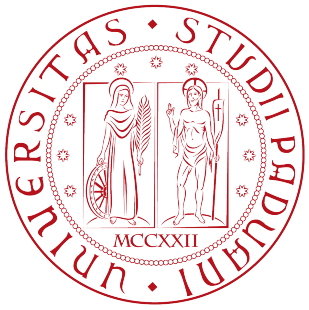
Università degli studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche



Analisi delle partite di Premier League: previsione dei risultati con un approccio bayesiano

Relatore: Prof. Bernardi Mauro
Correlatrice: Prof.ssa Cattelan Manuela
Dipartimento di Scienze Statistiche

Laureando: Mazzarolo Simon
Matricola N. 1241997

Anno Accademico 2021/2022

Indice

Introduzione

Il presente lavoro si pone l'obiettivo di introdurre una metodologia poco vista nell'ambito della previsione dei risultati nel calcio, che cerca di sfruttare delle variabili che non sono disponibili prima dello svolgimento della gara che si vuole prevedere, utilizzando i modelli bayesiani per questo obiettivo.

In letteratura, la previsione dei risultati nel calcio è stata trattata tramite l'utilizzo di variabili disponibili prima dell'inizio della gara, come per esempio il sistema di valutazione *Elo*[1], o il numero di goal segnato dalle squadre negli ultimi match. Per questo tipo di analisi si veda Aslan e Inceoglu (2007) oppure Carpita et al. (2019) per ulteriori dettagli. Altri approcci utilizzano, oltre a queste variabili appena citate, anche altre informazioni derivanti dalle quote delle scommesse, le quali si rilevano essere molto utili e dettagliate. In questo senso si può vedere il lavoro di Wunderlich e Memmert (2018) che sfrutta proprio queste informazioni.

Con lo scopo di presentare questa metodologia, il presente lavoro è suddiviso in quattro parti così distinte.

Nel primo capitolo viene presentato il *dataset* utilizzato per l'analisi, contenente le variabili che si utilizzeranno per la previsione. In particolare si effettueranno delle previsioni per dei valori futuri di queste variabili trattandole come delle serie storiche cercando quindi di modellare la loro dipendenza temporale.

[1] Si tratta di un sistema di valutazione nato per il gioco degli scacchi e attualmente molto usato, che si basa sui risultati delle partite giocate in precedenza.

Nel secondo capitolo viene introdotto il modello logistico bayesiano ed in particolare l'approccio di *data augmentation* tramite variabili latenti *Pòlya-Gamma* come proposto da Polson et al. (2013). Questo approccio permette di simulare catene di Markov che hanno come distribuzione limite la distribuzione a posteriori dei parametri tramite l'utilizzo del *Gibbs sampler*. In questo capitolo si vedrà inoltre una prima applicazione ai dati, utilizzando la variabile risposta del risultato come binaria.
Nel terzo capitolo è stata descritta l'estensione multinomiale dello stesso modello bayesiano presentanto da Polson et al. (2013), che mantiene l'approccio con le variabili latenti *Pòlya-Gamma*. Anche con questa estensione, mantenendo questa metodologia è possibile costruire un *Gibbs sampler* per simulare dei valori dalla distribuzione a posteriori dei parametri.
Infine nel quarto capitolo vengono effettuate le previsioni dei risultati e vengono discussi i risultati ottenuti sulla base dei modelli stimati nel capitolo precedente.
Seguono poi alcune considerazioni finali sulle analisi svolte e gli sviluppi futuri perseguibili.

Capitolo 1

Analisi esplorativa

I dati qui analizzati sono stati presi da *Kaggle*[1], una piattaforma *online* che mette a disposizioni svariati *dataset* con i quali si possono anche fare competizioni, e riguardano la *Premier League*, cioè il massimo campionato inglese, per le stagioni sportive che vanno dal 2010-11 al 2019-20.
Nel *dataset* analizzato sono presenti, tra le altre, le seguenti variabili:

- `Season`: stagione a cui fa riferimento la gara;
- `Date`: data della gara;
- `Home_team`: nome della squadra di casa;
- `Away_team`: nome della squadra ospite;
- `Result_full`: risultato finale della gara;
- `Goal_home_ft`: reti segnate dalla squadra di casa a fine gara;
- `Goal_away_ft`: reti segnate dalla squadra ospite a fine gara;
- `Sg_match_ft`: differenza tra le reti segnate dalla squadra di casa e quella ospite a fine gara.

[1] https://www.kaggle.com/pablohfreitas/all-premier-league-matches-20102021

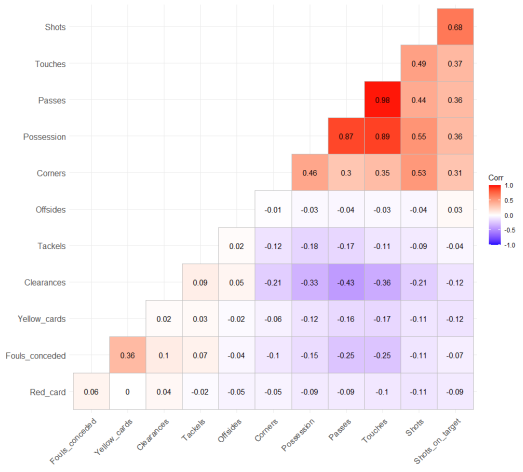
Oltre a queste appena citate, sono presenti una serie di variabili che saranno utilizzate per la modellazione nei prossimi capitoli. Queste sono riportate nel *dataset* due volte, in quanto vengono rilevate sia per la squadra che gioca la partita in casa che per quella ospite. Queste sono:

- `Clearances`: numero di volte che i calciatori della squadra calciano il pallone lontano dalla propria porta come azione difensiva;
- `Corners`: numero di calci d'angolo della gara;
- `Fouls_conceded`: numero di falli concessi dalla squadra;
- `Offsides`: numero di fuorigioco per la squadra;
- `Passes`: numero di passaggi effettuati dalla squadra;
- `Possession`: percentuale di possesso della squadra;
- `Red_cars`: numero di cartellini rossi ricevuti;
- `Shots`: numeri di conclusioni verso la porta avversaria;
- `Shots_on_target`: numero di conclusioni nello specchio della porta avversaria;
- `Tackles`: numero di contrasti in scivolata effettuati dalla squadra;
- `Touches`: numero di tocchi del pallone nella gara;
- `Yellow_cards`: numero di cartellini gialli ricevuti.

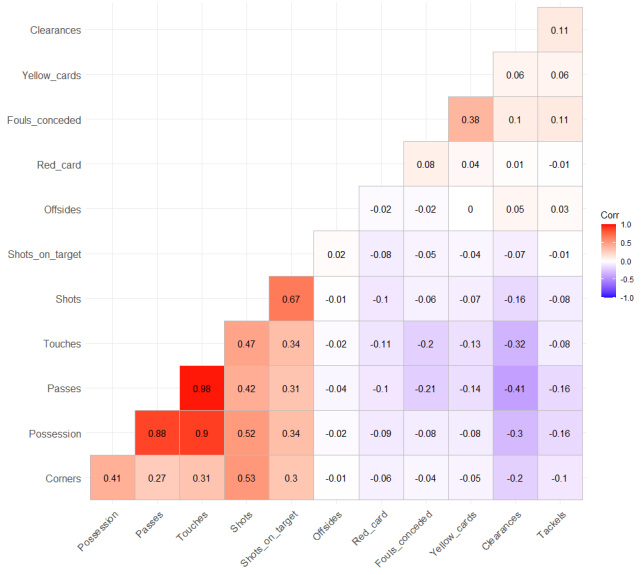
Come prima cosa si effettua un'analisi delle correlazioni delle variabili che saranno usate come esplicative nel modello. Come si può vedere in Figura 1.1 le variabili maggiormente tra loro correlate sono `Possession`, `Touches` e `Passes`, e a livello intuitivo potrebbero effettivamente rappresentare una informazione molto simile. Data la loro natura è facile aspettarsi che la

squadra che fa più passaggi e tocchi del pallone faccia anche più possesso palla.

Queste supposizioni derivanti dalla fase esplorativa andranno poi confermate con la modellazione, valutando quindi quali variabili togliere e quali tenere per migliorare l'adattamento del modello ai dati.



(a) Correlazioni tra le variabili relative alla squadra di casa.



(b) Correlazioni tra le variabili relative alla squadra ospite.

Figura 1.1: Grafico delle correlazioni tra variabili esplicative.

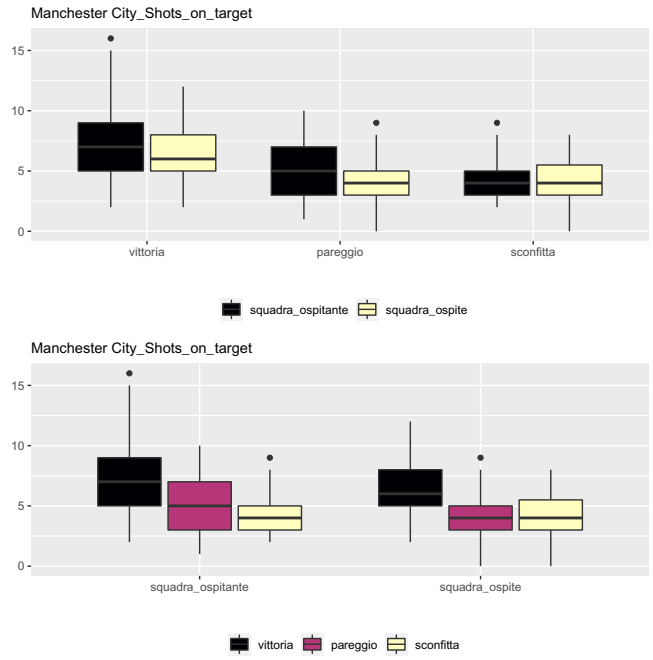


Figura 1.2: Numero di tiri in porta della squadra *Manchester City*, stratificato per casa/fuori casa e per le modalità della variabile risposta.

Fatto ciò, si è poi passati ad analizzare le variabili nello specifico, guardando, almeno in fase esplorativa, se fosse presente un effetto dovuto al disputare le gare in casa o fuori casa, oppure una dipendenza con le modalità della variabile risposta.

Per prima cosa si sono guardate le squadre che hanno disputato tutte e dieci le stagioni di *Premier League* presenti nel *dataset*, cioè le formazioni mai retrocesse e che quindi, intuitivamente, sono state tra le più forti del campionato in questo decennio. Nello specifico queste squadre sono: *Arsenal*, *Chelsea*, *Everton*, *Liverpool*, *Manchester City*, *Manchester United* e *Tottenham Hotspur*.

A titolo esemplificativo si riportano qui alcuni esempi per qualche variabile relativi alla squadra *Manchester City*, per altri grafici si vedano le figure riportate in Appendice.

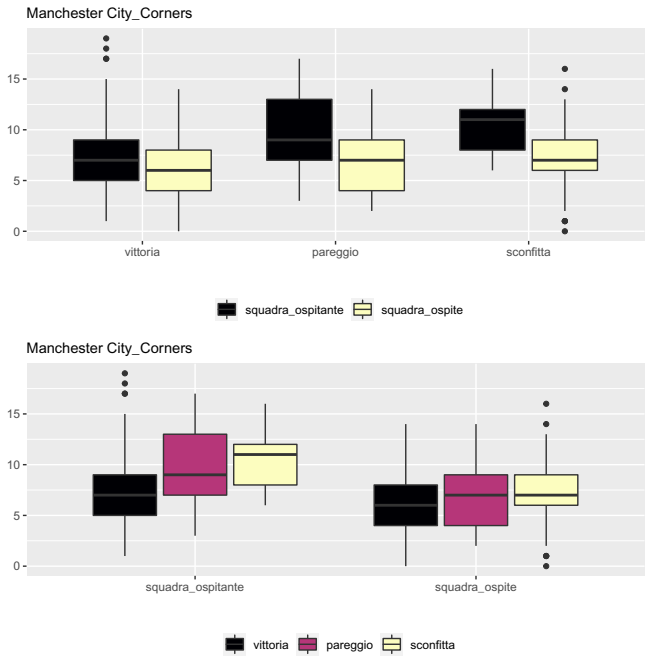


Figura 1.3: Numero di calci d'angolo della squadra *Manchester City*, stratificato per casa/fuori casa e per le modalità della variabile risposta.

Come si può vedere in Figura 1.2, per la squadra *Manchester City* sembra esserci un effetto relativo all'essere in casa o fuori casa per quanto riguarda il numero di tiri in porta. Si nota che in media la squadra ha tirato in porta di più nelle partite casalinghe. Si può fare un ragionamento analogo anche volendo analizzare l'effetto legato alle modalità della variabile risposta: quando il *Manchester City* ha vinto, mediamente ha tirato di più in porta, mentre quando ha perso ha tirato, nello specchio della porta, di meno.

Allo stesso modo in Figura 1.3 si può vedere come nei match casalinghi ci siano più calci d'angolo per la squadra ospitante, come ad indicare che ci sia una spinta maggiore da parte del pubblico di casa nell'attaccare la squadra avversaria. Guardando alla variabile risposta, si nota che mediamente quando la squadra vince effettua meno calci d'angolo, che si può interpretare in questo modo: in una situazione di vantaggio, la squadra in vantaggio ha una

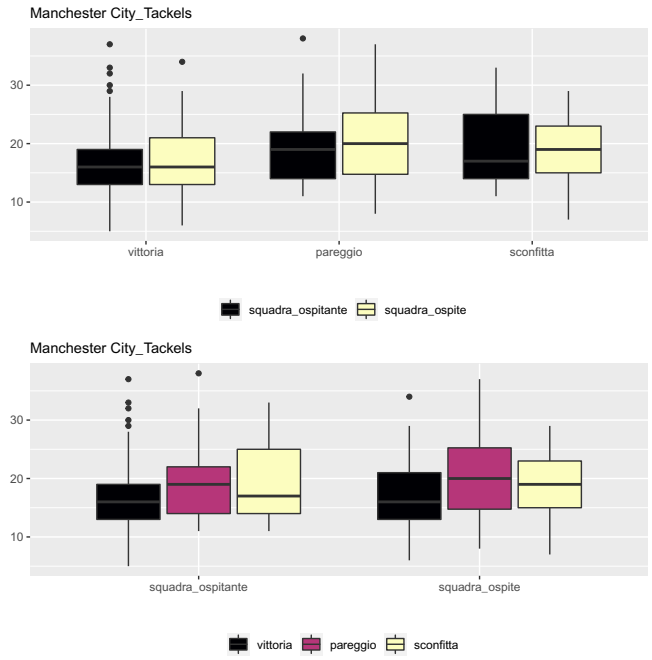


Figura 1.4: Numero di scivolate effettuate dalla squadra *Manchester City*, stratificato per casa/fuori casa e per le modalità della variabile risposta.

tendenza a difendersi anziché continuare ad attaccare.

Di contro, alcune variabili non sembrano essere influenzate né dal risultato della partita, né dall'essere in casa o fuori casa. Infatti come si vede in Figura 1.4, la squadra *Manchester City* effettua in media lo stesso numero di scivolate, indipendentemente da tutto il resto.

Per fare un confronto con le squadre che, invece, si possono considerare minori, cioè quelle che nel decennio 2010-2020 hanno disputato poche volte la massima categoria inglese, si decide di esaminare la squadra *Sheffield United*, che ha disputato il massimo campionato inglese nella sola stagione 2019-20. Questa parte di analisi ha lo scopo di verificare se, almeno in prima battuta, c'è una differenza tra squadre più e meno forti del campionato in termini di interpretazione delle variabili esplicative, oppure se queste si comportano alla stessa maniera.

