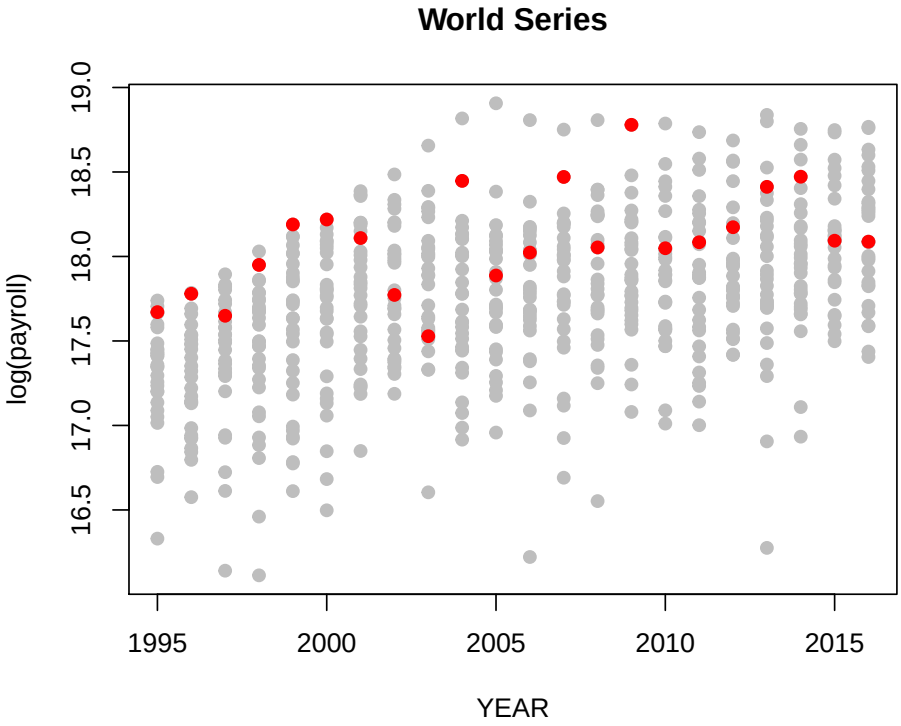
In un modo o nell'altro i soldi contano, anche nel mondo dello sport. Ci sono volte in cui l'abbondanza di denaro da spendere non ha nulla da offrire al pubblico se non un risultato quasi scontato, ma ci sono dei casi in cui la scarsità di budget disponibile regala storie difficili da scordare: questa è la doppia faccia dei payrolls (monte ingaggi). Si può credere che gli improbabili successi raccontati nelle storie sportive come "Moneyball: The Art of Winning an Unfair Game" siano il frutto di un caso, oppure si può cercare di capire i motivi che stanno alla base delle vittorie. Indubbiamente avere un grosso budget a disposizione aiuta, ma da solo non basta. Infatti risulta spesso più importante il "come" si spende rispetto al "quanto" si spende. Valutiamo ora l'evidenza empirica di questa considerazione andando ad analizzare il legame che sussiste tra payrolls[11] e capacità di vittoria. La prima variabile esplicativa che consideriamo è l'accesso ai playoff: dal primo grafico presente nella pagina seguente osserviamo che generalmente le squadre con payrolls elevati hanno una probabilità più alta di accedere ai playoff. Ci sono però dei casi in cui squadre con budget limitati sono riuscite comunque a raggiungere questo obiettivo. Nel grafico, in corrispondenza degli Oakland A's negli anni in cui la squadra ha raggiunto la postseason, è stato apposto un pallino verde. È evidente che è una delle poche squadre ad ottenere risultati simili con un payroll sotto la media della MLB.

[11] In realtà utilizziamo il logaritmo dei payrolls per smorzare alcuni effetti generati dagli outliers.



In rosso le squadre che sono riuscite ad accedere ai playoff

Per quanto riguarda le World Series il grafico corrispondente è il seguente:



In rosso le squadre che sono riuscite a vincere le World Series

L'effetto dei payrolls in questo caso è più evidente. Infatti, in particolar modo prima del 2002, anno in cui Billy Beane decise di cambiare il modo di operare sul mercato dei giocatori di baseball, le squadre vincenti risultano essere quasi esclusivamente nella parte alta della distribuzione annuale. Questo delinea il fatto che utilizzare dei metodi scientifici per colmare il salary gap può sicuramente portare a dei buoni risultati, ma quasi mai all'eccellenza. I modelli logistici corrispondenti ai due grafici appena presentati sono:

- Accesso ai playoff

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-30.1311	4.0722	-7.40	0.0000
log(payroll)	1.6384	0.2276	7.20	0.0000

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	780.96	(Modello nullo)		
2	652	719.31	1	61.65	0.0000

- Vittoria delle World Series

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-35.7200	9.6700	-3.69	0.0002
log(payroll)	1.8060	0.5350	3.38	0.0007

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	653	192.50	(Modello nullo)		
2	652	180.01	1	12.50	0.0004

Capitolo 2

Previsione

In questo capitolo verranno considerate contemporaneamente le statistiche di attacco e di difesa[1] delle singole squadre per prevedere l'accesso ai playoff e la vittoria delle World Series in una determinata stagione. Dopo aver strutturato adeguatamente i dati disponibili, verranno utilizzate due tecniche di classificazione: in primo luogo, per dare anche una visualizzazione grafica e quindi ottenere una più semplice interpretazione, verranno usati due alberi di classificazione, mentre in secondo luogo, data la maggior capacità previsiva, verranno utilizzate le random forests.

