

ALMA MATER STUDIORUM

UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI SCIENZE STATISTICHE

“PAOLO FORTUNATI”

Corso di Laurea in Scienze Statistiche

LA PREVISIONE NEL BASEBALL

Analisi di Modelli di Regressione Logistica

Presentata da:

Alessandro Montanari

matricola: 0000794489

Relatore:

Prof. Daniele Ritelli

APPELLO I

ANNO ACCADEMICO 2018 / 2019

Introduzione

«È difficile fare previsioni, soprattutto sul futuro». Questa frase, attribuita al celebre ex giocatore e allenatore di baseball - nonché prolifico aforista - Yogi Berra, sintetizza adeguatamente l'idea che gli eventi futuri dipendano da un certo grado di aleatorietà, di cui è particolarmente complicato tenere conto quando si effettuano previsioni di qualunque genere.

L'obiettivo di questo elaborato sarà quello di testare empiricamente questa teoria, analizzando l'effetto del caso sui risultati della corrente stagione di Major League Baseball. A tal fine, in seguito a un primo capitolo di natura descrittiva dove saranno illustrati i dataset, le variabili e le tecniche utilizzate, si procederà allo sviluppo e alla selezione di modelli di regressione logistica atti a stimare la probabilità di vittoria di una singola partita in funzione di determinati predittori. Questi modelli saranno infine utilizzati nel quarto capitolo per la simulazione di un elevato numero di stagioni, allo scopo di verificare il ruolo della componente aleatoria su una singola ripetizione della serie di partite e determinare la forza della relazione fra abilità della squadra e vittoria finale del campionato. Il software utilizzato nel corso di tutta la fase di analisi sarà R, per via della sua duttilità e dell'ampia disponibilità di bibliografia.

Indice

Capitolo 1 – <i>Baseball, dati e regressione logistica</i>	6
1.1 La Major League Baseball	6
1.2 Obiettivo dell’analisi	7
1.3 Fonti dei dati	8
1.4 Stima delle abilità e dello stato di forma	8
1.5 Dataset e variabili	11
1.6 La regressione logistica	12
1.6.1 <i>Il modello</i>	12
1.6.2 <i>La stima dei parametri</i>	14
1.6.3 <i>Interpretazione e test sui coefficienti</i>	14
1.6.4 <i>La scelta fra modelli non annidati</i>	15
1.6.5 <i>Diagnostica e valutazione del modello</i>	16
 Capitolo 2 – <i>Analisi descrittive e modelli bivariati</i>	18
2.1 Il talento delle squadre	18
2.1.1 <i>Statistiche descrittive</i>	18
2.1.2 <i>Modelli di regressione bivariata</i>	22
2.2 Il talento dei lanciatori partenti	25
2.2.1 <i>Statistiche descrittive</i>	25
2.2.2 <i>Modelli di regressione bivariata</i>	29
2.3 Il talento dei lanciatori di rilievo	31
2.3.1 <i>Statistiche descrittive</i>	31
2.3.2 <i>Modelli di regressione bivariata</i>	35
2.4 Lo stato di forma delle squadre	35
2.5 La variabile risposta	38
2.6 Riepilogo	39

Capitolo 3 – <i>La scelta del modello</i>	40
3.1 I modelli interi	40
3.2 Il modello stepwise	45
3.3 La capacità previsiva	47
3.4 Diagnostica	49
3.5 Riepilogo	53
<hr/>	
Capitolo 4 – <i>Causalità, previsioni e simulazioni</i>	54
4.1 Obiettivi e problematiche	54
4.2 Una simulazione della stagione in corso	56
4.3 1000 simulazioni della stagione in corso	57
4.4 Simulazione della Post Season	60
4.5 Conclusioni	63
<hr/>	
<i>Bibliografia</i>	65

Capitolo 1

Baseball, dati e regressione logistica

1.1 La Major League Baseball

La Major League Baseball (*MLB*) è la lega professionistica nord-americana di baseball che rappresenta la massima espressione di questo sport a livello mondiale. È costituita dall'unione della *National League* e dell'*American League*, entrambe formate da 15 squadre suddivise in 3 *divisioni* a seconda della località geografica (*Est*, *Centro*, *Ovest*). La tabella 1.1 riassume brevemente la struttura della MLB.

American League			National League		
Squadra	Località	Divisione	Squadra	Località	Divisione
Orioles	Baltimore	Est	Braves	Atlanta	Est
Red Sox	Boston	Est	Marlins	Miami	Est
Yankees	New York	Est	Mets	New York	Est
Rays	Tampa Bay	Est	Phillies	Philadelphia	Est
Blue Jays	Toronto	Est	Nationals	Washington	Est
White Sox	Chicago	Centro	Cubs	Chicago	Centro
Indians	Cleveland	Centro	Reds	Cincinnati	Centro
Tigers	Detroit	Centro	Brewers	Milwaukee	Centro
Royals	Kansas City	Centro	Pirates	Pittsburgh	Centro
Twins	Minnesota	Centro	Cardinals	St. Louis	Centro
Angels	Los Angeles	Ovest	Diamondbacks	Arizona	Ovest
Astros	Houston	Ovest	Rockies	Colorado	Ovest
Athletics	Oakland	Ovest	Dodgers	Los Angeles	Ovest
Mariners	Seattle	Ovest	Padres	San Diego	Ovest
Rangers	Texas	Ovest	Giants	San Francisco	Ovest

Tabella 1.1: Struttura delle squadre associate alla Major League Baseball

Nel corso della *Regular Season* ogni squadra disputa 162 partite. Di queste, generalmente 76 sono contro squadre della stessa divisione (19 per ognuna delle 4 avversarie), 66 contro squadre da altre divisioni della stessa lega e 20 contro avversarie militanti nella lega opposta. Al termine, le vincitrici di ogni divisione unitamente alle due migliori escluse di ogni lega (che si scontrano in una partita secca per ottenere la *Wild Card*), si qualificano alla *Postseason*. I playoff sono articolati con una formula a eliminazione diretta: le squadre di National e American League si scontrano in serie di rispettivamente 5 e 7 partite (*Division Series* e *Championship Series*) per decretare i due campioni di lega. Questi, infine, si affrontano

nelle *World Series*, un’ulteriore serie al meglio delle 7 partite che determina il vincitore finale del campionato.

Il baseball, rispetto ad altri sport, ha una struttura relativamente *discreta*: ogni partita è composta da una serie di eventi concatenati, classificabili entro un insieme finito e relativamente ridotto di situazioni possibili. Questo, unitamente all’elevato numero di partite disputate ogni anno (2.430 solo nella stagione regolare, contro le 380 di un campionato di Serie A di calcio) lo rende particolarmente adatto ad analisi quantitative e, in generale, ad essere rappresentato in numeri: ne sono testimonianza la mole di dati e letteratura disponibile sull’argomento. Infine, un altro aspetto importante da considerare è che nel baseball - così come nella maggior parte degli sport americani - non esiste il pareggio: se al termine delle riprese regolamentari le squadre hanno segnato lo stesso numero di punti, si procede ad oltranza finché una di queste non ha la meglio. Ciò permette di trasformare il risultato di una partita in una variabile dicotomica (1 = vittoria, 0 = sconfitta), con un notevole guadagno di semplicità per lo studio in questione.

1.2 *Obiettivo dell’analisi*

L’obiettivo della prima parte di analisi sarà quello di determinare un modello per spiegare la probabilità di vittoria di una partita di MLB in funzione di dati conosciuti a priori (ovvero, prima dell’inizio dell’incontro). In generale, ci possiamo aspettare che la probabilità di vittoria di una squadra A in una partita contro un’altra squadra B dipenda da almeno **tre fattori principali**:

- L’abilità della squadra A ;
- L’abilità della squadra avversaria B ;
- Il fattore casa.

In secondo luogo, potrebbero sicuramente giocare un ruolo importante:

- L’abilità dei lanciatori partenti di A e di B ;
- L’abilità dei lanciatori di rilievo di A e di B ;
- Lo stato di forma di A e di B .

Il problema evidente è che i valori di 8 di queste 9 variabili (ossia tutte ad eccezione del fattore casa) sono incogniti, e necessitano dunque di essere stimati. La fase di ricerca ed elaborazione dei dati sarà fondamentalmente incentrata su questo aspetto.

1.3 *Fonti dei dati*

Il database utilizzato per l’analisi è il risultato dell’unione ed elaborazione dei dati provenienti da 3 fonti: il *Lahman Database*, il *Retrosheet Database* e la piattaforma *Fangraphs*. Tutte le librerie sono liberamente consultabili online: per gli indirizzi completi si rimanda alla bibliografia.

Il Lahman Database contiene statistiche e informazioni generali raccolte stagione per stagione su battitori, lanciatori e difensori, sia a livello individuale che a livello di squadra. Per l’analisi sono stati utilizzati i dati sui lanciatori e i dati complessivi di squadra relativi alle stagioni 2015, 2016, 2017 e 2018.

Il Retrosheet Database raccoglie i riassunti di ogni singola partita disputata dal 1871 al 2018, completi di punteggio, giocatori schierati, lanciatori partenti, e informazioni quali orario d’inizio, durata e luogo della partita. Sono stati selezionati i dati sulle 9718 partite di regular season disputate fra il 2015 e il 2018.

Fangraphs, infine, contiene alcune statistiche avanzate sulla produzione offensiva delle squadre, e permette inoltre di dividere i dati riguardanti i lanciatori partenti e i rilievi anche a livello di squadra.

1.4 *Stima delle abilità e dello stato di forma*

I potenziali stimatori delle variabili descritte nel paragrafo 1.2 sono stati selezionati considerando sia indicatori semplici, aventi il vantaggio della facile interpretazione, che statistiche avanzate in grado di isolare le abilità individuali e la produzione complessiva dal possibile confondimento di altre componenti. Le abilità sia di squadra che individuali sono state considerate fisse su base stagionale, ma variabili da una stagione all’altra¹.

Nel dettaglio, gli stimatori presi in considerazione per l’abilità di squadra sono:

¹ Mentre appare logico considerare variabile nel tempo l’abilità di una squadra, poiché il talento complessivo del gruppo può essere modificato drasticamente da una stagione all’altra tramite l’acquisto, la cessione o lo scambio dei giocatori, si potrebbe obiettare che lo stesso non sia valido per le abilità individuali. Tuttavia, sia per motivi di praticità che per mancanza di prove a supporto dell’ipotesi di costanza nel tempo del talento individuale, anch’esso sarà ritenuto fisso solo su base stagionale. Inoltre, anche se questa ipotesi fosse valida, il suo effetto dovrebbe riflettersi sulle performance del giocatore, tendendo a stabilizzarle nel tempo (l’elevato numero di partite aiuterebbe sicuramente a ridurre l’effetto di fluttuazioni casuali) e, dunque, non costituirebbe in ogni caso un problema.

- **Differenza punti per partita (RDg)** =
$$\frac{\text{Punti segnati} - \text{Punti subiti}}{\text{N}^{\circ} \text{ di incontri disputati}}$$

La differenza punti per partita è una misura del talento complessivo di una squadra, includendo sia la fase offensiva che quella difensiva. Una $\text{RDg} > 0$ implica che la squadra in questione segna mediamente più punti rispetto a quelli che subisce, e viceversa. È misurata a partire dai dati del *Lahman Database*.

- **Punti segnati per partita (RSg)** =
$$\frac{\text{Punti segnati}}{\text{N}^{\circ} \text{ di incontri disputati}}$$

I punti segnati per partita sono un indicatore dell'abilità esclusivamente offensiva di una squadra. Può risultare utile ai fini dell'analisi centrare la variabile, così che un valore di $\text{RSg} > 0$ indichi una squadra che solitamente segna più punti rispetto alla media, e viceversa. Anche questo stimatore è ricavato dai dati *Lahman*.

- **Weighted Run Created + (wRC+)**

I wRC+ sono i *punti creati* dalla squadra, derivati dalle relative statistiche offensive avanzate, aggiustati per il fattore di campo e la lega di appartenenza². È un indicatore centrato su 100, e un incremento o decremento unitario implica un incremento o decremento di un punto percentuale rispetto alla media di lega nella produzione offensiva. Si tratta dunque di una stima dell'abilità offensiva di una squadra, al netto di fattori confondenti o casuali. È ottenuta da *Fangraphs*.

Per la stima dell'abilità dei lanciatori sono state utilizzate:

- * **Earned Run Average (ERA)** =
$$\frac{\text{Punti guadagnati}}{\text{Riprese lanciate}} \times 9$$

L'ERA è una media dei punti guadagnati da un lanciatore nell'arco di 9 riprese, la durata tradizionale di una partita di baseball. Per punti guadagnati si intendono i punti segnati dagli avversari che non derivano da errori difensivi.

- * **Fielding Independent Percentage (FIP)** =
$$\frac{13 \times \text{HR} + 3 \times (\text{BB} + \text{HBP}) - 2 \times \text{K}}{\text{IP}} + c$$

² Per la formula completa si rimanda a <https://library.fangraphs.com/offense/wrc/>

Dove HR sono i *fuoricampi concessi*, BB le *basi ball concesse*, HBP i *battitori colpiti*, K gli *strikeout effettuati*, IP le *riprese lanciate* e c è un valore costante su base stagionale³ che centra la FIP nello stesso intervallo di valori dell'ERA. La FIP è un tentativo di isolare la performance del lanciatore da quella della difesa, o in generale da eventi che non sono sotto il suo diretto controllo. ERA e FIP, sia individuali che di squadra, sono state ottenute da Lahman.

$$* \text{ Expected FIP (xFIP)} = \frac{13 \times \left(FB \times \text{Lg} \frac{HR\%}{FB} \right) + 3 \times (BB + HBP) - 2 \times K}{IP} + c$$

La *expected FIP* (FIP attesa) è una versione avanzata della FIP, dove il totale di fuoricampi concessi è rimpiazzato da una stima dei fuoricampi attesi data dal numero di *fly ball* (palle al volo) concesse e il rapporto medio di fuoricampi per fly ball nella lega di appartenenza. L'obiettivo è quello di eliminare l'effetto di ulteriori elementi casuali sulla misura della vera abilità del lanciatore. I valori utilizzati di xFIP sono stati scaricati da Fangraphs.

Appare evidente dall'osservazione delle formule che, mentre per quanto riguarda gli estimatori delle abilità di squadra valori maggiori sono indicatori di squadre con talento maggiore, ERA FIP e xFIP sono tanto inferiori quanto più il lanciatore riesce a limitare la produzione avversaria, e sono dunque negativamente correlate con la sua abilità.

Per la stima dello stato di forma della squadra, infine, sono stati ricavati:

- **RDg nelle 5 partite precedenti** = $\frac{(P. \text{ segnati nelle 5 partite precedenti} - P. \text{ subiti nelle 5 partite precedenti})}{5}$
- **RDg nelle 10 partite precedenti** = $\frac{(P. \text{ Segnati nelle 10 partite precedenti} - P. \text{ subiti nelle 10 partite precedenti})}{10}$

Entrambe le variabili consistono in una misura dello stato di forma complessivo della squadra, misurato in termini di differenza punti media rispettivamente nell'arco delle 5 e 10 partite precedenti a quella considerata. Sono state ottenute a partire dai dati partita per partita di *Retrosheet*.

³ Per le stagioni considerate, i valori di c sono i seguenti: 3.134 (2015), 3.147 (2016), 3.158 (2017), e 3.161 (2018). Per maggiori informazioni si rimanda a <https://www.fangraphs.com/guts.aspx?type=cn>

- **Percentuale di vittorie nelle 10 partite precedenti** = $\frac{\text{N}^{\circ} \text{ di vittorie nelle 10 partite precedenti}}{10}$

Quest'ultimo consiste in un tentativo di misurare lo stato di forma di una squadra in base ai suoi risultati nelle 10 partite precedenti a quella considerata. Anch'esso è stato derivato a partire dal database *Retrosheet*.

1.5 Dataset e variabili

Il database finale è costituito da 9717 osservazioni (le partite di regular season disputate fra il 2015 e il 2018⁴) di 48 variabili, di cui due sono gli ID delle squadre, una è la variabile risposta (vittoria/sconfitta della squadra di casa), 5 sono relative a informazioni spazio-temporali (giorno, mese, anno, orario e stadio della partita) e le rimanenti consistono negli estimatori precedentemente descritti, riportati per entrambe le squadre sia in forma ordinaria che in termini di differenza dalla media (per questioni di interpretabilità che saranno approfondite in seguito).

La tabella 2 contiene un riassunto delle variabili inserite nel dataset, completo di breve descrizione, obiettivo e fonte di provenienza.

N°	Variabile	Descrizione	Stima di	Fonte
1	hteam	Squadra in casa	/	Retrosheet
2	vteam	Squadra in trasferta	/	Retrosheet
3	W (variabile risposta)	1 = Vittoria squadra in casa 0 = Vittoria squadra in trasferta	/	Retrosheet
4	day	Giorno della partita	/	Retrosheet
5	month	Mese della partita	/	Retrosheet
6	year	Anno della partita	/	Retrosheet
7	time	N = Partita notturna D = Partita diurna	/	Retrosheet
8	park	Stadio	/	Retrosheet
9-10	RDg_hteam RDg_vteam	Differenza punti per partita delle due squadre (casa e trasferta)	Abilità complessiva delle squadre	Lahman

⁴ Le partite non corrispondono a 9720 (=2430*4) dal momento che può capitare che una squadra dispiuti in una stagione 1-2 partite in più (se sono necessari degli spareggi in seguito ad arrivo a pari merito) o in meno (se vengono rinviate delle partite ritenute ininfluenti ai fini di classifica) rispetto alle 162 previste. Inoltre, dai dati è stata eliminata una partita sospesa a causa del maltempo e conclusa in pareggio (Cubs – Pirates del 29/09/2016, per maggiori informazioni si consulti <https://www.mlb.com/news/cubs-pirates-game-suspended-ends-in-tie/c-204121484>)

11-12	RSg_hteam RSg_vteam	Punti segnati per partita delle due squadre (casa e trasferta)	Abilità offensiva delle squadre	Lahman
13-14	wRC+_hteam wRC+_vteam	Weighted Run Created+ delle due squadre (casa e trasferta)	Abilità offensiva delle squadre	Fangraphs
15-16	ERA_hs ERA_vs	Earned Run Average dei lanciatori partenti delle due squadre	Abilità dei lanciatori partenti	Lahman
17-18	FIP_hs FIP_vs	Fielding Independent Pitching dei partenti delle due squadre	Abilità dei lanciatori partenti	Lahman
19-20	xFIP_hs xFIP_vs	Expected Fielding Independent Pitching dei partenti delle due squadre	Abilità dei lanciatori partenti	Fangraphs
21-22	ERA_bullpen_hs ERA_bullpen_vs	Earned Run Average dei rilievi delle due squadre	Abilità dei rilievi	Fangraphs
23-24	FIP_bullpen_hs FIP_bullpen_vs	Fielding Independent Pitching dei rilievi delle due squadre	Abilità dei rilievi	Fangraphs
25-26	xFIP_bullpen_hs xFIP_bullpen_vs	Expected Fielding Independent Pitching dei rilievi delle squadre	Abilità dei rilievi	Fangraphs
27-42	Variabili 11-26 centrate nella media			
43-44	RDg_l5_h RDg_l5_v	Media della differenza punti nelle ultime 5 partite delle due squadre	Stato di forma delle squadre	Retrosheet / elaborazione
45-46	RDg_l10_h RDg_l10_v	Media della differenza punti nelle ultime 10 partite delle due squadre	Stato di forma delle squadre	Retrosheet / elaborazione
47-48	W_l10_h W_l10_v	Percentuale di vittorie nelle ultime 10 partite delle due squadre	Stato di forma delle squadre	Retrosheet / elaborazione

Tabella 1.2: Riassunto delle variabili contenute nel dataset finale

L’obiettivo del capitolo 2 sarà quello di studiare singolarmente la distribuzione delle variabili descritte e il loro effetto sulla probabilità di vittoria della squadra di casa, ignorando quello degli altri predittori. Il capitolo 3, invece, si occuperà di considerarne la relazione con Y al netto degli effetti delle altre variabili, e di selezionare quelle maggiormente in grado di spiegare la variabile risposta per arrivare alla specificazione del modello di regressione multivariata ottimale. Prima di passare alla fase analitica, il paragrafo seguente sarà dedicato all’introduzione dei concetti di base della tecnica di regressione utilizzata.

1.6 La regressione logistica

1.6.1 Il modello

L’obiettivo dei modelli di regressione uniequazionali è quello di spiegare la distribuzione di una variabile risposta (Y) in funzione dei valori assunti da una o più variabili

indipendenti (X_1, X_2, \dots, X_k). Nella regressione lineare semplice, si assume che il valore atteso condizionato di Y sia una combinazione lineare delle variabili X :

$$E[Y | X_1, X_2, \dots, X_k] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

Dove β_0 è il valore assunto da Y quando tutte le X sono pari a 0 (*intercetta*), e β_i ($i \neq 0$) esprime la variazione media di Y corrispondente a una variazione unitaria di X_i , a parità delle altre variabili indipendenti (*coefficiente di regressione*).

Tuttavia, quando la variabile risposta ha natura dicotomica (ovvero, può assumere solo due valori, convenzionalmente 0 per la non occorrenza e 1 per l'occorrenza dell'evento⁶), la regressione lineare risulta inadatta in quanto può portare a stimare valori di Y al di fuori dell'intervallo consentito. In questo caso, il modello generalmente preferito è quello di *regressione logistica*, in cui si assume la seguente relazione fra il valore atteso condizionato della variabile risposta e i predittori:

$$E[Y | X_1, X_2, \dots, X_k] = \text{Prob}[Y=1 | X_1, X_2, \dots, X_k] = \frac{e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$$

Infatti, per le variabili dicotomiche il valore atteso di Y equivale alla probabilità che Y assuma valore 1. Il rapporto della probabilità di un evento con il suo complemento a 1 è definito *odds*, ed esprime di quanto l'occorrenza di tale evento è più probabile della non occorrenza. Nel caso del modello logistico:

$$\text{Odds}[Y|X] = \frac{\text{Prob}[Y=1|X]}{1 - \text{Prob}[Y=1|X]} = e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}$$

La trasformata logaritmica dell'odds si definisce *trasformata logit*:

$$\text{logit}(Y|X) = \ln \left[e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)} \right] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

La trasformata logit svolge un ruolo centrale nell'ambito della regressione logistica. In primo luogo, costituisce la funzione legame del modello nell'ambito del modello lineare generalizzato: essa stabilisce, infatti, la relazione lineare fra il valore atteso condizionato di Y e le variabili indipendenti X_1, X_2, \dots, X_k . In aggiunta, il modello logistico è un modello probabilistico diretto, poiché specificato in termini di $P[Y=1|X]$. Questo implica che l'unica

⁶ Nel caso in questione, 1 indica la vittoria della squadra di casa e 0 la vittoria della squadra in trasferta.

assunzione alla base del modello sia relativa proprio alla linearità della relazione fra funzione logit e regressori⁷.

