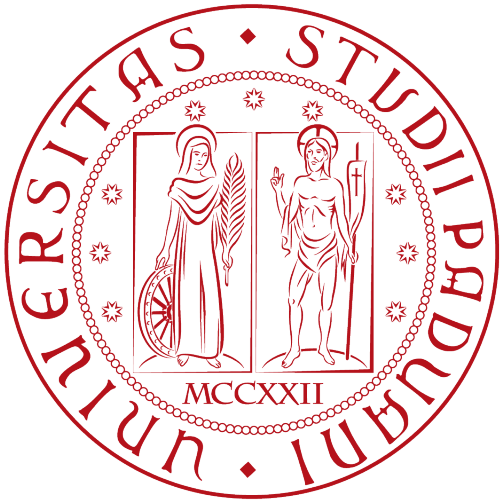
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea Magistrale in Informatica



Il modello Bradley-Terry per l'analisi delle partite della Serie A italiana di calcio

Tesi di laurea magistrale

Relatore
Prof. Annamaria Guolo

Laureando
Federico Perin

Anno Accademico 2022-2023

Abstract

Come sappiamo viviamo nell'era dei cosiddetti *Big Data*, dove grazie all'interconnessione; un grande flusso di informazioni e di dati può essere ricavato da ogni possibile attività.
Non fa eccezione il calcio in cui da un paio d'anni, le società calcistiche si affidano a sistemi di analisi per produrre tattiche di gioco ma anche per effettuare *scouting* di giocatori emergenti. Nel calcio moderno perciò, numerose variabili ad esempio il possesso palla, il numero di tiri effettuati da una squadra ecc. vengono raccolte durante una partita di calcio.
Tale fatto scaturisce l'attenzione su un ulteriore tematica d'analisi: dato che si hanno ha disposizione un gran numero di dati sulle prestazione delle squadre nelle loro partite, è possibile individuare quali variabili vanno ad influenzare in modo significativo il successo o il fallimento sportivo delle singole squadre?
Da questo quesito nasce la tesi qui presentata che ha come obbiettivo di presentare un'analisi che prova a rispondere a tale quesito, attraverso l'utilizzo di tecniche di *Data Mining*, in particolare lo sfruttamento di un modello a comparazione a coppie per le partite di calcio che sia in grado di tenere conto delle covariate specifiche per le partite. Nella nostra analisi tale modello sarà il *Bradley-Terry model*, il quale verrà esteso includendo possibili covariate significative e l'utilizzo di valori di risposta ordinati. Lo studio prenderà in considerazione i dati relativi alle partite della Serie A italiana della stagione 2021/2022.
TO DO + POSSIBLE ADDITIONS

"If something's important enough, you should try. Even if the probable outcome is failure."

— Elon Musk

Ringraziamenti

Innanzitutto, vorrei esprimere la mia gratitudine al Prof.Annamaria Guolo, relatrice della mia tesi, per l'aiuto ed il sostegno fornitomi durante tutto il lavoro.

Desidero ringraziare con affetto i miei genitori per il sostegno, per il grande aiuto che mi hanno dato e per essermi stati vicini in ogni momento durante gli anni di studio.

Voglio inoltre ringraziare i miei amici per questi tre bellissimi anni trascorsi assieme e per avermi sempre sostenuto anche nei momenti più difficili.

Padova, Febbraio 2023 Federico Perin

Indice

Elenco delle figure

Elenco delle tabelle

1 | Introduzione

MEMO: Spiegazione del problema affrontato (il suo dominio) alcune applicazioni fatte nell'ambito delle comparazioni sportive, con maggior attenzione a qui studi con approccio statistico, esporre tecnologie usate e tools (Packages R ecc), motivazione scelta argomento della tesi e esposizione struttura della tesi(capitoli) TO DO

1.1 Dominio del problema

1.2 Applicazione

1.3 Tecnologie e Tools usati

1.3.1 Tecnologie

1.3.2 Tools

1.4 Motivazioni personali

1.5 Struttura della tesi

2 | Serie A 2021/2022 dataset

Nel seguente capitolo verrà descritta la raccolta dati effettuata per costruire il dataset riguardante le partite di calcio della Serie A italiana della stagione 2021/2022 e la struttura di tale dataset.

2.1 Serie A 2021/2022

L'analisi effettuata ha preso in considerazione le partite della Serie A italiana della stagione 2021/2022. La Serie A è un torneo che comprende 20 squadre sparse per tutta l'Italia, alcune anche della stessa città, come ad esempio Milan e Inter per Milano.
Tale torneo è organizzato con una struttura Double-Round-Robin, dove ogni squadra affronta due volte le altre 19 avversarie del torneo. Vi è quindi una partita di andata e una di ritorno. In base al sorteggio necessario alla creazione del calendario delle partite si decide quale delle due partite sarà giocata in casa oppure fuori casa (in casa dell'avversario).
Il torneo della stagione 2021/2022 è iniziato il 22 Agosto con Inter - Genoa e si è concluso il 22 Maggio con le partite Salernitana - Udinese e Venezia - Cagliari, per un totale 380 partite giocate, suddivise in 38 turni, ciascuno composto da 10 partite.

2.1.1 Ranking

Le squadre di calcio sono classificate in base all'ordine dei punti che hanno totalizzato al termine della stagione. In un torneo calcistico, per ogni partita, la squadra vincitrice guadagna tre punti, la squadra sconfitta guadagna un punto, mentre, in caso di pareggio, entrambe le squadre guadagnano un punto. Nel torneo della Serie A chi guadagna più punti vince il campionato, mentre chi si classifica tra le ultime tre retrocede alla lega inferiore, la Serie B. Il posto delle tre squadre retrocesse verrà preso da tre squadre della Serie B che hanno guadagnato la promozione alla Serie A.
La classifica della stagione 2021/2022 è riportata nella Tabella 2.1.