2.1 Pre-processing dei dati

Il primo problema riscontrato è la suddivisione in training set e test set. Dato che nell'ultimo capitolo di questo elaborato verranno analizzati due casi celebri riguardanti l'utilizzo della statistica nel mondo del baseball, utilizzeremo gli anni in questione come test set, vale a dire il 2002 e il 2004. Definito questo, quali dati possiamo utilizzare come training set? Per rispondere a questa domanda dobbiamo innanzitutto risolvere un altro problema. Bisogna considerare che i dati disponibili vanno dal 1995 al 2016 e dunque che è presente un fattore temporale (inoltre ogni stagione avrà un determinato numero di qualificati alla postseason e un solo vincitore delle World Series). Potrebbe quindi verificarsi, per assurdo, la situazione in cui una squadra ha ad esempio la media battuta più bassa di tutte le altre squadre durante la stagione considerata, ma comunque più alta di molte squadre di un'altra stagione. Ciò causerebbe non pochi problemi nella classificazione. Per questo motivo verrà eliminato il fattore temporale, cosa che ci consente di prendere come training set tutti gli anni dal 1995 al 2016 (esclusi il 2002 e il 2004)[2].

[1] Per evitare problemi di multicollinearità, si utilizzeranno le statistiche disaggregate quali: B1 (singoli effettuati), B2 (doppi effettuati), B3 (tripli effettuati), HR (fuoricampo effettuali), BB (basi su ball ricevute), SO (strike out subiti), HA1 (singoli+doppi+tripli concessi), HRA (fuoricampo concessi), BBA (basi su ball concesse), SOA (strike out effettuati) e E (errori commessi).

[2] In sostanza, depurando i dati dall'effetto temporale, è accettabile utilizzare sia il "passato" (dal 1995 al 2001) sia il "futuro" (dal 2005 al 2016) come training set.

2.1.1 Training set

Ora che abbiamo definito il training set, vediamo come è stato risolto il problema temporale appena esposto. Per ogni stagione e per ogni lega viene calcolato il massimo e il minimo di ogni variabile considerata. Vengono poi costruite le variabili normalizzate nel seguente modo:

$$\frac{VAR_{anno,lega} - min(VAR_{anno,lega})}{max(VAR_{anno,lega}) - min(VAR_{anno,lega})}$$

Così facendo è possibile confrontare la qualità di due squadre (espressa in termini di una singola variabile) appartenenti a anni e leghe diverse.

2.1.2 Test set

Per definire il test set non è sufficiente prendere in considerazione le statistiche finali degli anni 2002 e 2004 perchè la previsione è basata sulle statistiche disponibili a inizio stagione. Bisogna perciò trovare un modo appropriato per aggregare le statistiche degli anni precedenti dei singoli giocatori che compongono la squadra in un determinato anno. La procedura utilizzata è la seguente:

- Per ogni giocatore si considerano i 5 anni precedenti a quello da prevedere e si sommano le singole statistiche (comprese AB e IPouts) con pesi differenti[3]
- Per ogni squadra si aggregano le statistiche dei giocatori che la compongono in quell'anno
- Le statistiche di battuta vengono trasformate con la seguente formula:

$$\frac{VAR_{team}}{AB_{team}} * AB_{medio}$$

- Le statistiche di lancio vengono trasformate con la seguente formula:

$$\frac{VAR_{team}}{IPouts_{team}} * IPouts_{medio}$$

- Gli errori difensivi vengono trasformati con la seguente formula:

$$\frac{E_{team}}{(PO + A)_{team}} * (PO + A)_{medio}$$

- Escludendo le variabili AB, IPouts, PO e A si effettua la stessa normalizzazione illustrata nella sezione riguardante il training set

Così facendo, ci siamo ricondotti a variabili normalizzate in sintonia con quelle presenti nel training set.

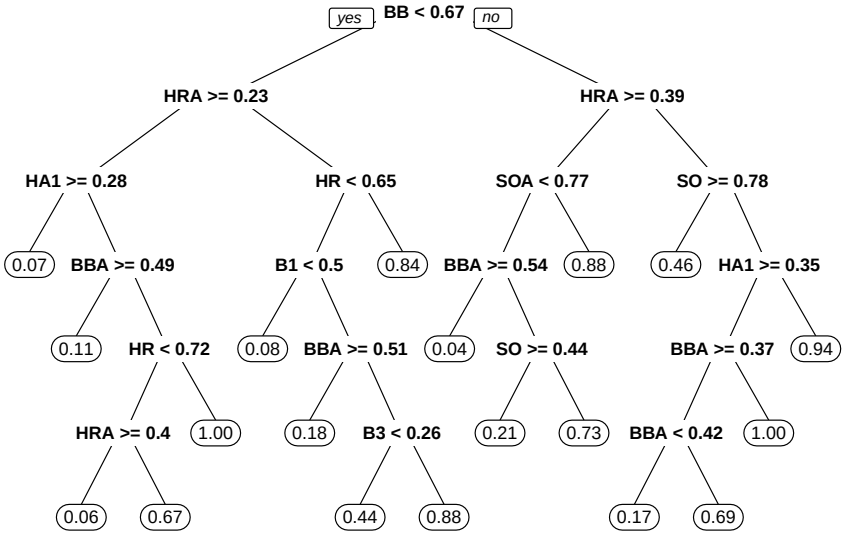
[3]Il primo anno pesa 2, il secondo 1.5, il terzo 1.2, il quarto 1.1 e il quinto 1.

2.2 Alberi di classificazione

Definendo un albero di decisione come «un grafo ad albero in cui il cosiddetto nodo radice è l'unico nodo senza rami entranti, le foglie sono tutti i nodi senza rami uscenti e i nodi che hanno un solo ramo entrante e più rami uscenti sono detti nodi interni o nodi di test»[**1, pag. 7**], possiamo specificare un albero di classificazione come un caso particolare in cui la variabile risposta è qualitativa o quantitativa discreta. Per costruire un albero di classificazione utile a scopi previsivi è necessario settare alcuni parametri come il numero minimo di osservazioni che devono essere presenti in un nodo affinchè questo possa essere ulteriormente suddiviso (nel nostro caso verrà settato a 18) e il numero minimo di osservazioni contenute in una foglia (nel nostro caso 6).