1.6.2 La stima dei parametri

Il miglior metodo di stima dei parametri nel caso di modello lineare è quello dei minimi quadrati (*LS*), che consiste nel trovare il vettore $\hat{\beta}$ che minimizza la somma dei quadrati dei residui fra valori osservati e attesi di Y. Nel caso di variabile risposta dicotomica, tuttavia, le proprietà degli estimatori LS vengono a meno ed è preferibile utilizzare il metodo della massima verosimiglianza. Questo consiste nel ricercare i valori di $\hat{\beta}$ che massimizzano la funzione di verosimiglianza, definita in base alla probabilità di osservazione della realizzazione campionaria. Dal momento che la soluzione di questo problema di massimo porta a equazioni non lineari nei parametri, per arrivarvi è necessario ricorrere a speciali metodi iterativi (implementati nei software statistici come R), per il cui approfondimento si rimanda a *Hosmer, Lemeshow e Sturdivant* (2013) e a *Harrel* (2016).

1.6.3 Interpretazione e test sui coefficienti

Nel modello logistico multivariato i coefficienti β sono interpretabili come segue:

$$\begin{aligned} e^{\beta_1} &= \frac{e(\beta_0 + \beta_1 \times (\mathbf{X}_1 = \alpha + 1) + \beta_2 X_2 + \dots + \beta_k X_k)}{e(\beta_0 + \beta_1 \times (\mathbf{X}_1 = \alpha) + \beta_2 X_2 + \dots + \beta_k X_k)} = \\ &= \frac{\text{Odds } [Y | \mathbf{X}_1 = \alpha + 1]}{\text{Odds } [Y | \mathbf{X}_1 = \alpha]} = \text{Odds Ratio } [Y, X_1 = \alpha + 1 / X_1 = \alpha] \\ \beta_1 &= \ln \{ \text{Odds Ratio } [Y, X_1 = \alpha + 1 / X_1 = \alpha] \} \end{aligned}$$

Per *odds ratio* si intende il rapporto tra gli odds di Y relativi a due determinati valori delle variabili indipendenti. Generalmente, questo indica l'aumento relativo di odds di Y derivante da una certa variazione delle X.

Il coefficiente di regressione β_1 corrisponde al logaritmo dell'odds ratio di Y relativo a un incremento unitario della variabile X_1 . In altre parole, β_1 esprime la variazione nel logaritmo dell'odds della variabile risposta che è causata da un incremento unitario della variabile X_1 , a parità degli altri regressori. Se questa non è continua ma dicotomica, β_1 indica l'incremento nel

⁷ *Harrel, 2016*

log-odds di Y relativo a $X_1 = 1$ (presenza) rispetto a $X_1 = 0$ (assenza). In generale, valgono le seguenti relazioni:

- **Odds Ratio > 1:** Un incremento unitario di X provoca un incremento significativo nell'odds di Y [$P(Y)/1-P(Y)$]. L'entità di tale incremento è pari all'(OR-1) %.
- **Odds Ratio = 1:** Una variazione di X non ha alcun effetto sull'odds della variabile risposta. Corrisponde al caso in cui $e^{\hat{\beta}} = 1$ e dunque $\hat{\beta} = 0$.
- **Odds Ratio < 1:** Un aumento unitario di X provoca una diminuzione significativa nell'odds di Y. L'entità di tale riduzione è pari all'(1-OR) %.

I due test principalmente utilizzati per verificare la significatività dei coefficienti sono il test di Wald e il metodo del rapporto delle verosimiglianze. Il primo prevede di mettere a rapporto la stima di massima verosimiglianza del coefficiente con la stima del suo errore standard asintotico, ottenendo una statistica test che si distribuisce come una normale standardizzata per dimensioni campionarie sufficienti. Il secondo, invece, consiste nella trasformata logaritmica del rapporto del massimo della funzione di verosimiglianza relativa al modello ridotto (privo della variabile in questione), e il massimo relativo al modello completo (contenente la variabile in questione). Quest'ultimo test può inoltre essere utilizzato per verificare l'ipotesi di non significatività congiunta di più coefficienti, quella di significatività generale del modello e per il confronto fra modelli annidati.

$$\begin{aligned} * \text{ Test di Wald} &= \frac{\hat{\beta}}{\text{S.E. asintotico } (\hat{\beta})} \sim N(0,1) \\ * \text{ Test del rapporto delle verosimiglianze} &= -2\ln \frac{L(\text{mod k parametri})}{L(\text{mod. k+r parametri})} \sim \chi^2_{(r)} \end{aligned}$$

Definite le caratteristiche della regressione logistica, appare evidente come questa rappresenti la tecnica ideale per modellare il fenomeno trattato, dal momento che mette in relazione diretta la probabilità di successo dell'evento con i valori assunti dai regressori. Nel caso delle partite di Major League Baseball, sarà oggetto dell'analisi la relazione fra la probabilità di vittoria della squadra di casa ($\text{Prob } [Y=1]$) e i valori assunti dalle variabili indipendenti X descritte nei paragrafi 1.4 e 1.5.

1.6.4 La scelta fra modelli non annidati

Quando si hanno a disposizione più modelli di regressione logistica *non annidati* (ovvero, le cui variabili indipendenti non sono sottoinsiemi l'uno dell'altro) uno dei principali

indicatori per scegliere il migliore tra quelli disponibili è l'*Akaike Information Criterion* (AIC). L'AIC consiste in una stima della quantità di informazione persa quando si utilizza un certo modello per rappresentare il fenomeno reale: tanto minore è questa quantità, ovviamente, tanto maggiore è la qualità del modello. La stima è ottenuta tramite un compromesso fra massimizzazione della bontà di adattamento ai dati del modello e minimizzazione della complessità dello stesso. La formula per il calcolo dell'AIC è:

$$\text{AIC} = 2k - 2\ln(L)$$

Dove k è il numero di parametri e L è il massimo della funzione di verosimiglianza del modello.

1.6.5 Diagnostica e valutazione del modello

Nell'ambito della regressione logistica, l'unica assunzione che richiede di essere verificata è la linearità della relazione fra variabili indipendenti e trasformata logit, come menzionato nel paragrafo 1.6.1. Gli altri accertamenti che devono essere effettuati sul modello e sulle osservazioni riguardano⁸:

- *L'assenza di osservazioni influenti* (cioè osservazioni che, se eliminate, modificano notevolmente le stime dei parametri).

Questo è verificato tramite l'analisi dei residui studentizzati, dei valori di leverage (h) e della distanza di Cook. Alti valori dei residui studentizzati sono indicatori di possibili *outliers*, ovvero osservazioni la cui variabile risposta si discosta particolarmente da quanto previsto in base ai valori dei regressori. Leverage elevati, invece, identificano osservazioni per cui le variabili indipendenti assumono valori estremi, mentre la distanza di Cook misura l'influenza complessiva della singola osservazione sulle stime dei parametri.

- *L'eventuale multicollinearità fra le variabili indipendenti* (quando fra alcuni predittori vi è una forte correlazione, le stime dei parametri e dei corrispondenti errori standard possono risultare distorte).

La presenza di multicollinearità è testata tramite il fattore di inflazione della varianza (VIF), che misura l'aumento della varianza del singolo coefficiente di regressione dovuta alla correlazione con le altre variabili. Si considerano problematici valori di $VIF > 10$.

- *La capacità previsiva del modello.*

⁸ Per approfondimenti sulla diagnostica e la valutazione del modello si rimanda a Harrel (2016), Hosmer, Lemeshow e Sturdivant (2013), Zhang (2016).

Questa, infine, è quantificata per mezzo dell'area sottostante la curva ROC (Receiver Operating Characteristic). L'idea alla base della curva ROC è che un modello di regressione logistica può essere utilizzato come metodo di classificazione se si seleziona un valore k di cut-off della probabilità predetta per cui se $\text{Prob}_i > k$ si assegna i alla classe 1, e viceversa se $\text{Prob}_i \leq k$ si assegna i alla classe 0. Si definisce *sensibilità* (*o proporzione di veri positivi*) del metodo di classificazione il rapporto fra osservazioni correttamente classificate in 1 e totale di osservazioni appartenenti a 1. La curva ROC consiste nella rappresentazione dell'insieme delle coppie *sensibilità* e *1-specificità* (*proporzione di falsi positivi*) al variare del valore di cut-off k . L'area ad essa sottostante (AUC, *Area Under Curve*) è una misura della concordanza fra probabilità previste e valori osservati della variabile risposta. Quest'area è compresa fra 0.5 (caso in cui il metodo di classificazione equivale alla classificazione casuale) e 1 (ovvero, completa separazione delle classi in base a un determinato valore di cut-off). Tanto più l'AUC si avvicina a 1, chiaramente, tanto più il modello di regressione logistica ha capacità previsive ottimali. Sebbene il fine dell'elaborato non sia quello di ottenere un metodo di classificazione dei risultati futuri in base a un semplice valore di cut-off⁹, l'area sottostante la curva ROC rimane una valida tecnica per quantificare la capacità previsiva del modello.

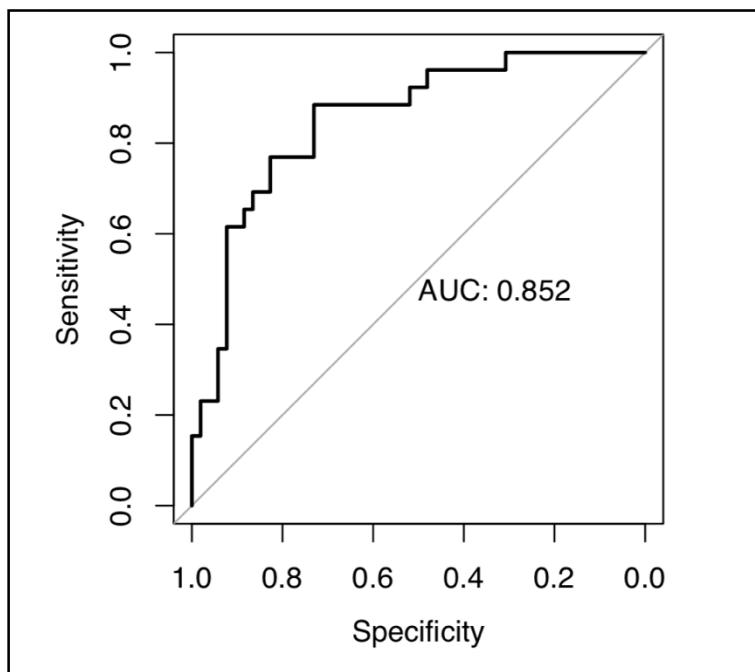


Figura 1.1: Esempio di curva ROC tratto da Kassambara, 2017. L'area sottostante pari a 0.852 evidenzia la bontà previsiva del metodo di classificazione utilizzato.

⁹ Dal momento che l'obiettivo è quello di analizzare il ruolo del caso sui risultati delle partite, non avrebbe senso classificare i risultati futuri in vittoria/sconfitta semplicemente in base a un valore soglia delle probabilità previste.

Capitolo 2

Analisi descrittive e modelli bivariati

2.1 Il talento delle squadre

2.1.1 Statistiche descrittive

La tabella 2.1 contiene le statistiche descrittive degli stimatori utilizzati per l'abilità delle squadre.

	RDg	RSg	wRC+
Minimo	-1.6667	3.5370	85
1° Quartile	-0.5525	4.1638	94
Mediana	-0.0278	4.4352	100
Media	-0.0002	4.4557	99.9583¹⁰
3° Quartile	0.5031	4.7006	104.25
Massimo	1.6235	5.5309	123
D. Standard	0.7212	0.4193	7.543
C.V.	/ ¹¹	0.0941	0.0755

Tabella 2.1: Statistiche descrittive delle variabili utilizzate per la stima dell'abilità delle squadre (stagioni 2015-2018)

Si può osservare come la differenza punti per partita (RDg) presenti una media quasi uguale a 0: questo è in linea con l'assunzione che squadre di media abilità tendano a segnare e subire lo stesso numero di punti, squadre con un talento maggiore siano propense a segnare più punti rispetto a quelli subiti, e squadre mediocri subiscano mediamente un maggior numero di punti rispetto a quelli segnati. I punti segnati per partita (RSg) sembrano distribuirsi con minore variabilità rispetto alla RDg, come evidenziato da un valore inferiore di deviazione standard¹². Le squadre di medio talento, nell'ipotesi che questo sia un valido indicatore della loro abilità, segneranno mediamente 4.46 p/p (punti a partita), le squadre nel 25% migliore almeno 4.70 p/p, e le squadre nel 25% peggiore segneranno in media meno di 4.16 p/p. I punti creati (wRC+),

¹⁰ La media non è esattamente uguale a 100 poiché non è pesata per i turni offensivi (PA) delle diverse squadre.

¹¹ Il coefficiente di variazione non è utilizzabile per variabili con media ≈ 0 , dal momento che si ottiene come rapporto fra la deviazione standard e la media.

¹² Le due variabili hanno uguale unità di misura (vedasi formule del paragrafo 1.4) e sono dunque direttamente comparabili in termini di variabilità tramite la deviazione standard. Nel caso contrario, invece, è preferibile utilizzare per il confronto una misura adimensionale come il coefficiente di variazione (C.V.).

infine, sono centrati sul valore 100 e presentano una variabilità inferiore rispetto ai punti segnati per partita. I grafici 2.1, 2.2 e 2.3 sintetizzano la distribuzione delle variabili in questione nelle stagioni considerate, mostrandone la densità di frequenza in funzione dei valori assunti.

Grafico 2.1

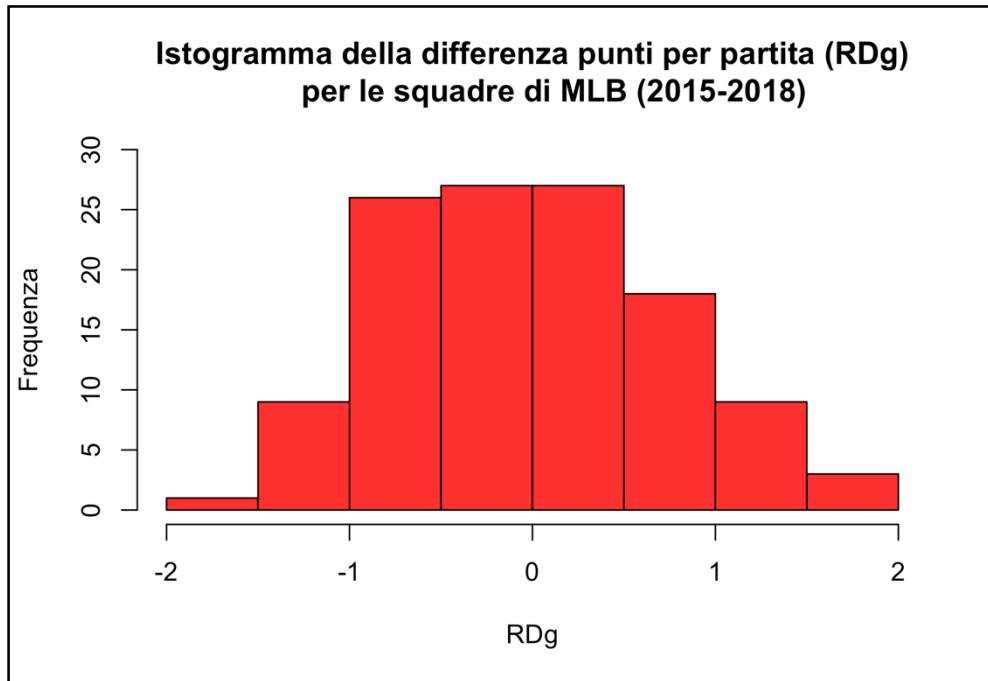


Grafico 2.2

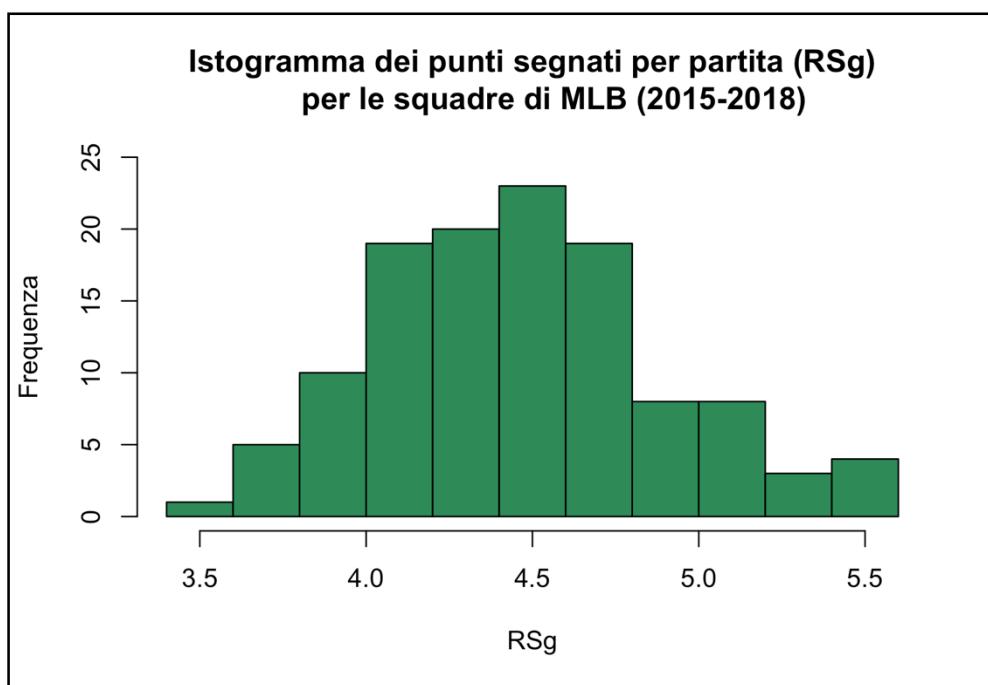
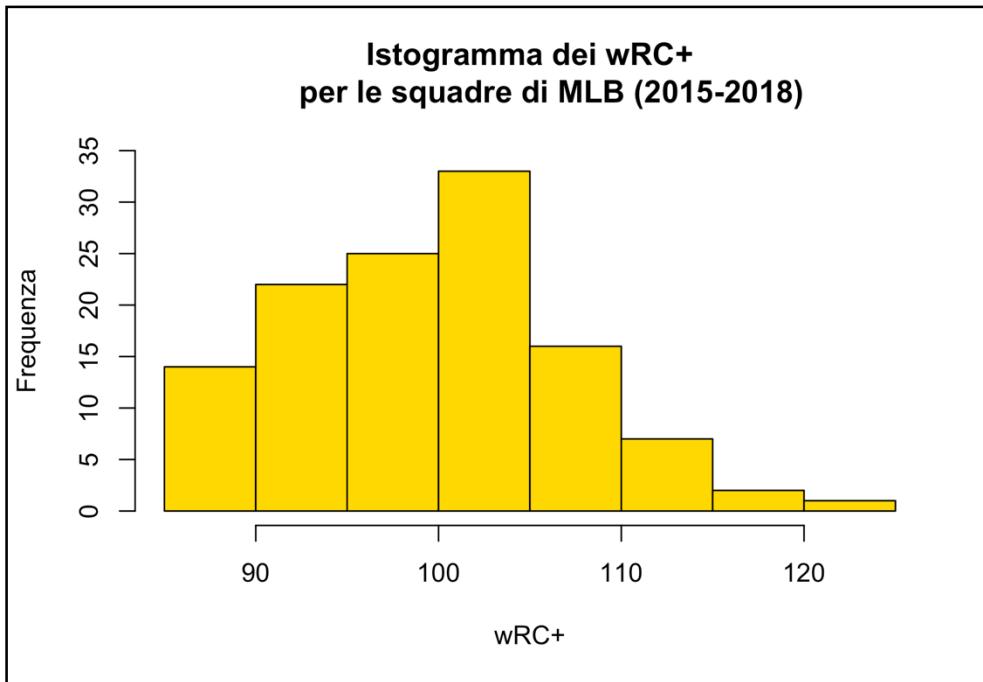


Grafico 2.3



RDg e RSg appaiano distribuite abbastanza simmetricamente rispetto ai corrispondenti valori centrali: questo implica che il numero di squadre con abilità sopra la media tende ad essere uguale a quello delle squadre sotto la media. I wRC+, invece, sembrano più concentrati nei valori inferiori della distribuzione.

Un altro aspetto che può risultare utile valutare è come varia la distribuzione delle variabili nel dataset delle partite, suddivise per squadra in casa e squadra in trasferta, in funzione dei valori assunti dalla variabile risposta. Ciò permette di testare graficamente la presenza di una relazione fra la distribuzione dei singoli predittori e il valore assunto da Y (vittoria o sconfitta della squadra di casa). Tale relazione sarà successivamente quantificata dalla specificazione dei modelli bivariati. I boxplot rappresentati di seguito (grafici 2.4, 2.5, 2.6) costituiscono un buon metodo per visualizzare sinteticamente quanto descritto: i bordi della *scatola* rappresentano il primo e il terzo quartile della distribuzione, la linea centrale costituisce il valore mediano e i *baffi* (le linee esterne alla scatola) sono di lunghezza pari a 1.5 volte il range interquartile e indicano il grado di dispersione delle osservazioni. Eventuali *outliers* (punti al di fuori dei baffi) sono evidenziati.