2.2 Costruzione del dataset

Al giorno d'oggi, nelle partite di calcio professionistico viene raccolta un'enorme quantità di variabili. Ad esempio, per ogni squadra è noto il tempo in percentuale del possesso della palla e il numero di tiri in porta in una determinata partita. L'obbiettivo principale di questo lavoro è determinare l'influenza che queste variabili hanno sull'esito della partita.
A tale scopo, sono state raccolte un gran numero di variabili che si suppone essere associate all'esito della partita.
Tali dati sono stati offerti dal sito web FBref(`https://fbref.com`), un sito web dedicato al tracciamento delle statistiche relative ai calciatori e alle squadre di calcio

Posizione	Squadra	Punti	% casa
1	Milan	86	0.47
2	Inter	84	0.54
3	Napoli	79	0.46
4	Juventus	70	0.50
5	Lazio	64	0.56
6	Roma	63	0.57
7	Fiorentina	62	0.66
8	Atalanta	59	0.33
9	Hellas Verona	53	0.57
10	Torino	50	0.58
11	Sassuolo	50	0.48
12	Udinese	47	0.53
13	Bologna	46	0.61
14	Empoli	41	0.42
15	Sampdoria	36	0.58
16	Spezia	36	0.50
17	Salernitana	31	0.48
18	Genoa	30	0.50
19	Cagliari	28	0.61
20	Venezia	27	0.52

Tabella 2.1: La tabella mostra i punti guadagnati da ogni squadra con il loro piazzamento. Inoltre viene mostrata la percentuale di punti guadagnati in casa.

di tutto il mondo. FBref mette a disposizione i dati sotto forma di tabelle che possono essere modificate per mantenere solo i dati di nostro interesse.

Dunque, per ogni squadra che ha partecipato alla stagione 2021/2022 di Serie A, sono state esportate le variabili di interesse per ogni partita giocata, selezionando le macro aree opportune e adattando le tabelle per ottenere solo i dati utili. Le varie tabelle hanno composto un file Excel divenuto il dataset per le analisi svolte nelle tesi

Figura 2.1: Logo di FBref.
Source: `https://fbref.com`

2.3 Struttura del dataset

Il dataset risultante dalla raccolta dati è composto da 760 righe e 35 colonne. Ogni riga riguarda una specifica partita di calcio giocata dalla squadra indicata nella colonna Team contro la squadra indicata nella colonna Vs. Ogni riga contiene informazioni riguardati solo la squadra indicata in Team fatta eccezione per la data della partita (Date), il turno (Round) e gli spettatori (Spec). Quindi, per ogni partita esistono due righe, una per ciascuna squadra coinvolta. Come risultato finale, ogni squadra appare nella colonna Team 38 volte e, siccome il numero totale di squadre è 20, si ottengono 760 righe. La Tabella 2.2 mostra un breve estratto dei dati riguardanti le prime tre partite della stagione.

Date	**AtHome**	**Res**	**GF**	**GA**	**Team**	**Vs**	**Poss**	**...**
21/08/2021	TRUE	1	4	0	Inter	Genoa	0,59	...
...	...	...	...	...	...	...	...	...
22/08/2021	TRUE	1	2	0	Napoli	Venezia	0,56	...
...	...	...	...	...	...	...	...	...
23/08/2021	FALSE	1	1	0	Milan	Sampdoria	0,51	...
...	...	...	...	...	...	...	...	...
21/08/2021	FALSE	-1	0	4	Genoa	Inter	0,41	...
...	...	...	...	...	...	...	...	...
22/08/2021	FALSE	-1	0	2	Venezia	Napoli	0,44	...
...	...	...	...	...	...	...	...	...
23/08/2021 1	TRUE	1	0	1	Sampdoria	Milan	0,49	...
...	...	...	...	...	...	...	...	...

Tabella 2.2: La tabella mostra un estratto del dataset utilizzato i cui dati sono stati ricavati da FBref.

Come scritto precedentemente all'interno del dataset sono presenti 35 colonne. Oltre

alle già citate Date, Round e Spec che hanno solo un valore di completezza dei dati, le restanti 32 colonne sanno le possibili variabili che possono influenzare l'esito della partita. Le covariate sono state raggruppate nelle seguenti cinque macro-aree:

* dati generali,
* dati relativi ai tiri,
* dati possesso,
* dati passaggi,
* dati difensivi,

che sono illustrate di seguito.

2.3.1 Dati generali

In questo gruppo sono presenti le variabili legate a statistiche che non fanno parte di una precisa macro-area ma che descrivono più genericamente la partita giocata.
Le possibili covariate sono le seguenti:

* AtHome: indica se la squadra specificata della variabile Team gioca nel proprio stadio, quindi in casa oppure fuori casa. Per indicare se la squadra gioca in casa viene messo come valore `TRUE` altrimenti `FALSE`.

  Come mostrato nella terza colonna della Tabella 2.1, la quale indica in percentuale quante partite sono state vinte in casa per ogni singola squadra, ci sono 11 squadre che hanno avuto un leggero vantaggio nel giocare in casa le partite di calcio rispetto a altre sei squadre che hanno avuto l'effetto opposto, mentre le rimanti tre hanno avuto un effetto nullo.

* Res: indica se la squadra specificata della variabile Team ha vinto, pareggiato o perso la partita. Per indicare se ha vinto viene inserito il valore 1, se ha pareggiato 0, altrimenti se ha perso -1. Res sarà la variabile risposta.

* GF: indica il numero di gol fatti dalla squadra specificata della variabile Team.

  È stata inserita perché può permettere di valutare la qualità della fase offensiva della squadra e quindi ci si aspetta che possa essere utile ai fini dell'analisi.

* GA: Indica il numero di gol subiti dalla squadra specificata della variabile Team e quindi fatti dalla squadra indicata nella variabile Vs.

  Essa può essere utile perché subire pochi gol incide positivamente sull'esito della partita, limitando l'esposizione della squadra ad uno sbilanciamento in attacco per recuperare lo svantaggio e quindi rischiando maggiormente di subire ulteriori gol dagli avversari. Inoltre, è un fatto riconosciuto che aver la miglior difesa del campionato è associato ad una maggiore probabilità di vittoria del campionato.

* Team: indica il nome della squadra a cui i dati della riga fanno riferimento.

* Vs: indica il nome della squadra avversaria.