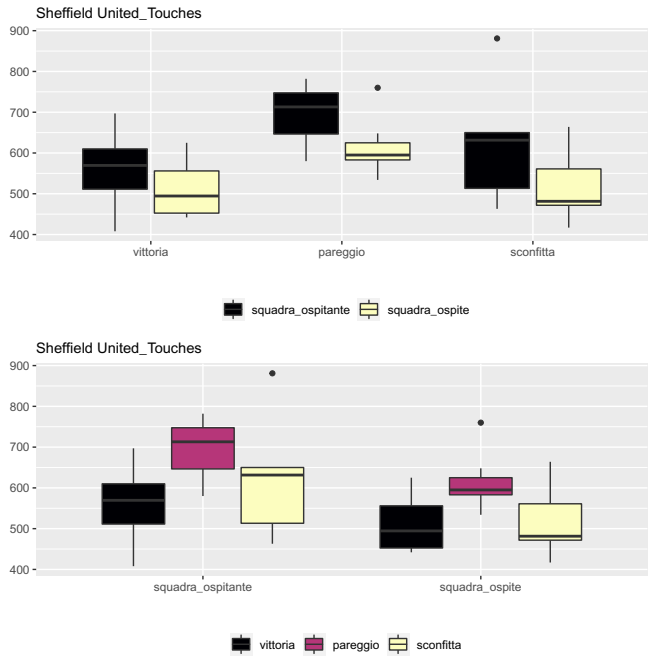


Figura 1.5: Numero di tocchi del pallone della squadra *Manchester City*, stratificato per casa/fuori casa e per le modalità della variabile risposta.

Come si vede in Figura 1.5, la squadra mediamente fa più tocchi, e quindi più possesso, nelle gare in casa, come a voler dominare il gioco in queste partite. Contestualmente si riscontra che mediamente lo *Sheffield United* fa più possesso nelle partite che pareggia rispetto a quelle dove vince o perde. Questo si può interpretare, almeno in prima analisi, come una tendenza ad accontentarsi del risultato quando pareggia e a subire di più gli avversari quando sta vincendo oppure perdendo. Essendo questa una squadra che ha giocato una singola stagione di *Premier League*, è possibile aspettarsi che le partite nelle quali ha vinto abbia toccato meno palloni perché ha poi subito il gioco avversario che cercava di recuperare la partita. Allo stesso modo però, si rileva che, in partite nelle quali ha perso ha subito maggiormente il gioco avversario.

Per concludere la fase esplorativa, si passa ad un'analisi della dipendenza

temporale al variare delle diverse stagioni e si inizia prendendo a riferimento il *Liverpool*. Come si vede in Figura 1.6, il numero di passaggi medi che effettua questa squadra cambia al variare delle stagioni sportive, e in particolare negli ultimi anni, quando la squadra è tornata a lottare per il titolo (poi vinto nella stagione 19/20), questo è aumentato considerevolmente. Altra nota interessante che si evince dal grafico è l'effettiva importanza del fattore campo, in quanto negli ultimi anni il *Liverpool* in casa non ha mai perso, concedendo solo qualche sporadico pareggio (addirittura solo uno nella stagione del titolo).

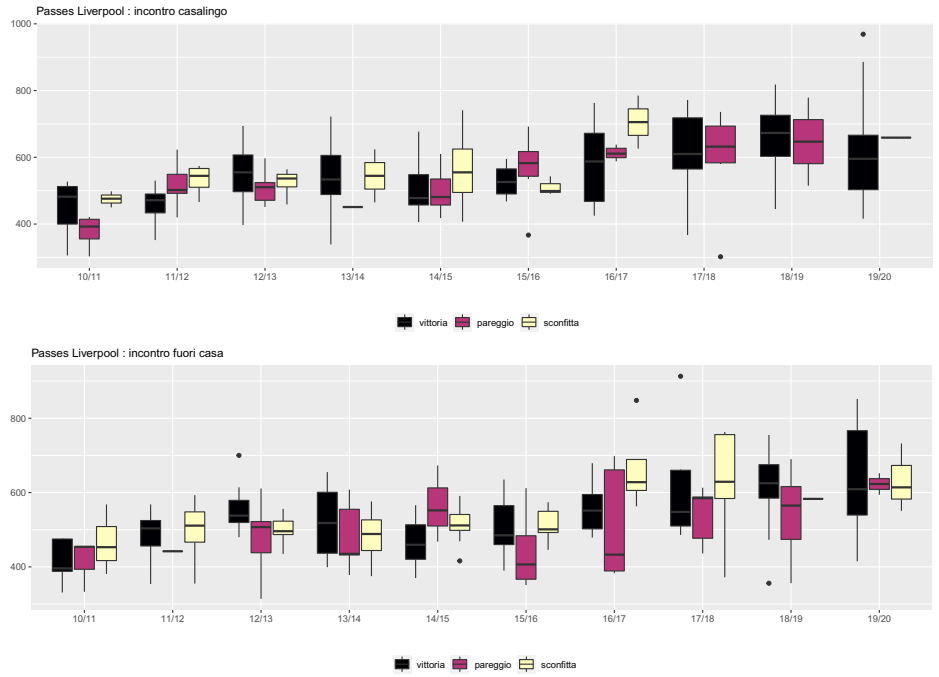


Figura 1.6: Numero di passaggi effettuati dalla squadra *Liverpool*, al variare delle stagioni.

Guardando invece la Figura 1.7, si può notare come sia presente un *trend* decrescente per quanto riguarda il numero di volte che il *Chelsea* allontana il pallone dalla propria area di rigore per difendersi, evidenziando uno spirito più volto all'attacco come si può evincere dal fatto che negli ultimi anni

questa squadra sia tornata a competere per le prime posizioni in classifica. Anche in questo caso si può notare l'importanza del fattore campo, infatti si può notare come in casa ci sia una presenza nettamente maggiore della frequenza relativa alla vittoria rispetto alle altre due. Questo si può notare in tutto l'arco temporale, e in particolare nelle stagioni 2014-15 e 2016-17, quelle nelle quali il *Chelsea* ha vinto il titolo, le modalità che si sono verificate in casa sono solo due su tre.

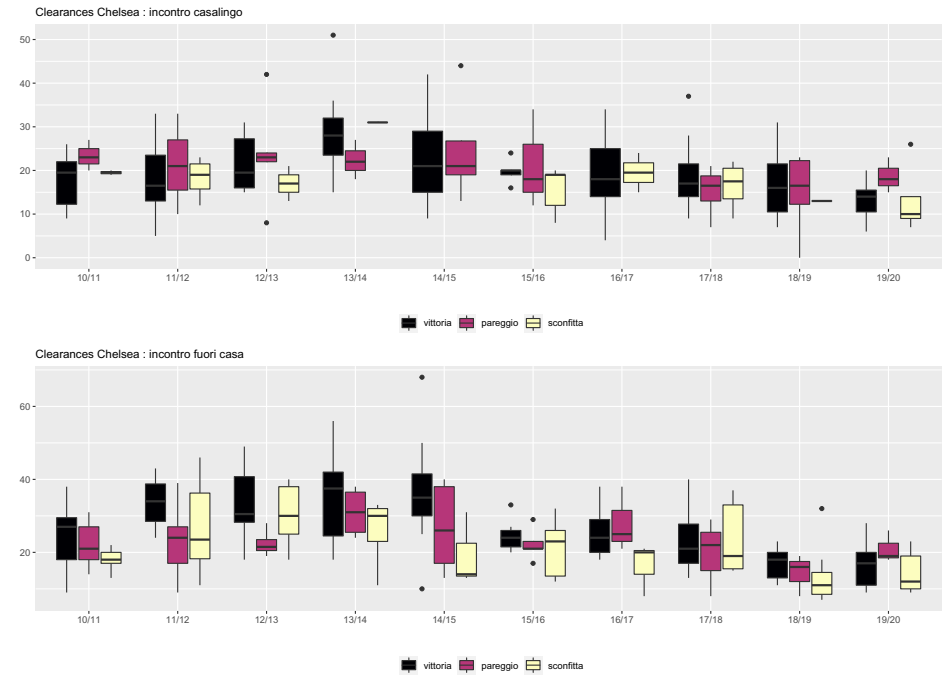


Figura 1.7: Numero di *clearences* effettuate dalla squadra *Chelsea*, al variare delle stagioni.

1.1 Previsione di variabili esplicative

L'obiettivo dell'analisi sarà quella di prevedere i risultati di un'intera stagione sportiva prima che questa sia effettivamente giocata. Per questo motivo si decide di analizzare le covariate per ogni squadra come se fossero delle serie storiche e fare poi le previsioni di quella variabile trentotto passi in avanti, cioè il numero di giornate di una stagione sportiva. È da notare che queste variabili esplicative sono osservate solo dopo che la gara si è conclusa e non sono disponibili a priori. Per ovviare a ciò, si tratta ciascuna variabile esplicativa con un modello `ARIMA` mettendo assieme tutte le osservazioni disponibili per ogni squadra senza considerare la suddivisione in singole stagioni, con il quale poi si faranno le previsioni necessarie per la stagione oggetto di studio. A titolo esemplificativo si riporta il grafico di autocorrelazione e autocorrelazione parziale di una squadra, in questo caso l'*Arsenal*, per fare un esempio di analisi proposta. In generale per tutte le squadre ci si affiderà al comando di `R`, `auto.arima()` che stima in automatico il modello `ARIMA` migliore sulla base dei criteri di informazione, quali `AIC, BIC e AICC`, confrontato i modelli stimati sulla base di tutti i *lag* che si decide di testare. Fatto ciò si si procede alle previsioni.

Analizzando la serie storica dei passaggi effettuati durante la partita della squadra *Arsenal*, riportata in Figura 1.8, si nota subito una non stazionarietà della serie, e analizzando i grafici di autocorrelazione e autocorrelazione parziale della differenza prima della stessa, Figura 1.9, si può notare la presenza di autocorrelazione per diversi *lag* significativamente diversi da zero. In questo caso, è ragionevole usare un modello `ARIMA(2,1,1)`, che cattura la non stazionarietà e per la serie, una relazione di ordine due per la parte autoregressiva del modello e di ordine uno per la parte a media mobile, in base a quanto visto nei grafici. In linea con questa analisi, la funzione `auto.arima()` di `R` propone di usare questo modello che si conferma anche il migliore in termini di criteri di informazione.

Figura 1.8: Serie storica di `Passes` della squadra *Arsenal* e valori stimati modello `ARIMA(2,1,1)`.

Con la linea blu della Figura 1.8 si possono vedere i valori stimati dal modello appena citato.
Procedendo con la selezione del modello `ARIMA` più adeguato per ciascuna variabile esplicativa di ciascuna squadra presente nella stagione da prevedere, si ottengono i risultati presenti in Tabella 1.1. Si può notare come alcune variabili indipendenti vengano modellate direttamente con la media (tramite un modello `ARIMA(0,0,0)`) e altri abbiano una parte autoregressiva o a media mobile da modellare. Come visto nell'esempio appena trattato, in alcuni casi sono presenti anche effetti di non stazionarietà.

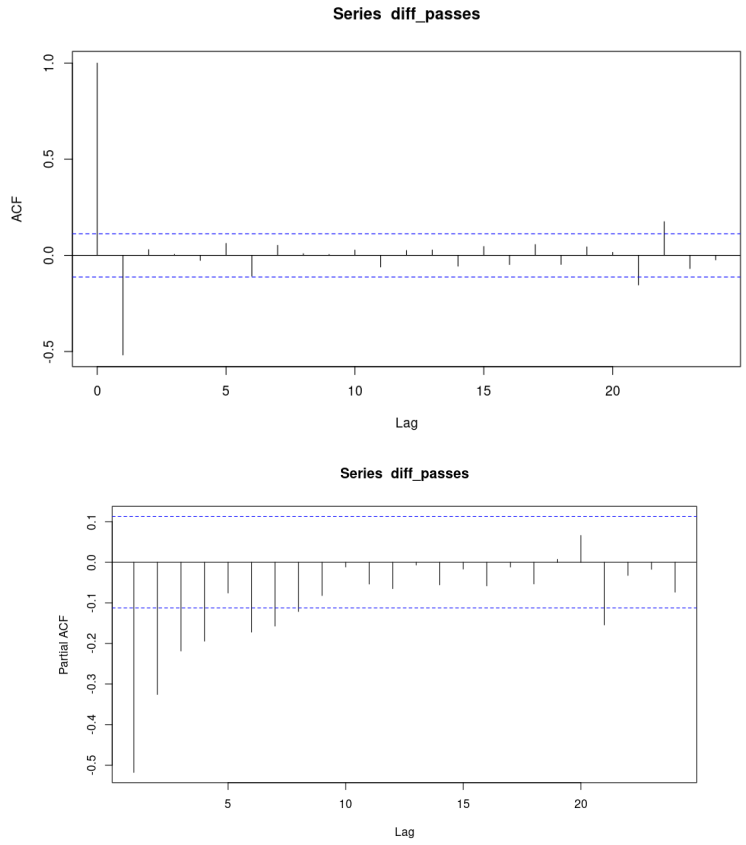


Figura 1.9: Grafico di autocorrelazione (sopra) e autocorrelazione parziale (sotto) dell serie storica di `Passes` della squadra *Arsenal*.

Tabella 1.1: Ordine `(p,d,q)` dei modelli `ARIMA` stimati per ciascuna variabile esplicativa di ogni squadra.

	clearances	corners	fouls conceded	offsides	passes	possession	red c.	shots	shots_target	tackels	touches	yellow c.
AFC Bournemouth	(1, 0, 1)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(0, 1, 1)	(0, 0, 0)
Arsenal	(6, 0, 0)	(1, 0, 0)	(1, 1, 2)	(1, 0, 1)	(2, 1, 1)	(0, 0, 0)	(1, 0, 0)	(2, 0, 0)	(1, 0, 2)	(3, 1, 1)	(0, 0, 0)	(0, 0, 0)
Brighton and Hove Albion	(1, 0, 0)	(3, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(0, 0, 0)	(0, 0, 0)	(1, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(0, 0, 1)
Burnley	(0, 1, 1)	(0, 0, 0)	(0, 0, 0)	(1, 0, 0)	(0, 0, 0)	(0, 0, 0)	(2, 0, 0)	(0, 0, 0)	(2, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)
Cardiff City	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(1, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(1, 0, 0)	(1, 0, 0)	(1, 0, 0)	(0, 0, 0)	(0, 0, 0)
Chelsea	(1, 1, 1)	(0, 0, 0)	(0, 1, 1)	(0, 0, 0)	(1, 1, 2)	(1, 0, 2)	(0, 0, 0)	(2, 1, 1)	(1, 1, 2)	(2, 1, 2)	(1, 1, 2)	(0, 0, 0)
Crystal Palace	(1, 1, 1)	(0, 0, 0)	(1, 0, 1)	(2, 0, 2)	(0, 1, 1)	(0, 1, 1)	(0, 1, 1)	(0, 0, 0)	(0, 0, 0)	(2, 1, 2)	(0, 1, 1)	(0, 0, 0)
Everton	(0, 0, 0)	(0, 1, 1)	(2, 1, 1)	(0, 1, 1)	(0, 1, 1)	(1, 1, 2)	(0, 0, 0)	(1, 1, 1)	(0, 1, 2)	(1, 0, 1)	(1, 1, 1)	(2, 0, 2)
Fulham	(2, 0, 2)	(0, 0, 0)	(0, 1, 1)	(0, 0, 0)	(0, 1, 1)	(0, 1, 1)	(0, 0, 0)	(2, 1, 1)	(0, 1, 2)	(1, 0, 2)	(0, 1, 1)	(1, 0, 0)
Huddersfield Town	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)
Leicester City	(0, 1, 1)	(0, 0, 0)	(1, 1, 1)	(1, 1, 1)	(0, 1, 1)	(0, 1, 1)	(2, 0, 2)	(0, 0, 0)	(0, 0, 0)	(2, 1, 1)	(0, 0, 2)	(0, 0, 0)
Liverpool	(0, 1, 1)	(0, 0, 0)	(1, 1, 1)	(1, 1, 2)	(1, 1, 2)	(1, 1, 1)	(1, 0, 0)	(1, 0, 0)	(0, 0, 0)	(1, 1, 2)	(2, 1, 1)	(0, 0, 0)
Manchester City	(0, 1, 1)	(0, 0, 0)	(1, 1, 1)	(0, 0, 1)	(1, 1, 2)	(1, 1, 1)	(0, 0, 0)	(2, 0, 1)	(1, 0, 1)	(4, 1, 3)	(1, 1, 1)	(0, 0, 0)
Manchester United	(1, 1, 1)	(0, 0, 0)	(0, 1, 1)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(1, 1, 2)	(3, 1, 1)	(0, 0, 0)	(0, 1, 1)
Newcastle United	(2, 0, 2)	(0, 0, 0)	(3, 0, 1)	(1, 1, 2)	(0, 0, 1)	(2, 1, 1)	(0, 0, 2)	(2, 1, 1)	(2, 0, 0)	(1, 0, 0)	(3, 0, 2)	(1, 1, 2)
Southampton	(0, 1, 1)	(1, 0, 4)	(0, 0, 0)	(0, 0, 0)	(2, 0, 1)	(1, 0, 1)	(0, 0, 1)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(1, 1, 1)	(0, 0, 0)
Tottenham Hotspur	(2, 1, 1)	(4, 0, 0)	(0, 0, 0)	(0, 0, 1)	(1, 1, 1)	(1, 1, 1)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(1, 1, 1)	(1, 1, 1)	(1, 1, 2)
Watford	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(0, 0, 0)	(4, 1, 1)	(0, 0, 1)	(0, 0, 3)	(0, 0, 0)	(3, 1, 1)	(0, 0, 0)	(0, 0, 0)
West Ham United	(0, 1, 2)	(0, 0, 0)	(1, 1, 3)	(2, 0, 1)	(1, 1, 1)	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(1, 0, 1)	(0, 1, 1)	(1, 1, 1)	(0, 0, 0)
Wolverhampton Wanderers	(0, 0, 0)	(0, 0, 0)	(0, 1, 1)	(0, 0, 0)	(1, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(2, 0, 2)	(0, 0, 1)