2.2.1 Accesso ai playoff

Considerando le variabili citate nel primo paragrafo di questo capitolo il risultato che si ottiene costruendo un albero di classificazione avente come risposta il passaggio alla postseason è il seguente:



Spostandoci verso la sinistra dell'albero si seguono i rami che hanno verificata la condizione imposta. Per esempio la prima condizione che si incontra è BB<0.67. Se una squadra avrà ottenuto poche basi su ball (meno del 67% del team con più basi si ball nella lega in quell'anno), allora percorrerà il ramo di sinistra, altrimenti quello di destra. Le foglie alla fine di ogni percorso contengono la probabilità di passaggio alla postseason di una

squadra che rispecchia le condizioni che hanno portato a quella foglia. Proviamo ora a percorrere i rami che portano a una probabilità di successo >0.5:

- BB<0.67 ⇒ HRA≥0.23 ⇒ HA1<0.28 ⇒ BBA<0.49 ⇒ HR<0.72 ⇒ HRA<0.4
- BB<0.67 ⇒ HRA≥0.23 ⇒ HA1<0.28 ⇒ BBA<0.49⇒ HR≥0.72
- BB<0.67 ⇒ HRA<0.23 ⇒ HR<0.65 ⇒ B1≥0.5 ⇒ BBA<0.51 ⇒ B3≥0.26
- BB<0.67 ⇒ HRA<0.23 ⇒ HR≥0.65
- BB≥0.67 ⇒ HRA≥0.39 ⇒ SOA<0.77 ⇒ BBA<0.54 ⇒ SO<0.44
- BB≥0.67 ⇒ HRA≥0.39 ⇒ SOA≥0.77
- BB≥0.67 ⇒ HRA<0.39 ⇒ SO<0.78 ⇒ HA1≥0.35 ⇒ BBA≥0.37 ⇒ BBA≥0.42
- BB≥0.67 ⇒ HRA<0.39 ⇒ SO<0.78 ⇒ HA1≥0.35 ⇒ BBA<0.37
- BB≥0.67 ⇒ HRA<0.39 ⇒ SO<0.78 ⇒ HA1<0.35

Nel nostro caso però questo non è l'unico modo di classificare una squadra come "partecipante ai playoff". Infatti ogni anno c'è un numero prestabilito di squadre che accedono ai playoff in ogni lega. Una soluzione alternativa sarebbe quella di selezionare le quattro squadre[4] con la probabilità stimata più alta. Visualizziamo allora le previsioni effettuate per gli anni 2002 e 2004 per l'American League[5]:

2002			2004		
Prob	Post	Team	Prob	Post	Team
0.94	0	BOS	1.00	1	ANA
0.94	1	NYA	0.94	1	NYA
0.88	0	CLE	0.94	0	OAK
0.73	0	SEA	0.67	1	BOS
0.11	0	TBA	0.07	0	BAL
0.08	1	OAK	0.07	0	CHA
0.07	1	ANA	0.07	0	CLE
0.07	0	BAL	0.07	0	DET
0.07	0	DET	0.07	0	KCA
0.07	0	KCA	0.07	1	MIN
0.07	1	MIN	0.07	0	TBA
0.07	0	TEX	0.07	0	TEX
0.07	0	TOR	0.07	0	TOR
0.06	0	CHA	0.06	0	SEA

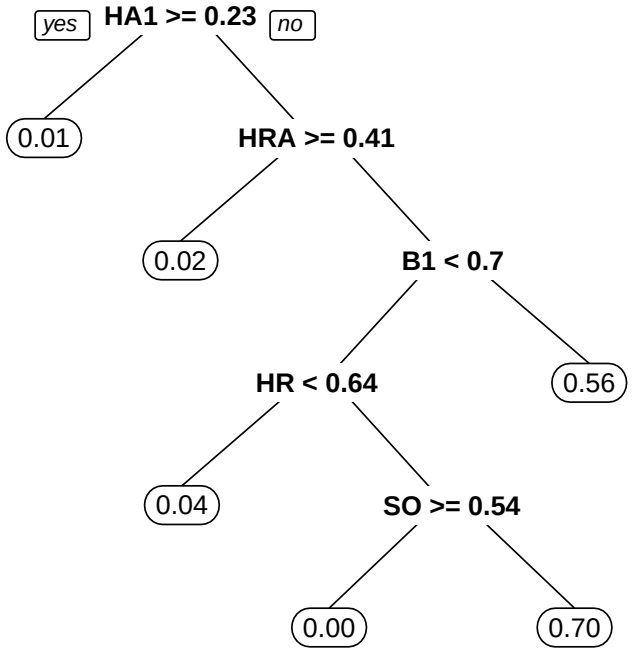
In questo caso le prime quattro squadre di ogni anno coincidono con le uniche ad avere assegnata una probabilità di successo >0.5, pertanto la scelta del metodo di decisione è ininfluente. Vediamo che per l'anno 2002 viene prevista come "partecipante ai playoff" una sola squadra che vi abbia preso parte effettivamente (New York Yankees), mentre nel 2004 questo accade per tre squadre (Los Angeles Angels, New York Yankees e Boston Red Sox).

[4]Nel 2002 e nel 2004 passavano quattro squadre per ogni lega.

[5]Lega in cui giocano Boston Red Sox e Oakland Athletics.

2.2.2 Vittoria delle World Series

Considerando invece come variabile risposta la vittoria delle World Series, l'albero che ne consegue è il seguente:



Come è facile notare, risulta essere molto più semplice in quanto i "successi" osservati sono molto inferiori rispetto a quelli del caso precedente. Di conseguenza si rivela più complicato effettuare delle suddivisioni e quindi costruire un numero elevato di ramificazioni. Esistono due percorsi che portano a una probabilità di vittoria >0.5:

- HA1<0.23 $\Rightarrow$ HRA<0.41 $\Rightarrow$ B1<0.7 $\Rightarrow$ HR$\geq$0.64 $\Rightarrow$ SO<0.54
- HA1<0.23 $\Rightarrow$ HRA<0.41 $\Rightarrow$ B1$\geq$0.7