Grafico 2.4: Distribuzione assunta da RDg della squadra di casa e RDg della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

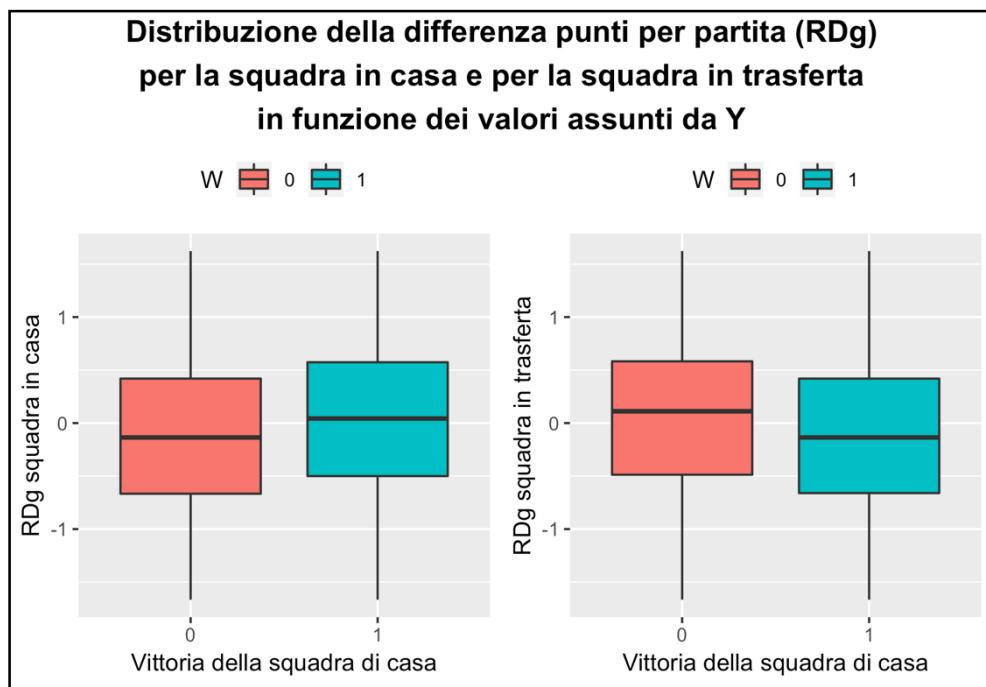


Grafico 2.5: Distribuzione assunta da RSg della squadra di casa e RSg della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

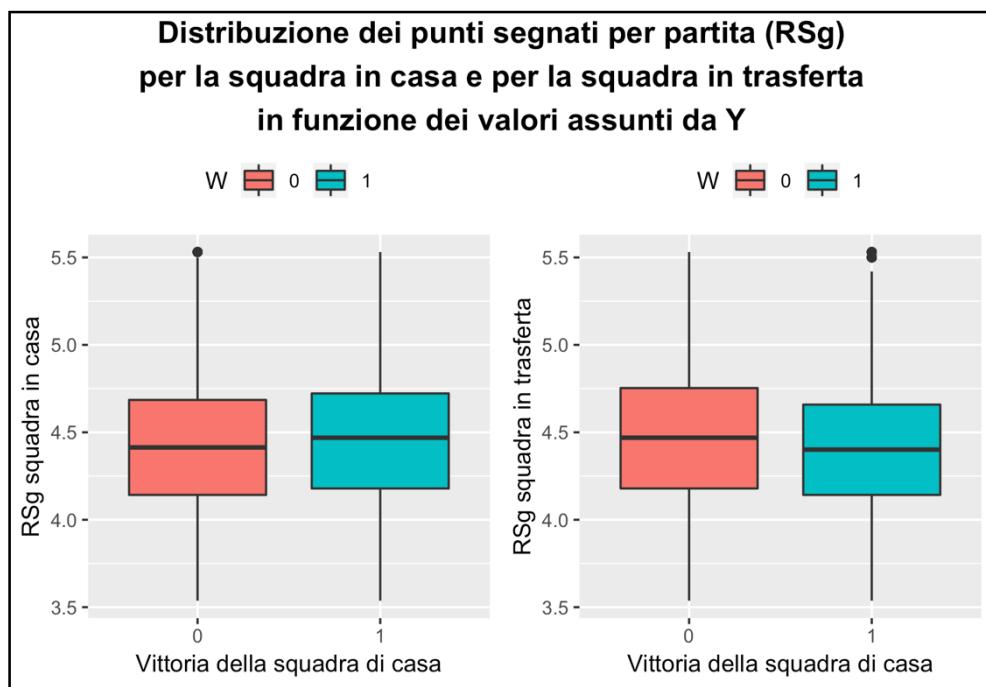
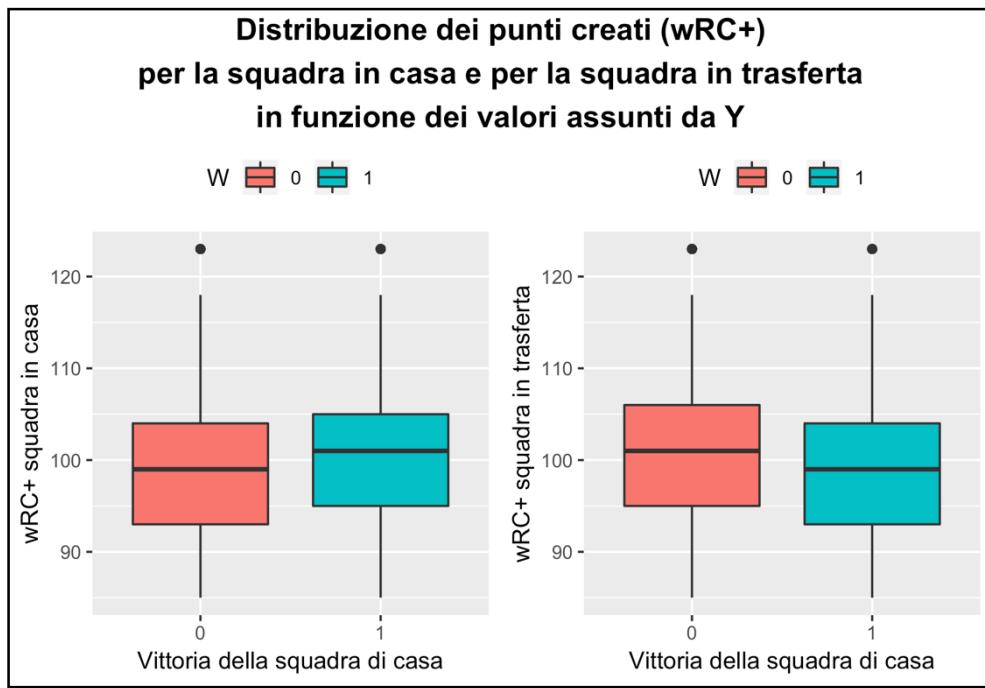


Grafico 2.6: Distribuzione assunta da wRC+ della squadra di casa e wRC+ della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)



Il pattern comune è che gli stimatori riferiti alla squadra di casa (lato sinistro) assumono mediamente valori maggiori quando $Y = 1$ (vittoria della squadra di casa) e minori quando $Y = 0$ (vittoria della squadra in trasferta). Allo stesso modo, gli stimatori riferiti alla squadra in trasferta (lato destro) sono tendenzialmente più elevati in corrispondenza di $Y = 0$ e più ridotti quando $Y = 1$. In poche parole, sembra almeno in parte essere presente una relazione di tipo positivo fra abilità della squadra e risultato della partita, come sarebbe d'altronde ovvio aspettarsi. Appare altresì evidente come le distribuzioni degli indicatori di abilità non differiscano in maniera netta al variare dei due livelli di Y , ma tendano piuttosto ad essere alquanto simili. Questo indica che una quota non indifferente della variabilità di Y risulta ancora non spiegata dai regressori, e potrebbe essere attribuibile alla relazione con altre variabili non ancora considerate.

2.1.2 Modelli di regressione bivariata

I modelli di regressione bivariati quantificano la relazione fra la variabile risposta Y e due variabili indipendenti X , considerando l'effetto simultaneo di entrambi i regressori. Nell'ambito della regressione logistica, questo corrisponde ad analizzare la dipendenza della probabilità di outcome positivo dati i valori delle X ($\text{Prob } [Y=1|X_1, X_2]$). Il motivo per cui si passa direttamente alla regressione con due predittori, sorvolando sui modelli univariati, è che per ogni stimatore di abilità e di stato di forma ha più senso, sia dal punto di vista logico che da

quello interpretativo, considerare contemporaneamente il valore da esso assunto sia per la squadra di casa che per quella in trasferta, anziché analizzarne gli effetti singolarmente.

Dopo aver verificato graficamente tramite i boxplot che le abilità delle due squadre sembrano essere correlate all'esito finale della partita, questo paragrafo sarà dedicato alla stima dell'entità di tali relazioni. Le tabelle 2.2-2.4 contengono in ordine: i valori dei coefficienti $\hat{\beta}$ stimati con il metodo della massima verosimiglianza, i relativi errori standard, il valore Z del test di Wald, il corrispondente *p-value*¹³, e infine la trasformata esponenziale di $\hat{\beta}$, che si ricorda essere pari all'aumento relativo degli odds di Y derivante da una variazione unitaria di X (*Odds Ratio*, si veda il paragrafo 1.6.3).

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1443	0.0207	6.9720	0	1.1552
RDg casa	0.3210	0.0292	10.9848	0	1.3784
RDg trasferta	-0.3926	0.0293	-13.4120	0	0.6753

Tabella 2.2: Output del modello logistico *Vittoria squadra casa ~ RDg squadra casa + RDg squadra trasferta*

I p-value pari a zero indicano che per entrambi i coefficienti è rifiutata l'ipotesi di non-significatività. In altre parole, i valori stimati di $\hat{\beta}$ risultano significativamente diversi da zero, e di conseguenza gli Odds Ratio diversi da $e^0 = 1$. Ciò conferma la presenza di una relazione fra le variabili in questione e la variabile dipendente Y. Nello specifico, un aumento di un punto di RDg della squadra di casa genera mediamente un aumento del 37.84% (1.3784-1) negli odds di vittoria della squadra di casa; un aumento di un punto di RDg della squadra in trasferta, invece, ne causa mediamente una riduzione del 32.47% (1-0.6753). Interessante è anche l'interpretazione dell'OR dell'intercetta, che si ricorda esprimere la variazione iniziale negli odds di risposta quando tutte le variabili X sono uguali a 0. Poiché la variabile RDg ha media nulla (si veda la tabella 2.1), l'intercetta rappresenta in questo caso l'aumento relativo degli odds di vittoria della squadra di casa quando sia quest'ultima che la sua avversaria hanno un'abilità nella media (differenza punti casa = differenza punti trasferta = 0). In altri termini, l'intercetta è un'approssimazione dell'effetto *fattore campo*: il vantaggio iniziale negli odds per la squadra ospitante sembra essere in media pari al 15%.

¹³ Ovvero, la probabilità di ottenere un risultato uguale o più estremo di quello osservato, sotto l'ipotesi nulla. In questo caso, poiché il test di Wald verifica l'ipotesi $H_0: \hat{\beta} = 0$, il p-value indica la probabilità di ottenere un valore di Z maggiore o uguale a quello osservato ipotizzato che il valore di $\hat{\beta}$ sia 0. Scelto un certo livello di significatività α (solitamente pari a 0.05), l'ipotesi nulla si rifiuta se risulta $p\text{-value} < \alpha$.

Passiamo ora a considerare il secondo stimatore dell'abilità di squadra, i punti segnati per partita (RSg). Per favorire l'interpretazione dei coefficienti e dell'intercetta, si è considerata la variabile RSg scalata rispetto alla propria media stagionale. Infatti, il valore zero riferito ai punti segnati per partita non avrebbe alcun senso, poiché nessuna squadra segnerà mai zero punti per tutta la stagione. Scalando la variabile, il valore zero diventa riferito alla media, e di conseguenza un valore pari a 1 o -1 è relativo a una squadra che segna mediamente 1 punto a partita in più o in meno rispetto alla media stagionale.

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1420	0.0205	6.9158	0	1.1526
RSg casa	0.4042	0.0527	7.6725	0	1.4981
RSg trasferta	-0.5650	0.0528	-10.6967	0	0.5683

Tabella 2.3: Output del modello logistico $Vittoria \text{ squadra casa} \sim RSg \text{ squadra casa} + RSg \text{ squadra trasferta}$

Anche in questo caso, poiché i coefficienti sono significativi, appare esservi una relazione rilevante fra Y e le X. Nel dettaglio, una squadra in casa che ha un rendimento medio offensivo di 1 punto a partita superiore alla media vede aumentare i suoi odds di vittoria del 49%, a parità dell'abilità della squadra avversaria. Una squadra in trasferta che segna un punto a partita in più rispetto alla media, invece, genera di norma una riduzione del 43% negli odds di vittoria della squadra di casa, a parità dell'abilità di quest'ultima. L'odds ratio riferito al fattore campo è ancora una volta mediamente uguale a 1.15.

Infine, è presentato l'output del modello relativo alla variabile punti creati (wRC+), considerata anch'essa in termini di differenza dalla media stagionale:

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1421	0.0205	6.9193	0	1.1527
wRC+ casa	0.0211	0.0028	7.6458	0	1.0213
wRC+ trasf.	-0.0293	0.0028	-10.6048	0	0.9711

Tabella 2.4: Output del modello logistico $Vittoria \text{ squadra casa} \sim wRC+ \text{ squadra casa} + wRC+ \text{ squadra trasferta}$

Il terzo stimatore dell'abilità di squadra appare a sua volta in relazione con il risultato della partita. Gli effetti dei predittori sulla variabile risposta potrebbero a prima vista apparire inferiori rispetto a quelli esercitati da RDg e RSg, ma questo è semplicemente dovuto all'utilizzo di un'unità di misura differente. Nello specifico, un punto creato in più rispetto alla media da parte della squadra di casa produce un aumento medio del 2.13% negli odds di vittoria della partita, e viceversa un wRC+ in più della squadra in trasferta ne causa mediamente una riduzione del 2.89%. L'odds ratio dell'intercetta appare stabile sul valore di 1.15.

2.2 Il talento dei lanciatori partenti

2.2.1 Statistiche descrittive

La tabella 2.5 contiene le statistiche descrittive relative alle variabili ERA, FIP e xFIP dei 1353 lanciatori partenti di MLB nel periodo 2015-2018.

	ERA (partenti)	FIP (partenti)	xFIP (partenti)
Minimo	0.0000	0.1300	-0.1500 ¹⁴
1° Quartile	3.4500	3.6300	3.7100
Mediana	4.1500	4.1800	4.1900
Media	4.2777	4.2551	4.2009
3° Quartile	4.8200	4.7400	4.6700
Massimo	189.0000	36.1500	29.1600
D. Standard	1.4880	1.0242	0.7579
C.V.	0.3479	0.2407	0.1804

Tabella 2.5: *Statistiche descrittive delle variabili utilizzate per la stima dell'abilità dei lanciatori partenti, pesate per il numero di inning lanciati (stagioni 2015-2018)*

Le osservazioni sono state pesate per il numero di inning lanciati, al fine di ridurre l'effetto di valori anomali tendenti allo zero o estremamente elevati (si veda il valore massimo dell'ERA, 189), poco informativi poiché riferiti a lanciatori con un numero decisamente ridotto di presenze¹⁵. Le medie globali di ERA, FIP e xFIP risultano alquanto simili fra loro; ciò nonostante, le distribuzioni sono piuttosto differenti in termini di variabilità. Confrontando i valori di minimo-massimo e i coefficienti di variazione, appare infatti evidente che l'ERA tende ad assumere valori più estremi, mentre FIP e xFIP sono maggiormente concentrate intorno ai valori centrali.

¹⁴ FIP e xFIP contengono a numeratore il termine (-2*K), dunque possono teoricamente assumere valori negativi in presenza di quantità elevate di strikeout effettuati e ridotte di HR, BB e HBP concessi. Ciò è riscontrato nella realtà solo in presenza di dimensioni campionarie molto ridotte (ovvero, un numero di riprese lanciate molto basso).

¹⁵ Nel dettaglio, tutti i lanciatori partenti del periodo 2015-2018 con ERA o FIP o xFIP stagionale > 15 hanno lanciato non più di 9.1 inning nella corrispondente stagione; tutti quelli con un valore di ERA, FIP o xFIP < 1, invece, hanno disputato al massimo 5.2 riprese. A scopo di riferimento, si riporta che la media di inning lanciati in una stagione per un lanciatore partente (2015-2018) è di 79.97, e il 1° quartile è 19.1.

Grafico 2.7

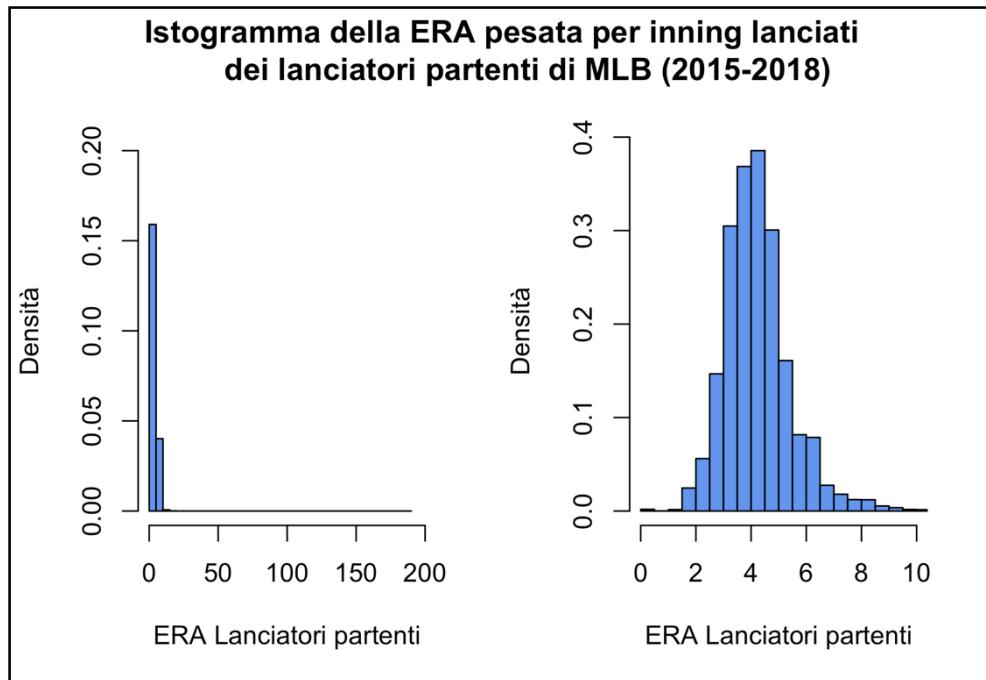


Grafico 2.8

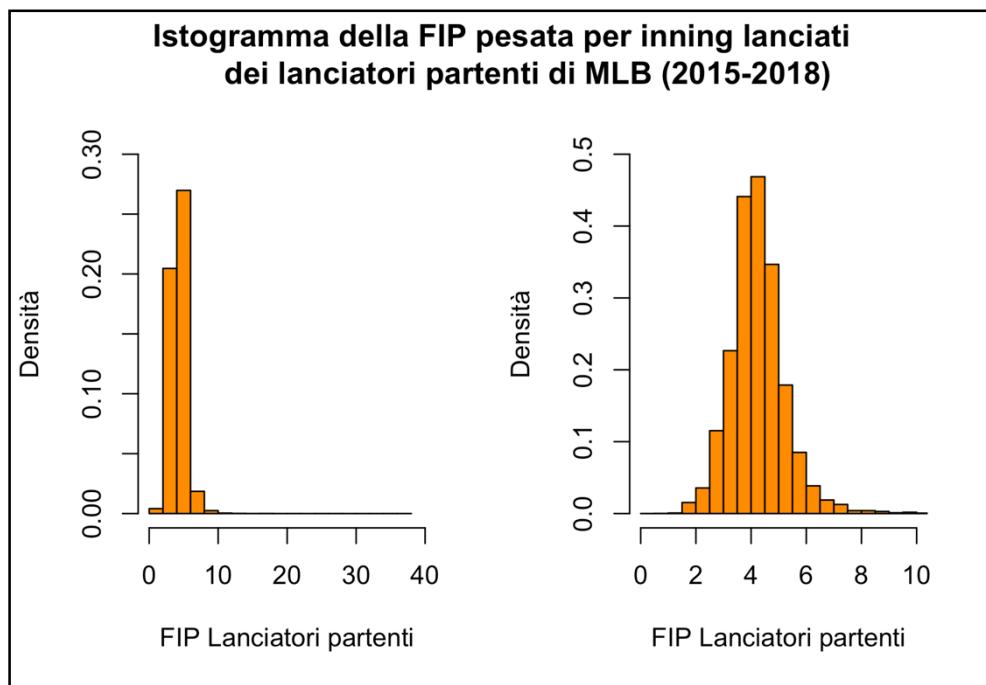
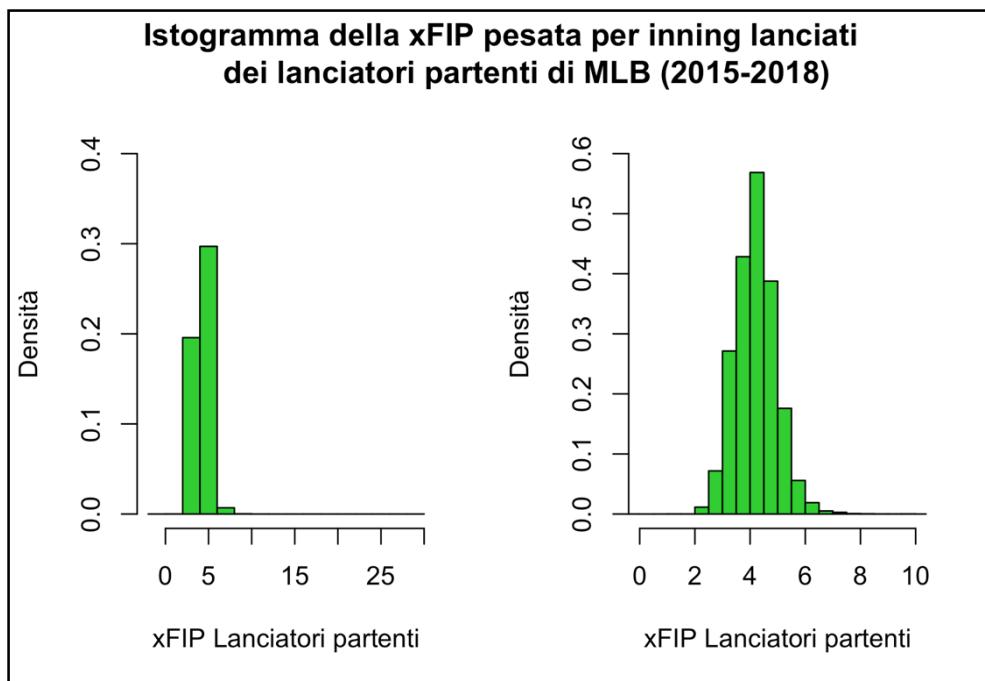


Grafico 2.9



I grafici 2.7-2.9 affiancano l’istogramma completo delle variabili e quello riferito al solo intervallo di valori 0-10. In generale, le distribuzioni appaiono asimmetriche con uno sbilanciamento verso i valori inferiori. Tuttavia, se si considera un intervallo non anomalo (analizzando valori compresi fra 0 e 10, si escludono solo 91 osservazioni per l’ERA, 45 per la FIP e 9 per l’xFIP, rispettivamente il 7%, 3% e 1% del totale) le densità appaiono tendenti alla curva normale.

Di seguito, come già realizzato per le abilità di squadra, sono riportati i boxplot che mostrano la differenza nella distribuzione di ERA, FIP e xFIP del lanciatore della squadra di casa e del lanciatore della squadra in trasferta al variare del valore di Y (grafici 2.10-2.12). Per facilitare la visualizzazione sono stati nascosti gli outliers, ovvero i valori anomali al di sopra e al di sotto dei *baffi* dei boxplot, che rappresentano una quota irrilevante del totale. Anche in questo caso appare presente una qualche forma di relazione, benché le distribuzioni non siano considerevolmente differenti in base al risultato finale della partita. La tendenza generale è che a una vittoria della squadra di casa sono associati valori mediamente inferiori di ERA, FIP e xFIP (e dunque un maggiore talento) del partente della squadra di casa, e valori mediamente superiori delle variabili riferite al partente avversario. Viceversa, quando Y = 0 (vittoria della squadra in trasferta) i valori di ERA, FIP e xFIP del partente di casa sono orientativamente superiori e quelli del lanciatore in trasferta minori.