Capitolo 2

Modello logistico bayesiano

2.1 Inferenza bayesiana

In questo approccio statistico la probabilità è intesa come una valutazione dell'incertezza che riguarda sia quantità osservabili come i dati $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{Y}$ con $\mathcal{Y}$ spazio campionario, sia quantità non osservabili come il parametro che regola il modello $\boldsymbol{\theta} \in \Theta$, con Θ spazio parametrico.
Con questa formulazione si assume che il vero e ignoto valore del parametro è una realizzazione di una distribuzione di probabilità sullo spazio parametrico Θ, con densità a-priori $\pi(\boldsymbol{\theta})$, che rappresenta un riassunto dell'informazione preliminare su $\boldsymbol{\theta}$. Nella scelta della distribuzione a-priori ci sono sostanzialmente due possibili strade perseguibili:

- distribuzioni a-priori non informative: sono distribuzioni praticamente piatte sullo spazio parametrico, che assegnano simile probabilità a tutti i punti e rappresentano assenza di informazione a-priori sul parametro, lasciando quindi la possibilità ai dati di "parlare" influenzando la distribuzione a-posteriori (di cui si discuterà più avanti);

- distribuzioni a-priori informative: al contrario sono invece distribuzioni concentrate in qualche valore a seguito di informazioni che si pensa

avere a disposizione sul parametro di interesse. Di conseguenza questo tipo di distribuzioni a-priori lasciano meno spazio all'informazione apportata dai dati;

Una scelta che è possibile fare per le a-priori in generale sono le distribuzioni coniugate. Queste distribuzioni sono utili perché matematicamente convenienti, in quanto lasciano inalterata la forma funzionale della distribuzione a-posteriori rispetto a quella a-priori. In altre parole con una distribuzione coniugata la verosimiglianza rappresenta per la distribuzione a-posteriori un aggiornamento dei parametri ipotizzati a priori.

Nella specificazione bayesiana, indichiamo con $p(\mathbf{y} \mid \boldsymbol{\theta})$ la distribuzione del vettore di dati $\mathbf{y} = (y_1, \ldots, y_n)$ condizionatamente al parametro del modello. Chiaramente, il campione casuale semplice di dati, una volta osservati, sono dei valori fissati, quindi definiamo la verosimiglianza come funzione esclusivamente del parametro:

$$L(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}). \tag{2.1}$$

Lo scopo ultimo dell'inferenza bayesiana è quello di aggiornare la conoscenza a-priori sul parametro alla luce dei dati osservati, e per fare ciò si ricorre al teorema di *Bayes*, grazie al quale si ottiene la distribuzione a-posteriori:

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{2.2}$$

Di fatto il denominatore è una costante di normalizzazione, perché non dipende dal parametro, e rende la distribuzione a-posteriori proporzionale al prodotto tra verosimiglianza (nella quale entrano i dati osservati) e distribuzione a-priori:

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta})\pi(\theta). \tag{2.3}$$

La difficoltà della statistica bayesiana è proprio il calcolo dell'integrale della costante di normalizzazione, cioè il denominatore della formula (2.2), che fatto salvo il caso di distribuzioni a-priori coniugate, non è sempre disponibile in forma esplicita, soprattutto per elevate dimensioni del parametro di interesse. Questa costante di normalizzazione è importante perché permette di calcolare le statistiche descrittive (come la media e la mediana a-posteriori) e gli indici di posizione (per calcolare gli intervalli di credibilità) che rappresentano l'informazione che si vuole ottenere e che è racchiusa nella distribuzione a-posteriori.

2.2 Markov Chain Monte Carlo (MCMC)

La difficoltà del calcolo della costante di normalizzazione, può essere in parte superato grazie ai metodi *Monte Carlo* applicati alla generazione con Catene di Markov.
I metodi Monte Carlo permettono l'approssimazione di integrali del tipo:

$$\psi = E_g[f(y)] = \int_{\mathcal{Y}} f(y)g(y)dy \tag{2.4}$$

tramite l'utilizzo di un campione casuale semplice di numerosità R dalla distribuzione generatrice dei dati e usando come stima la relativa media empirica:

$$\hat{\psi} = \frac{1}{R}\sum_{r=1}^{R} f(Y_r)$$

che converge in probabilità a ψ per la legge debole dei grandi numeri.
Generare valori in modo esatto dalla distribuzione d'interesse non è sempre possibile o facilmente perseguibile, e quindi si può ovviare al problema usando delle approssimazioni tramite Catene di Markov. Gli algoritmi di MCMC permettono di generare valori tramite una Catena di Markov ergodica e invariante che ha come distribuzione limite una distribuzione di interesse, nel nostro caso la a-posteriori in equazione 2.3. Lo svantaggio di questo

approccio è che, data la dipendenza temporale all'ultimo stato delle Catene di Markov, i valori generati saranno tra loro correlati, non rappresentando più un campione *i.i.d* come sarebbe desiderabile. Di conseguenza i valori conterranno meno informazione rispetto ad un campione casuale semplice, a parità di numerosità campionaria, nonostante la distribuzione marginale dei dati sia quella limite di interesse. Sarà necessario di conseguenza generare catene più lunghe. Per superare il problema derivante dalla correlazione dei dati si può applicare un filtraggio alla serie, per esempio prendendo solo un valore ogni dieci generati.

Resta ora da capire come costruire una catena che abbia la distribuzione limite corretta per generare il campione dalla a-posteriori.

2.2.1 Algoritmo Metropolis - Hastings

L'algoritmo di *Metropolis - Hastings* (1953-1970) serve per generare catene di Markov che abbiano come distribuzione limite quella di interesse. Fissati gli stati della catena, in questo specifico caso saranno i punti contenuti nello spazio parametrico Θ, si vuole costruire un algoritmo che partendo da uno stato iniziale esplori gli altri avendo una certa probabilità di spostarsi o meno in un nuovo stato. In altre parole, si parte dallo stato $\boldsymbol{\theta}$ della catena, si propone di muoversi nello stato $\boldsymbol{\theta}^*$, generato da una densità scelta $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ (dalla quale si sa generare), e accetto di muovermi in quello stato con probabilità $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. La probabilità di accettazione verrà scelta in modo tale che la sua distribuzione limite sia proprio $\pi(\boldsymbol{\theta} \mid \mathbf{y})$.

Tramite la proprietà di reversibilità della catena (necessaria affinché si arrivi a convergenza):

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \pi(\boldsymbol{\theta}^* \mid \boldsymbol{y}) q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*) \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \tag{2.5}$$

si possono definire le probabilità di accettazione come segue:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^* \mid \boldsymbol{y}) q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{y})}\right\}. \tag{2.6}$$

Basterà iterare per R volte i passi dell'algoritmo per generare i valori dalla catena ergodica con distribuzione $\pi(\boldsymbol{\theta} \mid \mathbf{y})$.

Dato che la catena parte da uno stato iniziale scelto a priori, prima di arrivare a convergenza saranno necessarie delle iterazioni di *burm-in*, nelle quali ci si avvicinerà sempre più alla vera distribuzione. Superata questa fase si considereranno i valori generati come provenienti dalla vera distribuzione limite e verranno scartati i primi di *burn-in*.

2.2.2 Gibbs sampler

Il *Gibbs sampler* è una variante dell'algoritmo *Metropolis-Hastings* nel caso multivariato, infatti si può fare una particolare scelta per le generazioni proposte la cui probabilità di accettazione è pari a uno.

Data la distribuzione $p(\boldsymbol{\theta})$ del parametro p-dimensionale di interesse $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, se si conosce la distribuzione condizionata per la j-esima componente θ_j di $\boldsymbol{\theta}$: $p(\theta_j \mid \boldsymbol{\theta}_{(j)})$, dove $\boldsymbol{\theta}_{(j)}$ indica il vettore $\boldsymbol{\theta}$ senza l'elemento in posizione j, si può usare questa come densità di proposta.

$$q(\theta_j^* \mid \boldsymbol{\theta}) = p(\theta_j^* \mid \boldsymbol{\theta}_{(j)}).$$

Dovendo quindi generare R valori, ad ogni passo si procederà a generare dalla densità di proposta condizionatamente ai valori più recenti a disposizione del vettore dei parametri esclusa quella componente.

2.3 Modello logistico

Come primo modello proposto si prevede una risposta dicotomica per la previsione dei risultati delle partite di calcio.

Sia $C_T = \big(c_1, \ldots, c_{38}\big)$ l'insieme delle trentotto giornate giocate in un intero campionato per $T = 1, 2, \ldots,$ e sia c_t la singola giornata per $t = 1, \ldots, 38$. Fissato un singolo campionato, si definisce A_t come l'insieme delle coppie (a_i, a_j) di squadre comparate nella giornata c_t, per $t = 1, \ldots, 38$, per $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ e $i \neq j$, con n numero di squadre nel campionato osservato (solitamente avremo $n = 20$), e sia $\boldsymbol{y}_t$ il vettore contenente le singole risposte dicotomiche y_{tij} per l'intera giornata di campionato.
Definiamo quindi l'insieme Ω_t dei risultati delle partite giocate nel turno di campionato t:

$$\Omega_t = \{y_{tij} \mid (a_i, a_j) \in A_t\}, \quad t = 1, \ldots, 38,$$

e l'insieme degli indici associati agli elementi dell'insieme Ω_t:

$$S_t = \{(i, j) \mid y_{tij} \in \Omega_t\}, \quad t = 1, \ldots, 38.$$

L'interpretazione delle risposte sarà del tipo: $y_{tij} = 1$ per indicare la vittoria della squadra a_i su a_j, mentre $y_{ij} = 0$ il viceversa.
La probabilità della singola risposta è modellata come segue:

$$P\big(y_{tij} = 1\big) = \frac{e^{\eta_{tij}}}{1 + e^{\eta_{tij}}},$$
$$\eta_{tij} = \mu_i + \alpha_i - \alpha_j + \mathbf{d}_{tij}^{\intercal}\boldsymbol{\gamma},$$

per ogni $y_{tij} \in \Omega_t$. Con $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\intercal}$ vettore $n \times 1$ delle abilità delle squadre $a_1, \ldots, a_n$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\intercal}$ vettore $n \times 1$ delle intercette per ogni squadra che gioca in casa, dove il singolo μ_i indica il vantaggio di giocare in casa derivante dal "fattore campo", $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^{\intercal}$ vettore $m \times 1$ di parametri relativi alle covariate, e $\mathbf{d}_{tij} = \big(\mathbf{d}_{tij,1}, \ldots, \mathbf{d}_{tij,m}\big)^{\intercal}$ vettore $m \times 1$ contiene le variabili esplicative, che possono essere esogene o endogene, anch'esse definite come differenza dei valori associati alle squadre a riferimento, mettendo per prima la squadra che gioca in casa:

$$\mathbf{d}_{tij,k} = \mathbf{d}_{ti,k} - \mathbf{d}_{tj,k}, \qquad k = 1, \ldots, m.$$

Per applicare la modellazione logistica bayesiana con l'introduzione della variabili latente Pòlya-Gamma, dobbiamo riscrivere in forma diversa il vettore dei parametri nel modello. Indichiamo con $\boldsymbol{\beta} = \big(\boldsymbol{\mu}^\intercal, \boldsymbol{\alpha}^\intercal, \boldsymbol{\gamma}^\intercal\big)^\intercal$ il vettore $2n + m \times 1$ dei parametri contenente $\boldsymbol{\mu} = \big(\mu_1 \dots \mu_n\big)^\intercal$, vettore $n \times 1$ di intercette per ogni squadra che gioca in casa, $\boldsymbol{\alpha} = \big(\alpha_1 \dots \alpha_n\big)^\intercal$, vettore $n \times 1$ delle abilità delle squadre, e $\boldsymbol{\gamma} = \big(\gamma_1 \dots \gamma_m\big)^\intercal$ vettore $m \times 1$ dei parametri associati alle covariate. Riscriviamo quindi η_{tij} come segue:

$$\eta_{tij} = \mathbf{x}_{tij}^\intercal \boldsymbol{\beta},$$

con $\mathbf{x}_{tij} = \big(\mathbf{b}_{tij}^\intercal, \mathbf{c}_{tij}^\intercal, \mathbf{d}_{tij}^\intercal\big)^\intercal$ di dimensione $2n + m \times 1$. In particolare: $\mathbf{b}_{tij}^\intercal = (0, \dots, 1, \dots, 0)$ è un vettore $n \times 1$ di selezione con valore 1 in posizione i e zero altrove, $\mathbf{c}_{tij}^\intercal = (0, \dots, 1, \dots, -1, \dots, 0)$ è un vettore $n \times 1$ di selezione con valore 1 in posizione i, -1 in posizione j e 0 altrove, mentre $\mathbf{d}_{tij}$ è il vettore $m \times 1$ di covariate.

Come si è detto, l'inferenza Bayesiana richiede la specificazione di una distribuzione a priori per l'ignoto vettore dei parametri $\boldsymbol{\beta}$. Per lasciare spazio ai dati osservati si possono usare a priori non informative per il vettore da stimare, del tipo: $\pi\big(\boldsymbol{\beta}\big) \propto 1$. Alternativamente, si può usare l'usuale specificazione Gaussiana: $\boldsymbol{\beta} \sim \mathcal{N}\big(\mu_0, \Sigma_0\big)$, che in questo caso torna utile.