Si può vedere come per la vittoria delle World Series, a differenza dell'accesso alla postseason, i lanciatori sono determinanti. Infatti le prime due condizioni da verificare per raggiungere una foglia contente una probabilità >0.5 riguardano le valide subite e i fuoricampo subiti (che ovviamente dovranno essere minori di una certa soglia). Nell'albero precedente, per ottenere una probabilità di successo sufficientemente alta, bastava rispettare condizioni molto meno restrittive sui lanciatori (sempre se questa mancanza veniva colmata da una buona fase di battuta). Questo potrebbe essere uno dei motivi per cui per una squadra con un payroll basso è molto difficile vincere le World Series. Infatti, come detto in precedenza, mentre i battitori vengono spesso valutati male e quindi pagati per delle abilità non utili al fine della vittoria, i lanciatori vengono valutati sufficientemente bene anche dai team che non seguono questa filosofia improntata sull'analisi dei dati. Ciò significa che squadre con budget elevati si aggiudicano i lanciatori più forti (e utili), ren-

dendo il salary gap difficile da colmare. Valutiamo ora la capacità predittiva dell'albero appena costruito:

2002			2004		
Prob	WS	Team	Prob	WS	Team
0.04	0	OAK	0.70	1	BOS
0.02	0	ARI	0.04	0	OAK
0.02	0	SEA	0.04	0	PHI
0.01	1	ANA	0.02	0	SEA
0.01	0	ATL	0.01	0	ANA
0.01	0	BAL	0.01	0	ARI
0.01	0	BOS	0.01	0	ATL
0.01	0	CHA	0.01	0	BAL
0.01	0	CHN	0.01	0	CHA
0.01	0	CIN	0.01	0	CHN
0.01	0	CLE	0.01	0	CIN
0.01	0	COL	0.01	0	CLE
0.01	0	DET	0.01	0	COL
0.01	0	FLO	0.01	0	DET
0.01	0	HOU	0.01	0	FLO
0.01	0	KCA	0.01	0	HOU
0.01	0	LAN	0.01	0	KCA
0.01	0	MIL	0.01	0	LAN
0.01	0	MIN	0.01	0	MIL
0.01	0	MON	0.01	0	MIN
0.01	0	NYA	0.01	0	MON
0.01	0	NYN	0.01	0	NYA
0.01	0	PHI	0.01	0	NYN
0.01	0	PIT	0.01	0	PIT
0.01	0	SDN	0.01	0	SDN
0.01	0	SFN	0.01	0	SFN
0.01	0	SLN	0.01	0	SLN
0.01	0	TBA	0.01	0	TBA
0.01	0	TEX	0.01	0	TEX
0.01	0	TOR	0.01	0	TOR

Per l'anno 2002 gli Oakland A's risulta la squadra con la probabilità di successo più alta, anche se nettamente inferiore a 0.5; detto ciò il modello sbaglia a prevedere il vincitore delle World Series nel primo anno. Nel 2004 invece l'unica squadra con probabilità >0.5 sono in Boston Red Sox che nell'anno in questione si aggiudicano effettivamente il titolo.

2.3 Random forests

Per aumentare le prestazioni previsive degli alberi di classificazione sono state introdotte delle tecniche leggermente più complesse basate sugli alberi stessi, come ad esempio il bagging[**1, pag. 29**], un metodo in cui vengono prodotti più alberi a partire da training set diversi[6] e l'output viene selezionato per votazione a maggioranza. Un passo successivo è stato fatto con l'utilizzo del metodo random forest[**1, pag. 30**]. Questa tecnica si basa sui concetti del bagging, ma con una caratteristica aggiuntiva: per ogni albero si utilizza un sottoinsieme casuale delle variabili di partenza. Le random forests generate successivamente saranno composte da 500 alberi, ognuno costruito a partire da tre variabili scelte casualmente.

2.3.1 Accesso ai playoff

Seguiamo la stessa procedura adottata nella sezione precedente: partendo dalle variabili normalizzate costruiamo prima di tutto una random forest avente come risposta l'accesso ai playoff. In questo caso non è possibile ottenere un grafico interpretativo, ma otteniamo una tabella contenente il valore dell'indicatore "importance"[7] per ogni esplicativa utilizzata:

	Importance
BB	38.10
HRA	29.59
BBA	28.66
HR	25.97
HA1	24.97
SOA	22.73
B1	17.84
B2	14.50
SO	14.21
E	13.58
B3	11.85

Le quattro variabili che incidono di più sulla classificazione sono le basi su ball (ottenute e concesse) e i fuoricampo (effettuati e subiti). Questo fatto è molto interessante in quanto evidenzia come la media battuta[8] è un indicatore piuttosto povero per spiegare la capacità di vittoria di una squadra. Vediamo ora se la capacità previsiva di questo modello è migliore rispetto all'albero di classificazione precedentemente costruito:

[6] Estratti con campionamento casuale con ripetizione dal training set originale.

[7] L'algoritmo stima, in modo piuttosto complesso, l'importanza di ogni variabile basandosi sulle variazioni dell'errore di previsione.

[8] Si ricorda che AVG è condizionata dalle valide nel loro complesso e non solo dai fuoricampo. Inoltre le basi su ball non rientrano nel calcolo di questa statistica.

2002			2004		
Voti	Post	Team	Voti	Post	Team
0.77	1	NYA	0.85	1	NYA
0.62	0	BOS	0.76	1	BOS
0.59	0	SEA	0.72	0	OAK
0.50	1	OAK	0.34	1	ANA
0.27	0	CHA	0.21	0	SEA
0.27	0	CLE	0.16	0	TOR
0.12	0	TOR	0.16	0	CHA
0.10	0	TBA	0.09	0	BAL
0.08	0	TEX	0.08	0	TBA
0.06	1	MIN	0.05	1	MIN
0.05	0	KCA	0.02	0	KCA
0.05	1	ANA	0.02	0	TEX
0.04	0	DET	0.01	0	DET
0.01	0	BAL	0.01	0	CLE

In questo caso risulta fondamentale decidere il metodo di classificazione da utilizzare. Si è ritenuto opportuno classificare come "partecipante ai playoff" le prime quattro squadre di ogni anno (anche se queste sono state classificate come tali da meno del 50% degli alberi della random forest). Così facendo, nel 2002 si classificano correttamente due team su quattro come "partecipanti ai playoff" (New York Yankees e Oakland A's), mentre nel 2004 si ripresenta una situazione analoga a quella dell'albero di classificazione della sezione precedente. Possiamo perciò affermare che la capacità previsiva è aumentata, anche se questo accade solo per il primo anno.