Grafico 2.10: Distribuzione assunta dalla ERA del partente della squadra di casa e dalla ERA del partente della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

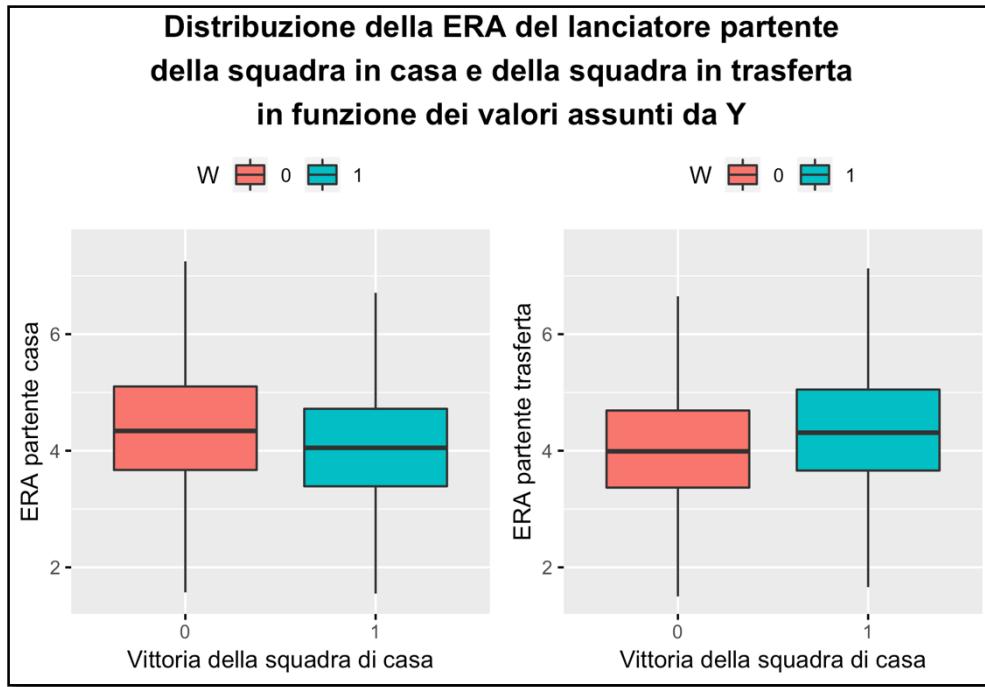


Grafico 2.11: Distribuzione assunta dalla FIP del partente della squadra di casa e dalla FIP del partente della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

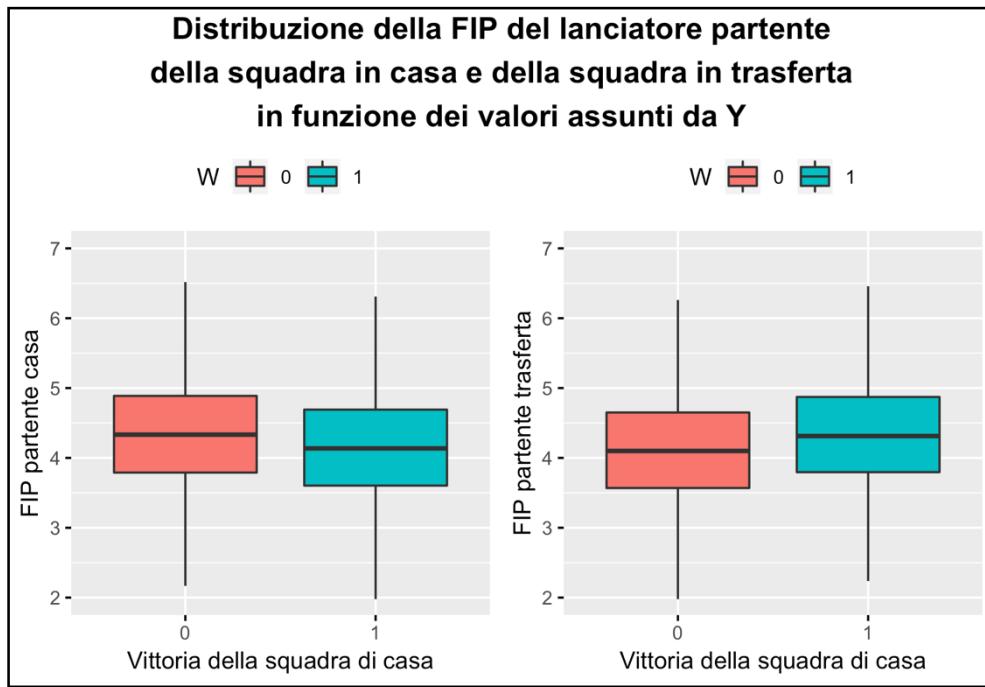
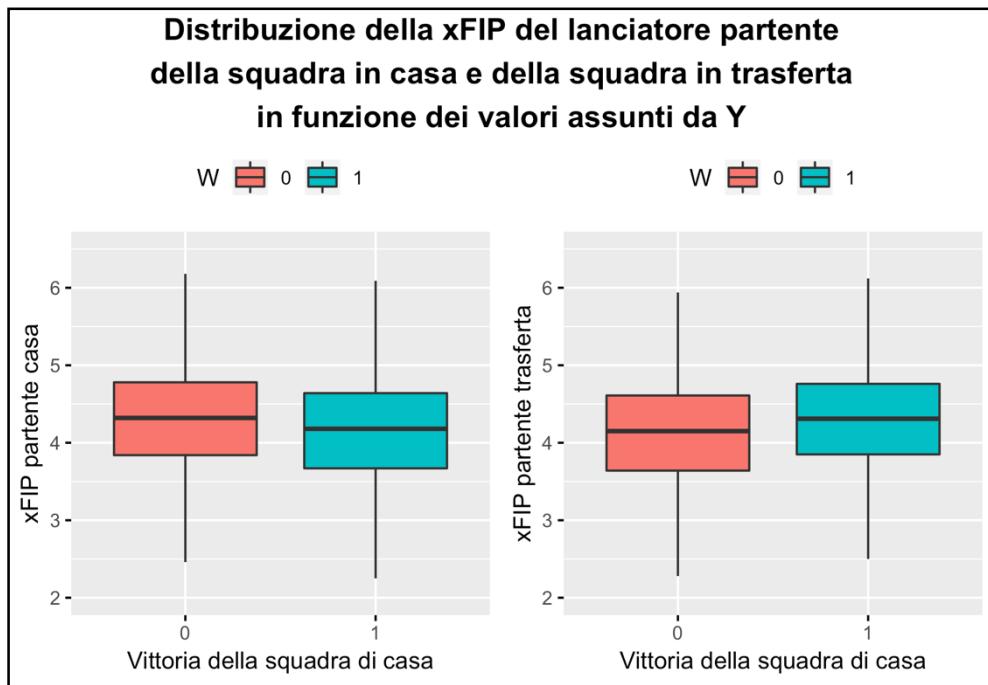


Grafico 2.12: Distribuzione assunta dalla xFIP del partente della squadra di casa e dalla xFIP del partente della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)



2.2.2 Modelli di regressione bivariata

Il paragrafo in questione è dedicato alla specificazione dei modelli logistici bivariati, al fine di quantificare l'effetto esercitato dal talento dei due lanciatori partenti sulla probabilità della squadra di casa di vincere la partita. Per le ragioni di interpretabilità precedentemente discusse, nella stima dei coefficienti le variabili saranno centrate sulle medie stagionali.

Sono di seguito presentati (tabelle 2.6-2.8) gli output di R contenenti, come descritto nel paragrafo 2.1.2, i $\hat{\beta}$ stimati, gli errori standard, le Z del test di Wald, i p-value e gli odds ratio.

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1506	0.0209	7.2186	0	1.1625
ERA P casa	-0.2425	0.0174	-13.9093	0	0.7847
ERA P trasf.	0.2871	0.0184	15.5925	0	1.3326

Tabella 2.6: Output del modello logistico Vittoria squadra casa ~ ERA partente **casa** + ERA partente **trasferta**

I p-value nulli evidenziano la presenza di una relazione. Dall'interpretazione degli odds ratio si evince che per ogni punto sopra alla media dell'ERA del lanciatore partente, gli odds di vittoria della squadra di casa diminuiscono del 21% se si tratta del proprio lanciatore, e al contrario aumentano del 33% se si parla del lanciatore avversario. In poche parole,

all'aumentare del talento del proprio lanciatore partente, la squadra in questione aumenta le proprie possibilità di vittoria. L'intercetta è ancora una volta interpretabile come l'aumento negli odds dovuto dal fattore campo, poiché si riferisce al valore iniziale degli odds di vittoria della squadra di casa quando entrambi i lanciatori hanno uguale abilità nella media. Questo aumento sembra essere pari al 16%.

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1473	0.0207	7.1125	0	1.1587
FIP P casa	-0.2387	0.0217	-11.0247	0	0.7877
FIP P trasf.	0.3194	0.0232	13.7678	0	1.3763

Tabella 2.7: Output del modello logistico Vittoria squadra casa ~ FIP partente **casa** + FIP partente **trasferta**

Per quanto riguarda la FIP, ogni punto superiore alla media stagionale del lanciatore avversario appare generare un aumento del 38% negli odds di vittoria per la squadra di casa, e un aumento del 21% per quella in trasferta. Quando entrambi i lanciatori hanno una FIP uguale alla media nella popolazione dei lanciatori, gli odds di vittoria della squadra di casa sono solitamente superiori del 16%.

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1414	0.0207	6.8317	0	1.1519
xFIP P casa	-0.2924	0.0274	-10.6895	0	0.7465
xFIP P trasf.	0.3360	0.0282	11.925	0	1.3994

Tabella 2.8: Output del modello logistico Vittoria squadra casa ~ xFIP partente **casa** + xFIP partente **trasferta**

Valori simili, come riassunto dalla tabella 2.8, sono riferiti alla xFIP. Un lanciatore partente della squadra di casa con un punto di xFIP sopra alla media diminuisce gli odds di vittoria dei suoi compagni di circa il 25%. Un partente della squadra in trasferta con un punto di xFIP sopra alla media stagionale, invece, genera solitamente un aumento del 40% negli odds di vittoria degli avversari.

I valori dei coefficienti di ERA, FIP e xFIP appaiono in tutti e tre i modelli superiori in valore assoluto per quanto riguarda il lanciatore della squadra in trasferta rispetto al partente della formazione di casa. Questo, unitamente alla significatività dell'intercetta, può evidenziare la presenza di un vantaggio iniziale della squadra di casa indipendente dai talenti individuali e complessivi.

2.3 Il talento dei lanciatori di rilievo

2.3.1 Statistiche descrittive

Il database Retrosheet riporta, per ogni partita disputata negli ultimi 150 anni, i nomi dei lanciatori partenti delle due squadre, ma non fa purtroppo menzione dei lanciatori di rilievo, che ricoprono un ruolo comunque importante nella determinazione del risultato finale di una partita. A tal proposito, basti pensare che la durata media dell’impiego dei partenti nelle stagioni 2015-2018 è stata di 5.57 inning. Poiché una partita completa è composta da (almeno) 9 riprese, ciò che accade nelle rimanenti 3.43 è in buona parte attribuibile all’efficacia del *bullpen*¹⁶. In mancanza di dati individuali, si considererà quindi l’abilità complessiva dei lanciatori di rilievo di entrambe le squadre.

	ERA (bullpen)	FIP (bullpen)	xFIP (bullpen)
Minimo	2.6700	3.1410	3.3285
1° Quartile	3.5600	3.7124	3.8668
Mediana	3.9200	3.9736	4.1196
Media	3.9763	4.0141	4.1050
3° Quartile	4.4000	4.3209	4.3418
Massimo	5.6300	5.3352	5.0143
D. Standard	0.5646	0.4134	0.3472
C.V.	0.1420	0.1030	0.0846

Tabella 2.9: Statistiche descrittive delle variabili utilizzate per la stima dell’abilità dei lanciatori di rilievo considerate a livello di squadra e pesate per il numero di riprese lanciate (stagioni 2015-2018)

La tabella 2.9 contiene il riassunto delle statistiche riguardanti l’insieme dei lanciatori di rilievo di ogni squadra per le 4 stagioni considerate, pesate ancora una volta per il numero di riprese lanciate. Trattandosi di dati aggregati anziché individuali, la variabilità è notevolmente ridotta rispetto alle stesse variabili riferite ai lanciatori partenti, come dimostrato dai valori di minimo e massimo e dai coefficienti di variazione. Osservando i valori medi e confrontandoli con quelli della tabella 2.5, si potrebbe pensare che i lanciatori di rilievo siano generalmente più abili dei loro colleghi partenti. Una teoria diffusa è che i numeri dei *bullpen* tendano ad essere migliori almeno in parte per via delle modalità di utilizzo dei lanciatori di rilievo, maggiormente basate sulle performance, rispetto a quelle dei partenti, che indipendentemente

¹⁶ Termine che indica l’insieme dei lanciatori di rilievo di una squadra, derivante dal nome dell’area in cui questi si riscaldano prima di fare il loro ingresso nella partita.

dalla situazione iniziale lanciano ogni 5-6 partite a seconda della rotazione¹⁷. Tuttavia, in mancanza di prove non si può negare che almeno una quota di questa differenza derivi da un talento medio superiore.

Grafico 2.13

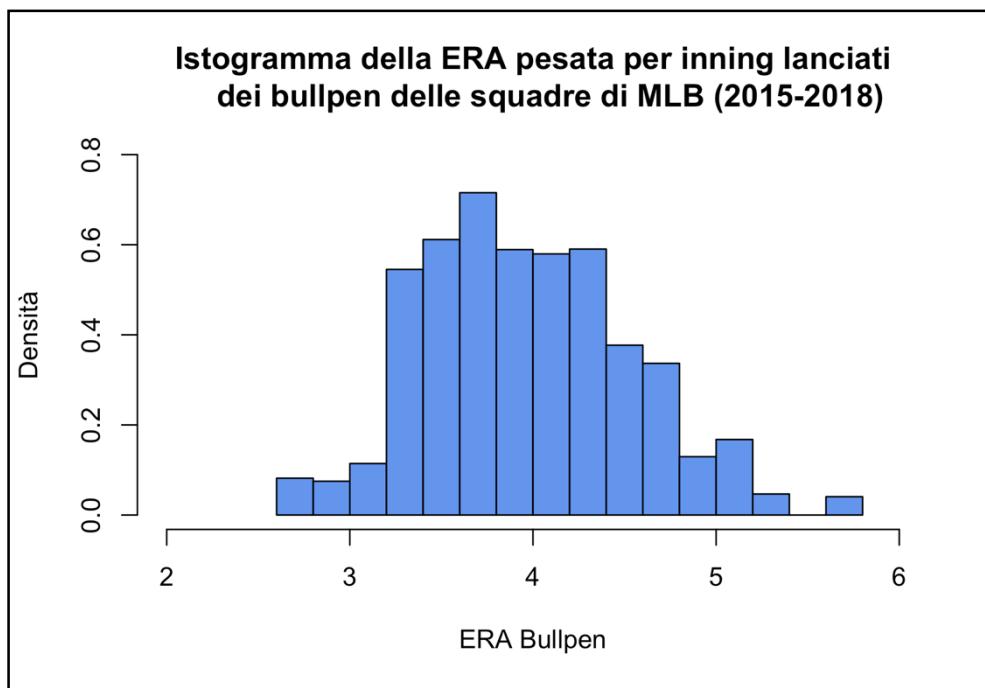
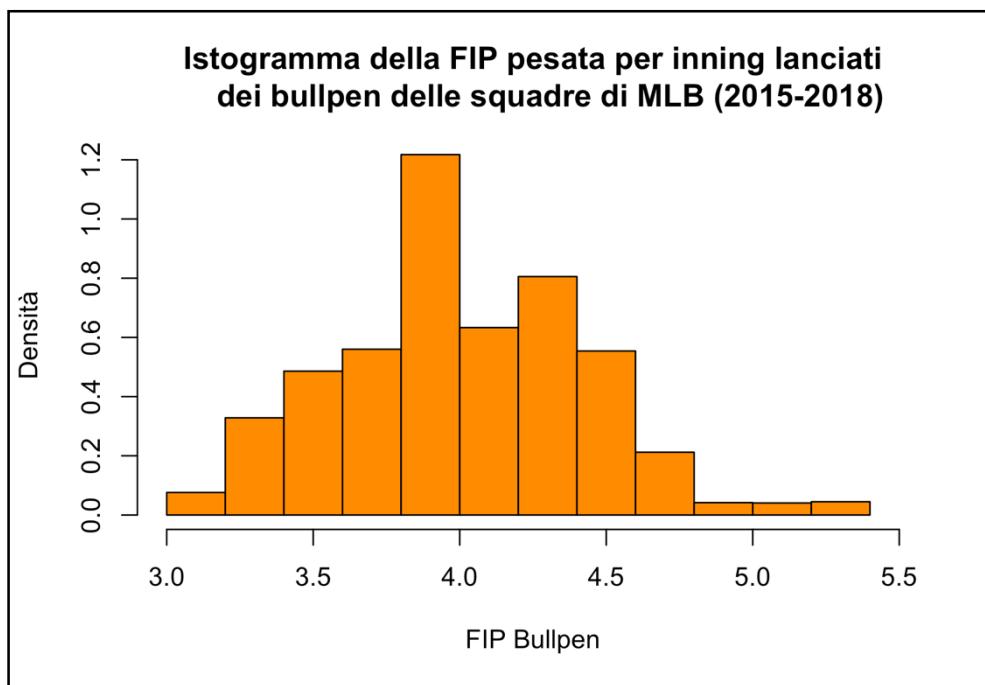
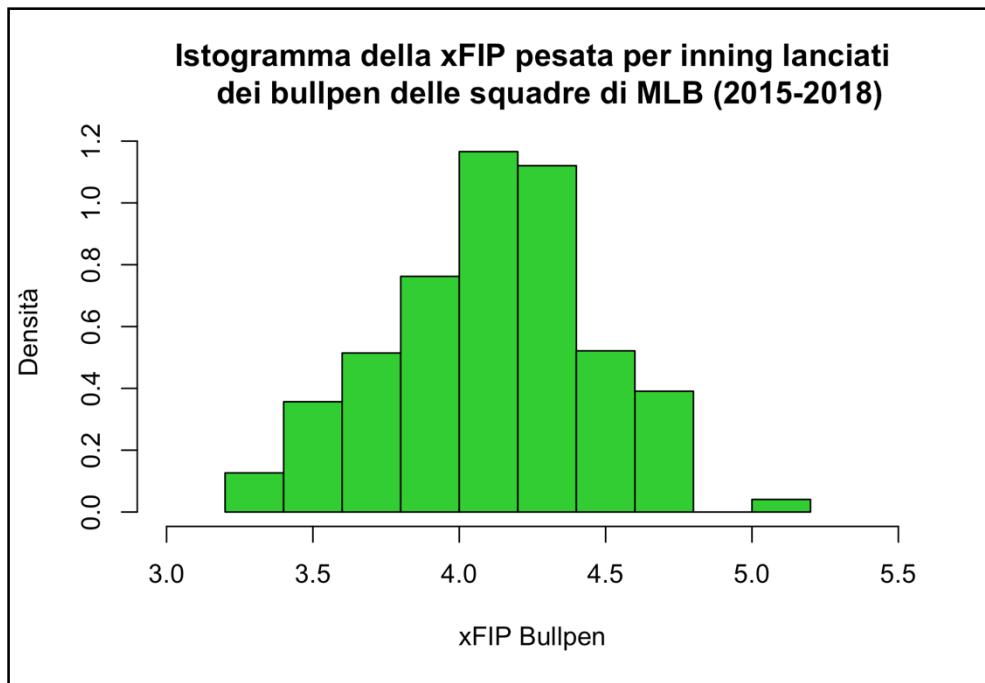


Grafico 2.14



¹⁷ Click, 2004

Grafico 2.15



I grafici 2.13-2.15 illustrano le distribuzioni delle variabili riferite ai lanciatori di rilievo, che appaiono tutto sommato simmetriche e particolarmente concentrate nei valori centrali. I successivi grafici 2.16-2.18, invece, rappresentano la tendenza di associazione fra valori del predittore e valore assunto dalla variabile risposta Y.