2.3.1 Estensione bayesiana

Per completare l'inferenza bayesiana nel modello logistico, bisogna aumentare il vettore delle osservazioni y_{tij} per ogni $y_{tij} \in \Omega_t$, con la variabile latente Pòlya-Gamma $\omega_{tij} \sim \mathcal{PG}\big(1,0\big)$ per $t = 1, 2, \ldots$. La funzione di verosimiglianza aumentata diventa quindi:

$$
\begin{aligned}
&L\big(\boldsymbol{\beta} \mid \mathbf{y}_t, \boldsymbol{\omega}_t, \mathbf{X}_t\big) = \\
&\quad \prod_{i,j \in S_t} \mathbb{P}\big(Y_{tij} = y_{tij} \mid \mathbf{x}_{tij}, \omega_{tij}, \boldsymbol{\beta}\big)\pi\big(\omega_{tij}\big) = \\
&\quad \prod_{i,j \in S_t} \Big(\frac{e^{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}}\Big)^{y_{tij}} \Big(1 - \frac{e^{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}}\Big)^{1 - y_{tij}} \pi\big(\omega_{tij}\big) = \\
&\quad \prod_{i,j \in S_t} \frac{\big(e^{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}\big)^{y_{tij}}}{1 + e^{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}} \pi\big(\omega_{tij}\big) = \\
&\quad \prod_{i,j \in S_t} \frac{1}{2} \exp\left\{ y_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}\big) - \frac{\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}}{2} - \frac{\omega_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}\big)^2}{2} \right\} \pi\big(\omega_{tij}\big) = \\
&\quad \frac{1}{2^n} \exp\left\{ \sum_{i,j \in S_t} \tilde{y}_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}\big) - \sum_{i,j \in S_t} \frac{\omega_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}\big)^2}{2} \right\} \pi\big(\omega_{tij}\big)
\end{aligned}
$$

dove $\tilde{y}_{tij} = y_{tij} - \frac{1}{2}$, $\mathbf{X}_t = \big[x_{tij}\big]_{i,j \in S_t}$ è la matrice contenente tutti i vettori x_{tij} di dimensione $m \times 1$ di covariate per ogni $y_{tij} \in \Theta_t$ e $\boldsymbol{\omega}_t = \big(\omega_{tij}\big)_{i,j \in S_t}$ è il vettore contenente tutti gli ω_{tij} per gli i e j associati alle partite giocate al tempo t.

Perciò la distribuzione a posteriori di $\boldsymbol{\beta}$ diventa:

$$
\begin{aligned}
\pi(\boldsymbol{\beta} \mid \mathbf{y}_t, \mathbf{X}_t, \boldsymbol{\omega}_t) &\propto \exp\left\{ \sum_{i,j \in S_t} \tilde{y}_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}\big) - \sum_{i,j \in S_t} \frac{\omega_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta}\big)^2}{2} \right\} \pi(\boldsymbol{\beta}) \\
&\propto \exp\left\{ -\frac{1}{2} \sum_{i,j \in S_t} \omega_{tij}\big(\mathbf{x}_{tij}^\intercal \boldsymbol{\beta} - \frac{\tilde{y}_{tij}}{\omega_{tij}}\big)^2 \right\} \pi(\boldsymbol{\beta}) \\
&\propto \exp\left\{ -\frac{1}{2} (\widetilde{\widetilde{\mathbf{y}}}_t - \mathbf{X}_t \widetilde{\boldsymbol{\beta}})^\intercal \mathbf{W}_t (\widetilde{\widetilde{\mathbf{y}}}_t - \mathbf{X}_t \widetilde{\boldsymbol{\beta}}) \right\} \pi(\boldsymbol{\beta}),
\end{aligned}
$$

dove $\widetilde{\widetilde{\mathbf{y}}}_t = \big(\frac{\tilde{y}_{tij}}{\omega_{tij}}\big)_{i,j \in S_t}$ è un vettore, $\mathbf{W}_t = \operatorname{diag}\big(\omega_{tij}\big)_{i,j \in S_t}$, è una matrice con i corrispondenti elementi ω_{tij} in diagonale e zero altrove, e $\widetilde{\beta}$ è un vettore

opportunamente definito, che è proporzionale ad una distribuzione Gaussiana multivariata del tipo $\pi(\boldsymbol{\beta} \mid \mathbf{y}_t, \mathbf{X}, \boldsymbol{\omega}_t) \propto \mathsf{N}(\widehat{\boldsymbol{\mu}}_{t\beta}, \widehat{\boldsymbol{\Sigma}}_{t\beta})$, con:

$$\widehat{\boldsymbol{\mu}}_{t\beta} = \widehat{\boldsymbol{\Sigma}}_{t\beta}\big(\mathbf{X}^\intercal \tilde{\mathbf{y}}_t + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\big),$$
$$\widehat{\boldsymbol{\Sigma}}_{t\beta} = \big(\mathbf{X}^\intercal \mathbf{W}_t \mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\big)^{-1},$$

dove $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ sono la media e la matrice di varianze e covarianze a priori del vettore $\boldsymbol{\beta}$, scelte in modo che la a priori sia sufficientemente sparsa. Inoltre, la distribuzione condizionata di $\boldsymbol{\omega}_t$ è ancora Pòlya-Gamma: $\pi(\boldsymbol{\omega}_{tij} \mid \mathbf{y}_t, \mathbf{X}, \boldsymbol{\beta}) \sim \mathcal{PG}(1, \mathbf{x}_{tij}^\intercal \boldsymbol{\beta})$.

2.3.2 Implementazione sui dati

Per implementare quanto detto nella sottosezione 2.3.1 ed ottenere la stima dei parametri bayesiani, con relativi intervalli di credibilità, bisogna scrivere un algoritmo che esegua il *Gibbs Sampling*. L'algoritmo in questione è quello riportato di seguito:

Algoritmo 1 *Gibbs Sampling* per la regressione logistica bayesiana

1. Inizializzazione di $\boldsymbol{\beta}^{(0)}$ facendo una simulazione dalla distribuzione a-priori, ipotizzata $N(\mu_0, \epsilon_0)$;
2. Per $b = 1, \ldots, B$:
   - Simulare dalla distribuzione condizionata:

     $$\hat{\omega}_{tij}^{(b)} \mid \mathbf{x}_{tij}, \boldsymbol{\beta} \sim PG(1, \mathbf{x}_{tij}^\intercal \boldsymbol{\beta}^{(b-1)}) \text{ per } (i,j) \in S_t;$$

   - aggiornare la distribuzione di $\boldsymbol{\beta}$

     $$\widehat{\boldsymbol{\mu}}_{t\beta} = \widehat{\boldsymbol{\Sigma}}_{t\beta}\big(\mathbf{X}^\intercal \tilde{\mathbf{y}}_t + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\big)$$
     $$\widehat{\boldsymbol{\Sigma}}_{t\beta} = \big(\mathbf{X}^\intercal \mathbf{W}_t \mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\big)^{-1};$$

   - simulare nuovi valori $\boldsymbol{\beta}^{(b)}$ dalla distribuzione condizionata:

     $$\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \hat{\omega}^{(b)} \sim N(\widehat{\boldsymbol{\mu}}_{t\beta}, \widehat{\boldsymbol{\Sigma}}_{t\beta}).$$

Utilizzando $B = 1000$ iterazioni, e scartando le prime 200 simulazioni considerandole di riscaldamento, otteniamo un campione dalla distribuzione a posteriori per procedere con le analisi.

Si considerino quindi i dati presentati all'inizio sulla *Premier League* per le dieci stagioni sportive dal 2010-11 al 2019-20. Come primo modello si considera una risposta binaria, quindi vengono tolte le partite che finiscono con un pareggio, e vengono ricodificate le modalità della variabile risposta in 1 e 0 per indicare la vittoria o la sconfitta della squadra in casa. In questa fase si provano due modelli, uno bayesiano e uno frequentista, senza covariate, col fine di confrontare le stime nei due approcci. Si crea quindi una matrice del disegno contenente una variabile *dummy* per ogni squadra nel dato campionato relativo al parametro di abilità di ciascuna squadra. Per ogni partita giocata si metterà, oltre al termine di intercetta diverso per ogni squadra che gioca in casa, il valore 1 per la squadra ospitante e -1 per relativa squadra ospite.

Il modello così specificato diventa:

$$\begin{aligned} P\big(y_{tij} = 1\big) &= \frac{e^{\eta_{tij}}}{1 + e^{\eta_{tij}}} \\ \eta_{tij} &= \mu_i + \alpha_i - \alpha_j \\ &= \mathbf{x}_{tij}^{\intercal}\boldsymbol{\beta} \qquad i, j \in S_t \text{ e } t = 1, 2, \ldots, 380, \end{aligned}$$

con $\boldsymbol{\beta} = \big(\mu_1, \ldots, \mu_n, \alpha_1, \ldots, \alpha_n\big)^{\intercal}$ e $\mathbf{x}_{tij}^{\intercal} = (\mathbf{b}_{tij}^{\intercal}, \mathbf{c}_{tij}^{\intercal})$ vettore $2n \times 1$ con $\mathbf{b}_{tij}^{\intercal} = (0, \ldots, 1, \ldots, 0)$ vettore di selezione $n \times 1$ con valore diverso da zero in posizione i e $\mathbf{c}_{tij}^{\intercal} = (0, \ldots, 1, \ldots, -1, \ldots, 0)$ vettore di selezione $n \times 1$ con valore uguale a 1 in posizione i, -1 in posizione j e 0 altrove. Questo è un classico modello lineare generalizzato per risposta binaria con funzione di legame logistica.

Scritto in forma matriciale il modello diventa:

$$\begin{bmatrix} P(y_{1ij}=1) \\ \vdots \\ P(y_{380ij}=1) \end{bmatrix} = \begin{bmatrix} f(\eta_{1ij}) \\ \vdots \\ f(\eta_{380ij}) \end{bmatrix}$$

$$\begin{bmatrix} \eta_{1ij} \\ \vdots \\ \eta_{380ij} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{b}_{1ij}^{\intercal}, \boldsymbol{c}_{1ij}^{\intercal}) \\ \vdots \\ (\boldsymbol{b}_{380ij}^{\intercal}, \boldsymbol{c}_{380ij}^{\intercal}) \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

e $f(\cdot)$ funzione logistica.

La stima viene ripetuta al variare di dieci stagioni sportive dal 2010-11 al 2019-20.

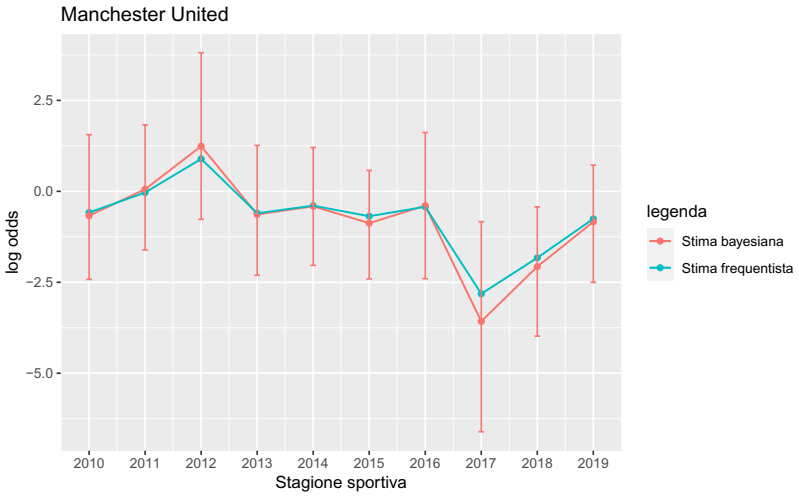


Figura 2.1: Stima del parametro α_i relativo all'abilità della squadra *Manchester United*.

Col fine di confrontare la presenza di un effettivo cambiamento nella stima dei parametri nei diversi anni, sono state confrontate le squadre che in tutto l'arco temporale hanno sempre giocato nel massimo campionato inglese senza mai retrocedere, facendo le stime sia con il modello frequentista che quello *bayesiano*. Come si può vedere dalle figure 2.1 e 2.2, i due approcci conducono

a stime simili (sia in termini di valori che di andamento), e in più si nota una variazione nel tempo, il che induce a pensare che anche estensioni tempo dipendenti siano sensate.

Passando all'interpretazione delle stime, si ricorda che queste vanno interpretate in rapporto alla squadra presa a riferimento nel modello, in questo caso il *Manchester City*, perché nel decennio in esame ha vinto più campionati ed è sempre stata nelle posizioni di testa della classifica. In Figura 2.1 è riportato l'andamento delle stime del parametro di abilità della squadra *Manchester United*. Si può vedere come ci sia un picco nella stagione 2012/2013, nella quale la squadra ha vinto per l'ultima volta il campionato per poi fare diversi anni nelle zone centrali della classifica, salvo poi riprendersi nell'ultimo periodo.

Figura 2.2: Stima del parametro α_i relativo all'abilità della squadra *Chelsea*.

In Figura 2.2 si vede invece l'andamento in campionato della squadra *Chelsea*, la quale è sempre stata abbastanza costante in campionato, finendo le stagioni nella parte alta della classifica, tranne negli anni 2011/2012 (visibile il peggioramento nel grafico) e nell'anno 2015/2016. Inoltre questa squadra ha vinto il campionato negli anni 2014/2015 e 2016/2017, che corrispon-

dono anche ai punti di massimo della serie, per poi calare in campionato salvo poi riprendersi in questi ultimi anni. Come si può vedere, l'andamento del *Chealsea* che è stato appena descritto si rispecchia abbastanza bene nel grafico 2.2.

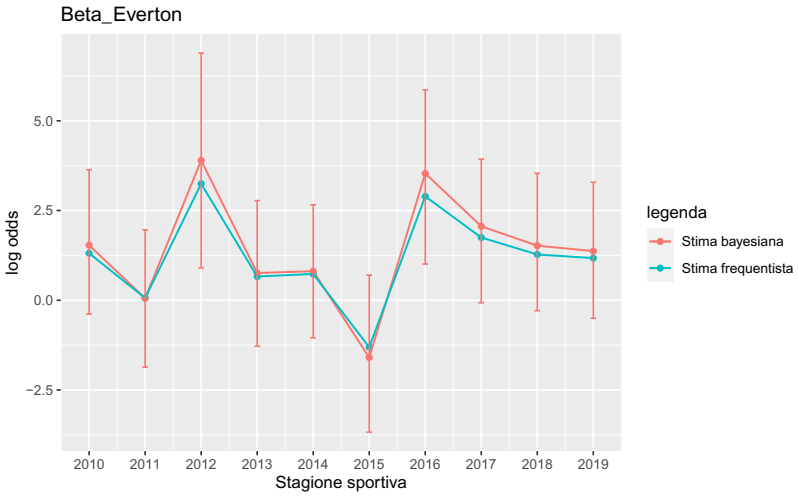


Figura 2.3: Stima del parametro relativo al vantaggio di giocare in casa per la squadra *Everton*.

Per quanto riguarda invece la stima del parametro del vantaggio di giocare in casa, si può vedere come questo venga effettivamente rilevato negli anni in cui le squadre concedono poche vittorie tra le mura amiche, come si vede in Figura 2.3, e in particolare nelle stagioni 2012/2013, dove non perde mai in casa, e 2016/2017, dove perde solo due volte nelle partite casalinghe. Un effetto ancora più marcato si vede in Figura 2.4, considerando il fatto che nelle stagioni dal 2017/2018 al 2019/2020, come si era notato in fase esplorativa, il *Liverpool* non perde mai in casa e vince quarantasette partite su cinquantasette giocate in quel triennio.

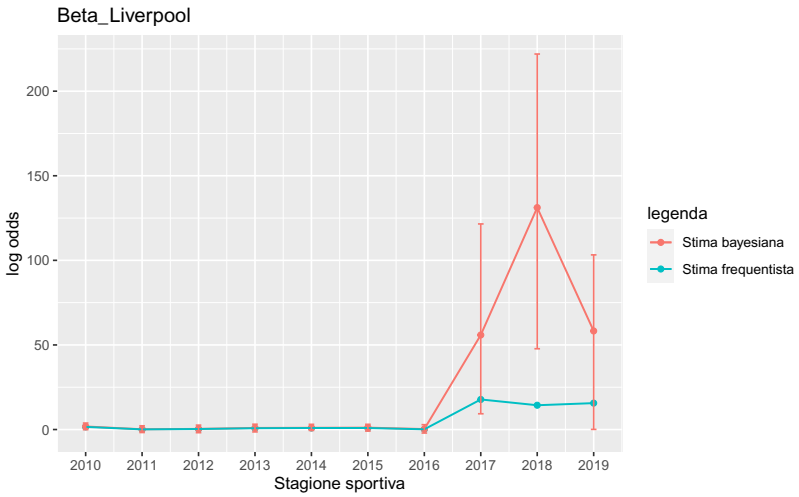


Figura 2.4: Stima del parametro relativo al vantaggio di giocare in casa per la squadra *Liverpool*.