2.3.2 Vittoria delle World Series

Procediamo dunque con la costruizione della seconda random forest. Utilizzando come risposta la vittoria delle World Series e mantenendo le esplicative invariate, la tabella contentente l'indicatore "importance" di questo modello è la seguente:

	Importance
HA1	5.41
SO	4.35
HR	4.01
HRA	3.72
BB	3.71
B1	3.23
BBA	3.03
B3	2.86
SOA	2.72
E	2.70
B2	2.64

In questo caso è facile notare che le basi su ball contano molto meno, mentre la variabile con maggior peso è rappresentata dalle valide subite, esclusi i fuoricampo. Possiamo di conseguenza affermare che i fattori che incidono sulla possibilità di accedere alla postseason

sono differenti rispetto a quelli che incidono sulla possibilità di vincere le World Series. Bisogna però considerare il fatto che la previsione di quest'ultimo obiettivo è molto più complicata dato il numero ristretto di partite in cui si determina il vincitore dei playoff. Detto questo, analizziamo la classificazione effettuata da questo modello predittivo:

2002			2004		
Voti	WS	Team	Voti	WS	Team
0.18	0	ATL	0.22	1	BOS
0.17	0	ARI	0.15	0	PHI
0.13	0	SEA	0.15	0	OAK
0.05	0	COL	0.13	0	SEA
0.04	0	OAK	0.12	0	HOU
0.04	0	KCA	0.07	0	FLO
0.03	0	HOU	0.06	0	NYA
0.02	0	CLE	0.04	0	ARI
0.02	0	PIT	0.04	0	CHA
0.02	0	BOS	0.04	0	CHN
0.02	0	SFN	0.04	0	BAL
0.01	0	MON	0.03	0	ATL
0.01	0	CHN	0.02	0	CIN
0.01	0	SDN	0.02	0	SLN
0.01	0	SLN	0.02	0	TBA
0.01	0	DET	0.01	0	MIL
0.01	0	FLO	0.01	0	SFN
0.00	0	CHA	0.01	0	COL
0.00	0	MIN	0.01	0	MON
0.00	0	NYA	0.01	0	SDN
0.00	0	TOR	0.00	0	ANA
0.00	0	NYN	0.00	0	CLE
0.00	0	PHI	0.00	0	MIN
0.00	0	TEX	0.00	0	NYN
0.00	1	ANA	0.00	0	TEX
0.00	0	BAL	0.00	0	TOR
0.00	0	CIN	0.00	0	KCA
0.00	0	LAN	0.00	0	DET
0.00	0	MIL	0.00	0	LAN
0.00	0	TBA	0.00	0	PIT

Decidendo di classificare come "vincitore delle World Series" la squadra che ha ricevuto più "voti positivi" dagli alberi della random forest, notiamo che la capacità previsiva non è cambiata rispetto all'albero di classificazione precedente. Infatti nel 2002 vengono classificati come vincitori gli Oakland A's anche se nella realtà il titolo è stato vinto dai Los Angeles Angels, mentre nel 2004 il modello classifica correttamente i Boston Red Sox come vincitori.

Capitolo 3

Casi celebri

Ora che sono stati individuati i fattori che incidono maggiormente sulla capacità di vittoria di una squadra, analizziamo due casi diversi in cui la statistica ha giocato un ruolo fondamentale durante la costruzione della rosa. Il primo esempio riportato è quello degli Oakland A's, oggetto del libro "Moneyball: The Art of Winning an Unfair Game". L'autore racconta dettagliatamente molti aspetti della strategia di mercato di Billy Beane, partendo dal Draft, passando per la gestione delle squadre della Minor League, fino ad arrivare agli acquisti e agli scambi di giocatori. Non avendo a disposizione le statistiche delle leghe minori, né tanto meno quelle riguardanti i college e le high school, il capitolo sarà focalizzato sui giocatori in entrata e in uscita considerando esclusivamente le rose della MLB.

3.1 Oakland 2002

Alla fine della stagione 2001, brillantemente conclusa con 102 vittorie su 162 partite durante la regular season e con una sconfitta al primo turno dei playoff contro gli Yankees (vista la differenza di budget è comunque un buon risultato), gli Athletics devono privarsi di tre giocatori appena diventati free agents e fino a quel momento considerati le "stelle" del team: Jason Giambi (prima base), Johnny Damon (esterno centro) e Jason Isringhausen (closer). La preoccupazione principale di Billy Beane e Paul DePodesta riguardava il possibile sostituto di Giambi dato che le sue statistiche di battuta erano ineguagliabili. Singolarmente rappresentava un grosso problema, ma guardando le statistiche complessive dei tre battitori da rimpiazzare (Giambi, Damon e il battitore designato Olmedo Saenz) non sembrava un ostacolo insormontabile. Rendendosi conto che la on-base percentage è l'indicatore di sintesi che meglio rappresenta l'utilità di un battitore e che le basi su ball, ovvero l'elemento che contraddistigue la OBP dalla semplice media battuta, vengono estremamente sottovalutate dal punto di vista salariale, si decise che i tre sostituti da trovare dovevano complessivamente eguagliare la OBP media dei tre giocatori sopracitati.