Grafico 2.16: Distribuzione assunta dalla ERA del bullpen della squadra di casa e dalla ERA del bullpen della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

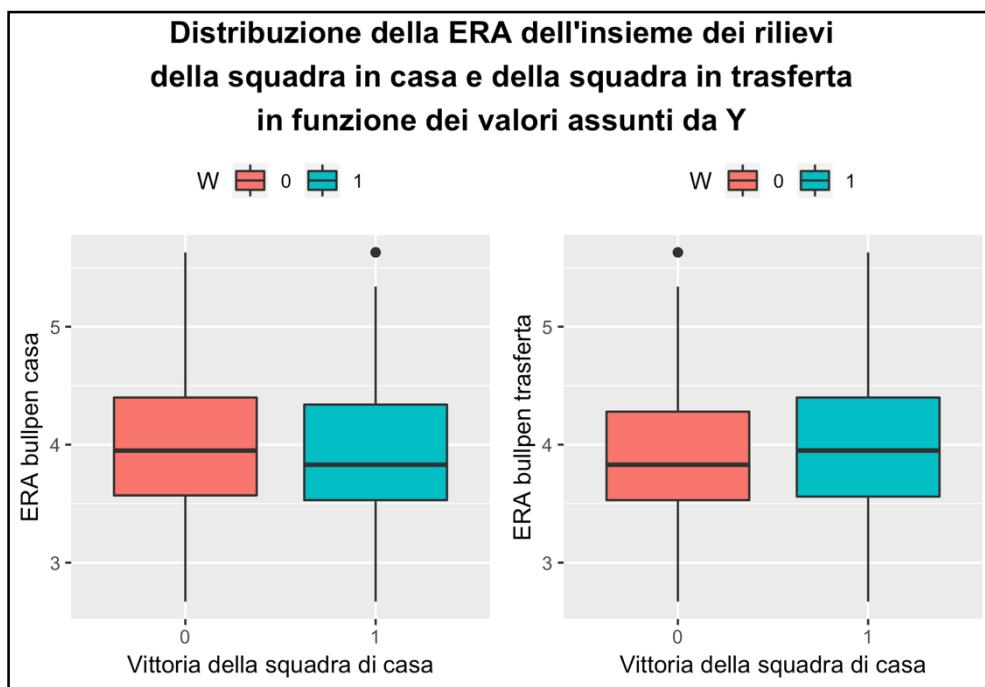


Grafico 2.17: Distribuzione assunta dalla FIP del bullpen della squadra di casa e dalla FIP del bullpen della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

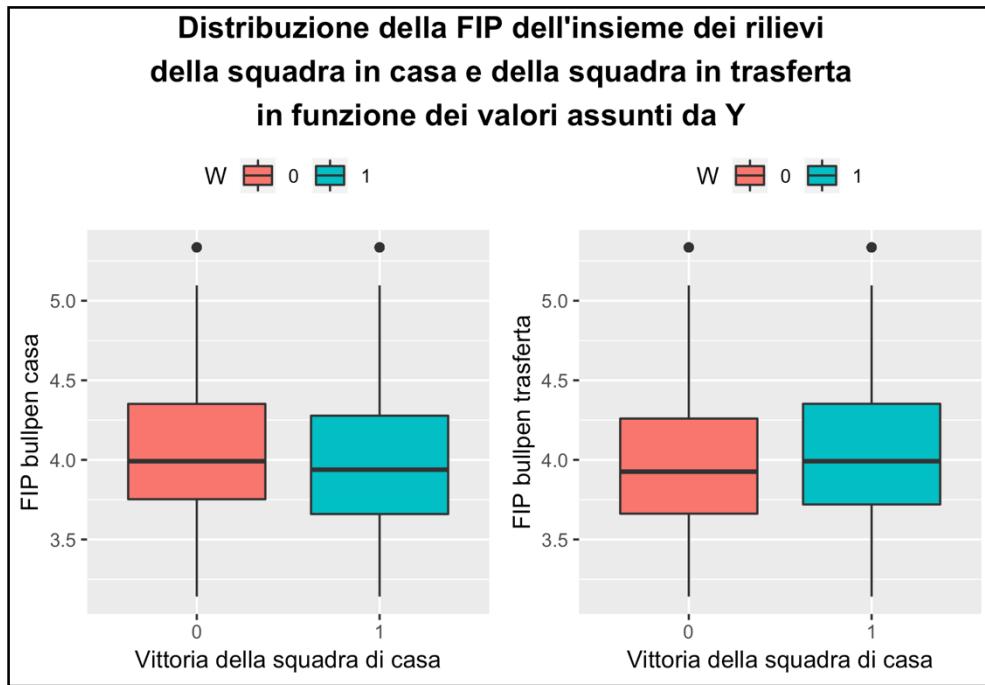
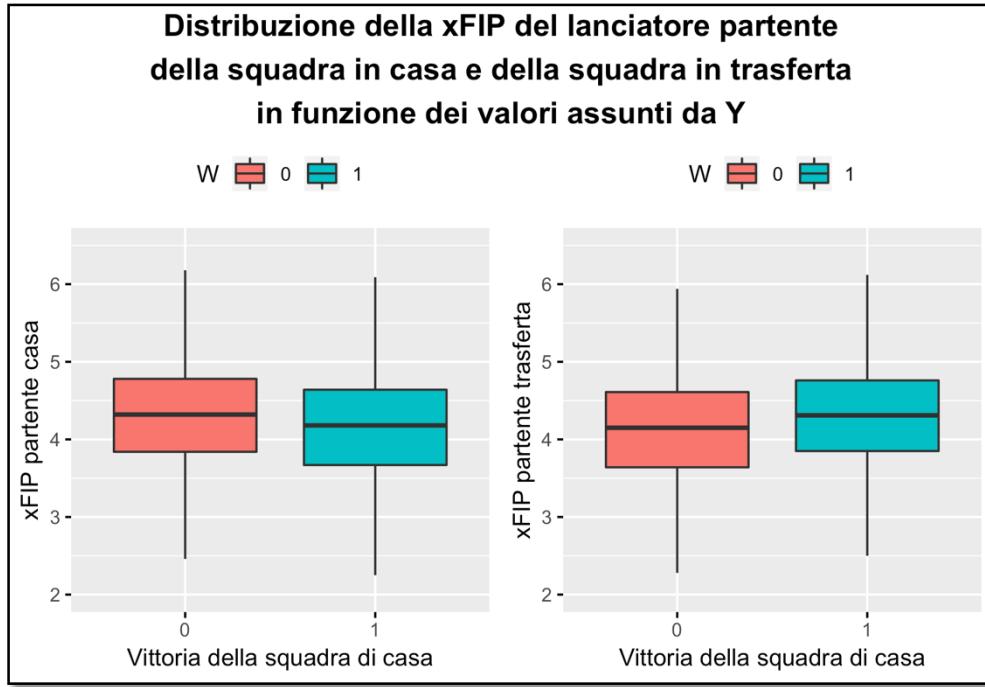


Grafico 2.18: Distribuzione assunta dalla xFIP del bullpen della squadra di casa e dalla xFIP del bullpen della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)



Così come osservato per i dati dei lanciatori partenti, si può concludere che alla vittoria di una squadra rispetto ad un'altra sono associati valori generalmente migliori di ERA, FIP e xFIP del relativo bullpen. Ancora una volta, la differenza tra le distribuzioni è però tutt'altro che netta e ben definita.

2.3.2 Modelli di regressione bivariata

Le tabelle 2.10-2.12 presentano gli output dei modelli di regressione logistica stimati per spiegare la variabilità di Y (vittoria della squadra di casa) in funzione dei valori dei bullpen delle due squadre. Al fine di dare un senso logico al valore zero, si sono ancora una volta considerate le trasformazioni delle variabili centrate nella media.

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1420	0.0205	6.9232	0	1.1526
ERA B casa	-0.3309	0.0381	-8.6762	0	0.7183
ERA B trasf.	0.3498	0.0384	9.1211	0	1.4189

Tabella 2.10: Output del modello logistico Vittoria squadra casa ~ ERA bullpen *casa* + ERA bullpen *trasferta*

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1418	0.0205	6.9214	0	1.1523
FIP B casa	-0.3796	0.052	-7.2943	0	0.6841
FIP B trasf.	0.4395	0.0523	8.4035	0	1.5519

Tabella 2.11: Output del modello logistico Vittoria squadra casa ~ FIP bullpen *casa* + FIP bullpen *trasferta*

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.142	0.0205	6.9268	0	1.1526
xFIP B casa	-0.5068	0.0644	-7.8735	0	0.6024
xFIP B trasf.	0.5678	0.0642	8.8454	0	1.7645

Tabella 2.12: Output del modello logistico Vittoria squadra casa ~ xFIP bullpen *casa* + xFIP bullpen *trasferta*

In tutti e tre i casi, i coefficienti risultano significativamente diversi da zero, evidenziando la presenza di relazione fra le variabili considerate e il risultato finale della partita. Come è logico pensare, un bullpen con un talento maggiore alla media genera un aumento negli odds di vittoria della propria squadra, e viceversa. Infine, appare ancora confermata la presenza del fattore campo: quando le due squadre hanno entrambe lanciatori di rilievo nella media ($X=0$) o con abilità abbastanza simili (i coefficienti tendono in ogni caso ad annullarsi fra loro), gli odds di vittoria della squadra di casa sono maggiori del 15%.

2.4 Lo stato di forma delle squadre

Per quanto riguarda gli stimatori dello stato di forma delle squadre, sarebbe di scarso interesse riportare le statistiche descrittive, trattandosi di medie mobili di differenza punti e percentuale di vittorie calcolate su intervalli di 5 o 10 partite (ovvero, della media dei valori in

questione nelle 5 o 10 partite precedenti a quella considerata). Si passerà dunque direttamente a considerare l'associazione di queste variabili con l'outcome della partita.

Grafico 2.19: Distribuzione assunta dalla RDg nelle 5 partite precedenti della squadra di casa e della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

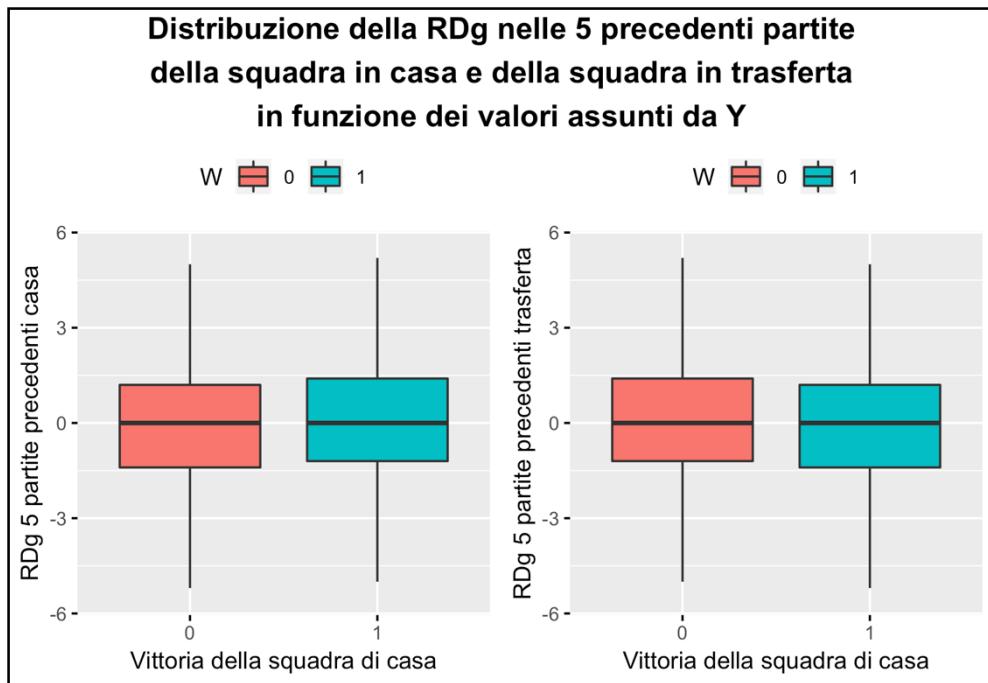


Grafico 2.20: Distribuzione assunta dalla RDg nelle 10 partite precedenti della squadra di casa e della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)

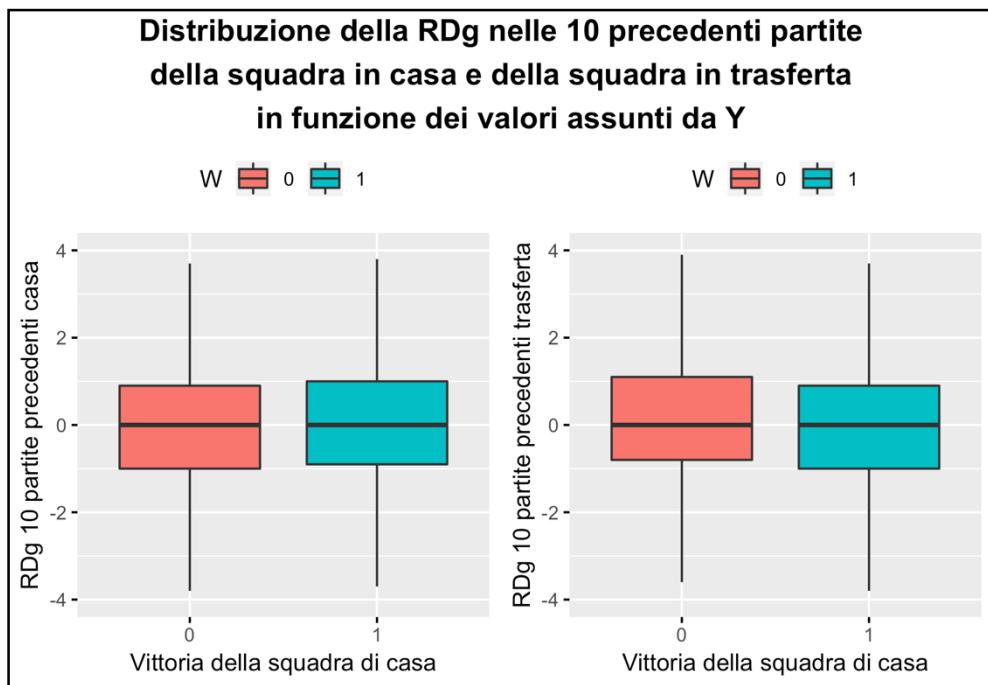
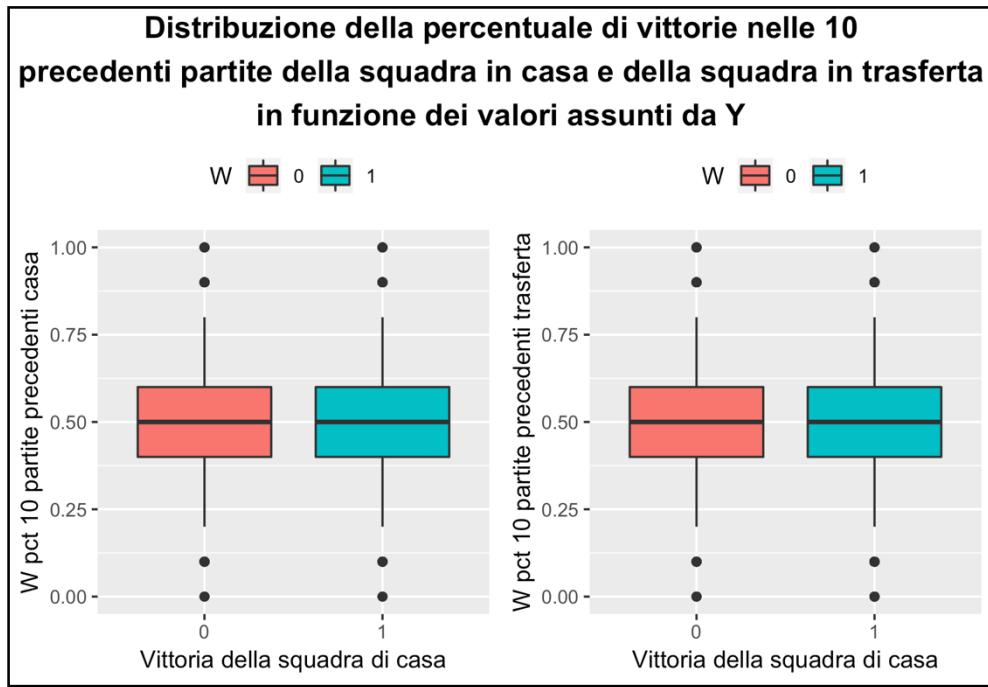


Grafico 2.21: Distribuzione assunta dalla percentuale di vittorie nelle 10 partite precedenti della squadra di casa e della squadra in trasferta quando $Y = 1$ (vittoria squadra di casa) e quando $Y = 0$ (sconfitta squadra di casa)



I boxplot 2.19-2.20 evidenziano la presenza di una leggera differenza fra le distribuzioni della RDg delle due squadre nelle 5 e 10 partite precedenti e il risultato finale. Non appare invece esservi alcuna diversità per quanto riguarda la distribuzione della percentuale di vittoria nelle 10 partite precedenti delle due squadre al variare di Y.

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1404	0.0204	6.8895	0	1.1507
RDg 5 casa	0.0290	0.0101	2.8749	0.004	1.0294
RDg 5 trasf.	-0.0470	0.0102	-4.6228	0	0.9541

Tabella 2.13: Output del modello logistico Vittoria squadra casa ~ RDg 5 prec. **casa** + RDg 5 prec. **trasferta**

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1418	0.0204	6.9544	0	1.1524
RDg 10 casa	0.0537	0.0137	3.9127	0.0001	1.0552
RDg 10 trasf.	-0.0736	0.0138	-5.3305	0	0.929

Tabella 2.14: Output del modello logistico Vittoria squadra casa ~ RDg 10 prec. **casa** + RDg 10 prec. **trasferta**

	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1515	0.0984	1.5388	0.1238	1.1636
W% 10 casa	0.4456	0.1269	3.5126	0.0004	1.5614
W% 10 trasf.	-0.4636	0.1267	-3.6583	0.0003	0.629

Tabella 2.15: Output del modello logistico Vittoria squadra casa ~ xFIP bullpen **casa** + xFIP bullpen **trasferta**

Le tabelle 2.13-2.15 mostrano gli output dei modelli in cui la vittoria della squadra di casa è spiegata esclusivamente dallo stato di forma delle due squadre. Dal momento che i p-value dei coefficienti di regressione risultano tutti inferiori al livello soglia di 0.05, sembra essere presente una relazione fra la variabile Y e i regressori. In particolare, all'aumentare della differenza punti nelle ultime 5 o 10 partite e della percentuale di vittorie nelle ultime 10 partite di una squadra, sembrano aumentare i suoi odds di vittoria rispetto all'avversaria.

Tuttavia, è bene specificare che squadre generalmente più forti (quindi con un talento, più che uno stato di forma, superiore) tenderanno ad avere una differenza punti e una percentuale di vittorie maggiore qualsiasi sia l'intervallo di partite considerato. Di conseguenza, non tenere conto di ciò nella specificazione dei modelli inserendo, oltre alle variabili riferite allo stato di forma, una misura dell'abilità media generale delle due squadre, non può che creare confondimento fino al punto di evidenziare relazioni che sono in realtà inesistenti. In altre parole, la presenza di una effettiva relazione fra stato di forma delle squadre e risultato della partita potrà essere verificata correttamente solo nella fase successiva dell'analisi, che si occuperà della ricerca del miglior modello multivariato considerando tutti i regressori necessari.

2.5 *La variabile risposta*

Terminate le analisi descrittive dei regressori, la tabella 2.16 riassume brevemente la distribuzione della variabile risposta Y nelle 9717 partite di Regular Season di MLB giocate fra il 2015 e il 2018. Come si deduce, la squadra in casa ha vinto il 53.48% delle volte, a ulteriore dimostrazione che giocare nel proprio stadio rappresenta un effettivo vantaggio. Questo potrebbe derivare dal fatto che i giocatori di casa, rispetto a quelli in trasferta, godono solitamente di maggiori comodità (ad esempio, poter alloggiare nelle rispettive abitazioni e non dover viaggiare centinaia di km) e del supporto del pubblico¹⁸.

Vittoria squadra in trasferta (Y=0)	Vittoria squadra in casa (Y=1)	Totale
4520	5197	9717
46.52 %	53.48 %	100%

Tabella 2.16: Distribuzione della variabile risposta Y nelle 9717 partite di Regular Season di MLB (2015-2018)

¹⁸ Albert e Bennet, 2003

2.6 Riepilogo

In questo secondo capitolo è stata analizzata la distribuzione delle variabili adoperate per la stima delle abilità e lo stato di forma, ed è stato quantificato il loro effetto isolato sulla variabile risposta tramite la specificazione di modelli di regressione logistica. Come logico attendersi, abilità mediamente maggiori sia dei singoli che della squadra nel complesso sono risultate legate a un aumento della probabilità di vittoria nell'ambito singola partita, e viceversa. Anche lo stato di forma delle squadre è apparso positivamente collegato al risultato finale. Generalmente, le relazioni sono state tutte riscontrate significative, sebbene sia doveroso puntualizzare che considerare alcune variabili singolarmente possa indurre alla presenza di confondimento. Infine, sia dall'analisi della distribuzione della variabile risposta che della sua relazione con i regressori è apparso significativo l'effetto del fattore campo: a parità di tutte le altre condizioni, la squadra di casa sembra avere una probabilità corrispondente a circa il 53% di vincere la partita.

Il prossimo capitolo, come precedentemente accennato, si occuperà di selezionare il miglior modello possibile per spiegare l'esito di una partita dati i regressori disponibili e di quantificarne la capacità previsiva.

Capitolo 3

La scelta del modello

3.1 I modelli interi

Con lo scopo di procedere alla selezione del modello finale, sono stati stimati i 27 modelli *interi* che costituiscono le combinazioni dei singoli regressori riguardanti l'abilità di squadra, il talento dei lanciatori e lo stato di forma. Per praticità sono state considerate le stesse variabili sia per quanto riguarda i partenti che i rilievi¹⁹.

La tabella 3.1 contiene una sintesi dei risultati della suddetta operazione: ogni riga corrisponde a uno dei modelli, per cui sono riportate le variabili indipendenti inserite per la stima delle 4 componenti. Ovviamente, ogni variabile è stata considerata contemporaneamente sia per la squadra di casa che per quella in trasferta, per un totale di 8 regressori per modello. Per ciascuno di questi, infine, è indicato se le stime sono risultate o no significativamente diverse da zero.

Abilità squadra	Abilità partenti	Abilità bullpen	Stato di forma
RDg <i>significativo</i>	ERA <i>significativo</i>	ERA non significativo	RDg 10 non significativo
RDg <i>significativo</i>	ERA <i>significativo</i>	ERA non significativo	RDg 5 non significativo
RDg <i>significativo</i>	ERA <i>significativo</i>	ERA non significativo	W pct 10 non significativo
RDg <i>significativo</i>	FIP <i>significativo</i>	FIP non significativo	RDg 10 non significativo
RDg <i>significativo</i>	FIP <i>significativo</i>	FIP non significativo	RDg 5 non significativo
RDg <i>significativo</i>	FIP <i>significativo</i>	FIP non significativo	W pct 10 non significativo
RDg <i>significativo</i>	xFIP <i>significativo</i>	xFIP non significativo	RDg 10 non significativo
RDg <i>significativo</i>	xFIP <i>significativo</i>	xFIP non significativo	RDg 5 non significativo
RDg <i>significativo</i>	xFIP <i>significativo</i>	xFIP non significativo	W pct 10 non significativo

¹⁹ Il numero di modelli, altrimenti, sarebbe passato da $3^3 = 27$ a $4^3 = 64$. Questi saranno comunque presi in considerazione nel paragrafo successivo.

Abilità squadra	Abilità partenti	Abilità bullpen	Stato di forma
RSg significativo	ERA significativo	ERA significativo	RDg 10 non significativo
RSg significativo	ERA significativo	ERA significativo	RDg 5 non significativo
RSg significativo	ERA significativo	ERA significativo	W pct 10 non significativo
RSg significativo	FIP significativo	FIP significativo	RDg 10 non significativo
RSg significativo	FIP significativo	FIP significativo	RDg 5 non significativo
RSg significativo	FIP significativo	FIP non significativo	W pct 10 non significativo
RSg significativo	xFIP significativo	xFIP significativo	RDg 10 non significativo
RSg significativo	xFIP significativo	xFIP significativo	RDg 5 non significativo
RSg significativo	xFIP significativo	xFIP significativo	W pct 10 non significativo
wRC+ significativo	ERA significativo	ERA significativo	RDg 10 non significativo
wRC+ significativo	ERA significativo	ERA significativo	RDg 5 non significativo
wRC+ significativo	ERA significativo	ERA significativo	W pct 10 non significativo
wRC+ significativo	FIP significativo	FIP significativo	RDg 10 non significativo
wRC+ significativo	FIP significativo	FIP significativo	RDg 5 non significativo
wRC+ significativo	FIP significativo	FIP non significativo	W pct 10 non significativo
wRC+ significativo	xFIP significativo	xFIP significativo	RDg 10 non significativo
wRC+ significativo	xFIP significativo	xFIP significativo	RDg 5 non significativo
wRC+ significativo	xFIP significativo	xFIP significativo	W pct 10 non significativo

Tabella 3.1: Riassunto della significatività delle variabili inserite nei modelli interi, consideranti contemporaneamente l'effetto di tutte e 4 le componenti (abilità delle due squadre, abilità dei lanciatori partenti, abilità dei lanciatori di rilievo, stato di forma) sulla variabile risposta Y.

I risultati riguardanti la significatività dei coefficienti permettono di giungere a due conclusioni rilevanti:

1. Per tutti i 27 modelli considerati, **l'effetto dello *stato di forma* delle due squadre sulla variabile risposta è risultato non significativo**, in qualsiasi modo questo fosse misurato.

Le relazioni isolate che erano state evidenziate nel paragrafo 2.4, infatti, risultano completamente spiegate dalle altre variabili inserite nel modello, cioè l'abilità generale della squadra e quella dei lanciatori. Si può quindi affermare che, in base all'evidenza presentata, non vi sono prove per decretare che la differenza punti o la percentuale di vittoria misurata nelle partite precedenti abbiano alcun effetto sulla probabilità di vittoria della singola partita.