Capitolo 3

Modello multinomiale bayesiano

3.1 Modello multinomiale

Si consideri ora una risposta multinomiale: $y_{tij} = \{y_{tij,k}\}_{k=1}^{K}$ con, in questo caso, $K = 3$ e dove $y_{tij,k} = 1$ se si è osservata la k-esima modalità e $y_{tij,k} = 0$ altrimenti. L'interpretazione delle modalità della variabile risposta è del tipo: $y_{tij,1} = 1$ per indicare la vittoria della squadra ospitante, quindi della squadra a_i su a_j, $y_{tij,2} = 1$ per indicare il pareggio tra le due e $y_{tij,3} = 1$ per indicare la vittoria della squadra ospite, quindi di a_j su a_i.
La probabilità della singola risposta è modellata come segue:

$$P\big(Y_{tij,k} = 1\big) = \frac{\exp(\eta_{tij,k})}{1 + \sum_{k=1}^{K-1} \exp(\eta_{tij,k})}, \quad k = 1, 2,$$

$$\eta_{tij,k} = \mu_{i,k} + \alpha_{i,k} - \alpha_{j,k} + \mathbf{d}_{tij}^{\intercal} \boldsymbol{\gamma_k},$$

con $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{n,k})^{\intercal}$ vettore $n \times 1$ di intercette per ogni squadra che gioca in casa, che stanno ad indicare il vantaggio derivante dal "fattore campo" per la squadra ospitante, $\boldsymbol{\alpha}_k = (\alpha_{1,k}, \ldots, \alpha_{n,k})^{\intercal}$ vettore delle abilità di ciascuna squadra, $\boldsymbol{\gamma}_k = (\gamma_{1,k}, \ldots, \gamma_{m,k})^{\intercal}$ vettore dei parametri relativi al-

le covariate e $\mathbf{d}_{tij} = \big(\mathbf{d}_{tij,1}, \dots, \mathbf{d}_{tij,m}\big)^\intercal$ il vettore contenente le covariate (esogene o endogene).

Per semplicità si riscrive il vettore dei parametri del modello in forma compatta. Si indica con $\boldsymbol{\beta}_k = \big(\boldsymbol{\mu}_k^\intercal, \boldsymbol{\alpha}_k^\intercal, \boldsymbol{\gamma}_K^\intercal\big)$ vettore $(K-1+n+m)\times 1$ dei parametri. Si riscrive quindi $\eta_{tij,k}$ come segue:

$$\eta_{tij,k} = \mathbf{x}_{tij}^\intercal \boldsymbol{\beta}_k,$$

con $\mathbf{x}_{tij} = \big(\mathbf{b}_{ti}, \mathbf{c}_{tij}^\intercal, \mathbf{d}_{tij}^\intercal\big)^\intercal$ di dimensione $(2n+m)\times 1$. In particolare: $\mathbf{b}_{ti} = (0, \dots, 1, \dots, 0)$ è un vettore $n\times 1$ di selezione che assume valore 1 in posizione i e 0 altrove, $\mathbf{c}_{ij}^\intercal = (0, \dots, 1, \dots, -1, \dots, 0)$ è un vettore $n \times 1$ di selezione con valore 1 in posizione i, -1 in posizione j e 0 altrove, mentre $\mathbf{d}_{tij}$ è il vettore $m \times 1$ di covariate.

Il vettore di interesse per l'inferenza è quindi: $\boldsymbol{\beta} = \big(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{K-1}\big)$, in questo caso:

$$\boldsymbol{\beta} = \big(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\big).$$

3.1.1 Estensione bayesiana

Come nel caso del modello logistico, per completare l'inferenza bayesiana bisogna ipotizzare una distribuzione a priori del parametro di interesse ed aumentare le osservazioni con la variabile latente Pòlya-Gamma. La distribuzioni a priori per ogni componente di $\boldsymbol{\beta}$ sarà ancora una volta la normale, quindi: $\boldsymbol{\beta}_k \sim \mathcal{N}\big(\mu_{0k}, \Sigma_{0k}\big)$ per $k = 1, \dots, K-1$. A questo punto si definisce: $\boldsymbol{\beta}_{-k} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}_{k+1}, \dots, \boldsymbol{\beta}_{K-1}\}$ e si scrive la verosimiglianza per il singolo vettore dei parametri $\boldsymbol{\beta}_k$ (Polson et al., 2013):

$$L\big(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_{-k}, \mathbf{y}_t, \mathbf{X}_{t,k}\big) = \prod_{i,j \in S_t} \Big(\frac{e^{\psi_{tij,k}}}{1+e^{\psi_{tij,k}}}\Big)^{y_{tij,k}} \Big(\frac{1}{1+e^{\psi_{tij,k}}}\Big)^{1-y_{tij,k}},$$

dove

$$\psi_{tij,k} = \eta_{tij,k} - C_{tij,k} \text{ con } C_{tij,k} = \log \sum_{k \neq c} \exp\, \eta_{tij,k},$$

che ha la forma della verosimiglianza discussa prima nel caso binario della regressione logistica.

Incorporando la variabile latente Pòlya-Gamma $\omega_{tij,k} \sim \mathcal{PG}\big(1,0\big)$, la verosimiglianza aumentata diventa:

$$
\begin{aligned}
&L\big(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_{-k}, \mathbf{y}_t, \mathbf{X}_{t,k}\big) = \\
&\quad \prod_{i,j \in S_t} \mathbb{P}\big(Y_{tij,k} = y_{tij,k} \mid \mathbf{x}_{tij,k}, \omega_{tij,k}, \boldsymbol{\beta}\big)\pi\big(\omega_{tij}\big) = \\
&\quad \prod_{i,j \in S_t} \Big(\frac{e^{\psi_{tij,k}}}{1+e^{\psi_{tij,k}}}\Big)^{y_{tij,k}} \Big(\frac{1}{1+e^{\psi_{tij,k}}}\Big)^{1-y_{tij,k}} \pi\big(\omega_{tij,k}\big) = \\
&\quad \prod_{i,j \in S_t} \frac{\big(e^{\psi_{tij,k}}\big)^{y_{tij,k}}}{1+e^{\psi_{tij,k}}} \pi\big(\omega_{tij,k}\big) = \\
&\quad \prod_{i,j \in S_t} \frac{1}{2} \exp\left\{ y_{tij,k}\big(\psi_{tij,k}\big) - \frac{\psi_{tij,k}}{2} - \frac{\omega_{tij,k}\big(\psi_{tij,k}\big)^2}{2} \right\} \pi\big(\omega_{tij,k}\big) = \\
&\quad \frac{1}{2^n} \exp\left\{ \sum_{i,j \in S_t} \tilde{y}_{tij,k}\big(\psi_{tij,k}\big) - \sum_{i,j \in S_t} \frac{\omega_{tij,k}\big(\psi_{tij,k}\big)^2}{2} \right\} \pi\big(\omega_{tij,k}\big)
\end{aligned}
$$

dove $\tilde{y}_{tij,k} = \big(y_{tij,k} - \frac{1}{2}\big)$, $\mathbf{X}_{t,k} = \big[x_{tij,k}\big]_{i,j \in S_t}$ è la matrice contenente tutti i vettori $x_{tij,k}$ di dimensione $m \times 1$ di covariate per ogni $y_{tij,k} \in \Theta_t$ e $\boldsymbol{\omega}_{t,k} = \big(\omega_{tij,k}\big)_{i,j \in S_t}$ è il vettore contenente tutti gli $\omega_{tij,k}$ per gli i e j associati alle partite giocate al tempo t.

Perciò la distribuzione a posteriori di $\boldsymbol{\beta}_k$ diventa:

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_{-k}, \mathbf{y}_t, \mathbf{X}_{t,k}, \boldsymbol{\omega}_{t,k}) \propto \\
&\quad \exp\left\{ \sum_{i,j \in S_t} \tilde{y}_{tij,k}\big(\psi_{tij,k}\big) - \sum_{i,j \in S_t} \frac{\omega_{tij,k}\big(\psi_{tij,k}\big)^2}{2} \right\} \pi(\boldsymbol{\beta}_k) \propto \\
&\quad \exp\left\{ -\frac{1}{2} \sum_{i,j \in S_t} \omega_{tij,k}\big(\psi_{tij,k} - \frac{\tilde{y}_{tij,k}}{\omega_{tij,k}}\big)^2 \right\} \pi(\boldsymbol{\beta}_k) \propto \\
&\quad \exp\left\{ -\frac{1}{2} \big[\widetilde{\widetilde{\mathbf{y}}}_{t,k} - (\mathbf{X}_{t,k}\widetilde{\boldsymbol{\beta}_k} - \mathbf{C}_{t,k})\big]^{\intercal} \mathbf{W}_t \big[\widetilde{\widetilde{\mathbf{y}}}_{t,k} - (\mathbf{X}_{t,k}\widetilde{\boldsymbol{\beta}_k} - \mathbf{C}_{t,k})\big] \right\} \pi(\boldsymbol{\beta}_k),
\end{aligned}
$$

dove $\widetilde{\widetilde{\mathbf{y}}}_{t,k} = \big(\frac{\tilde{y}_{tij,k}}{\omega_{tij,k}}\big)^{\intercal}_{i,j \in S_t}$ è un vettore, $\mathbf{W}_t = \operatorname{diag}\big(\omega_{tij,k}\big)_{i,j \in S_t}$, è una matrice con i corrispondenti elementi $\omega_{tij,k}$ in diagonale e zero altrove, $\mathbf{C}_{t,k} =$

$\big(C_{tij,k}\big)^{\intercal}_{i,j\in S_t}$ è il vettore contenente tutti gli elementi $C_{tij,k}$ per ogni i e j in S_t e $\widetilde{\beta}$ è un vettore opportunamente definito, che è proporzionale ad una distribuzione Gaussiana multivariata del tipo $\pi(\boldsymbol{\beta}_k \mid \mathbf{y}_t, \mathbf{X}_{t,k}, \boldsymbol{\omega}_{t,k}) \propto \mathsf{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$, con:

$$\widehat{\boldsymbol{\mu}}_k = \widehat{\boldsymbol{\Sigma}}_k\big(\mathbf{X}^{\intercal}_{t,k}\big(\tilde{\mathbf{y}}_{t,k} + \mathbf{W}_{t,k}\mathbf{C}_{t,k}\big) + \boldsymbol{\Sigma}^{-1}_{0k}\boldsymbol{\mu}_{0k}\big)$$
$$\widehat{\boldsymbol{\Sigma}}_k = \big(\mathbf{X}^{\intercal}_{t,k}\mathbf{W}_{t,k}\mathbf{X}_{t,k} + \boldsymbol{\Sigma}^{-1}_{0k}\big)^{-1},$$

dove $\boldsymbol{\mu}_{0k}$ and $\boldsymbol{\Sigma}_{0k}$ sono la media e la matrice di varianze e covarianze a priori del vettore $\boldsymbol{\beta}$, che come nel caso binomiale seguiranno una assunzione sufficientemente sparsa. Inoltre, la distribuzione condizionata di $\boldsymbol{\omega}_t$ è ancora Pòlya-Gamma: $\pi(\boldsymbol{\omega}_{tij,k} \mid \mathbf{y}_t, \mathbf{X}_{t,k}, \boldsymbol{\beta}_k) \sim \mathcal{PG}(1, \psi_{tij,k})$.

Per costruire il *Gibs sampler* basterà iterare la procedura usata nel caso binomiale per ogni $k = 1, \ldots, K-1$, facendo ad ogni passo un modello logistico utilizzando la modalità k come successo e tutte le altre modalità come insuccessi. Anche in questo caso si faranno $B = 1000$ iterazioni per ogni parametro da stimare e si scarteranno le prime 200 osservazioni considerate di riscaldamento.

3.1.2 Nota metodologica

La procedura di *data-augmentation* descritta è presa dall'articolo di Polson et al. (2013), e nei *Technical supplement* è presentata anche l'estensione multinomiale. Sulla base di quanto riportato nell'articolo verranno ora esplicitati tutti i conti che portano ai risultati utilizzati nella sezione precedente.

Si consideri una risposta multinomiale $y_i = \big\{y_{ij}\big\}_{j=1}^{J}$ per $i = 1, \ldots, N$, che registri il numero di risposte per ogni categoria $j = 1, \ldots, J$, e sia n_i il numero totale di risposte. La funzione di legame logistica per la regressione multinomiale prevede che la probabilità di estrazione per la singola risposta dalla modalità j-esima dal campione i-esimo sia:

$$p_{ij} = \frac{\exp\, \psi_{ij}}{\sum_{k=1}^{J} \exp\, \psi_{ik}},$$

dove il log-rapporto ψ_{ij} è modellata da $x_i^\intercal \boldsymbol{\beta}_j$ e sia, per motivi di identificabilità, $\boldsymbol{\beta}_J$ è vincolato a zero. Sulla base di quanto riportato nell'articolo di Held e Holmes (2006), la verosimiglianza per il parametro $\boldsymbol{\beta}_j$ condizionatamente a $\boldsymbol{\beta}_{-j}$, matrice di parametri nella quale viene rimossa la colonna del vettore $\boldsymbol{\beta}_j$, è:

$$L\big(\boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}, \mathbf{y}\big) = \prod_{i=1}^{N} \Big(\frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}}\Big)^{y_{ij}} \Big(\frac{1}{1+e^{\eta_{ij}}}\Big)^{1-y_{ij}},$$

dove

$$\eta_{ij} = x_i^\intercal \boldsymbol{\beta}_j - C_{ij} \text{ con } C_{ij} = \log \sum_{k \neq j} \exp\, x_i^\intercal \boldsymbol{\beta}_k,$$

che ha una forma simile a quella della regressione logistica trattata nello stesso articolo. Incorporando la variabile latente Pòlya-Gamma: $\omega_{ij} \sim \mathcal{PG}\big(n_i, 0\big)$, la verosimiglianza aumentata diventa:

$$\begin{aligned}
&L\big(\boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}, \mathbf{y}, \mathbf{X}\big) = \\
&\quad \prod_{i=1}^{N} \mathbb{P}\big(Y_{ij} = y_{ij} \mid \mathbf{x}_i, \omega_{ij}, \boldsymbol{\beta}\big) \pi\big(\omega_{ij}\big) = \\
&\quad \prod_{i=1}^{N} \Big(\frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}}\Big)^{y_{ij}} \Big(\frac{1}{1+e^{\eta_{ij}}}\Big)^{1-y_{ij}} \pi\big(\omega_{ij}\big) = \\
&\quad \prod_{i=1}^{N} \frac{\big(e^{\eta_{ij}}\big)^{y_{ij}}}{1+e^{\eta_{ij}}} \pi\big(\omega_{ij}\big) = \\
&\quad \prod_{i=1}^{N} \frac{1}{2} \exp\left\{ y_{ij}\eta_{ij} - \frac{\eta_{ij}}{2} - \frac{\omega_{ij}\eta_{ij}^2}{2} \right\} \pi\big(\omega_{ij}\big) = \\
&\quad \frac{1}{2^n} \exp\left\{ \sum_{i=1}^{N} \kappa_{ij}\eta_{ij} - \sum_{i=1}^{N} \frac{\omega_{ij}\eta_{ij}^2}{2} \right\} \pi\big(\omega_{ij}\big),
\end{aligned}$$

dove $\kappa_{ij} = (y_{ij} - n_i/2)$ per $i = 1, \ldots, N$, e $\mathbf{X}$ è la matrice contenente le variabili esplicative.