* Fls: indica il numero di falli fatti dai giocatori della squadra specificata della variabile Team.

  Questa variabile è stata inserita per capire se una squadra adotta un gioco più fisico/tattico. In questo caso sarà più propensa a interrompere il gioco della squadra avversaria e a commettere più falli. Si vuole perciò capire come questa variabile possa essere associata all'esito della partita, ricordando però che una squadra che commette molti falli è più soggetta a ricevere cartellini gialli o rossi che condizionano la prestazione dei giocatori.

* Fld: indica il numero di falli subiti ai giocatori della squadra specificata della variabile Team da parte della squadra avversaria specificata della variabile Vs.

  Si è deciso di inserire questa covariata perché un alto numero di falli può portare a molte interruzione della manovra di gioco e quindi permettere alla squadra avversaria di riorganizzarsi.

* Off: indica il numero di volte che la squadra specificata della variabile Team è finita in fuorigioco. Un calciatore si trova in posizione di fuorigioco quando una qualsiasi parte del suo corpo, fatta eccezione per braccia e mani, si trova nella metà campo avversaria ed è più vicina alla linea di porta avversaria, sia rispetto al pallone che rispetto al penultimo giocatore difendente avversario, portiere compreso nel caso in cui un compagno di questi è più vicino alla linea di porta. Una rappresentazione grafica del fuorigioco è mostrata nella Figura 2.2.

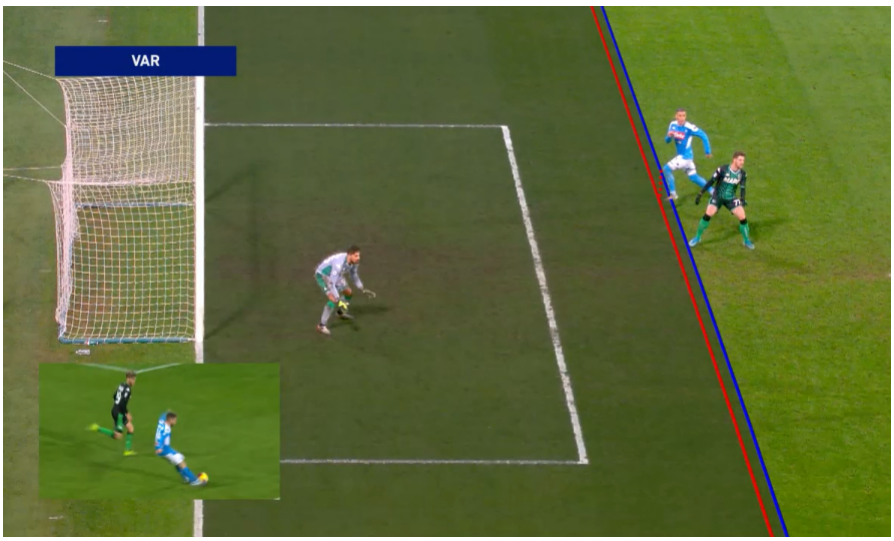


Figura 2.2: Rappresentazione del fuorigioco
Source: `https://sport.sky.it/calcio/2021/10/05/fifa-figc-var-fuorigioco`

  È stata inserita perché, se una squadra viene colta molte volte in fuorigioco allora il suo gioco sarà interrotto con vantaggio della squadra avversaria che farà ripartire la sua azione a proprio favore.

2.3.2 Dati relativi ai tiri

In questo gruppo sono presenti le variabili collegata alla fase offensiva della squadra in esame.

* Sh: indica il numero di tiri totali fatti dalla squadra specificata della variabile Team. Quindi vengono conteggiati il numero di tiri in porta più i tiri fuori dalla porta.

  Una squadra che effettua tanti tiri ha più probabilità di segnare un gol. Occorre però capire quanto è precisa una squadra nel centrare la porta.

* SoT: Indica il numero di tiri in porta totali fatti dalla squadra specificata della variabile Team.

  Una squadra con un alto valore di tiri in porta è più probabile che possa segnare un gol. `SoT` permette di capire quanto è precisa in combinazione con `Sh` la squadra di calcio nel centrare la porta.

* G/Sh: indica la proporzione tra gol e tiri fatti dalla squadra specificata della variabile Team.

  Questo può permettere di capire quanto la produzioni di tiri della squadra è efficace o meno. Con `Sh` e `SoT` si riesce a valutare quanto sia offensiva la squadra, cioè se essa gioca costantemente in attacco o utilizza la tattica "difesa e contropiede". Inoltre permette di capire quanto la squadra sia precisa nell'effettuare i tiri in porta.

2.3.3 Dati relati al possesso

In questo gruppo sono contenute le variabili collegate al possesso della palla

* Poss: indica la quantità di tempo (in percentuale) di possesso palla durante una partita di calcio per la squadra specificata della variabile Team. Nel gioco del calcio, con il termine “possesso palla” si intende un'azione manovrata di due o più giocatori che riescono a passarsi la palla evitando i contrasti degli avversari. Durante la partita, ogni volta che una squadra ha il dominio della palla si dice che questa squadra è in fase di “possesso palla”, quindi in questa variabile viene indicato quanto questa fase è durata nell'intera partita.
  Il metodo più comune utilizzato per calcolare il possesso palla di una squadra si basa sull'utilizzo di tre cronometri, uno per ciascuna formazione più uno per i tempi morti. Quando un giocatore della squadra A tocca un pallone che prima era in possesso della squadra B, il cronometro della squadra A parte e quello della squadra B si ferma e così via. Il terzo cronometro registra il tempo in tutte le situazioni di palla inattiva, ad esempio, rimesse laterali, calci di punizione ecc.. I tempi vengono poi trasformati in percentuali. Per una registrazione più sofisticata, si possono utilizzare ventidue cronometri, uno per ogni giocatore.

  La variabile è stata inserita perché, la supremazia nel possesso palla è solitamente desiderabile e utile, dati i seguenti vantaggi:

  – spingere l'avversario a muoversi verso la palla per allontanarlo dalla difesa della propria porta per poi sorprenderlo negli spazi lasciati incustoditi.