Giocatore	AB	HR	BB	AVG	OBP	SLG	OPS
Jason Giambi	520	38	129	.342	.477	.660	1.137
Johnny Damon	644	9	61	.256	.324	.363	.687
Olmedo Saenz	305	9	19	.220	.291	.384	.675
				OBP media	.364		

Analizzando il problema sotto questo punto di vista, la soluzione sembra piuttosto semplice e soprattutto poco onerosa. Per sostituire Johnny Damon venne acquistato David Justice (esterno destro) e spostato Terrence Long dall'esterno destro all'esterno centro. Questa mossa avrebbe fatto perdere qualche attitudine difensiva alla squadra, ma Paul DePodesta stimò questa perdita in quindici punti subiti in più durante tutta la stagione, un costo tutto sommato sopportabile data la situazione. Per il ruolo di battitore designato si scelse di promuovere Jeremy Giambi (fratello di Jason) a titolare, anche se poi venne ceduto a stagione in corso per motivi disciplinari. La scelta più complicata fu quella del prima base. Secondo il GM degli A's l'acquisto del giovane Carlos Pena non era sufficiente a colmare il vuoto lasciato da Jason Giambi. Dato lo scarso budget a disposizione però, la squadra non poteva permettersi di acquistare un degno sostituto se considerato il giocatore nel suo complesso. L'attenzione così ricadde su Scott Hatteberg, ex ricevitore dei Boston Red Sox, reduce da un grave infortunio al braccio che avrebbe compromesso per sempre la sua carriera. La squadra di Oakland decise di reclutarlo come prima base (capita raramente di sforzare il braccio in quel ruolo) anche se non aveva mai giocato in quella posizione; dopotutto, quello che importava era semplicemente la sua OBP. Data la riluttanza del Team Manager Art Howe a schierarlo in campo viste le sue discutibili abilità difensive, Beane decise a stagione in corso di vendere Carlos Pena, giocatore che spesso scendeva in campo al posto di Hatteberg. Il segnale dunque era chiaro: la fase difensiva non era importante, bisognava puntare tutto sugli arrivi in base, raprresentati dalla OBP.

Anno	Team	Giocatore	AB	HR	BB	AVG	OBP	SLG	OPS
2001	NYY	David Justice	381	18	54	.241	.333	.430	.763
2002	OAK		398	11	70	.266	.376	.410	.785
2001	BOS	Scott Hatteberg	278	3	33	.245	.332	.345	.678
2002	OAK		492	15	68	.280	.374	.433	.807
2001	TEX	Carlos Pena*	62	3	10	.258	.361	.500	.861
2002	OAK		124	7	15	.218	.305	.419	.724
2001	OAK	Jeremy Giambi*	371	12	63	.283	.391	.450	.841
2002	OAK		157	8	27	.274	.390	.471	.862
2001	CWS	Ray Durham	611	20	64	.267	.337	.466	.804
2002	CWS		345	9	49	.299	.390	.446	.836
2002	OAK		219	6	24	.274	.350	.457	.806

**Giocatore ceduto a stagione in corso*

Inoltre, come è possibile notare dalla tabella soprastante, a stagione in corso venne acquistato Ray Durham, ufficialmente seconda base, che giocò spesso anche come battitore designato. Passando al reparto dei lanciatori, la partenza di Jason Isringhausen non sembrava essere molto preoccupante. Secondo Billy Beane infatti, i closer vengono spesso sopravvalutati dato che si ritrovano a lanciare molte volte in situazioni estremamente favorevoli. Sempre secondo il General Manager, i lanciatori su cui vale la pena investire sono i partenti, ma non era questo il caso.

Giocatore	IP	BB	HR	ERA
Jason Isringhausen	71.1	23	5	2.65

Come appurato nel primo capitolo, i lanciatori vengono già pagati sufficientemente bene per le giuste qualità e dunque il mercato presenta poche inefficienze da sfruttare. Con un budget ristretto da investire, si optò per l'acquisto di Billy Koch. Successivamente si scelse di rinforzare il bullpen, fino a metà stagione piuttosto deludente, acquistando Ricardo Rincon.

Anno	Team	Giocatore	IP	BB	HR	ERA
2001	TOR	Billy Koch	69.1	33	7	4.80
2002	OAK		93.2	46	7	3.27
2001	CLE	Ricardo Rincon	54.0	21	3	2.83
2002	CLE		35.2	8	3	4.79
2002	OAK		20.1	3	1	3.10

Nella regular season 2002 gli A's portarono a casa 103 vittorie, una in più dell'anno precedente; anche questa stagione però si concluse con la sconfitta nel primo turno dei playoff, questa volta per mano dei Minnesota Twins. Come constatato nei capitoli precedenti, il successo durante la regular season è più semplice da prevedere e da programmare rispetto alla vittoria delle World Series. Infatti giocando poche partite durante i playoff il fattore "fortuna" risulta essere spesso rilevante.

3.2 Boston 2004

Il caso dei Boston Red Sox è completamente diverso rispetto a quello appena esposto. Questa società non ha problemi finanziari, ma fino al 2002, anno in cui John W. Henry diventa proprietario, ha condotto campagne acquisti senza l'ausilio della statistica. La nuova proprietà, vedendo l'ottimo lavoro svolto da Billy Beane con la squadra di Oakland, nel 2003 decise di contattarlo per offrigli il posto di General Manager, con uno stipendio adeguato e una consistente possibilità di spesa sul mercato. Nonostante l'offerta venne declinata, Henry impose comunque una nuova linea dirigenziale: l'analisi dei dati doveva essere al centro di ogni decisione. La stagione seguente i Red Sox vinsero le World Series dopo 86 anni dall'ultima volta e ne vinsero altre due nel 2007 e nel 2013. Vediamo ora quali strategie di mercato hanno portato a questo successo.