Perciò la distribuzione a posteriori di $\boldsymbol{\beta}_j$, ipotizzata con densità a prioi $\mathsf{N}(\boldsymbol{m}_{0j}, \boldsymbol{V}_{0j})$, diventa:

$$\begin{aligned}
&\pi(\boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}, \mathbf{y}, \mathbf{X}, \boldsymbol{\omega}) \propto \\
&\quad \exp\left\{\sum_{i=1}^{N} \kappa_{ij}\eta_{ij} - \sum_{i=1}^{N} \frac{\omega_{ij}\eta_{ij}^2}{2}\right\} \pi(\boldsymbol{\beta}_j) \propto \\
&\quad \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \omega_{ij}\Big(\eta_{ij} - \frac{\kappa_{ij}}{\omega_{ij}}\Big)^2\right\} \pi(\boldsymbol{\beta}_j) \propto \\
&\quad \exp\left\{-\frac{1}{2}\big[\widetilde{\boldsymbol{\kappa}}_j - (\mathbf{X}\widetilde{\boldsymbol{\beta}_j} - \mathbf{C}_j)\big]^\intercal \boldsymbol{\Omega}_j \big[\widetilde{\boldsymbol{\kappa}}_j - (\mathbf{X}_{t,j}\widetilde{\boldsymbol{\beta}_j} - \mathbf{C}_{t,j})\big]\right\} \pi(\boldsymbol{\beta}_j),
\end{aligned}$$

dove $\widetilde{\boldsymbol{\kappa}}_j = \big(\frac{\kappa_{1j}}{\omega_{1j}}, \ldots, \frac{\kappa_{Nj}}{\omega_{Nj}}\big)$, $\boldsymbol{\Omega}_j = \operatorname{diag}\big(\omega_{1j}, \ldots, \omega_{Nj}\big)$, $\mathbf{C}_j = \big(C_{1j}, \ldots, C_{Nj}\big)$ e $\widetilde{\beta}$ è un vettore opportunamente definito. La distribuzione a posteriori per $\widetilde{\beta}$ è proporzionale ad una Gaussiana multivariata del tipo $\pi(\boldsymbol{\beta}_j \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\omega}) \propto \mathsf{N}(\boldsymbol{m}_j, \boldsymbol{V}_j)$, con:

$$\begin{aligned}
\boldsymbol{V}_j^{-1} &= \mathbf{X}^\intercal \boldsymbol{\Omega}_j \mathbf{X} + \boldsymbol{V}_{0j}^{-1}, \\
\boldsymbol{m}_j &= V_j\big(\mathbf{X}^\intercal\big(\boldsymbol{\kappa}_j + \boldsymbol{\Omega}_j \mathbf{C}_j\big) + \boldsymbol{V}_{0j}^{-1}\boldsymbol{m}_{0j}\big).
\end{aligned} \tag{3.1}$$

Inoltre, la distribuzione condizionata di $\boldsymbol{\omega}$ è ancora Pòlya-Gamma:

$$\pi(\boldsymbol{\omega}_{ij} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}_j) \sim \mathcal{PG}(n_i, \eta_{ij}), \quad i = 1, \ldots, N$$

.

Questi risultati ottenuti differiscono da quelli riportati nell'articolo di Polson et al. (2013), in quanto l'aggiornamento dei parametri della distribuzione a posteriori vengono riportati come segue:

$$\begin{aligned}
\boldsymbol{V}_j^{-1} &= \mathbf{X}^\intercal \boldsymbol{\Omega}_j \mathbf{X} + \boldsymbol{V}_{0j}^{-1}. \\
\boldsymbol{m}_j &= V_j\big(\mathbf{X}^\intercal\big(\boldsymbol{\kappa}_j - \boldsymbol{\Omega}_j \mathbf{C}_j\big) + \boldsymbol{V}_{0j}^{-1}\boldsymbol{m}_{0j}\big).
\end{aligned} \tag{3.2}$$

L'Equazione 3.1 e 3.2 differiscono per un segno, riportato in rosso nella seconda equazione, il quale determina la buona riuscita della stima dei parametri.

Allo stato attuale, il pacchetto del *software* `R` implementato dagli autori e rilasciato in modo ufficiale, presenta solo le funzioni per la stima del modello logistico, e non riporta nulla sulle funzioni per il modello multinomiale.
Ad onor del vero, va riportato che nella *repository GitHub*[1] originale relativa al pacchetto sono presenti i file che implementano in `R` le funzioni per il modello multinomiale, nei quali è stata effettuata la correzione rilevata in questa sottosezione.

[1] https://github.com/jwindle/BayesLogit-Thesis

Capitolo 4

Previsioni con il modello multinomiale

In questo capitolo si presentano i risultati derivanti dalla previsione del campionato 2018-19 di *Premier League* sulla base delle stime ottenute con i campionati precedenti a partire dalla stagione 2010-11.
Si decide di escludere dal *dataset* il campionato 2019-20 per due motivi. Il primo è che la stagione è stata viziata dalla pandemia di Covid-19, che ha portato alla conclusione del campionato nei mesi estivi e giocando molte partite a distanza ravvicinata. In questa analisi si decide di ovviare al problema derivante da questo evento eccezionale fermandosi una stagione prima. Il secondo motivo è che in questa stagione, sale per la prima volta nella massima categoria inglese lo *Shellield United*, squadra quindi per la quale non sono mai stati stimati i parametri del modello che si è usato in questa analisi.
Per le previsioni si useranno in questo capitolo due approcci diversi.
Nel primo approccio si stimeranno i parametri delle squadre che partecipano alla *Premier League* anno per anno con i soli dati di quella stagione. Per prevedere i risultati della stagione successiva, si useranno i parametri più aggiornati relativi alle squadre del campionato.
Nel secondo approccio si stimeranno i parametri di ogni squadra che ha

partecipato alla *Premier League* negli anni che vanno dal 2010 al 2018, stimandoli con tutti i dati assieme. Per fare la previsione nella stagione 2018-19 si selezioneranno solo i parametri relative alle squadre di quel campionato, costruendo, anche in questo caso, una matrice di disegno *ad hoc*.

Ovviamente, in ogni nuova stagione ci sono tre squadre neo promosse, nuove rispetto alla stagione precedente, e tre squadre che invece retrocedono nella categoria precedente. Per la stagione 2018-19, le tre squadre retrocesse sono: *Stoke City*, *Swansea City* e *West Bromwich Albion*; mentre le neo promosse sono: *Cardiff City*, *Fulham* e *Wolverhampton Wandereres*.

Nel secondo approccio proposto questo problema è relativo, perché si dispone di tutte le stime dei parametri necessari, basterà appunto selezionare i parametri corretti. Per quanto riguarda invece il primo approccio, per ovviare al problema, si prendono, per le squadre nuove, le ultime stime disponibili dall'ultima stagione che ha visto le squadre neo immesse nella massima categoria, considerando che, l'anno in cui sono retrocesse, si può assumere che queste abbiano avuto dei parametri di "forza" paragonabili ad una neo immessa. Salvo casi eccezionali, le squadre neo promosse, non concludono la loro prima stagione nelle zone alte della classifica, e anzi spesso retrocedono (come nel caso della stagione 2018-19, nella quale retrocedono due neo immesse). Si decide di utilizzare questo approccio anche perché la squadra presa a riferimento in tutti i campionati è sempre il *Manchester City*, quindi le stime, seppur per anni diversi, restano paragonabili senza forzature eccessive.

Di conseguenza, il modello specificato per ogni stagione diventa:

$$
\begin{aligned}
P\big(Y_{tij,k} = 1\big) &= \frac{\exp(\eta_{tij,k})}{1 + \sum_{k=1}^{K-1} \exp(\eta_{tij,k})}, \quad k = 1, 2, \\
\eta_{tij,k} &= \mu_{i,k} + \alpha_{i,k} - \alpha_{j,k} + \mathbf{d}_{tij}^{\intercal} \boldsymbol{\gamma_k} \\
&= \mathbf{x}_{tij,k}^{\intercal} \boldsymbol{\beta}_k, \qquad i, j \in S_t, t = 1, 2, \ldots, 380,
\end{aligned}
$$

con $\boldsymbol{\beta}_k = \big(\mu_{1,k}, \ldots, \mu_{n,k}, \alpha_{1,k}, \ldots, \alpha_{n,k}, \gamma_{1,k}, \ldots, \gamma_{m,k}\big)^{\intercal}$ e $\mathbf{x}_{tij}^{\intercal} = (\mathbf{b}_{tij}^{\intercal}, \mathbf{c}_{tij}^{\intercal}, \mathbf{d}_{tij}^{\intercal})$

vettore $(2n+m)\times 1$ con $\mathbf{b}_{tij}^{\intercal} = (0,\ldots,1,\ldots,0)$ vettore di selezione $n\times 1$ con valore diverso da zero in posizione i, $\mathbf{c}_{tij}^{\intercal} = (0,\ldots,1,\ldots,-1,\ldots,0)$ vettore di selezione $n \times 1$ con valore uguale a 1 in posizione i, -1 in posizione j e 0 altrove e $\mathbf{d}_{tij} = \left(\mathbf{d}_{tij,1},\ldots,\mathbf{d}_{tij,m}\right)^{\intercal}$ vettore $m \times 1$ contiene le variabili esplicative, anch'esse definite come differenza dei valori associati alle squadre a riferimento, mettendo per prima la squadra che gioca in casa:

$$\mathbf{d}_{tij,r} = \mathbf{d}_{ti,r} - \mathbf{d}_{tj,r}, \qquad r = 1,\ldots,m.$$

I relativi parametri stimati tramite le medie a posteriori sono quelli riportati nelle tabelle 4.1, 4.2, 4.3.

L'interpretazione dei parametri avviene confrontando le due modalità stimate con quella presa a riferimento, in questo caso la sconfitta della squadra ospitante, e ricordando che in ogni caso i confronti delle singole squadre vanno fatti con quella presa a riferimento, in questo caso il *Manchester City*. Come si può vedere in Tabella 4.1, i parametri relativi al vantaggio di giocare in casa sono quasi sempre di segno positivo sia per la stima di vittoria della squadra locale che per la stima di pareggio. Questo parametro è negativo per *Burnley*, *Cardiff*, *Fulham* e *Southampton*, che sono le squadre che chiudono il campionato nelle posizioni che vanno dalla sedicesima alla diciannovesima, quindi le ultime. Per queste squadre ci si può aspettare che non ci sia un vero e proprio vantaggio nel giocare in casa, in quanto subiscono soprattutto sconfitte nell'arco dell'intero campionato.

Due eccezioni in questo senso sono rappresentate da *Huddersfield* e *Chelsea*, in quanto la prima termina il campionato in ultima posizione, ma nonostante ciò si rileva una certa importanza per questa squadra del fattore campo, mentre la seconda che finisce il campionato in terza posizione, ma non sembra trarre vantaggio dal giocare in casa.

Per quanto riguarda i parametri di forza di ciascuna squadra riportati in Tabella 4.2, questi sono tutti negativi. Il ragionamento è plausibile in quanto la squadra presa a riferimento è il *Manchester City*, la quale nell'anno previsto

vince il campionato, dopo averlo vinto anche nell'anno precedente, cioè quello con cui si stimano quasi tutti i parametri del modello. Va notato che per le squadre neopromosse in *Premier League* nell'anno 2018-2019, sono stati usati i parametri relativi alla loro ultima apparizione nel massimo campionato inglese e in particolare il 2013-2014 per le squadre *Cardiff* e *Fulham* e il 2011-2012 per il *Wolverhampton*. In entrambi questi campionati, la squadra vincitrice è stata quella presa a riferimento, cioè il *Manchester City*.
Per l'interpretazione del parametri riportati in Tabella 4.3, si faranno dei ragionamenti in termini di differenze dei valori di queste variabili esplicative. È bene ricordare che la differenza viene fatta mettendo come primo termine il valore della variabile esplicativa relativa alla squadra ospitante. Quindi differenze grandi con segno positivo indicano valori della variabile nettamente a favore della squadra di casa rispetto a quella ospite, mentre valori negativi indicano il contrario. Dalla Tabella 4.3 si può notare che la stima del parametro relativo alla differenza di numero di cartellini rossi è di segno negativo, e di conseguenza come ci si può aspettare, valori grandi della variabile `diff_red_card` diminuiranno la probabilità di vittoria della squadra di casa, viceversa in caso di valori negativi. In maniera inversa, la stima del parametro relativo a `diff_shots_on_target` ha segno positivo, quindi come ci si può aspettare se la squadra di casa tira di più verso la porta aumenta le proprie probabilità di vittoria, viceversa rischia di perdere. In maniera analoga si può procedere all'interpretazione di tutti gli altri parametri. È interessante soffermarsi su due parametri che hanno un'interpretazione contro intuitiva, cioè `diff_clearances` e `diff_possession`. Come visto in fase di presentazione del dataset, una *clearances* è un'azione difensiva volta ad allontanare il pallone dalla propria porta verso quella avversaria, e una stima del parametro con segno positivo sta ad indicare che se la squadra di casa effettua più giocate di questo tipo aumenta le proprie probabilità di vittoria rispetto alla sconfitta. In modo analogo, il valore negativo della

variabile `diff_possession` indica che non è aumentando la percentuale di possesso palla che si favorisce la probabilità di vincere.

Dopo aver simulato i risultati dell'intero campionato, si calcola l'errore di errata classificazione confrontando i risultati previsti con quelli effettivamente realizzati. In questo caso il modello prevede bene nel 53.15% dei casi, e di conseguenza, la classifica che si genererebbe a fine stagione dal modello è quella riportata in Tabella 4.4.

Confrontandola con quella reale, si può notare che il modello prevede correttamente le prime sette del campionato, invertendo solo la posizione di terza e quinta classificata. Da notare che il modello prevedere correttamente la settima posizione raggiunta dalla squadra *Wolverhampton Wanderes* che è una neo promossa, che non saliva nella massima categoria dalla stagione 2011-12.

Per quanto riguarda invece le squadre che nella stagione prevista sono retrocesse, queste non sono ben individuate dalla simulazione. Le squadre in questione dovrebbero essere, nell'ordine, *Cardiff City*, *Fulham* e *Huddersfield*, ma nessuna di queste tre è correttamente individuata da questa classifica, e anzi il *Cardiff City* è stimato nella parte medio alta della classifica finale.

In Tabella 4.5 è stato calcolato anche il *ranking* di ciascuna squadra facendo cento estrazioni per ogni parametro dalla propria distribuzione a posteriori. Per ogni estrazione sono stati calcolati i risultati previsti dal modello con i parametri campionati ed è stata calcolata la conseguente classifica. È stata quindi calcolata la frequenza con cui ciascuna squadra viene stimata in tutte e venti le posizioni per le cento simulazioni. I *ranking* così calcolati confermano in parte quanto visto dalla classifica riportata in Tabella 4.4. In particolare sono stati segnati con il colore verde le frequenze maggiori relative alle squadre che si sono classificate tra le prime sei in campionato, le quali sono state tutte correttamente individuate (senza inversione tra terza e quinta classificata), e in rosso le frequenze maggiori delle squadre che

Tabella 4.1: Stima dei parametri relativi al vantaggio di giocare in casa per ciascuna squadra del campionato 2018/2019 con relativo intervallo di credibilità.