– modulare il ritmo della gara, ad esempio, se una squadra sta vincendo con un gol di scarto, "congela" il risultato mantenendo il possesso della palla in modo da non ricevere attacchi da parte della squadra avversaria.

Il possesso palla però non garantisce la vittoria. Produrre un possesso palla "sterile", cioè senza che questo porti alla produzioni di azioni offensive, può esporre la squadra in possesso della palla a contropiedi nel caso in cui la palla venga persa e quindi all'alto rischio di subire gol perché sbilanciata e non ben posizionata. Vedremo di seguito quali variabili possono essere utili per capire se il possesso palla fatto dalla squadra è "sterile" oppure no.

* ToDefPen: indica il numero di tocchi fatti dai giocatori della squadra specificata della variabile Team nella propria area di rigore.

  Questa variabile è stata inserita perché può essere utile per capire come venga gestito il possesso della palla. Se vi è un alto numero di tocchi, vuol dire che la squadra subisce molto la pressione della squadra avversaria, viceversa cerca di fare un gioco più offensivo. Questa variabile, in combinazione con le variabili ToDef3rd, ToMid3rd, ToAtt3rd e ToAttPen permette di capire se il possesso della palla fatto della squadra sia utile e porti benefici ai fini del risultato oppure sia sterile. Inoltre, si vuole capire in che misura come ToDefPen influenza il risultato della partita con un alto o un basso valore di numero di tocchi nella propria area di rigore, la cui area è indicata nella Figura 2.3.

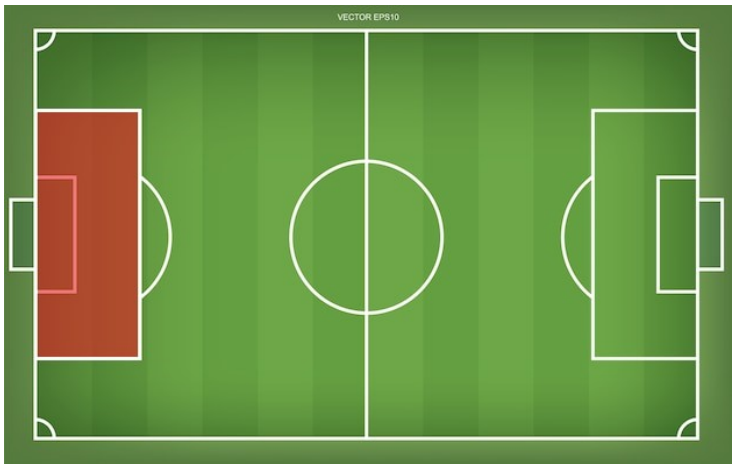


Figura 2.3: In rosso l'area di rigore in un campo da calcio.
Source: `https://it.freepik.com/foto-vettori-gratuito/campo-da-calcio`

* ToDef3rd: indica il numero di tocchi fatti dai giocatori della squadra specificata della variabile Team nella propria mediana o trequarti difensiva.

  Questa variabile è stata inserita perché può essere utile per capire come venga gestito il possesso della palla. Se vi è un alto numero di tocchi, vuol dire che la squadra cerca di mantenere il possesso palla creando poche azioni offensive, viceversa cerca di fare un gioco più offensivo. Questa variabile, in combinazione con ToDefPen, ToMid3rd, ToAtt3rd e ToAttPen, permette di capire se il possesso della palla fatto della squadra sia utile e porti benefici ai fini del risultato oppure

sia sterile. Inoltre, si vuole capire in che misura ToDef3rd influenza il risultato della partita con un alto o un basso valore di numero di tocchi nella propria mediana la cui area, è indicata nella Figura 2.4.

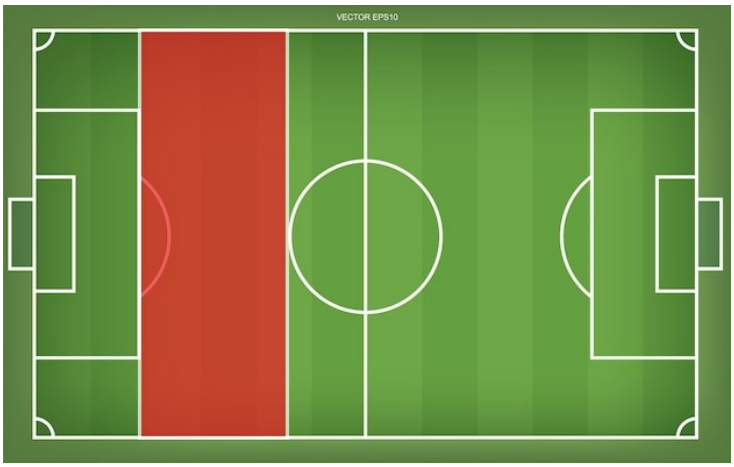


Figura 2.4: In rosso la mediana nel campo da calcio.
Source: `https://it.freepik.com/foto-vettori-gratuito/campo-da-calcio`

* ToMid3rd: indica il numero di tocchi fatti dai giocatori della squadra specificata della variabile Team a centrocampo.

  Questa variabile è stata inserita perché può essere utile per capire come venga gestito il possesso della palla. Se vi è un alto numero di tocchi, vuol dire che la squadra cerca di mantenere il possesso palla cercando di creare delle azioni offensive, viceversa cerca di fare un gioco più difensivo. Questa variabile, in combinazione con le variabili ToDefPen, ToDef3rd, ToAtt3rd e ToAttPen, permette di capire se il possesso della palla fatto dalla squadra sia utile e porti benefici ai fini del risultato oppure sia sterile. Inoltre, si vuole capire in che misura ToMid3rd influenza il risultato della partita con un alto o un basso valore di numero di tocchi a centrocampo la cui area, è indicata nella Figura 2.5.

* ToAtt3rd: indica il numero di tocchi fatti dai giocatori della squadra specificata della variabile Team a nella trequarti dell'avversario.