Anno	Giocatore	AB	HR	BB	AVG	OBP	SLG	OPS
2002	Brian Daubach	444	20	51	.266	.348	.464	.812
2002	Rey Sanchez	357	1	17	.286	.318	.345	.662
2002	Tony Clark	275	3	21	.207	.265	.291	.556
2002	Shea Hillenbrand*	634	18	25	.293	.330	.459	.789
2003		185	3	7	.303	.335	.443	.778
2003	Nomar Garciaparra*	658	28	39	.301	.345	.524	.870
2004		156	5	8	.321	.367	.500	.867

**Giocatore ceduto a stagione in corso*

È palese come i giocatori ceduti tra il 2002 e il 2003 abbiamo una OBP non esalatante, soprattutto se messa a confronto con il caso precedentemente esposto. Per quanto riguarda Garciaparra invece, la cessione è stata dettata esclusivamente da problemi fisici. Per sostituire questi battitori sono stati effettuati i seguenti acquisti:

Anno	Team	Giocatore	AB	HR	BB	AVG	OBP	SLG	OPS
2002	CIN	Todd Walker	612	11	50	.299	.353	.431	.785
2003	BOS		587	13	48	.283	.333	.428	.760
2002	FLA	Kevin Millar	438	16	40	.306	.366	.509	.875
2003	BOS		544	25	60	.276	.348	.472	.820
2004	BOS		508	18	57	.297	.383	.474	.857
2002	(TOT)	Bill Mueller	366	7	52	.262	.350	.393	.743
2003	BOS		524	19	59	.326	.398	.540	.938
2004	BOS		399	12	51	.283	.365	.446	.811
2002	MIN	David Ortiz	412	20	43	.272	.339	.500	.839
2003	BOS		448	31	58	.288	.369	.592	.961
2004	BOS		582	41	75	.301	.380	.603	.983
2003	(TOT)	Mark Bellhorn	249	2	50	.221	.353	.293	.646
2004	BOS		523	17	88	.264	.373	.444	.817
2003	MON	Orlando Cabrera	626	17	52	.297	.347	.460	.807
2004	MON		390	4	28	.246	.298	.336	.634
2004	BOS		228	6	11	.294	.320	.465	.785
2004	BOS	Kevin Youkilis	208	7	33	.260	.367	.413	.780

Todd Walker, preso nel 2002 come seconda base al posto di Rey Sanchez, non avendo reso come ci si poteva aspettare in termini di arrivi in base è stato ceduto nel 2003 per acquistare Mark Bellhorn. Kevin Millar andò a sostituire il prima base Tony Clark, Bill Mueller prese il posto di Shea Hillenbrand e David Ortiz divenne battitore designato subentrando a Brian Daubach (entrambi registrarono presenze anche come prima base). Infine nel 2004, data la stagione disturbata dagli infortuni che stava vivendo Nomar Garciaparra, si optò per l'arrivo di un nuovo interbase, Orlando Cabrera. Va segnalata inoltre la presenza del roster di Kevin Youkilis, proveniente da una squadra minore affiliata ai Boston. Questo giocatore viene spesso citato anche nel libro di Michael Lewis, in cui viene definito da Paul DePodesta "il dio greco delle basi su ball". Per quanto riguarda i lanciatori, tra le uscite più importanti si registrano quelle di Rolando Arrojo (fine carriera) e Ugueth Urbina (venduto) nel 2002 e John Burkett (fine carriera) nel 2003.

Anno	Giocatore	IP	BB	HR	ERA
2002	Rolando Arrojo	81.1	27	7	4.98
2002	Ugueth Urbina	60.0	20	8	3.00
2003	John Burkett	181.2	47	20	5.15

Il primo anno vennero acquistati Mike Timlin e Byung-Hyun Kim rispettivamente come rilievo e closer. Quest'ultimo però l'anno seguente subì un grave infortunio che lo constrinse a saltare quasi tutta la stagione. Venne perciò sostituito da Keith Foulke con ottimi risultati. Tuttavia il lanciatore in entrata più importate per la vittoria delle World Series è il partente Curt Schilling. Confrontando le sue statistiche con quelle di John Burkett il salto di qualità è evidente.

Anno	Team	Giocatore	IP	BB	HR	ERA
2002	(TOT)	Mike Timlin	96.2	14	15	2.98
2003	BOS		83.2	9	11	3.55
2004	BOS		76.1	19	8	4.13
2002	ARI	Byung-Hyun Kim	84.0	26	5	2.04
2003	ARI		43.0	15	6	3.56
2003	BOS		79.1	18	6	3.18
2004	BOS		17.1	7	1	6.23
2003	ARI	Curt Schilling	168.0	32	17	2.95
2004	BOS		226.2	35	23	3.26
2003	OAK	Keith Foulke	86.2	20	10	2.08
2004	BOS		83.0	15	8	2.17

Ciò che distinge principalmente la gestione dei Boston Red Sox da quella degli Oakland Athletics sono i rispettivi payrolls. Abbiamo visto che entrambe le squadre hanno puntato sulla fase di battuta concentrandosi sull'indicatore OBP. Come evidenziato più volte durante l'elaborato, questa strategia risulta vincente nella regular season, ma non è altrettanto valida nei playoff. Allora per quale motivo i Red Sox sono riusciti ad ottenere risultati migliori rispetto agli A's? Inizialmente si potrebbe pensare che la qualità dei lanciatori sia stata determinante nella postseason. Analizzando però le statistiche si nota che nel 2002 gli A's avevano una ERA di squadra di 3.68[1], mentre lo stesso indicatore risultava essere 4.18 per i Red Sox nel 2004. L'attenzione quindi potrebbe ricadere sulla OBP e la SLG, che sommate formano la statistica OPS, oggetto di studio nel primo capitolo.

	OAK 2002	BOS 2004
OBP	.339	.360
SLG	.432	.472
OPS	.771	.832

Osservando questa tabella si può concludere che le statistiche di battuta dei Boston Red Sox nel 2004 sono nettamente migliori rispetto a quelle degli Oakland A's nel 2002. Tuttavia questa informazione non è sufficiente per poter rispondere adeguatamente alla domanda precedente. Si può affermare invece che un alto budget permette di acquistare giocatori di un certo livello, che verosimilmente a fine stagione avranno totalizzato statistiche migliori rispetto agli altri e che di conseguenza aumenteranno le probabilità di vittoria della squadra.

[1]Nonostante il budget limitato, la squadra di Oakland riusciva ad aggiudicarsi alcuni dei lanciatori migliori selezionandoli fin da giovani senza spendere somme ingenti.