	Stima vittoria	Stima pareggio
AFC Bournemouth	2.62	2.13
	(-0.41, 5.28)	(-0.37, 4.34)
Arsenal	7.27	3.9
	(4.15, 10.8)	(1.08, 7.27)
Brighton	4.45	4.79
	(1.36, 7.33)	(2.02, 7.57)
Burnley	-0.9	1.19
	(-3.86, 2.44)	(-1.05, 3.44)
Cardiff	-0.03	0.46
	(-2.96, 2.8)	(-1.85, 3.19)
Chelsea	-1.07	-0.39
	(-4.22, 2.04)	(-3.28, 2.47)
Crystal Palace	1.7	1.9
	(-1.37, 4.6)	(-0.26, 4.28)
Everton	5.63	3.8
	(2.5, 9.08)	(0.58, 6.18)
Fulham	-0.21	-1.02
	(-3.45, 2.62)	(-3.73, 1.94)
Huddersfield	4.41	1.74
	(1.13, 7.42)	(-0.65, 4.7)
Leicester	1.27	1.14
	(-1.13, 3.99)	(-1.04, 3.37)
Liverpool	9.78	12.08
	(0.99, 18.1)	(3.43, 20.39)
Manchester United	2.42	-0.26
	(-0.4, 5.49)	(-3.22, 3)
Newcastle	2.62	0.91
	(-0.16, 5.31)	(-1.46, 3.18)
Southampton	-1.73	1.78
	(-5.32, 1.48)	(-0.56, 4.34)
Tottenham	1.2	1.14
	(-1.96, 4.51)	(-2.12, 4.18)
Watford	3.75	1.52
	(0.76, 6.5)	(-1.21, 4.53)
West Ham	3.86	3.77
	(0.57, 6.93)	(0.95, 6.3)
Wolverhampton	1.45	1.84
	(-1.35, 4.66)	(-0.96, 4.46)

Tabella 4.2: Stima dei parametri relativi alla forza di ciascuna squadra del campionato 2018/2019 con realitivi intervalli di credibilità.

	Stima vittoria	Stima pareggio
AFC.Bournemouth	-6.91	-3.7
	(-10.52, -3.4)	(-6.83, -0.72)
Arsenal	-5.9	-2.44
	(-9.03, -3.02)	(-4.9, 0.31)
Brighton	-7.79	-4.49
	(-11.7, -3.7)	(-7.98, -1.04)
Burnley	-3.78	-1.43
	(-7.64, 0.45)	(-4.98, 1.52)
Cardiff	-5.22	-2.56
	(-8.41, -2.33)	(-5.3, 0.08)
Chelsea	-2.14	-0.03
	(-4.97, 1.42)	(-2.75, 2.66)
Crystal.Palace	-6.52	-3.51
	(-10.31, -2.43)	(-7.3, -0.65)
Everton	-7.76	-4.2
	(-11.76, -4.03)	(-7.28, -0.32)
Fulham	-6.8	-2.57
	(-9.8, -3.76)	(-5.49, 0.04)
Huddersfield	-7.22	-3.01
	(-10.94, -3.32)	(-6.53, -0.01)
Leicester	-4.26	-1.72
	(-7.89, -0.75)	(-4.75, 1.41)
Liverpool	-1.46	-1.43
	(-4.26, 2.06)	(-4.08, 0.99)
Manchester.United	-1.65	-0.37
	(-4.74, 1.21)	(-3.24, 2.16)
Newcastle	-8.28	-3.14
	(-12.42, -4.62)	(-6.57, 0.17)
Southampton	-3.72	-2.98
	(-7.87, -0.39)	(-5.95, 0.26)
Tottenham	-1.04	-0.59
	(-3.79, 2.16)	(-3.28, 2.24)
Watford	-5.83	-1.81
	(-9.63, -1.96)	(-5.03, 2.17)
West.Ham	-6.36	-3.27
	(-10.78, -2.81)	(-6.77, 0.26)
Wolverhampton	-6.31	-5.59
	(-9.23, -3.54)	(-8.44, -3.19)

Tabella 4.3: Stima parametri relativi alle covariate con relativi intervalli di credibilità.

	Stima vittoria	Stima pareggio
diff_clearances	0.21	0.09
	(0.14,0.27)	(0.04,0.14)
diff_corners	0.14	0.09
	(-0.01,0.31)	(-0.04,0.23)
diff_fouls_conceded	-0.04	0.15
	(-0.16,0.1)	(0.04,0.26)
diff_offsides	-0.12	-0.23
	(-0.38,0.14)	(-0.45,0.01)
diff_passes	0.03	0
	(-0.01,0.08)	(-0.03,0.03)
diff_possession	-0.52	-0.22
	(-0.74,-0.3)	(-0.37,-0.04)
diff_red_card	-3.58	-0.86
	(-5.96,-1.3)	(-2.35,0.67)
diff_shots	0.03	-0.02
	(-0.1,0.15)	(-0.12,0.07)
diff_shots_on_target	0.84	0.46
	(0.62,1.06)	(0.25,0.63)
diff_tackels	-0.07	-0.09
	(-0.16,0.03)	(-0.17,-0.01)
diff_touches	0.02	0.02
	(-0.02,0.05)	(-0.01,0.05)
diff_yellow_cards	-0.25	-0.3
	(-0.58,0.1)	(-0.6,-0.01)

nella stagione prevista sarebbero dovute retrocedere, le quali appunto sono stimate in posizioni differenti rispetto a quelle reali. L'unica eccezione è la squadra *Fulham* che quanto meno tramite queste simulazioni viene stimata maggiormente in zona retrocessione.

Vediamo ora il modello stimato sul *dataset* composto da tutte e otto le stagioni dal 2010-11 al 2017-18, dobbiamo ridefinire alcune quantità specificate all'inizio.

Quindi, sia $C = \big(c_1, \ldots, c_{304}\big)$ l'insieme delle trecentoquattro giornate giocate nell'arco degli otto campionati e sia c_t la singola giornata per $t = 1, \ldots, 304$. Definiamo A_t come l'insieme delle coppie (a_i, a_j) di squadre comparate nella giornata c_t, per $(i, j) \in \{1, \ldots, N\} \times \{1, \ldots, N\}$ e $i \neq j$ con N numero di squadre che hanno partecipato alla *Premier League* negli otto campionati presi a riferimento, qui $N = 35$. Sia $\boldsymbol{y}_t$ il vettore contenente le singole risposte dicotomiche y_{tij} per l'intera giornata di campionato c_t.

Definiamo quindi l'insieme Ω_t dei risultati delle partite giocate nel turno di campionato t:

$$\Omega_t = \{y_{tij} \mid (a_i, a_j) \in A_t\}, \quad t = 1, \ldots, 304,$$

e l'insieme degli indici associati agli elementi dell'insieme Ω_t:

$$S_t = \{(i, j) \mid y_{tij} \in \Omega_t\}, \quad t = 1, \ldots, 304.$$

Il modello diventa quindi:

$$\begin{aligned} P\big(Y_{tij,k} = 1\big) &= \frac{\exp(\eta_{tij,k})}{1 + \sum_{k=1}^{K-1} \exp(\eta_{tij,k})} \quad \text{per } k = 1, 2 \\ \eta_{tij,k} &= \mu_{i,k} + \alpha_{i,k} - \alpha_{j,k} + \mathbf{d}_{tij}^{\mathsf{T}} \boldsymbol{\gamma_k} \\ &= \mathbf{x}_{tij,k}^{\mathsf{T}} \boldsymbol{\beta}_k \qquad i, j \in S_t \text{ e } t = 1, 2, \ldots, 304 \end{aligned}$$

con

$$\boldsymbol{\beta}_k = \big(\mu_{1,k}, \ldots, \mu_{N,k}, \alpha_{1,k}, \ldots, \alpha_{N,k}, \gamma_{1,k}, \ldots, \gamma_{m,k}\big)^{\mathsf{T}}$$

Tabella 4.4: Classifica stimata con medie a posteriori delle distribuzioni di ciascun parametro.

	Squadra	Punti	Classifica reale
1	Manchester City	114	1°
2	Liverpool	106	2°
3	Arsenal	97	5°
4	Tottenham Hotspur	97	4°
5	Chelsea	93	3°
6	Manchester United	90	6°
7	Wolverhampton Wanderers	66	7°
8	Cardiff City	65	18°
9	Southampton	59	16°
10	Burnley	53	15°
11	AFC Bournemouth	52	14°
12	Crystal Palace	47	12°
13	Fulham	47	19°
14	Everton	45	8°
15	West Ham United	45	10°
16	Newcastle United	37	13°
17	Huddersfield Town	35	20°
18	Leicester City	26	9°
19	Brighton and Hove Albion	23	17°
20	Watford	23	11°

Tabella 4.5: Probabilità di posizione in classifica di ogni squadra facendo 100 simulazioni dalla distribuzione a posteriori.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
AFC Bournemouth	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.12	0.15	0.10	0.05	0.12	0.06	0.10	0.04	0.05	0.06	0.03	0.01	0.00
Arsenal	0.00	0.14	0.18	0.15	0.28	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brighton and Hove Albion	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.03	0.03	0.09	0.09	0.11	0.18	0.13	0.25	0.07
Burnley	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.18	0.12	0.10	0.15	0.06	0.08	0.06	0.06	0.03	0.01	0.02	0.01	0.00
Cardiff City	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.14	0.07	0.09	0.08	0.06	0.02	0.03	0.06	0.01	0.01	0.02	0.01	0.00
Chelsea	0.02	0.24	0.26	0.16	0.18	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Crystal Palace	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.04	0.04	0.08	0.14	0.09	0.20	0.09	0.10	0.02	0.09	0.03	0.02	0.01
Everton	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.11	0.14	0.11	0.07	0.11	0.10	0.10	0.06	0.08	0.02	0.01	0.01	0.00
Fulham	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.09	0.07	0.07	0.07	0.04	0.03	0.04	0.02	0.09	0.12	0.30
Huddersfield Town	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.06	0.04	0.07	0.10	0.07	0.15	0.10	0.10	0.10	0.08	0.07	0.01
Leicester City	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.02	0.02	0.04	0.02	0.03	0.04	0.16	0.08	0.22	0.14	0.19
Liverpool	0.02	0.37	0.22	0.19	0.09	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Manchester City	0.94	0.04	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Manchester United	0.01	0.09	0.15	0.16	0.25	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Newcastle United	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.10	0.06	0.09	0.14	0.04	0.09	0.04	0.07	0.02	0.09	0.03	0.05	0.07
Southampton	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.09	0.10	0.08	0.08	0.06	0.05	0.12	0.11	0.08	0.04	0.06	0.06
Tottenham Hotspur	0.01	0.12	0.19	0.33	0.20	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Watford	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.02	0.03	0.03	0.06	0.05	0.11	0.10	0.10	0.17	0.29
West Ham United	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.05	0.04	0.07	0.13	0.10	0.13	0.16	0.10	0.07	0.08	0.01	0.00
Wolverhampton Wanderers	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.19	0.14	0.09	0.04	0.04	0.07	0.03	0.02	0.06	0.09	0.12	0.07	0.00

e

$$\mathbf{x}_{tij}^{\mathsf{T}} = (\mathbf{b}_{tij}^{\mathsf{T}}, \mathbf{c}_{tij}^{\mathsf{T}}, \mathbf{d}_{tij}^{\mathsf{T}}),$$

vettori $(2N + m) \times 1$, con $N = 35$ numero totale di squadre distinte che hanno partecipato alla *Premier League* negli otto anni presi a riferimento. Ovviamente $\mathbf{b}_{tij}^{\mathsf{T}}$, $\mathbf{c}_{tij}^{\mathsf{T}}$ e $\mathbf{d}_{tij}$ vettori definiti esattamente come prima.

Le stime dei parametri sono riportate nelle tabelle: 4.6, 4.7 e 4.8.

Le interpretazioni sono analoghe alle precedenti, con la differenza che i parametri relativi al vantaggio di giocare in casa ripostati in Tabella 4.6 sono tutti positivi, eccetto quello per il *Southampton*, che come prima resta negativo. I parametri relativi all'abilità restano sempre negativi confrontati con il *Manchester City* e le stime delle covariate relative alla partita restano immutate in termini di interpretazione rispetto a prima.

Il modello in questo caso ha un tasso di errata classificazione leggermente migliore, cioè del 54.47%, ma la classifica derivante, che è riportata in Tabella 4.9, è meno precisa rispetto a quella precedente. Infatti le prime sei classificate sono ancora individuate correttamente, ma in ordine completamente sbagliato rispetto alla classifica reale. Per quanto riguarda le tre squadre che retrocedono, ancora una volta non sono individuate tutte dal modello. Rispetto al modello precedente ora si individua correttamente una squadra in zona retrocessione, cioè il *Cardiff* (che prima era prevista nelle zone medio-alte della classifica), mentre le altre due sono stimate a metà classifica. Tramite il *ranking* simulato con cento estrazioni dalle distribuzioni a posteriori dei parametri si coferma quanto visto dalla singola classifica. Si possono vedere nella Tabella 4.10 segnate in verde le squadre correttamente individuate e in arancione le altre che vengono stimate maggiormente in posizioni sbagliate rispetto a quella reale rimanendo sempre nelle prime sei posizioni della classifica. In rosso ancora una volta la frequenza maggiore relativa alle tre squadre ultime classificate in campionato.

Tabella 4.6: Stima dei parametri relativi al vantaggio di giocare in casa per ciascuna squadra del campionato 2018/2019 con relativi intervalli di credibilità mettendo assieme tutte le stagioni di stima.

	Stima vittoria	Stima pareggio
AFC Bournemouth	0.77	0.49
	(-0.36, 1.94)	(-0.48, 1.73)
Arsenal	1.19	0.69
	(0.33, 2.05)	(-0.09, 1.48)
Brighton	3.04	2.66
	(0.77, 5.27)	(0.65, 4.82)
Burnley	1.26	0.86
	(0.09, 2.38)	(-0.21, 1.9)
Cardiff	0.18	0.15
	(-1.87, 2.62)	(-1.95, 2.54)
Chelsea	0.67	0.46
	(-0.23, 1.45)	(-0.34, 1.16)
Crystal Palace	0.01	-0.43
	(-0.93, 0.95)	(-1.4, 0.39)
Everton	1.1	1.34
	(0.33, 1.8)	(0.74, 2)
Fulham	0.53	0.61
	(-0.42, 1.74)	(-0.35, 1.58)
Huddersfield	3.02	1.73
	(0.76, 5.33)	(-0.21, 3.6)
Leicester	1.27	0.71
	(0.1, 2.14)	(-0.23, 1.6)
Liverpool	0.92	0.78
	(0.25, 1.72)	(0.08, 1.51)
Manchester United	0.65	0.36
	(-0.13, 1.39)	(-0.34, 1.07)
Newcastle	1.31	0.51
	(0.6, 2.16)	(-0.23, 1.3)
Southampton	-0.13	0.23
	(-1, 0.62)	(-0.48, 1.02)
Tottenham	0.26	0.35
	(-0.63, 0.94)	(-0.42, 1.01)
Watford	1.57	0.33
	(0.28, 2.74)	(-0.88, 1.51)
West Ham	0.84	0.9
	(0.09, 1.56)	(0.13, 1.51)
Wolverhampton	1.49	0.74
	(-0.09, 2.96)	(-0.66, 2.13)

Tabella 4.7: Stima dei parametri relativi alla forza di ciascuna squadra del campionato 2018/2019 con relativi intervalli di credibilità mettendo assieme tutte le stagioni di stima.

	Stima vittoria	Stima pareggio
AFC Bournemouth	-2.61	-1.48
	(-3.59, -1.73)	(-2.24, -0.53)
Arsenal	-0.86	-0.46
	(-1.49, -0.2)	(-1.01, 0.15)
Brighton	-4.16	-2.42
	(-6.19, -2.26)	(-4.26, -0.64)
Burnley	-3.07	-1.66
	(-4.02, -1.99)	(-2.6, -0.68)
Cardiff	-2.9	-1.32
	(-4.48, -1.26)	(-2.85, 0.57)
Chelsea	-0.32	-0.18
	(-0.92, 0.31)	(-0.7, 0.4)
Crystal Palace	-2.39	-1.19
	(-3.27, -1.68)	(-1.92, -0.53)
Everton	-1.78	-1.45
	(-2.41, -1.16)	(-1.98, -0.9)
Fulham	-2.82	-1.74
	(-3.71, -1.89)	(-2.5, -0.94)
Huddersfield	-3.46	-2.06
	(-5.18, -1.64)	(-3.86, -0.66)
Leicester	-1.93	-0.78
	(-2.71, -1.06)	(-1.53, -0.1)
Liverpool	-1.24	-0.38
	(-1.86, -0.58)	(-0.97, 0.14)
Manchester United	-0.35	-0.44
	(-1.02, 0.25)	(-0.99, 0.03)
Newcastle	-2.71	-1.14
	(-3.31, -1.89)	(-1.8, -0.4)
Southampton	-1.43	-0.86
	(-2.15, -0.75)	(-1.47, -0.18)
Tottenham	-0.5	-0.37
	(-1.16, 0.07)	(-0.89, 0.16)
Watford	-2.64	-0.72
	(-3.68, -1.61)	(-1.68, 0.28)
West.Ham	-2.45	-1.53
	(-3.17, -1.81)	(-2.22, -0.93)
Wolverhampton	-2.57	-1.59
	(-3.76, -1.43)	(-2.64, -0.48)

Tabella 4.8: Stima parametri relativi alle covariate con relativi intervalli di credibilità mettendo assieme tutte le stagioni di stima.