  Questa variabile è stata inserita perché può essere utile per capire come venga gestito il possesso della palla. Se vi è un alto numero di tocchi, vuol dire che la squadra cerca di mantenere il possesso palla per effettuare una pressione sulla squadra avversaria affinché si possano creare degli spazi per delle azioni offensive, viceversa cerca di fare un gioco molto più difensivo. Questa variabile, in combinazione con le variabili ToDefPen, ToDef3rd, ToMid3rd e ToAttPen, permette di capire se il possesso della palla fatto della squadra sia utile e porti benefici ai fini del risultato oppure sia sterile. Inoltre, si vuole capire in che misura ToAtt3rd influenza il risultato della partita con un alto o un basso valore di numero di tocchi nella trequarti dell'avversario la cui area, è indicata nella Figura 2.6.

* ToAttPen: indica il numero di tocchi fatti dai giocatori della squadra specificata della variabile Team a nell'area di rigore dell'avversario.

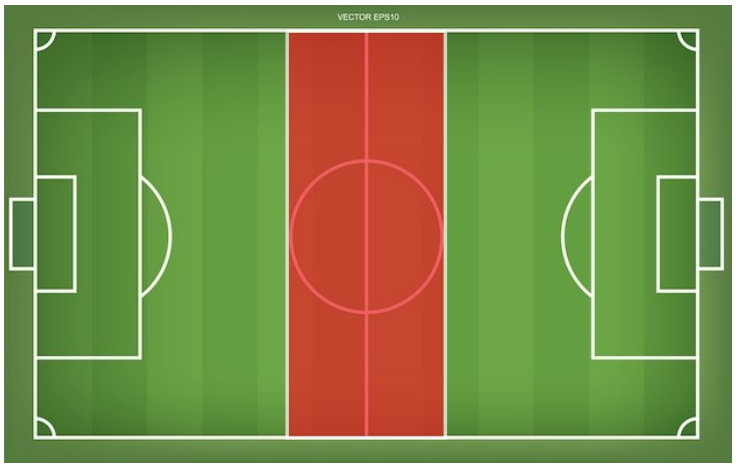


Figura 2.5: In rosso il centrocampo nel campo da calcio.
Source: `https://it.freepik.com/foto-vettori-gratuito/campo-da-calcio`

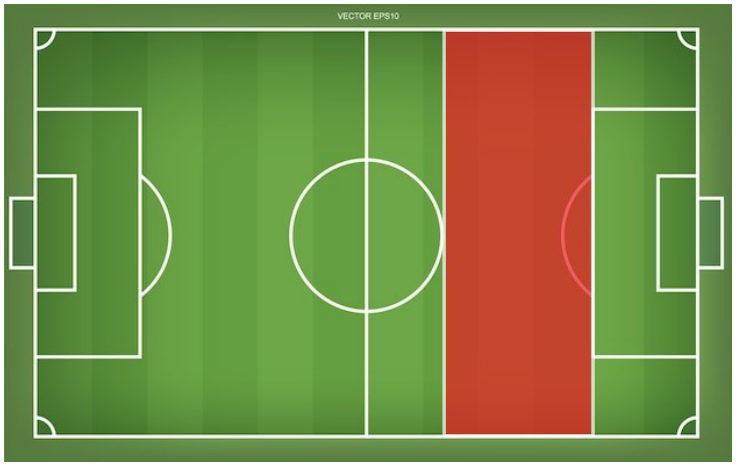


Figura 2.6: In rosso la trequarti dell'avversario nel campo da calcio.
Source: `https://it.freepik.com/foto-vettori-gratuito/campo-da-calcio`

Questa variabile è stata inserita perché può essere utile per capire come venga gestito il possesso della palla. Se vi è un alto numero di tocchi, vuol dire che la squadra cerca di mantenere il possesso palla applicando un'alta pressione sulla squadra avversaria affinché si possano creare molte occasioni da gol in area, viceversa o la squadra subisce troppo la pressione dell'avversario oppure tende ad avere un gioco molto difensivo. Questa variabile, in combinazione con le variabili ToDefPen, ToDef3rd, ToMid3rd e ToAtt3rd permette di capire se il possesso della palla fatto della squadra sia utile e porti benefici ai fini del risultato oppure sia sterile. Inoltre, si vuole capire in che misuraToAttPen influenza il risultato della partita con un alto o un basso valore di numero di tocchi nell'area di rigore dell'avversario.

* ToDist: Indica la distanza totale, espressa in metri, in cui un giocatore della squadra specificata della variabile Team si è mosso con la palla in qualsiasi direzione, controllandola con i piedi.

  Questa variabile è stata inserita perché permette di comprendere se il possesso della palla sia stato statico, ovvero i giocatori si sono mossi poco senza avanzare, oppure no. Sarà di interesse analizzare se un alto valore di metri percorsi con palla al piede possa essere utile ad ottenere la vittoria.

2.3.4 Dati relativi ai passaggi

In questo gruppo vi sono raggruppate le variabili collegate ai passaggi della palla.

* PAtt: Indica il numero di tutti i passaggi tentati dai giocatori della squadra specificata della variabile Team.

  Utile a capire quanto la squadra sia incline a tentare i passaggi.

* PCmp%: Indica la percentuale di passaggi riusciti ai giocatori della squadra specificata della variabile Team.

  È stata inserita perché permette di capire quanti passaggi siano andati a buon fine tra tutti quelli tentati e quindi la precisione dei giocatori della squadra.

* SPAtt: Indica il numero di passaggi corti tentati dai giocatori della squadra specificata della variabile Team. Per passaggi corti si intendono tutti quelli effettuati all'interno di una lunghezza tra i tre e quattordici metri.

  È stata inserita per capire se un alto numero di passaggi corti possa essere determinanti ai fini dell'esito della partita.

* SPCmp%: Indica la percentuale di passaggi corti riusciti ai giocatori della squadra specificata della variabile Team.

  È stata inserita perché permette di capire quanti passaggi andati a buon fine tra tutti quelli tentati e quindi la precisione dei giocatori della squadra.