Conclusioni

Come spesso evidenziato durante l'elaborato, la prima conclusione che si può trarre è che l'utilizzo delle statistiche ha un limite evidente: quando la numerosità campionaria si abbassa, l'accuratezza dei risultati si riduce. In una competizione come la Major League Baseball, in cui 162 partite di regual season servono a determinare solo le otto squadre su trenta che hanno diritto a partecipare ai playoff, durante la singola stagione gli sforzi fatti per costruire una squadra vincente tramite l'analisi dei dati potrebbero venire vanificati in pochissime partite di postseason. Se si giudica la bontà di una squadra e il lavoro svolto dalla dirigenza esclusivamente dalla vittoria finale però, si rischia di commettere un errore. Il caso degli Oakland Athletics è il più eclatante: come si può dire che l'approccio usato da Billy Beane è sbagliato e non funziona soltanto perchè non è riuscito a conquistare le World Series? Indubbiamente nessuno assicura la certezza nel prevedere un evento, tantomeno nel mondo dello sport dove un piccolo fatto imprevisto, come l'infortunio di un giocatore chiave, potrebbe compromettere l'intera stagione. L'utilizzo di tecniche statistiche in questo ambito mira a ridurre l'incertezza dovuta al fattore umano e a supportare le decisioni che altrimenti sarebbero basate esclusivamente su sensazioni e supposizioni di scout e manager. Inoltre, tramite un'accurata analisi, è possibile colmare la differenza di valori in campo con squadre aventi payrolls più elevati. Basandoci ancora sulla stagione 2002, la classifica della "division" di cui fanno parte gli A's è la seguente:

	W	L	Payroll
Oakland	103	59	$ 41,942,665
Anaheim	99	63	$ 62,757,041
Seattle	93	69	$ 86,084,710
Texas	72	90	$ 106,915,180

È evidente come l'ordine delle squadre sia crescente nei payrolls; ciò significa che in questo caso il budget disponibile a inizio stagione non solo non è stato rilevante, ma addirittura sembra essere un ostacolo per le squadre. Anche se può sembrare un'affermazione assurda, il senso è che certe società, sicure che un payroll elevato possa assicurare una buona possibilità di vittoria, badano poco al modo in cui vengono effettuati gli investimenti. Fino a quando sussisterà questa situazione, le squadre più povere riusciranno ancora ad ottenere buoni risultati nonostante il salary gap.

Come già esposto nel terzo capitolo, i dati utilizzati non sono completi e rispondo solo agli interrogativi riguardanti le squadre della MLB. Nel libro di Michael Lewis vengono esposte anche le strategie adottate dal GM degli A's durante il draft e i criteri da seguire per valutare i giocatori della Minor League. Ad esempio, secondo Billy Beane e Paul DePodesta, le abilità di un battitore variano da una lega all'altra. Così un giocatore che aveva una media battuta molto alta al college, potrebbe non avere ottimi rendimenti

in una squadra della Minor League. Lo stesso potrebbe accadere per il passaggio dalla Minor League alla Major League. L'unica abilità che secondo lo staff di Oakland rimane invariata è la "disciplina al piatto", ovvero, semplificando, la capacità di capire se un lancio sarà ball o strike. L'indicatore che meglio rappresenta questa abilità è il numero di basi su ball ottenute. Un discorso analogo è stato fatto anche per i lanciatori. Dopo un'attenta analisi, Paul DePodesta ha notato che l'unica abilità costante per i lanciatori risulta essere la capacità di far battere la pallina a terra piuttosto che al volo, rappresentata dalla statistica AO/GO (rapporto fra gli out effettutati dopo una battuta al volo e gli out effettuati dopo una battuta a terra). Questo indicatore è importante in quanto un lanciatore che induce a battere prevalentemente la pallina a terra, difficilmente subirà valide che permettono all'avversario di arrivare oltre la prima base. Inoltre l'approccio basato sugli indicatori classici adottato in questo elaborato non è l'unico possibile. Nel sesto capitolo del libro "Moneyball: The Art of Winning an Unfair Game" viene citata la società AVM Systems, fondata nel 1994 da Ken Mauriello e Jack Armbruster con lo scopo di analizzare più accuratamente ogni evento generato durante una partita di baseball. A tal fine decisero di registrare ogni palla battuta in gioco come un insieme di tre fattori: velocità, traiettoria e settore di campo. In questo modo, ad esempio, un doppio battuto sull'esterno destro diventa una palla battuta con una certa velocità e una certa traiettoria in un determinato settore di campo. Collezionando dati per molti anni, è possibile stimare per ogni genere di palla battuta in gioco la quantità di punti da essa generati. Ragionando in termini di punti attesi, è facile capire il contributo di ogni giocatore al netto del fattore "fortuna". Si pensi ad un esterno che prende in tuffo una palla battuta con una specifica modalità, tale da avere come valore 0.7 punti attesi. Si può affermare che il lanciatore è stato "fortunato", il battitore "sfortunato" e l'esterno è da considerarsi "bravo". Volendo formalizzare, al lanciatore dovrebbero essere addebitati 0.7 punti, stesso punteggio che andrebbe accreditato al battitore e all'esterno. Il metodo appena esposto è chiaramente più completo e preciso rispetto a quello utilizzato nei precedenti capitoli. Tuttavia, non essendo applicabile senza i dati registrati da AVM Systems, si è scelto di procedere con uno studio più classico che rimane comunque un valido spunto da cui partire per costruire una squadra vincente.

Bibliografia

[1] Marocchi, Fabio, *Apprendimento di alberi di decisione*

[2] James G., Witten D., Hastie T., Tibshirani R., *An Introduction to Statistical Learning*, Springer, 2013

[3] Nolan D., Lang D. T., *Data Science in R, A Casa Studies Approach to Computional Reasoning and Problem Solving*, CRC Press, 2015

[4] Lewis, Michael, *Moneyball: The Art of Winning an Unfair Game*, W. W. Norton & Company, 2003

Sitografia

[S1] Grabiner, David, *The Sabermetric Manifesto*, URL `http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto`

[S2] *Moneyball, un nuovo approccio per la caccia ai campioni*, URL `http://www.baseball.it/leggi_articolo.asp?id=15125`

[S3] `http://www.seanlahman.com`

[S4] `http://fred.stlouisfed.org`

[S5] `http://www.baseball-reference.com`