2. **Quando l'abilità di squadra è misurata in termini di differenza punti per partita (RDg), il talento dei lanciatori di rilievo non è risultato avere un effetto significativo sulla variabile risposta, mentre quello dei lanciatori partenti sì.**

La RDg, si ricorda, comprende sia i punti segnati che quelli subiti da una squadra: evidentemente, quest'ultima informazione include indirettamente anche l'abilità del bullpen di una squadra (rendendone non significativo l'effetto su Y), ma non quella del corrispondente lanciatore partente. D'altronde, i dati sui lanciatori di rilievo sono aggregati a livello di squadra, mentre quelli dei lanciatori partenti sono individuali, dunque la conclusione appare plausibile. Al contrario, se come stima dell'abilità di squadra si utilizza una misura di produzione puramente offensiva come i RSg o i wRC+, il talento dei lanciatori di rilievo non è più incluso e l'effetto delle corrispondenti variabili risulta di conseguenza significativo.

Prima di procedere alla valutazione della qualità dei modelli è necessario eliminare le variabili il cui effetto è risultato non significativo e procedere a stimare i nuovi valori dei coefficienti: questo riduce i modelli interi da 27 a 9, di cui 3 sono composti da 4 variabili indipendenti e i rimanenti da 6. Nella tabella 3.2 i suddetti sono ordinati in base al valore dell'AIC, che si ricorda essere una stima della quantità di informazione persa quando si rappresenta il fenomeno reale per mezzo del modello. Valori minori dell'AIC sono indicatori di modelli di qualità superiore.

Il miglior modello fra quelli considerati sembrerebbe essere quello che utilizza i punti segnati a partita per la stima dell'abilità delle due squadre e l'ERA come indicatore del talento dei lanciatori partenti e di rilievo.

Modello	Nº variabili	AIC
RSg squadre + ERA partenti + ERA bullpen	6	12824.288
RDg squadre + ERA partenti	4	12833.854
wRC squadre + ERA partenti + ERA bullpen	6	12857.314
RDg squadre + FIP partenti	4	12944.862
RSg squadre + FIP partenti + FIP bullpen	6	12969.640
RDg squadre + xFIP partenti	4	12984.945
wRC squadre + FIP partenti + FIP bullpen	6	12997.664
RSg squadre + xFIP partenti + xFIP bullpen	6	13032.253
wRC squadre + xFIP partenti + xFIP bullpen	6	13052.457

Tabella 3.2: Modelli interi (privi di variabili con effetto non significativo) per la relazione fra Y (vittoria della squadra di casa) e i regressori disponibili, ordinati in base ai valori dell'AIC. L'AIC è un compromesso fra bontà di adattamento del modello ai dati e numero di parametri (variabili) al suo interno. Valori minori dell'AIC sono indicatori di una qualità maggiore del modello rispetto a un altro.

Un'altra considerazione è che sia per stimare le abilità di squadra che quelle dei lanciatori non sembra essere particolarmente vantaggioso utilizzare statistiche avanzate (wRC, xFIP) rispetto a indicatori semplici (RSg, RDg, ERA) quando l'obiettivo finale è determinare la probabilità di vittoria della partita. Al contrario, i modelli contenenti variabili semplici sono risultati migliori in termini di AIC di quelli che utilizzavano indicatori complessi.

La tabella 3.3 contiene l'output del modello intero che è risultato rappresentare il miglior compromesso fra adattamento ai dati e numero di parametri fra quelli considerati. Le variabili sono centrate rispetto alle medie stagionali.

MODELLO INTERO 1	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1520	0.0210	7.2370	0	1.1642
RSg casa	0.2317	0.0554	4.1811	0	1.2607
RSg trasferta	-0.3937	0.0555	-7.0969	0	0.6745
ERA partente casa	-0.2110	0.0177	-11.8892	0	0.8098
ERA partente trasf.	0.2463	0.0189	13.0069	0	1.2793
ERA bullpen casa	-0.1876	0.0408	-4.6029	0	0.8290
ERA bullpen trasf.	0.1543	0.0410	3.7619	0.0002	1.1668

Tabella 3.3: Output del modello logistico Vittoria squadra casa ~ RSg **casa** + RSg **trasferta** + ERA partente **casa** + ERA partente **trasferta** + ERA bullpen **casa** + ERA bullpen **trasferta**

Quando si specifica un modello multiplo, le singole variabili risentono dell'effetto concomitante esercitato sulla variabile risposta da tutti gli altri regressori considerati. Questo spiega le differenze significative nei valori dei coefficienti (e degli odds ratio) riportati nella tabella 3.3 e quelli stimati nel capitolo 2 nell'ambito dei modelli bivariati. A parità di tutte le altre variabili, una squadra di casa che segna un punto per partita in più rispetto alla media stagionale vede aumentare i propri odds di vittoria del 26%, e una squadra in trasferta del 33%. Per pari abilità di squadra e dei lanciatori di rilievo, invece, ogni punto sopra alla media dell'ERA dei lanciatori partenti causa una riduzione media degli odds del 19% per la squadra di casa e del 28% per la squadra in trasferta. Infine, sempre a parità di tutte le altre variabili, avere un bullpen con un'ERA di un punto superiore alla media corrisponde generalmente a una riduzione negli odds di vittoria del 17%. Per squadre con abilità nella media (tutte le X pari a zero), gli odds di vittoria iniziali della squadra di casa sono normalmente superiori di circa il 16% rispetto a quelli della squadra ospite.

Specificato il modello, a questo punto è possibile utilizzarlo per calcolare, dati i valori delle variabili indipendenti per le due squadre avversarie, la probabilità di vittoria della squadra di casa ($Y=1$) tramite la seguente formula:

$$P [Y=1 | X] = \frac{e^{(\beta_0 + \beta X)}}{1 + e^{(\beta_0 + \beta X)}} =$$

$$\frac{0.152 + 0.232 * RSg\ casa - 0.394 * RSg\ trasferta +}{\exp \left(-0.211 * ERA\ partente\ casa + 0.246 * ERA\ partente\ trasferta + \right)} \\ \frac{-0.188 * ERA\ bullpen\ casa + 0.154 * ERA\ bullpen\ trasferta}{1 + \exp \left(-0.211 * ERA\ partente\ casa + 0.246 * ERA\ partente\ trasferta + \right)} \\ \frac{0.152 + 0.232 * RSg\ casa - 0.394 * RSg\ trasferta +}{-0.211 * ERA\ partente\ casa + 0.246 * ERA\ partente\ trasferta +} \\ \frac{-0.188 * ERA\ bullpen\ casa + 0.154 * ERA\ bullpen\ trasferta}{}$$

Ovviamente, la probabilità di vittoria per la squadra in trasferta ($P [Y=0|X]$) è facilmente ottenibile come $1 - P [Y=1 | X]$.

A titolo di esempio, si consideri la partita (estratta a sorte dal dataset) Chicago Cubs vs Chicago White Sox, disputata in casa dei White Sox il 14 agosto 2015. I Cubs nel 2015 hanno registrato una media punti segnati a partita di 4.25, mentre il valore di RSg dei White Sox è risultato uguale a 3.84. I due lanciatori partenti, Jeff Samardzija per i White Sox e Kyle Hendricks per i Cubs, hanno vantato un'ERA stagionale rispettivamente pari a 4.96 e 3.95. L'ERA dei bullpen, infine, è stata di 3.67 per i White Sox e 3.38 per i Cubs. La tabella 3.4 riprende quanto elencato, presentando inoltre i valori delle medie stagionali e, di conseguenza,

le variabili espresse in forma di differenza dalla media (che, si ricorda, sono quelle utilizzate per la specificazione del modello).

	Valore (2015)	Media (2015)	Differenza
RSg White Sox	3.840	4.250	-0.410
RSg Cubs	4.253	4.250	0.003
ERA partente White Sox	4.960	4.094	0.866
ERA partente Cubs	3.950	4.094	-0.144
ERA bullpen White Sox	3.670	3.713	-0.043
ERA bullpen Cubs	3.380	3.713	-0.333

Tabella 3.4: Valori assunti dai regressori per la partita di MLB White Sox – Cubs del 14/08/2015

La probabilità di vittoria per la squadra di casa (Chicago White Sox) stimata tramite il modello intero è pari a:

$$\frac{\exp \left(\begin{array}{l} 0.152 + 0.232 * (-0.410) - 0.394 * (0.003) + \\ - 0.211 * (0.866) + 0.246 * (-0.144) + \\ - 0.188 * (-0.043) + 0.154 * (-0.333) \end{array} \right)}{1 + \exp \left(\begin{array}{l} 0.152 + 0.232 * (-0.410) - 0.394 * (0.003) + \\ - 0.211 * (0.866) + 0.246 * (-0.144) + \\ - 0.188 * (-0.043) + 0.154 * (-0.333) \end{array} \right)} = \frac{\exp (-0.20565)}{1 + \exp (-0.20565)} = \mathbf{0.4487}$$

E, di conseguenza, la probabilità di vittoria stimata per la squadra in trasferta (Chicago Cubs) è uguale a $1 - 0.4487 = \mathbf{0.5513}$. Il risultato finale della partita è stato di 6-5 per i Chicago Cubs.

3.2 *Il modello stepwise*

La tecnica usata all'inizio del paragrafo precedente per la selezione delle variabili da inserire nel modello intero ha previsto la scelta, volta per volta, di un singolo stimatore per ognuna delle 4 componenti (abilità di squadra, abilità dei lanciatori partenti, abilità dei rilievi). Per quanto questo procedimento possa sembrare logicamente ineccepibile, non è del tutto corretto dal punto di vista metodologico poiché esclude dal ragionamento un numero elevato di modelli che prevedono, ad esempio, combinazioni di più variabili per la stessa componente. In altre parole, possiamo affermare con certezza che il modello *RSg squadre + ERA partenti + ERA bullpen* sia il migliore fra i 9 modelli *interi* trattati, ma non abbiamo alcuna certezza sul

fatto che questo sia il migliore fra tutti i modelli che si possono creare a partire dalla serie dei regressori.

Per rispondere a questa domanda, viene fortunatamente in aiuto la funzione *stepAIC* di R, che evita di dover specificare singolarmente i 16'777'215 modelli possibili²⁰. La funzione utilizza la tecnica *stepwise*, che consiste in un metodo di selezione che parte dal modello nullo (privo di regressori) e inserisce volta per volta la variabile che genera la maggiore riduzione dell'AIC. In seguito ad ogni aggiunta, le variabili già presenti vengono controllate per testare se il loro apporto risulta ancora significativo. Il risultato della funzione *stepAIC* è la specificazione del seguente modello:

MODELLO STEPWISE	$\hat{\beta}_i$	S. E.	Z di Wald	P-value	OR ($e^{\hat{\beta}_i}$)
(Intercetta)	0.1513	0.0211	7.1700	0	1.1633
RSg casa	0.2192	0.0555	3.9462	0.0001	1.2451
RSg trasf.	-0.3852	0.0556	-6.9313	0	0.6803
ERA P. trasf.	0.2169	0.0221	9.8012	0	1.2423
ERA P. casa	-0.2237	0.0268	-8.3509	0	0.7996
FIP P. casa	0.0893	0.0439	2.0338	0.0420	1.0934
xFIP P. trasf.	0.0824	0.0326	2.5256	0.0116	1.0859
xFIP P. casa	-0.1163	0.0444	-2.6214	0.0088	0.8902
ERA B. trasf	0.1458	0.0412	3.5400	0.0004	1.1570
ERA B. casa	-0.1805	0.0410	-4.4004	0	0.8349

Tabella 3.5: Specificazione del modello stepwise ottenuto tramite la funzione *stepAIC* di R.

Il relativo AIC è pari a 12816, quindi appena inferiore a quello del migliore dei modelli interi trattati in precedenza (12824). Come quest'ultimo, anche il modello stepwise contiene i punti subiti a partita (RSg) come stimatore dell'abilità delle squadre e l'ERA dei bullpen come indicatore del talento dei lanciatori di rilievo. Per quanto riguarda i lanciatori partenti, invece, è utilizzata una sorta di combinazione lineare di ERA, FIP e xFIP per stimare il talento del lanciatore di casa, e di ERA e xFIP per quello in trasferta. Tornando all'esempio del paragrafo

²⁰ I regressori totali sono: 3 per l'abilità di squadra, 3 per l'abilità dei partenti, 3 per l'abilità dei rilievi e 3 per lo stato di forma, ovvero 12. Poiché ognuno è considerato per entrambe le squadre, le variabili indipendenti totali sono $12 \times 2 = 24$. A partire da 24 elementi, si possono creare $\sum_{k=1}^{24} \frac{n!}{k! \times (n-k)!} = 16777215$ modelli con $k = 1, \dots, 24$ parametri.

precedente, la probabilità di vittoria dei White Sox stimata tramite il modello stepwise risulta pari a **0.4289** (-0.0198), mentre quella dei Cubs uguale a **0.5711** (+0.0198).

3.3 *La capacità previsiva*

Il modello riportato nella tabella 3.5 rappresenta il migliore possibile in termini di AIC data la serie di variabili indipendenti contenute nel dataset. La tecnica di selezione *stepwise*, tuttavia, non è esente da critiche²¹. Inoltre, il guadagno in AIC nel passaggio dal migliore dei modelli interi al modello stepwise è minimo, mentre è notevole la perdita in interpretabilità dei coefficienti. Poiché lo scopo della ricerca del modello ottimale è quello di utilizzarlo per effettuare delle previsioni sulla stagione in corso, appare naturale analizzare quale dei modelli descritti possiede migliore capacità previsiva. Per la valutazione di questo aspetto, come anticipato nel paragrafo 1.6.5, si utilizzerà l'area compresa sotto la curva ROC.

A tale proposito, il dataset 2015-2018 è stato suddiviso in *training set* (80%, 7774 osservazioni) e *test set* (20%, 1943 osservazioni). Il *training set* è stato utilizzato per stimare nuovamente i valori dei coefficienti dei modelli interi (tabella 3.2) e del modello stepwise (tabella 3.5). Questi sono in seguito stati utilizzati per stimare le probabilità di vittoria della squadra di casa nelle partite comprese nel *test set*, e quindi riprodurre le relative curve ROC. La curva ROC, si ricorda, rappresenta le coppie di valori *proporzione veri positivi (sensibilità)* e *proporzione falsi positivi (1-specificità)* del metodo di classificazione al variare del cut-off di probabilità k . Nel caso trattato, la proporzione di veri positivi consiste nella frazione (vittorie della squadra di casa previste correttamente / totale vittorie squadra di casa), mentre la proporzione di falsi positivi equivale al rapporto (vittorie della squadra in trasferta previste come vittorie della squadra di casa / totale vittorie squadra in trasferta). Dal momento che si sono usati due dataset indipendenti per stimare i coefficienti dei modelli e valutarne la capacità previsiva, la possibile distorsione da selezione è annullata²².

Il seguente grafico 3.1 mostra le curve ROC relative ai modelli interi, mentre il grafico 3.2 quella rappresentata a partire modello stepwise. In entrambi i casi sono inclusi i valori dell'area compresa sotto la curva (AUC).

²¹ Harrel, 2015.

²² Cawley e Talbot, 2010.

Grafico 3.1: Curve ROC relative ai modelli interi riportati in tabella 3.2

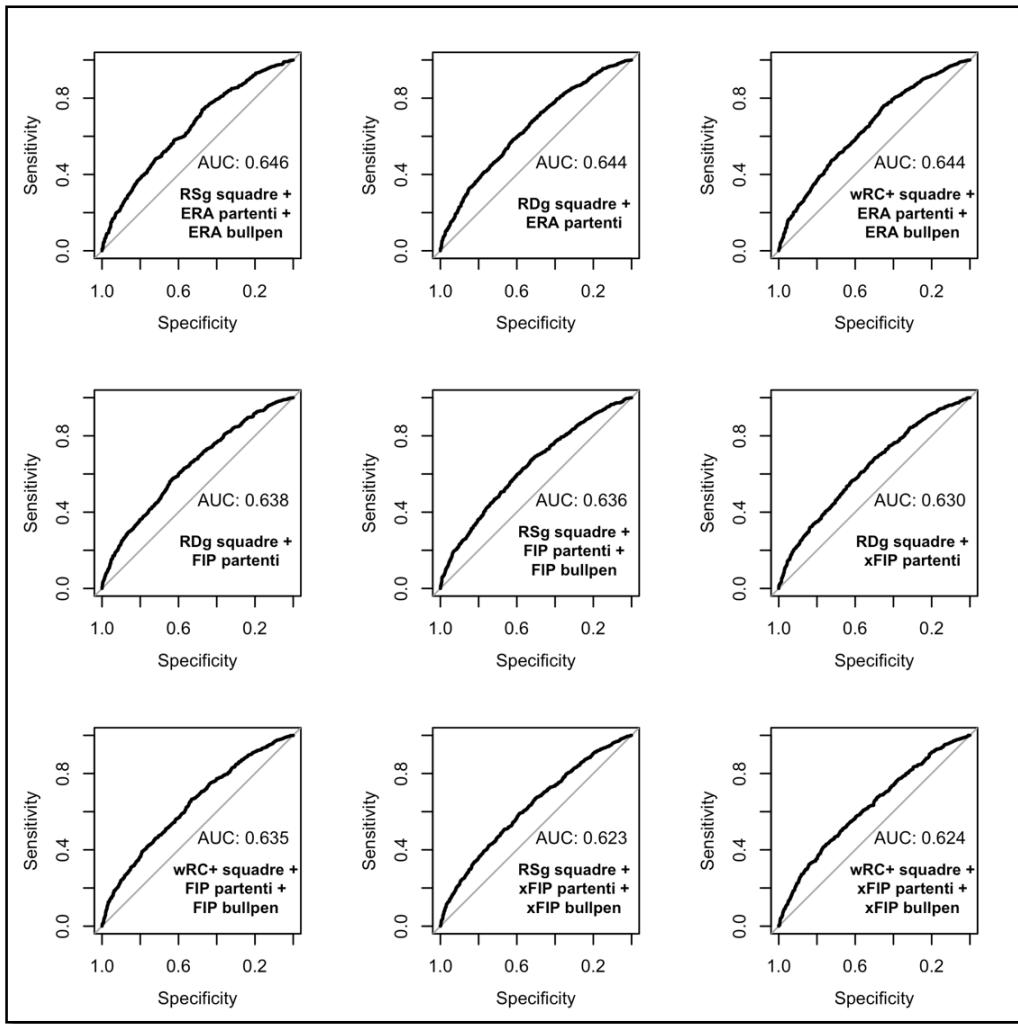
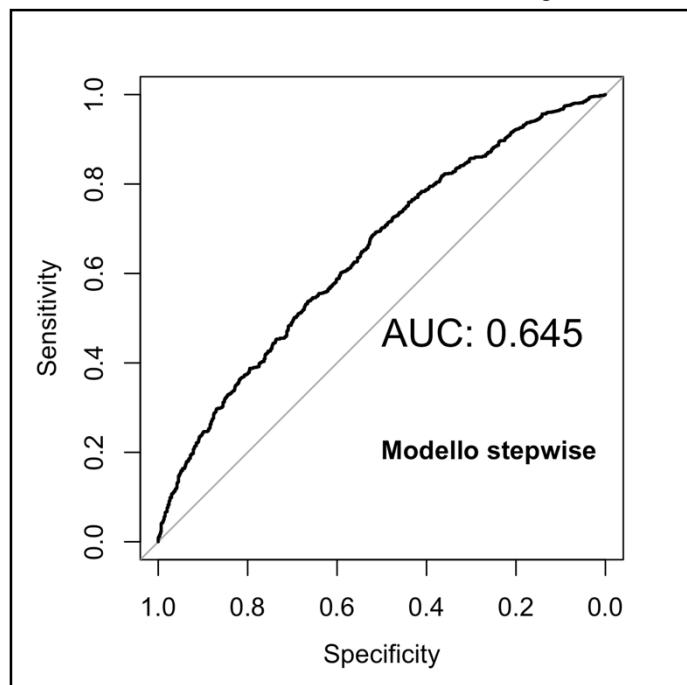


Grafico 3.2: Curva ROC del modello stepwise



Come si può notare, le aree sono tutte superiori a 0.5 e i modelli sono dunque da considerarsi quanto meno come un miglioramento rispetto alla scelta casuale di una delle due squadre. Tra i modelli interi (grafico 3.1), quello con la migliore capacità previsiva consiste nel modello *RSg squadre + ERA partenti + ERA bullpen* ($AUC = 0.646$, rappresentato in alto a sinistra), ovvero quello a cui corrisponde anche il minore valore di AIC. Il modello stepwise, come dimostrato in figura 3.2, pur rappresentando un (minimo) miglioramento in termini di AIC, possiede una capacità previsiva dell'outcome di interesse pressoché identica ($AUC = 0.645$) e comporta notevoli complicazioni nell'interpretazione dei parametri. Di conseguenza, il modello adottato sarà il modello intero avente AIC minore e AUC maggiore, i cui coefficienti stimati sono stati precedentemente riportati nella tabella 3.3.

3.4 *Diagnostica*

Il modello ritenuto ottimale in base ai regressori a disposizione, dunque, è quello che mette in relazione la probabilità di vittoria di una partita con i la media dei punti segnati delle due squadre, l'ERA dei corrispondenti lanciatori partenti e quella dei rispettivi lanciatori di rilievo. L'ultimo passo per completare la validazione del modello consiste nella fase diagnostica, che comprende la verifica dell'assunzione di linearità della trasformata logit rispetto ai valori delle variabili indipendenti, la verifica di assenza di multicollinearità fra i regressori e l'assenza di osservazioni influenti tramite analisi dei residui²³.