* MPAtt: Indica il numero di passaggi medi tentati dai giocatori della squadra specificata della variabile Team. Per passaggi medi si intendono tutti quelli effettuati all'interno di una lunghezza tra i tredici e ventisette metri. Questi passaggi possono essere considerati come passaggi filtranti, cioè non diretti al proprio compagno di squadra ma verso un area del campo dove il compagno di squadra deve andare a prendere la palla. Spesso questi passaggi vengono fatti per sorprendere la difesa avversaria e evitare che la palla venga intercettata. Nella Figura 2.7 viene mostrato l'esecuzione di un passaggio filtrante.

  È stata inserita per capire se un alto numero di passaggi medi possa essere determinante ai fini dell'esito della partita.

* MPCmp%: Indica la percentuale di passaggi medi riusciti ai giocatori della squadra specificata della variabile Team.

  È stata inserita perché permette di capire quanti passaggi siano andati a buon fine tra tutti quelli tentati e quindi la precisione dei giocatori della squadra.

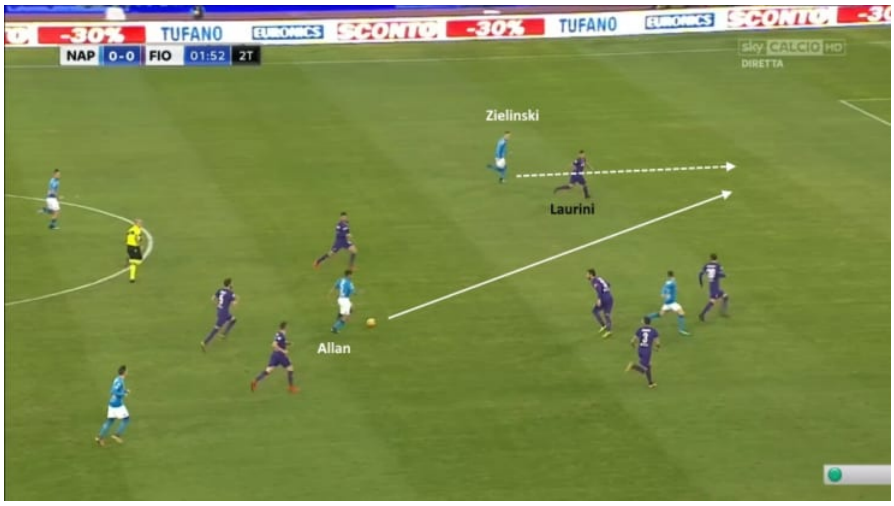


Figura 2.7: Esecuzione di un passaggio filtrante
Source: `https://www.ilmisterone.com/2019/01/16/passaggi-filtranti/`

∗ LPAtt: Indica il numero di passaggi lunghi tentati dai giocatori della squadra specificata della variabile Team. Per passaggi lunghi si intendono tutti quelli effettuati all'interno di una lunghezza superiore ai ventisette metri. Questi passaggi possono essere considerati come lanci lunghi per cambi di gioco o per lanciare le punte, cioè i giocatori che giocano come attaccanti, in profondità. Una rappresentazione di passaggio lungo è mostrata nella Figura 2.8.

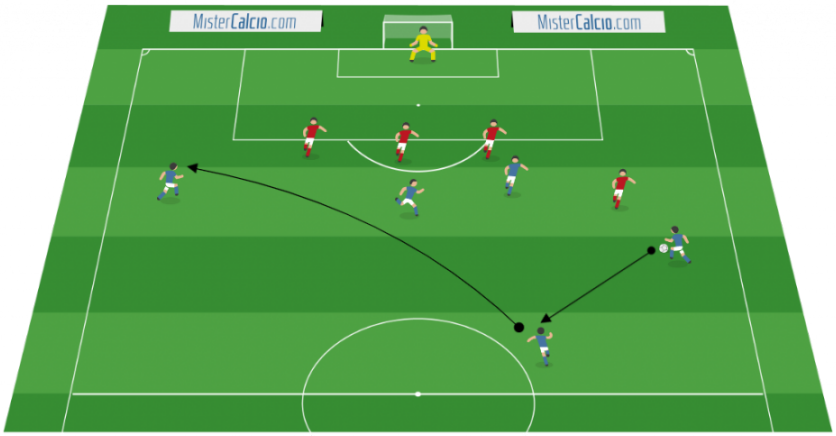


Figura 2.8: Esecuzione di un cambio di gioco
Source: `https://www.mistercalcio.com/tattica/il-cambio-di-gioco/`

È stata inserita per capire se un alto numero di passaggi lunghi possa essere determinante ai fini dell'esito della partita.

* LPCmp%: Indica la percentuale di passaggi lunghi riusciti ai giocatori della squadra specificata della variabile Team.

  È stata inserita perché permette di capire quanti passaggi sono andati a buon fine tra tutti quelli tentati e quindi qual'è la precisione dei giocatori della squadra.

* Crs: Indica il numero di cross effettuati dalla squadra specificata della variabile Team. Un cross (in italiano traversone) è un tipo di passaggio medio o lungo, solitamente effettuato sulle fasce laterali dell'area avversaria o comunque vicino all'area avversaria, che permette al compagno di squadra posizionato vicino alla porta avversaria di colpire la palla al volo di testa oppure di piede per segnare un possibile gol. Quindi, se eseguito correttamente, il cross può diventare un assist, cioè l'ultimo passaggio per la realizzazione del gol.

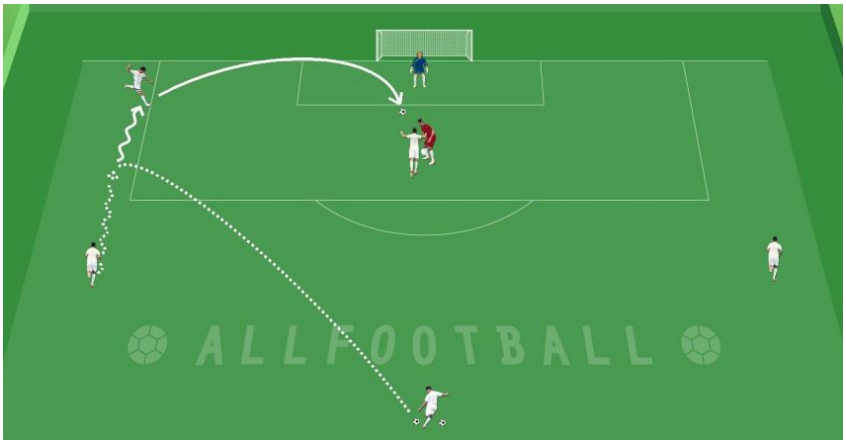


Figura 2.9: Rappresentazione di un cross
Source: `http://www.allfootball.it/blog/calcio-vincere-allenando-i-dettagli/27-2-2017/calcio-la-marcatura-a-uomo-su-cross-laterale/`

Una rappresentazione di cross è mostrata nella Figura 2.9.