	Stima vittoria	Stima pareggio
diff_clearances	0.08	0.03
	(0.07, 0.09)	(0.02, 0.04)
diff_corners	0.05	0.03
	(0.01, 0.09)	(-0.01, 0.07)
diff_fouls_conceded	-0.02	-0.01
	(-0.05, 0.01)	(-0.04, 0.01)
diff_offsides	0	-0.03
	(-0.05, 0.06)	(-0.08, 0.02)
diff_passes	0.02	0
	(0.02, 0.03)	(-0.01, 0.01)
diff_possession	-0.31	-0.12
	(-0.37, -0.26)	(-0.17, -0.07)
diff_red_card	-1.95	-1.07
	(-2.33, -1.61)	(-1.4, -0.77)
diff_shots	-0.05	-0.05
	(-0.07, -0.02)	(-0.07, -0.02)
diff_shots_on_target	0.57	0.29
	(0.52, 0.62)	(0.24, 0.33)
diff_tackels	0.02	0.02
	(0, 0.04)	(0.01, 0.04)
diff_touches	0.01	0.01
	(0, 0.01)	(0, 0.02)
diff_yellow_cards	-0.06	-0.03
	(-0.14, 0.01)	(-0.1, 0.04)

In conclusione il primo modello che sfrutta le informazioni derivanti dalla stima più aggiornata dei dati sembra stimare meglio la classifica finale, anche se in termini di tasso di errata classificazione è un pochino inferiore rispetto al secondo. In ogni caso questo sembra confermare l'utilità di modelli tempo-dipendente che aggiornano la stima dei parametri a frequenze maggiori rispetto all'intera stagione giocata.

Tabella 4.9: Classifica stimata con medie a posteriori delle distribuzioni di ciascun parametro, con il modello che mette assieme tutte le stagioni stimate.

	squadra	punti	Classifica reale
1	Manchester City	114	1°
2	Chelsea	106	3°
3	Arsenal	105	5°
4	Liverpool	99	2°
5	Manchester United	90	6°
6	Tottenham Hotspur	85	4°
7	Everton	76	8°
8	Newcastle United	74	13°
9	Southampton	64	16°
10	AFC Bournemouth	50	14°
11	Fulham	50	19°
12	Crystal Palace	45	12°
13	West Ham United	41	10°
14	Huddersfield Town	40	20°
15	Wolverhampton Wanderers	38	7°
16	Burnley	35	15°
17	Leicester City	24	9°
18	Cardiff City	22	18°
19	Watford	22	11°
20	Brighton and Hove Albion	17	17°

Tabella 4.10: Probabilità di posizione in classifica di ogni squadra facendo 100 simulazioni dalla distribuzione a posteriori con il modello che mette assieme tutte le stagioni di stima.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
AFC Bournemouth	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.06	0.12	0.21	0.16	0.06	0.07	0.11	0.04	0.04	0.05	0.00	0.00
Arsenal	0.03	0.44	0.37	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brighton and Hove Albion	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.07	0.06	0.08	0.07	0.09	0.13	0.16	0.17	0.10
Burnley	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.07	0.12	0.07	0.17	0.17	0.11	0.09	0.11	0.03
Cardiff City	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.02	0.03	0.04	0.02	0.05	0.04	0.07	0.13	0.11	0.12	0.34
Chelsea	0.00	0.48	0.49	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Crystal Palace	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.11	0.18	0.11	0.13	0.11	0.07	0.07	0.06	0.05	0.00	0.00	0.01
Everton	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.63	0.29	0.04	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fulham	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.15	0.16	0.12	0.18	0.16	0.10	0.06	0.02	0.01	0.02	0.00	0.00
Huddersfield Town	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.16	0.22	0.16	0.07	0.08	0.09	0.05	0.05	0.03	0.04	0.03	0.00
Leicester City	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.04	0.07	0.10	0.23	0.21	0.20	0.12
Liverpool	0.00	0.06	0.13	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Manchester City	0.97	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Manchester United	0.00	0.00	0.00	0.00	0.35	0.64	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Newcastle United	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Southampton	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.11	0.19	0.17	0.17	0.12	0.09	0.07	0.02	0.02	0.02	0.01	0.00	0.00
Tottenham Hotspur	0.00	0.00	0.00	0.00	0.65	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Watford	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.11	0.09	0.13	0.30	0.30
West Ham United	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.03	0.04	0.16	0.24	0.18	0.13	0.10	0.05	0.02	0.00
Wolverhampton Wanderers	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.10	0.11	0.10	0.12	0.14	0.06	0.13	0.05	0.10

Conclusioni

Il modello logistico multinomiale che fa uso di variabili esplicative relative a informazioni sulla gara prima che questa sia giocata è, sotto un certo punto di vista, una metodologia innovativa per la previsione dei risultati e, come si può vedere, porta a dei risultati discreti. Il tasso di errata classificazione del 53% (circa) è comunque considerevole se si pensa che un modello *baseline* che classifica in modo completamente casuale ha il 33.33% di probabilità di individuare correttamente il risultato.

Nonostante ciò, rimangono delle questioni aperte che si potranno trattare in futuro.

In primo luogo si possono pensare estensioni tempo dipendenti dei modelli proposti. Come si è più volte visto, la dipendenza temporale nei dati è riscontrabile e adattare modelli che tengono conto dell'aggiornamento dei parametri sulla base degli ultimi risultati può essere importante.

Un'altra estensione percorribile è l'utilizzo di un modello ordinale che tenga conto dell'ordinamento della variabile risposta. Questi modelli potrebbero essere più parsimoniosi in termini di parametri da stimare, ma bisognerebbe abbandonare il *data augmentation* tramite variabili *Pòlya-Gamma*, perché con questa metodologia non è possibile stimare modelli con queste ultime assunzioni.

Si potrebbe inoltre introdurre metodi bayesiani di selezione delle variabili, così da poter valutare, tramite opportuni test, quali siano le variabili più utili ai fini della previsione dei risultati.

Infine resta aperta la questione delle squadre che salgono nella massima categoria per la prima volta in assoluto nella loro storia. Questo evento è abbastanza raro ma non improbabile e in questi casi non si dispone di dati per stimare i parametri per fare le previsioni.

Bibliografia

Albert, J. H. e Chib, S. (1993). «Bayesian analysis of binary and polychotomous response data». *Journal of the american statistical association* 88(422), pp. 669–679.

All Premier League Matches 2010-2021. The most complete dataset of England Premier League! 4070 matches 113 features. https://www.kaggle.com/pablohfreitas/all-premier-league-matches-20102021.

Aslan, B. G. e Inceoglu, M. M. (2007). «A comparative study on neural network based soccer result prediction». *Seventh international conference on intelligent systems design and applications (isda 2007)*. IEEE, pp. 545–550.

Carpita, M., Ciavolino, E. e Pasca, P. (2019). «Exploring and modelling team performances of the kaggle european soccer database». *Statistical modelling* 19(1), pp. 74–101.

Fahrmeir, L. e Tutz, G. (1994). «Dynamic stochastic models for time-dependent ordered paired comparison systems». *Journal of the american statistical association* 89(428), pp. 1438–1449.

Fahrmeir, L. et al. (1994). *Multivariate statistical modelling based on generalized linear models*. Vol. 425. Springer.

Held, L. e Holmes, C. C. (2006). «Bayesian auxiliary variable models for binary and multinomial regression». *Bayesian analysis* 1(1), pp. 145–168.

Polson, N. G., Scott, J. G. e Windle, J. (2013). «Bayesian inference for logistic models using pólya–gamma latent variables». *Journal of the american statistical association* 108(504), pp. 1339–1349.

Salvan, A., Sartori, N. e Pace, L. (2020). «Modelli lineari generalizzati». *Modelli lineari generalizzati*. Springer, pp. 67–119.

Wunderlich, F. e Memmert, D. (2018). «The betting odds rating system: using soccer forecasts to forecast soccer». *Plos one* 13(6), e0198668.

Zens, G., Frühwirth-Schnatter, S. e Wagner, H. (2021). «Efficient bayesian modeling of binary and categorical data in r: the upg package». *Arxiv preprint arxiv:2101.02506*.

Immagini aggiuntive

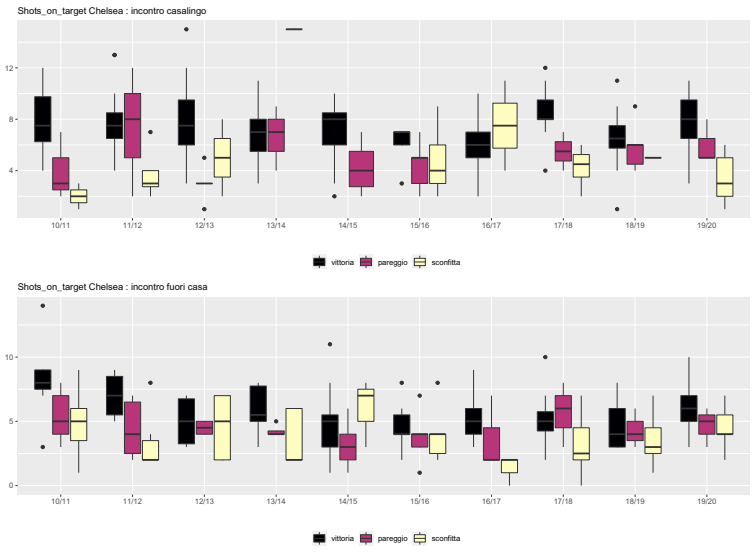


Figura 3: Numero di tiri in porta effettuati dalla squadra *Chelsea*, al variare delle stagioni.

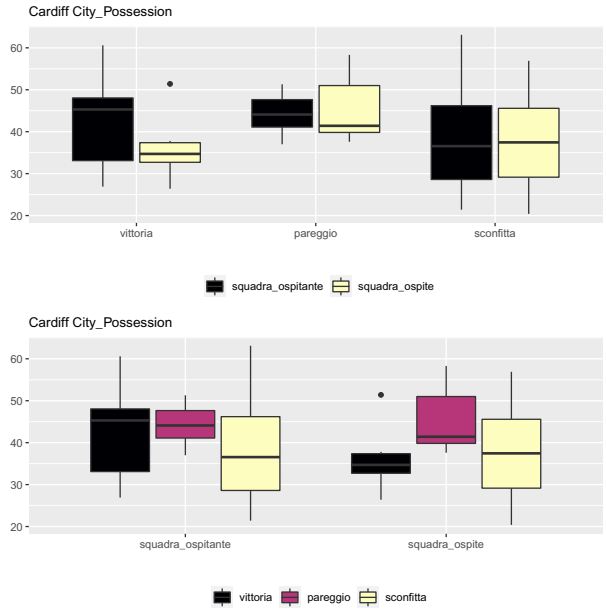


Figura 1: Percentuale di possesso del pallone della squadra *CardiffCity*, stratificato per casa/fuori casa e per le modalità della variabile risposta.

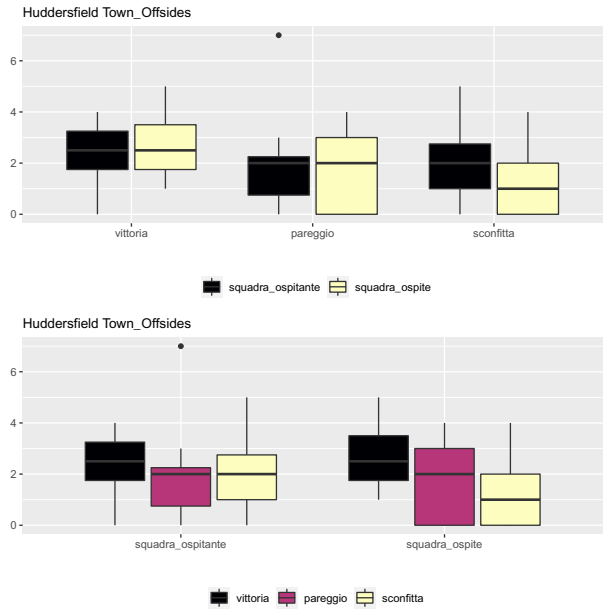


Figura 2: Numero di fuori gioco della squadra *Huddersfield*, stratificato per casa/fuori casa e per le modalità della variabile risposta.

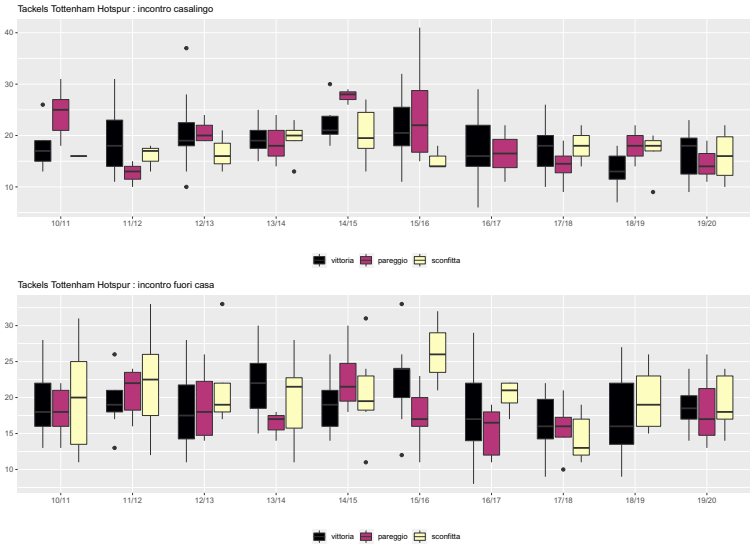


Figura 4: Numero di *tackels* effettuati dalla squadra *Tottenham Hotspur*, al variare delle stagioni.

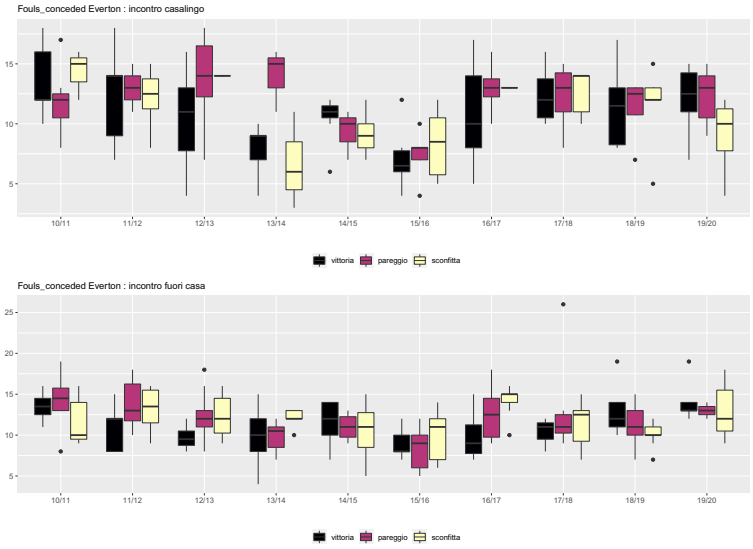


Figura 5: Numero di falli concessi dalla squadra *Everton*, al variare delle stagioni.