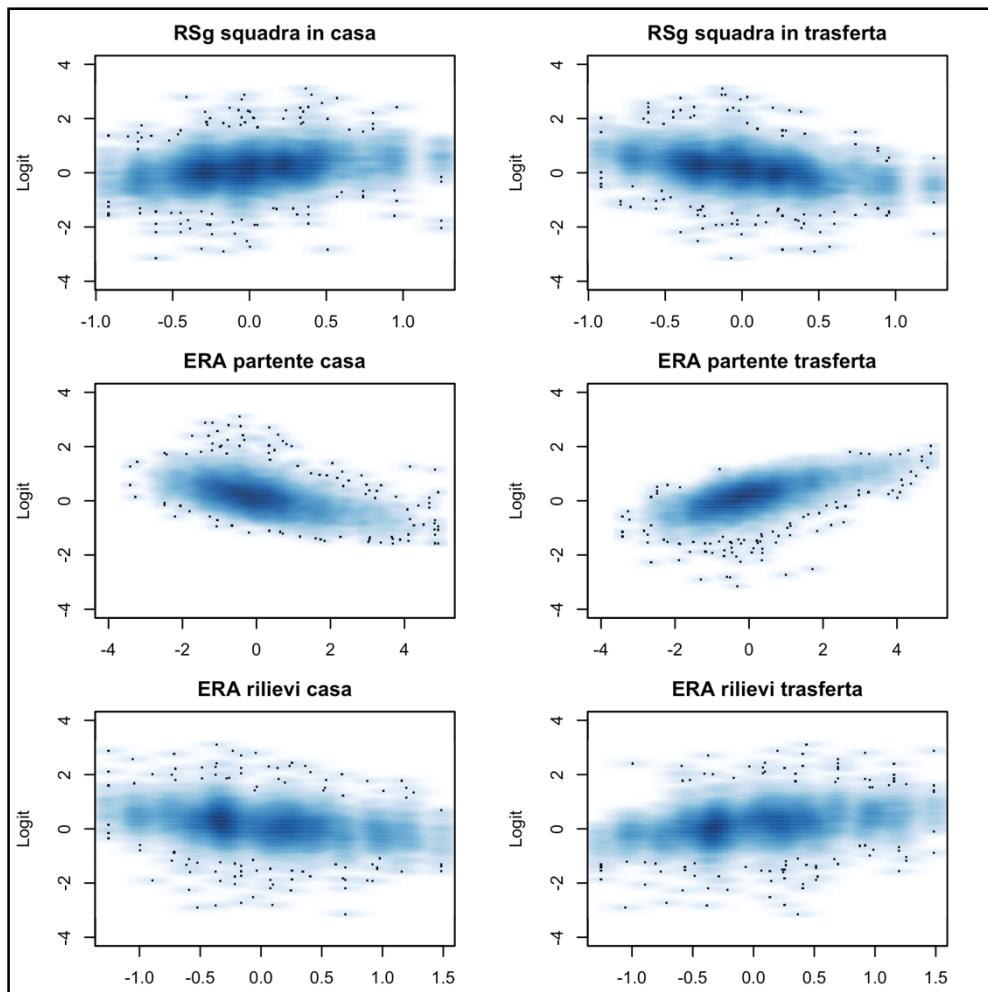
La linearità della relazione fra trasformata logit e variabili indipendenti, si ricorda, è di fatto l'unica assunzione alla base del modello di regressione logistica. La verifica può essere effettuata graficamente, rappresentando il logit, ovvero la trasformata logaritmica dell'odds ($\beta_0 + \beta_1 * X_1 + \dots + \beta_k * X_k$) in funzione di ognuno dei 6 regressori inclusi.

Il grafico 3.3 rappresenta quanto descritto tramite grafici di dispersione *lisciati*, ovvero tenenti conto della densità di distribuzione dei punti. Per l'ERA dei lanciatori partenti si è considerato un intervallo di valori compreso fra -4 e 5²⁴ allo scopo di eliminare i valori anomali, poco interessanti per quanto riguarda la relazione con il logit ma che avrebbero reso tuttavia incomprensibile il grafico.

²³ Per approfondimenti sulla diagnostica si rimanda a Harrel (2016), Hosmer et al. (2013), Zhang (2016).

²⁴ Si ricorda che le variabili sono espresse in forma di differenza rispetto alle medie stagionali. Un intervallo fra -4 e 5, di conseguenza, corrisponde a un'ERA compresa fra 0 e 9, dal momento che le medie stagionali per i partenti sono normalmente fra 4 e 4.5. Come già discusso, tale intervallo ricopre quasi il 95% delle osservazioni rilevate.

Grafico 3.4: Distribuzione della trasformata logit in funzione dei valori assunti dai regressori inclusi nel modello. L'assunzione alla base è quella di linearità di tale relazione.



Si può osservare come, seppure non molto accentuata e con un notevole livello di dispersione sull'asse verticale, la relazione fra logit e regressori sia indubbiamente di natura lineare. L'assunzione è dunque rispettata.

Per quanto riguarda la multicollinearità, la tabella 3.6 contiene i valori di VIF (Fattore di Inflazione della Varianza) calcolati per ogni variabile indipendente.

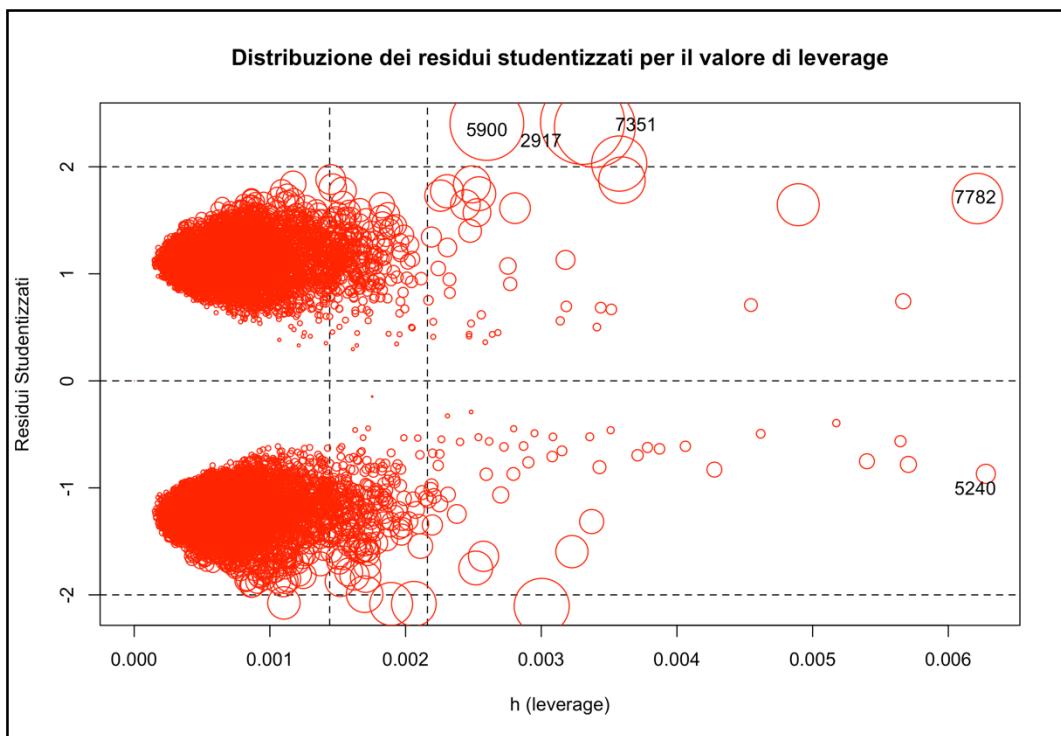
	RSg casa	RSg trasferta	ERA P casa	ERA P trasferta	ERA B casa	ERA B trasferta
VIF	1.0633	1.0627	1.0563	1.0683	1.0912	1.0992

Tabella 3.6: valori del VIF per le 6 variabili indipendenti inserite nel modello

Dal momento che tutti i valori sono largamente inferiori alla soglia di 10, si può concludere che non vi è alcuna relazione di multicollinearità fra i regressori.

L'assenza di osservazioni influenti, infine, può essere visualizzata dal grafico 3.4, ottenuto con la funzione *influencePlot* del pacchetto *Car*. Il grafico combina le tre misure principalmente utilizzate in ambito diagnostico: ogni punto è riferito a un'osservazione, di cui è riportato il *residuo studentizzato* sull'asse verticale e il *valore di leverage* sull'asse orizzontale, mentre la *distanza di Cook* è tanto maggiore in base alle dimensioni del corrispondente cerchio. Osservazioni estreme in termini di residui costituiscono possibili outliers, mentre valori elevati sull'asse delle X rappresentano potenziali punti di leverage.

Grafico 3.5



La funzione riporta automaticamente gli indici delle osservazioni che sono ritenute potenzialmente influenti. In particolare, le osservazioni 5900, 2917 e 7351 hanno residui studentizzati superiori alla soglia di 2 e distanza di Cook considerevoli, mentre la 5240 e la 7782 hanno valori di leverage estremi.

La tabella 3.7 riporta i valori assunti dalle variabili indipendenti, della variabile risposta ($Y =$ vittoria della squadra di casa) e della probabilità stimata di vittoria della squadra di casa per le osservazioni in questione. Come si può osservare, le tre osservazioni con residui elevati (*outliers*) sono relative a partite in cui la probabilità stimata era estremamente bassa (0.05-0.06) ma la squadra di casa ha vinto comunque ($Y=1$). In tutti e 5 i casi, la variabile ERA del lanciatore partente di casa assume valori anomali estremamente elevati (le variabili sono

espresse come differenza dalla media stagionale), influenzando fortemente la probabilità stimata e portando a considerare alcune osservazioni come potenziali leverage.

ID	Casa	Trasf.	data	RSg C.	RSg T.	ERA P. C.	ERA P. T.	ERA B. C.	ERA B. T.	Y	P [Y=1]
2917	CHC	SEA	2016-07-31	0.510	0.263	13.665	-0.515	-0.374	-0.384	1	0.056
5240	BOS	BAL	2017-05-04	0.199	-0.060	8.478	2.558	-0.995	-0.215	0	0.314
5900	KCR	MIN	2017-07-01	-0.313	0.384	12.398	-0.592	0.095	0.255	1	0.057
7351	ANA	OAK	2018-04-06	0.002	0.570	13.364	0.994	-0.165	-0.715	1	0.061
7782	CHC	MIA	2018-05-08	0.220	-0.790	9.314	-0.206	-0.735	1.255	1	0.237

Tabella 3.7: Valori dei regressori, della variabile risposta e della probabilità stimata per le potenziali osservazioni influenti del dataset. In rosso sono evidenziati valori inferiori al 5° o superiori al 95° percentile delle distribuzioni delle rispettive variabili.

Trattandosi di pochi casi anomali, la situazione non appare preoccupante. A conferma di ciò, la tabella 3.8 riporta i valori dei coefficienti e dei corrispondenti errori standard del modello, stimati da un lato sul dataset completo e dall'altro sul dataset da cui sono state rimosse le potenziali osservazioni influenti.

	Modello stimato su dataset completo	Modello stimato senza osservazioni influenti
(Intercetta)	0.1520	0.1518
Standard Error	0.0210	0.0210
RSg squadra casa	0.2317	0.2290
Standard Error	0.0554	0.0554
RSg squadra trasferta	-0.3937	-0.3937
Standard Error	0.0555	0.0555
ERA partente casa	-0.2110	-0.2151
Standard Error	0.0177	0.0178
ERA partente trasferta	0.2463	0.2463
Standard Error	0.0189	0.0189
ERA bullpen casa	-0.1876	-0.1852
Standard Error	0.0408	0.0408
ERA bullpen trasferta	0.1543	0.1544
Standard Error	0.0410	0.0410

Tabella 3.8: Confronto fra i coefficienti del modello stimato sul dataset completo e quello stimato senza le osservazioni influenti della tabella 3.7

Come previsto, le differenze fra i coefficienti sono inesistenti o minime, e si può dunque concludere che non vi sono osservazioni influenti nel dataset. Il modello, la cui bontà e le cui assunzioni sono state validate, è ora pronto all'utilizzo.

3.5 Riepilogo

Il capitolo appena concluso ha trattato la scelta delle variabili da inserire nel modello di regressione per ottenere quello che risultasse il migliore compromesso fra bontà di adattamento ai dati, numero ridotto di parametri utilizzati e capacità previsiva. La scelta è ricaduta sul modello che considera i punti segnati per partita come stimatore dell'abilità di squadra e l'ERA di partenti e rilievi come indicatore del talento dei lanciatori. Tutte le variabili sono da considerarsi come centrate sulle rispettive medie stagionali.

Il modello è stato in seguito sottoposto alla fase diagnostica, per valutare la linearità della relazione fra logit e regressori e l'assenza di multicollinearità e di osservazioni influenti. Tutte le verifiche hanno dato esito positivo.

Il passo conclusivo di questo elaborato sarà l'utilizzo del modello appena specificato e validato per scopi di previsione sulla stagione di Major League Baseball in corso (2019).

Capitolo 4

Causalità, previsioni e simulazioni

4.1 Obiettivi e problematiche

L’obiettivo principale del capitolo conclusivo di questo elaborato sarà stabilire il ruolo svolto dal caso sui risultati di una stagione di Major League Baseball. A tale scopo, il modello di regressione selezionato al termine della fase precedente sarà applicato al calendario della stagione 2019 per stimare le probabilità di vittoria di ciascuna squadra per ognuna delle partite in programma. Una volta ottenute le probabilità, queste saranno utilizzate per determinare i successi (vittoria della squadra di casa) o viceversa gli insuccessi (vittoria della squadra in trasferta) di una serie di *prove di Bernoulli* indipendenti. Ogni partita del calendario, in altri termini, sarà assunta come una variabile aleatoria di Bernoulli avente probabilità di successo p pari alla probabilità di vittoria della squadra di casa stimata dal modello ($P[Y=1|X]$), e probabilità di insuccesso $1-p$ pari alla probabilità stimata di vittoria della squadra in trasferta. La simulazione di una singola stagione consiste in una realizzazione di questa serie di prove.

Per fare un esempio, si riprendano i dati della partita Cubs – White Sox riportati alla fine del paragrafo 3.1. Una vittoria dei White Sox, squadra di casa, è risultata avere probabilità pari a 0.4487. Questo significa che giocando 1000 volte la partita in questione, è lecito attendersi per la Legge dei Grandi Numeri²⁵ che i White Sox risultino vincitori circa 449 volte. Nell’ambito di un singolo svolgimento, tuttavia, il risultato finale è fortemente dettato dal caso: è infatti paragonabile al lancio di una moneta “truccata” per cui l’esito *testa* (vittoria dei Cubs, $P = 0.5513$) è leggermente più probabile di *croce* (vittoria dei White Sox, $P = 0.4487$).

Simulando un’intera stagione, dal momento che le prove sono indipendenti²⁶ il numero di vittorie di ogni squadra sarà generato da 162 lanci di moneta, ognuno avente probabilità di *testa* e *croce* determinate dai valori assunti dalle variabili indipendenti per ogni partita. Chiaramente, squadre più forti tenderanno ad avere per ogni partita probabilità mediamente superiori di vittoria rispetto a squadre deboli, ed è dunque logico che vinceranno indicativamente un numero

²⁵ Routledge, 2016

²⁶ L’assunzione è plausibile, dal momento che è stato verificato nel capitolo 3 che un’elevata percentuale di vittorie nelle partite precedenti non ha effetto diretto sulla probabilità di vittoria della partita in questione.

maggior parte. È altresì evidente, tuttavia, che la fortuna o la sfortuna giochino un ruolo fondamentale nel determinare gli esiti di una stagione. A partire dalla simulazione di un numero elevato di ripetizioni dove le abilità sono mantenute costanti, è possibile stabilire una stima del numero di vittorie e sconfitte generalmente attribuibili al caso osservando la dispersione dei risultati finali per ogni squadra.

La principale problematica che si incontra nell'utilizzare il modello di regressione logistica selezionato in precedenza per prevedere i risultati futuri è che i valori delle abilità di squadra e individuali sono non solo ignoti, ma gli stimatori utilizzati (RSg, ERA, eccetera) devono inoltre essere calcolati o a partire da un numero ridotto di partite nel caso di stagione già iniziata, o addirittura interamente ipotizzati se questa deve ancora cominciare. Questo, ovviamente, inserisce un'ulteriore componente di variabilità e di potenziale distorsione. Fortunatamente, esistono numerose organizzazioni che si occupano di effettuare *proiezioni* del rendimento individuale, basate sulle performance passate, l'età, la regressione verso la media e una varietà di altri fattori. Le proiezioni delle performance complessive di squadra sono ottenute come combinazioni delle singole proiezioni dei giocatori. In generale, le proiezioni sono preferibili a dati derivanti da stime su un numero ridotto di partite o interamente ripresi dalla stagione precedente per due motivi principali:

1. Valori medi calcolati su un numero ridotto di partite risentono particolarmente di effetti casuali, e potrebbero quindi essere stime distorte del vero talento. Le proiezioni tengono conto di ciò regredendo i valori estremi in direzione delle rispettive medie.
2. I dati potrebbero essere non disponibili o variare considerevolmente da una stagione all'altra per alcuni giocatori. Gli algoritmi di proiezione riescono a risolvere il problema, considerando il fattore età per i giocatori in fase calante e adattando le performance al livello della Major League per quelli provenienti dalle leghe minori.

I sistemi di proiezione riconosciuti e utilizzati nel settore sono molteplici. Per la nostra simulazione, si utilizzeranno i dati *Steamer* per le performance individuali dei lanciatori partenti e le *Depth Charts* di *Fangraphs* per i dati aggregati di squadra. Il sistema Steamer è considerato uno dei metodi di proiezione più accurati²⁷ fra quelli disponibili, mentre le Depth Charts consistono in una combinazione di più sistemi²⁸.

²⁷ Si veda la bibliografia, *What is a Steamer? Glossary. (Major League Baseball)*

²⁸ *FanGraphs Baseball - Baseball Statistics and Analysis*

Un secondo problema è rappresentato dal fatto che il modello di regressione tiene conto dell'abilità dei lanciatori partenti, la cui sequenza futura è ignota²⁹. La soluzione adottata sarà quella di sorteggiare la rotazione di ogni squadra tramite una particolare tecnica di campionamento, con probabilità di estrazione per ciascun lanciatore proporzionale al numero di riprese che lancerà nel 2019 secondo le proiezioni *Steamer* e con il vincolo che un lanciatore sorteggiato per la partita n non possa essere selezionato nuovamente fino alla partita $n + 4$ (ovvero, tenendo conto del numero di partite di riposo normalmente concesso).

4.2 Una simulazione della stagione in corso

Le classifiche delle 6 *division* di Regular Season di MLB aggiornate al 4 giugno 2019 sono riportate nella tabella 4.1.

Squadra	G	V	P	% vittorie	Squadra	G	V	P	% vittorie
Yankees	58	38	20	0.655	Phillies	60	33	27	0.550
Rays	57	35	22	0.614	Braves	59	32	27	0.542
Red Sox	59	30	29	0.509	Mets	59	28	31	0.475
Blue Jays	59	21	38	0.356	Nationals	59	26	33	0.441
Orioles	59	18	41	0.305	Marlins	57	21	36	0.368
Twins	58	40	18	0.690	Brewers	60	34	26	0.567
White Sox	59	29	30	0.492	Cubs	58	32	26	0.552
Indians	59	29	30	0.492	Cardinals	58	30	28	0.517
Tigers	57	22	34	0.393	Pirates	58	28	30	0.483
Royals	59	19	40	0.322	Reds	59	27	32	0.458
Astros	61	41	20	0.672	Dodgers	61	42	19	0.689
Rangers	57	30	27	0.526	Rockies	58	31	27	0.535
Athletics	60	29	30	0.492	Padres	60	31	29	0.517
Angels	60	29	31	0.483	Diamondbacks	61	30	31	0.492
Mariners	63	25	38	0.397	Giants	58	24	34	0.414

Tabella 4.1: Classifiche delle division di MLB aggiornate al 4 giugno 2019. G = Partite Giocate, V = Partite Vinte, P = Partite Perse, % Vittorie = (Partite Vinte / Partite Giocate)

Come si può notare, tutte le squadre hanno disputato circa 60 partite su 162. Il quesito di interesse è: **quali saranno le classifiche finali al termine della Regular Season?** Per tentare di formulare una risposta, si è proceduto alla stima delle probabilità di vittoria di ciascuna delle rimanenti partite nel calendario tramite il modello di regressione specificato nel capitolo 3. Come discusso in precedenza, si sono utilizzati i dati delle proiezioni *Steamer* e *Fangraphs* aggiornati al 4 giugno e si è sorteggiata la futura rotazione dei lanciatori partenti. Ottenute le

²⁹ I lanciatori partenti di una partita vengono di norma programmati a non più di 3-4 giorni di distanza.

probabilità, il risultato di ogni partita è stato decretato tramite la realizzazione di un esperimento di Bernoulli, avente come probabilità di successo p la probabilità stimata per la vittoria della squadra di casa.

I risultati di una singola simulazione sono contenuti nella tabella 4.2. Per entrambe le Leghe, le vincitrici delle divisioni (evidenziate in rosso) accederebbero direttamente alla fase delle Division Series di Post Season, mentre quelle evidenziate in blu si scontrerebbero in una partita secca per ottenere la *Wild Card*.

Squadra	V	P	% vittorie	Squadra	V	P	% vittorie
Yankees	96	66	0.593	Mets	87	75	0.537
Rays	93	69	0.574	Nationals	87	75	0.537
Red Sox	88	74	0.543	Phillies	83	79	0.512
Blue Jays	62	100	0.383	Braves	80	82	0.494
Orioles	52	110	0.321	Marlins	71	91	0.438
Twins	92	70	0.568	Cardinals	92	70	0.568
Indians	84	78	0.519	Reds	88	74	0.543
White Sox	74	88	0.457	Brewers	88	74	0.543
Royals	73	89	0.451	Cubs	78	84	0.481
Tigers	67	95	0.414	Pirates	68	94	0.420
Astros	107	55	0.660	Dodgers	100	62	0.617
Rangers	79	83	0.488	Rockies	84	78	0.519
Angels	77	85	0.475	Padres	80	82	0.494
Athletics	77	85	0.475	Diamondbacks	75	87	0.463
Mariners	74	88	0.457	Giants	73	89	0.451

Tabella 4.2: Risultato di una simulazione della Regular Season 2019 a partire dalle proiezioni Fangraph e Steamer aggiornate al 4 giugno 2019, utilizzando le probabilità stimate tramite il modello di regressione logistica specificato nel capitolo 3 ($W \sim RSg + ERA\ partenti + ERA\ bullpen$)

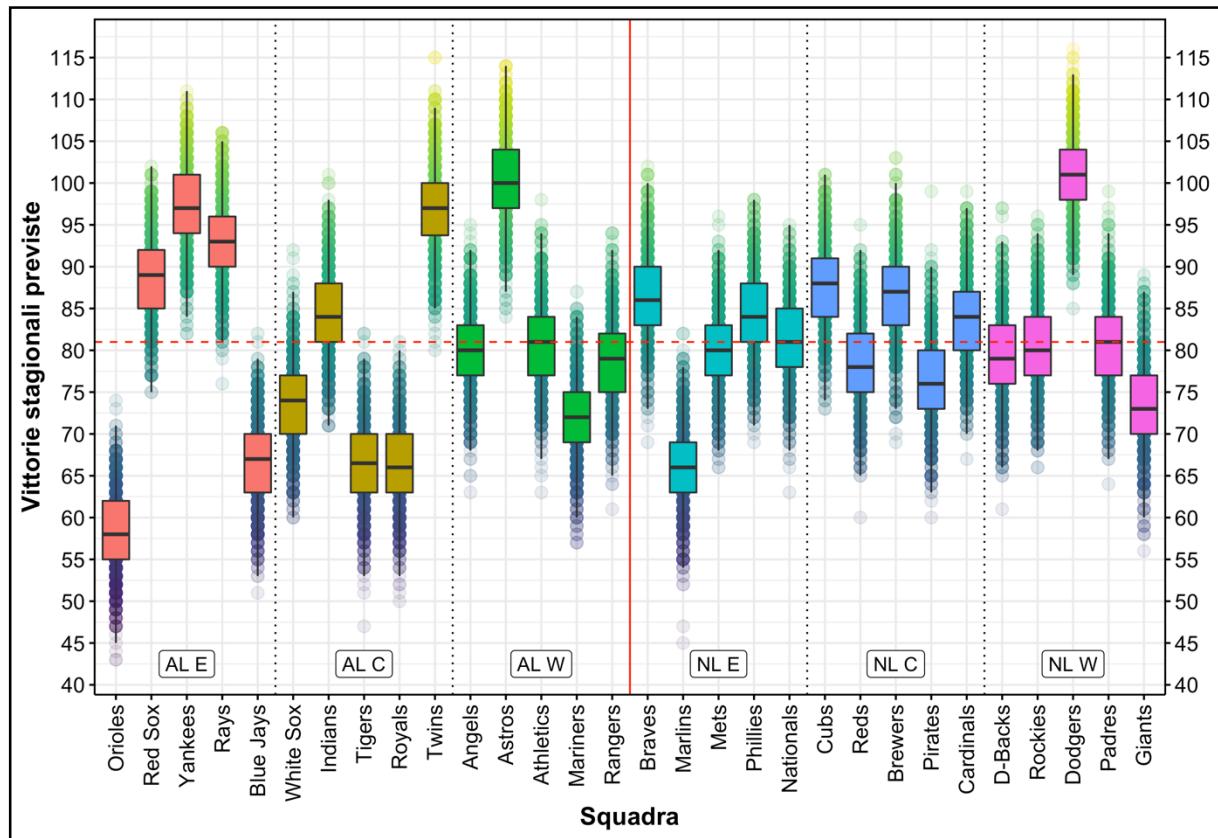
Tuttavia, come già discusso a inizio capitolo una singola ripetizione del calendario può risentire particolarmente di fluttuazioni casuali: alcune squadre potrebbero essere state molto fortunate durante tutto il proseguimento del campionato, mentre altre molto sfortunate. Per tentare di attutire e quantificare l'effetto del caso, la simulazione appena effettuata è stata ripetuta per 1000 volte, analizzando per ogni squadra la tendenza e la dispersione dei risultati previsti.