2.3.5 Dati difensivi

In questo gruppo sono contenute le variabili collegate alla fase difensiva.

* Saves: Indica il numero di parate fatte del portiere della squadra specificata della variabile Team.

  È stata inserita perché permette di valutare se la squadra subisce tanti tiri dagli avversari, cosi come la qualità del portiere nel salvare la squadra da un possibile gol subito.

* Int: Indica il numero di intercettazioni fatte dai giocatori della squadra specificata della variabile Team. Per intercettazione della palla si intende l'intercettazione di un passaggio della squadra avversaria entrando in possesso del pallone andando ad interrompere il passaggio avversario.

* TklWin: Indica il numero di contrasti vinti dai giocatori della squadra specificata della variabile Team. Per contrasto si intende il tentativo da parte di un giocatore

difendente di sottrarre il possesso della palla all'avversario. Quindi chi ha in possesso la palla viene attaccato da chi ne è privo. Se si riesce a prendere il pallone all'avversario allora si avrà vinto il contrasto. I contrasti vengono effettuati anche per allontanare l'avversario dalle zone pericolose. La Figura 2.10 mostra un contrasto di gioco.

Figura 2.10: Rappresentazione di un contrasto in scivolata
Source: `https://www.ilmisterone.com/2022/01/24/partita-solo-tackle/`

Visto che tale intervento senza palla modifica il gioco dell'avversario, si è deciso di inserire i contrasti vinti come variabile.

* Recov: Indica il numero di palle vaganti recuperate dalla squadra specificata della variabile Team. Per palle vaganti si intendono quei palloni che, a seguito di un contrasto di gioco, non sono stati recuperati dalla squadra che ha effettuato il contrasto ma chi ha subito il contrasto, ne ha comunque perso il controllo. Quindi nessuno ha in possesso il pallone e la palla viene detta vagante.

  Dato che questa variabile sembra essere legata al possesso del pallone, potrebbe essere interessante per l'analisi.

Nella Tabella 2.3 è riassunto l'insieme delle variabili presenti e le loro macro-aree di appartenenza.

Di seguito nella Tabella 2.4 è mostrato per ogni variabile il nome che ha all'interno del dataset.

Statistiche generali	**Tiri**	**Possesso**	**Passaggi**	**Difensive**
AtHome	Sh	Poss	PAtt	Saves
Res	SoT	ToDefPen	PCmp%	Int
GF	G/Sh	ToDef3rd	SPAtt	TklWin
GA		ToMid3rd	SPCmp%	Recov
Team		ToAtt3rd	MPAtt	
VS		ToAttPen	MPCmp%	
Fls		ToDist	LPAtt	
Fld			LPCmp%	
Off			Crs	

Tabella 2.3: La tabella riassuntiva variabili presenti nel dataset.

Originale	Rinominate
AtHome	AtHome
Res	Res
GF	GF
GA	GF
Team	Team
VS	Vs
Poss	Poss
Sh	Sh
SoT	SoT
G/Sh	G.Sh
Saves	Saves
PAtt	PAtt
PCmp%	PCmp.
SPAtt	SPAtt
SPCmp%	SPCmp.
MPAtt	MPAtt
MPCmp%	MPCmp.
LPAtt	LPAtt
LPCmp%	LPCmp.
ToDefPen	ToDefPen
ToDef3rd	ToDef3rd
ToMid3rd	ToMid3rd
ToAtt3rd	ToAtt3rd
ToAttPen	ToAttPen
ToDist	ToDist
Fls	Fls
Fld	Fld
Off	Off
Crs	Crs
Int	Int
TklWin	TklWin
Recov	Recov

Tabella 2.4: Tabella corrispondenza nomi originali e nomi nel dataset

3 | Analisi dei dati

Nel seguente capitolo verrà illustrata la fase di preprocessing e le analisi grafiche dei dati. Le analisi verranno svolte usando il linguaggio di programmazione di (R Core Team, 2022).

3.1 Preprocessing dei dati

Dopo aver importato il dataset utilizzando il linguaggio di programmazione R (R Core Team, 2022), il primo step da effettuare durante il prepocessing è individuare e risolvere possibili anomalie nei dati. Il dataset è stato importato in modo che la prima riga contenga l'intestazione, mentre le restanti righe tutte le osservazioni. Il comando usato per importare il dataset è il seguente:

```
> soccer<-read.xlsx("SerieA.xlsx", 1, header=TRUE)
```

Il dataset non ha valori mancanti. Questo è stato possibile grazie a FBref che ha messo a disposizione dati quasi sempre completi; in quei rari casi di mancanza di dati sono stati reperiti manualmente da altre fonti altrettanto attendibili.
Sono state inoltre tolte le variabili `Date` e `Round`.
Il passo successivo è stato controllare che le variabili fossero interpretate correttamente. `Team` e `Vs` vengono interpretate erroneamente come tipo `character`. `Team` e `Vs` devono essere interpretate come un fattore cioè è un valore non numerico, espresso in termini verbali, ad esempio una categoria; quindi ogni squadra sarà un livello del fattore. Analogamente, `AtHome` è stata fatta trasformata in un fattore a due livelli. Invece, `Res` è stata trasformata in un fattore ordinato con i livelli: -1 = sconfitta $<$ 0 = pareggio $<$ 1 = vittoria.