4.3 1000 simulazioni della stagione in corso

Il grafico 4.1 permette di visualizzare la distribuzione dei risultati finali per le 30 squadre di MLB nell'arco di 1000 simulazioni della fase di stagione seguente il 4 giugno. Sull'asse orizzontale sono riportate le squadre suddivise per *division*, mentre sull'asse verticale ogni

punto rappresenta il numero di vittorie registrato in una delle 1000 simulazioni. Tanto maggiore la densità di frequenza dei punti, tanto meno questi sono trasparenti. Per facilitare la visualizzazione della distribuzione dei risultati, sopra ai punti sono stati raffigurati i relativi boxplot. È infine stata tracciata una linea orizzontale in corrispondenza del valore 81, che rappresenta una percentuale di vittorie del 50%. Osservazioni al di sopra di questa linea rappresentano stagioni in cui la squadra ha ottenuto più vittorie che sconfitte, e viceversa.

Grafico 4.1: Distribuzione del numero di vittorie delle squadre di MLB in 1000 simulazioni della stagione 2019.



Dal grafico si può osservare che vi è sicuramente una tendenza a ottenere risultati migliori o peggiori a seconda del talento maggiore o inferiore della squadra, ma il livello di dispersione dovuto esclusivamente al caso è ad ogni modo piuttosto elevato. Per tutte le squadre escluse Yankees, Astros, Dodgers, Orioles e Royals si è osservata almeno una stagione su 1000 al di sopra e una al di sotto del livello di 50% di vittorie.

Le fluttuazioni aleatorie emergono in maniera evidente anche dalla tabella 4.3, che riassume in forma numerica i dati del grafico precedente.

Squadra	Min	2.5%	Mediana	Media	97.5%	Max	97.5%-2.5%	Max-Min
Yankees	82	88	97	97.31	107	111	19	29
Rays	76	83	93	92.96	103	106	20	30
Red Sox	75	80	89	88.86	98	102	18	27
Blue Jays	51	57	67	66.60	76	82	19	31
Orioles	43	49	58	58.13	67	74	18	31
Twins	80	86	97	96.99	107	115	21	35
Indians	71	75	84	84.35	95	101	20	30
White Sox	60	63	74	73.73	83	92	20	32
Tigers	47	57	66.5	66.50	76	82	19	35
Royals	50	57	66	66.34	76	81	19	31
Astros	84	91	100	100.32	110	114	19	30
Athletics	63	71	81	80.80	91	98	20	35
Angels	63	70	80	80.05	89	95	19	32
Rangers	61	70	79	78.83	89	94	19	33
Mariners	57	62	72	71.80	82	87	20	30
Braves	69	77	86	86.50	97	102	20	33
Phillies	69	74	84	84.30	94	98	20	29
Nationals	63	72	81	81.22	90	95	18	32
Mets	66	70	80	80.07	89	96	19	30
Marlins	45	56	66	66.09	76	82	20	37
Cubs	73	78	88	87.65	97	101	19	28
Brewers	69	76	87	86.56	96	103	20	34
Cardinals	67	74	84	83.60	94	99	20	32
Reds	60	68	78	78.30	89	95	21	35
Pirates	60	67	76	76.57	87	99	20	39
Dodgers	85	91	101	100.84	110	116	19	31
Padres	64	71	81	80.90	90	99	19	35
Rockies	66	70	80	80.39	91	96	21	30
Diamondbacks	61	69	79	79.03	90	97	21	36
Giants	56	63	73	73.44	84	89	21	33

Tabella 4.3: Riassunto dei risultati delle 1000 simulazioni per ogni squadra. All'interno di ogni division, le squadre sono ordinate in base alla media di vittorie ottenute.

È lecito attendersi che ogni squadra, al termine della stagione 2019, ottenga un numero di vittorie simile ai valori medi evidenziati in grassetto nella tabella (ammesso, ovviamente, che le proiezioni e i coefficienti del modello siano stime corrette delle rispettive quantità in popolazione). Tuttavia, all'interno delle 1000 simulazioni è occasionalmente successo che squadre particolarmente talentuose come i Los Angeles Dodgers e gli Houston Astros ottengessero appena 85 e 84 vittorie, così come è accaduto che squadre di medio livello quali gli Oakland Athletics e i San Diego Padres sfiorassero i 100 successi stagionali. Generalmente, la differenza casuale fra massimo e minimo di vittorie stagionali ottenute da ogni squadra si aggira

intorno alle 30 partite, corrispondente a circa il 30% dei risultati stimati³⁰. La differenza fra il percentile 97.5 e il percentile 2.5 della distribuzione è invece generalmente pari a 20 vittorie. Questo significa che è lecito attendersi con un livello di fiducia del 95% che il numero di vittorie ottenute da una squadra sia compreso entro un intervallo di ± 10 da quelle attese in base al suo talento. Tuttavia, squadre molto fortunate o molto sfortunate possono arrivare a ottenere anche 15 vittorie in più o in meno. Considerando una stagione intera da 162 partite, l'effetto massimo esercitato dal caso si traduce in 20-25 vittorie di differenza rispetto al valore atteso.

In definitiva, la quantità di vittorie dipendente dal caso è decisamente elevata, e può avere effetti determinanti sull'esito della stagione di una squadra. A dimostrazione di ciò, si esamini la singola simulazione presentata nella tabella 4.2: in molti casi, la differenza fra il qualificarsi o il non qualificarsi per i playoff è stata questione di una manciata di partite. Nella National League Centro, per esempio, i Cardinals sono stati decisamente fortunati, ottenendo 92 vittorie contro le 83.6 attese (cfr. tabella 4.3) e vincendo la propria *division*, mentre gli apparentemente favoriti Cubs non hanno conquistato neanche un posto per la partita di Wild Card.

4.4 Simulazione della Post Season

Sebbene la Regular Season costituisca oltre il 97% delle partite di MLB disputate ogni anno, è indiscutibile che il momento *clou* della stagione sia rappresentato dalla Post Season, competizione ad eliminazione diretta in cui le 5 migliori squadre di ogni lega si sfidano per decretare il vincitore finale del campionato. L'obiettivo di questo paragrafo sarà quello di aggregare alle 1000 simulazioni precedenti la previsione dei risultati dei Play Off, a partire dalle relative classifiche di Regular Season stimate. Per ogni simulazione, sono state selezionate le tre vincitrici delle divisioni e le due migliori escluse di entrambe le leghe, in base al numero previsto di vittorie. In caso di arrivi a pari merito, la squadra qualificata è stata estratta a sorte (nella realtà si procede a uno spareggio su partita secca, ma il sorteggio è un'approssimazione valida). Queste sono poi state inserite nel calendario dei Play Off seguendo il regolamento della Major League Baseball, a cui si rimanda per eventuali approfondimenti. Per ogni partita, si è quindi proceduto a stimare la probabilità di vittoria della squadra di casa e a generare il risultato tramite il solito esperimento di Bernoulli. Le vincitrici sono state fatte avanzare alla fase successiva, e così via fino ad ottenere il nome del campione delle World Series. La tabella 4.4 riassume i risultati del procedimento descritto, raccogliendo per ogni squadra le percentuali di

³⁰ Si ricorda che per ogni squadra circa 60 partite su 162 (ovvero, quelle precedenti al 4 giugno 2019) sono considerate già disputate e dunque non devono essere stimate. I valori di range, dunque, sono da considerarsi come riferiti esclusivamente alle circa 100 partite rimaste.

frequenza (su 1000 ripetizioni) di qualificazione alla post season, vittoria della propria division, vittoria delle Division Series, vittoria delle Championship Series e infine di trionfo nelle World Series. Queste sono assimilabili a stime delle probabilità di accadimento di ciascun evento.

Probabilità di: (%)	Qualificazione Post Season	Vittoria Division	Vittoria DS	Vittoria CS	Vittoria World Series
Orioles	0	0	0	0	0
Red Sox	67.4	8.4	19.0	9.5	4.7
Yankees	98.1	65.5	42.0	22.5	10.7
Rays	87.3	26.1	29.4	14.5	7.2
Blue Jays	0	0	0	0	0
White Sox	0.4	0.3	0.1	0	0
Indians	27.7	5.8	6	1.8	0.9
Tigers	0	0	0	0	0
Royals	0	0	0	0	0
Twins	96.8	93.9	46.5	23.2	12.1
Angels	6.9	0.2	1.1	0.5	0.2
Astros	99.6	99.0	53.8	26.9	13.9
Athletics	10.0	0.6	1.5	0.6	0.1
Mariners	0.1	0	0	0	0
Rangers	5.7	0.2	0.6	0.5	0.1
Braves	64.6	49.0	32.1	14.4	7.0
Marlins	0	0	0	0	0
Mets	17.7	8.7	3.5	0.7	0.2
Phillies	48.4	30.6	15.4	5.7	2.4
Nationals	24	11.7	5.0	1.9	0.7
Cubs	71.5	44.9	31.2	15.5	8.3
Reds	11.1	3.3	3.4	1.9	0.4
Brewers	62.8	34.8	21.9	8.4	2.5
Pirates	6.0	1.6	1.3	0.4	0
Cardinals	39.8	15.4	13.7	6.3	2.8
Diamondbacks	12.6	0.2	2.6	1.2	0.4
Rockies	17.8	0.5	2.4	0.8	0.3
Dodgers	100.0	99.2	64.9	42.0	24.8
Padres	21.5	0.1	2.4	0.8	0.3
Giants	2.2	0	0.2	0	0

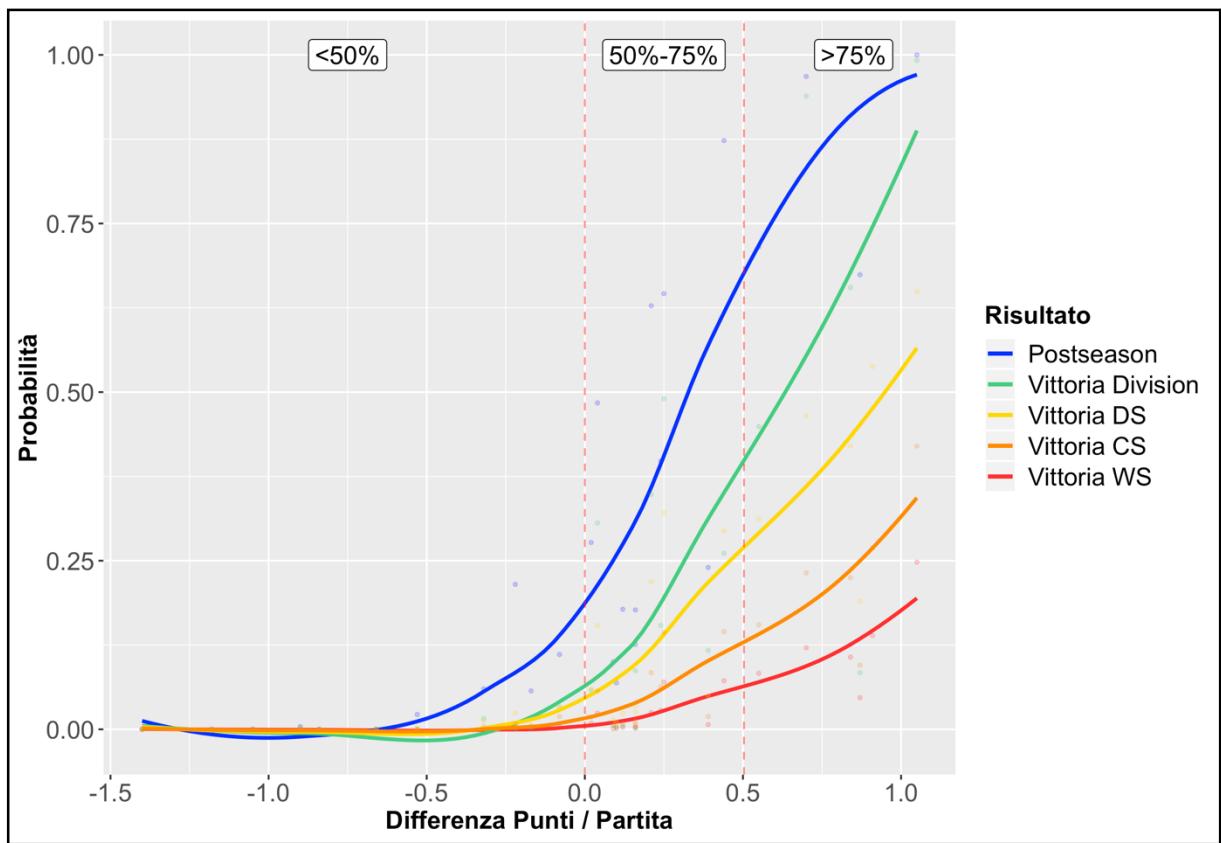
Tabella 4.4: Percentuali di frequenza di qualificazione alla post season, di vittoria della propria division, delle Division Series, delle Championship Series e delle World Series per ogni squadra di MLB ottenute simulando 1000 volte Regular Season e Postseason.

Come si può notare, le probabilità tendono a decrescere con l'avanzare delle fasi della Post Season. Infatti, le squadre più forti presentano un'elevata probabilità di qualificazione ai Play Off e di vittoria della propria division, ma una probabilità molto ridotta di vittoria finale delle World Series. In particolare, solo 4 (Yankees, Twins, Astros e Dodgers, evidenziate in rosso

nella tabella) sembrerebbero avere una probabilità superiore al 10% di aggiudicarsi il titolo, e di queste solo i Los Angeles Dodgers risultano superare il 20%. Inoltre, apparentemente solo 9 squadre su 30 non hanno alcuna possibilità di trionfo finale. Questo evidenzia come la vittoria delle World Series sia altamente imprevedibile a priori, e determinata più dal caso che dal talento della squadra. In altre parole, è probabile che a vincere il campionato non sia la *squadra migliore*, ed è possibile che sia una squadra con abilità poco superiore alla media.

Questo concetto è ben rappresentato nel grafico 4.2, che illustra come varia la probabilità di successo nelle diverse fasi di Post Season a seconda dell'abilità complessiva di squadra, in questo caso stimata tramite la RDg³¹ (*Differenza Punti per partita*). Le curve sono adattate ai punti tramite il metodo di regressione non parametrica *loess* implementato in R.

Grafico 4.2: Relazione fra talento della squadra (RDg) e probabilità di successo nelle varie fasi dei playoff, rappresentata tramite scatterplot e curve adattate con il metodo non parametrico loess. Le etichette <50%, 50%-75% e >75% sono riferite alla distribuzione di RDg, dunque squadre entro le due linee tratteggiate hanno un'abilità compresa fra la media e il terzo quartile, mentre quelle a destra si trovano nel top 25%.



³¹ Si è preferito usare la RDg anziché la RSg in quanto la prima è una misura più complessiva dell'abilità della squadra (comprende sia i punti segnati che quelli subiti).

4.6 Conclusioni

L’obiettivo dell’elaborato era quello di elaborare una tecnica di previsione dei risultati della Major League Baseball tramite la regressione logistica, e analizzare l’effetto del caso su di essi per mezzo della ripetizione di un numero elevato di simulazioni.

La prima fase è stata dedicata alla specificazione e alla valutazione di un modello che mettesse in relazione la probabilità di vittoria della squadra di casa con i valori di determinate covariate, utilizzate per stimare le abilità complessive e individuali delle due squadre. Il modello impiegante la *media punti segnati a partita* (RSg), l’*ERA dei lanciatori partenti* e l’*ERA complessiva dei lanciatori di rilievo* ha meglio soddisfatto i requisiti di significatività dei coefficienti, adattamento ai dati, capacità previsiva e semplicità interpretativa, ed è stato dunque selezionato per lo svolgimento della fase successiva. In generale, indicatori semplici sono risultati preferibili a variabili complesse in termini di effetto sull’outcome della partita.

La seconda fase è costituita nell’utilizzo del modello precedentemente specificato per scopi di previsione. Si sono utilizzati i dati finali delle proiezioni Steamer e delle *Depth Charts* di Fangraphs per la stima delle abilità di squadra e individuali da applicare ai coefficienti, insieme a una tecnica di campionamento proporzionale con determinati vincoli per la generazione della rotazione dei lanciatori partenti. Le probabilità così stimate sono poi state utilizzate come parametri per una serie di esperimenti di Bernoulli indipendenti, al fine di determinare i risultati delle singole partite. La simulazione è stata ripetuta 1000 volte e i risultati sono stati riportati sia direttamente che in forma grafica. La dispersione dovuta al caso del numero di vittorie per ogni squadra è risultata particolarmente elevata, sebbene fossero comunque presenti dei trend in funzione del talento. In particolare, il 95% delle osservazioni è risultato incluso in un intervallo di circa ± 10 partite rispetto alla media, ma in casi estremi si sono registrate differenze di natura aleatoria di oltre 15 partite.

Alla simulazione della Regular Season è seguita, per ogni stagione, la simulazione dei Play Off al fine di stimare le probabilità di vittoria finali per ogni squadra e decretare la relazione fra trionfo nelle World Series e talento della squadra. L’effetto del caso è risultato in questa circostanza ancora più incisivo: la probabilità maggiore osservata è apparsa pari al 24.8% (Los Angeles Dodgers), e solo 9 squadre su 30 hanno riportato probabilità nulle. Analizzando la relazione fra talento e probabilità di vittoria finale, si è osservato che la seconda sembra aumentare solo minimamente in funzione della prima. In altre parole, squadre molto forti risultano avere una probabilità molto alta di accedere ai Play Off, ma una probabilità di vittoria

delle World Series modesta e limitatamente superiore rispetto a squadre con un'abilità intorno alla media.

In conclusione, è possibile affermare che il caso ha un effetto decisamente rilevante sugli esiti della Major League Baseball, sia a livello della Regular Season che, in maniera ancora più marcata, della Post Season. Squadre con abilità superiori hanno certamente una tendenza a ottenere risultati migliori, ma è piuttosto comune che squadre sopra la media registrino stagioni scadenti e viceversa squadre mediocri ottengano risultati sorprendenti, arrivando addirittura al trionfo finale. Quando ciò accade, è frequente che giornalisti sportivi e tifosi provino a fornire una spiegazione logica dell'accaduto: la verità è che molto spesso questa motivazione è tutt'altro che razionale, ma essenzialmente da ricollegarsi alla fortuna o alla sfortuna. Giocatori che mettono a segno prestazioni al di sopra o al di sotto delle loro abilità, chiamate arbitrali discutibili, rimbalzi sbagliati della pallina, errori nella corsa sulle basi, qualità di determinati lanci, condizioni meteorologiche avverse o addirittura interferenze del pubblico sono tutti esempi di eventi casuali di cui è impossibile tenere conto a priori, ma che possono influenzare fortemente il risultato finale di una partita.

Yogi Berra non aveva tutti i torti: fare previsioni è difficile, e lo è ancora di più in un ambito particolarmente soggetto all'imponderabilità come il baseball. Ma in fondo, è proprio questa imponderabilità che contribuisce a renderlo uno sport affascinante.

Bibliografia

Albert, J., & Bennett, J. (2003). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. New York, NY: Springer New York.

Albert, J. (2017). *Teaching Statistics Using Baseball*. Washington, DC: The Mathematical Association of America.

Gavin C. Cawley and Nicola L.C. Talbot. 2010. *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation*. J. Mach. Learn. Res. 11 (August 2010), 2079-2107.

Click, J. (2004, October 14). *Starters vs. Relievers: The Changing Distribution of Pitching Performance*.

Retrieved July 2, 2019, from <https://www.baseballprospectus.com/news/article/3561/starters-vs-relievers-the-changing-distribution-of-pitching-performance/>

FanGraphs Baseball | Baseball Statistics and Analysis. Retrieved July 2, 2019, from <https://www.fangraphs.com/>

Harrel, F. (2016). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York, NY: Springer New York.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken, NJ: Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*. New York: Springer.

Kassambara, A. (2018). *Machine learning essentials*. Frankreich: STHDA.

Lahman, S. (2018). *Lahman's Baseball Database*. Retrieved July 2, 2019, from <http://www.seanlahman.com/baseball-archive/statistics>

Marchi, M., & Albert, J. (2014). *Analyzing Baseball Data with R*. Boca Raton: CRC Press.

MLB Official Playing Rules Committee (2019). *Official Baseball Rules: 2019 Edition*. Retrieved July 2, 2019, from:
content.mlb.com/documents/2/2/4/305750224/2019_Official_Baseball_Rules_FINAL_.pdf

Retrosheet Game Logs. (2018). Retrieved July 2, 2019, from <https://www.retrosheet.org/>

Routledge, R. (2016, October 12). Law of large numbers. Retrieved from
<https://www.britannica.com/science/law-of-large-numbers>

Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2014). *The book: Playing the Percentages in Baseball*. North Charleston, SC: CreateSpace.

What is a Steamer? | Glossary. (Major League Baseball). Retrieved July 2, 2019, from
<http://m.mlb.com/glossary/projection-systems/steamer>

Zhang, Zhongheng. (2016). *Residuals and Regression Diagnostics: Focusing on Logistic Regression*. Annals of Translational Medicine. 4. 195-195. 10.21037/atm.2016.03.36.