3.2 Analisi grafica dei dati

In questa sezione attraverso il supporto di grafici, si analizzerà graficamente i dati disponibili e le loro relazione per avere una prima visione dei dati raccolti. Si valuteranno le relazione tra covariate e la variabile di risposta, le relazioni tra due covariate. Tutto ciò per individuare quali covariate possano essere significative per la variabile risposta e quali interazioni emergono dall'analisi grafica.

Come primo passo, è stata valutata la distribuzione della variabile risposta `Res`, come è mostrato in Figura 3.1.
Si può notare come le classi sembrino ben distribuite, dato che abbiamo 196 pareggi e 282 vittorie e altrettante sconfitte. Si ha quindi un campione abbastanza ampio, distribuito e privo di classi povere.

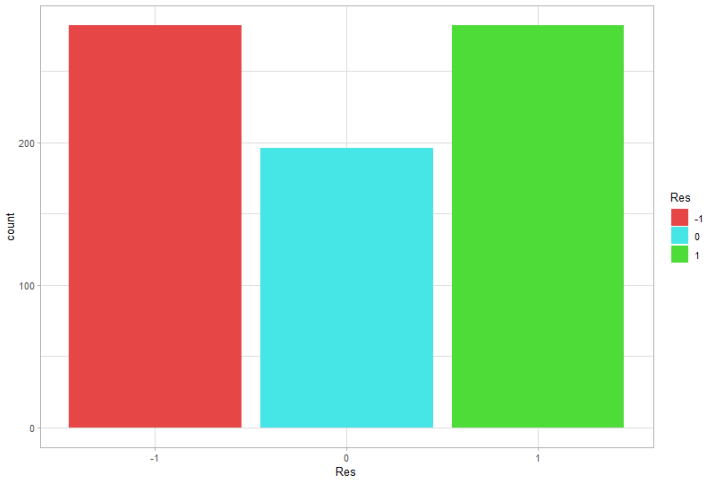


Figura 3.1: Barplot della distribuzione della variabile di risposta `Res`

La Figura 3.2 mostra la distribuzione delle vittorie, dei pareggi e delle sconfitte per ogni squadra.

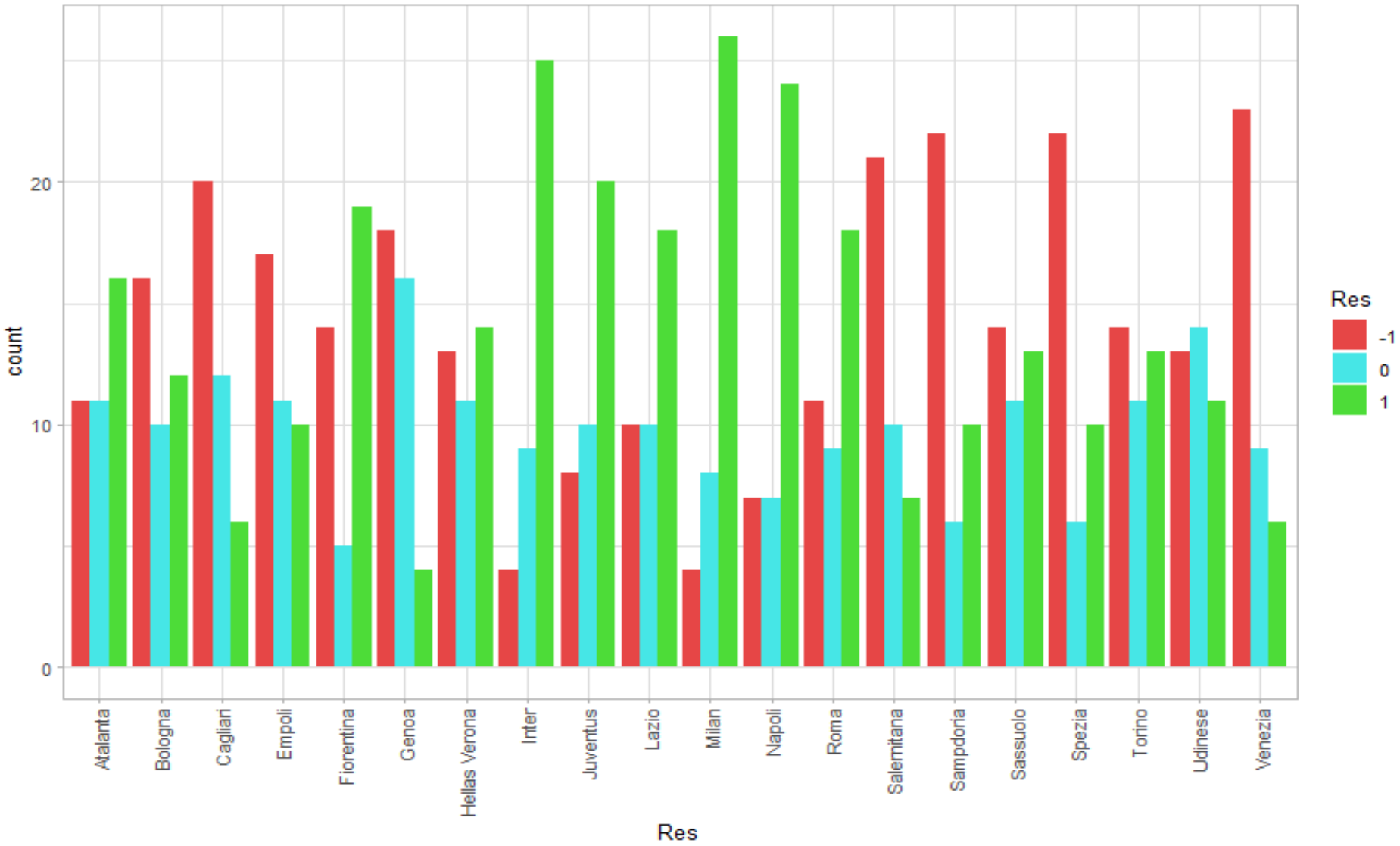


Figura 3.2: Barplot della distribuzione della variabile di risposta per squadra`Res`

3.2.1 Relazione tra la variabile risposta e le covariate

La prima relazione che si analizza riguarda la variabile categorica `AtHome`. Nella Figura 3.3 viene riportato il mosaicplot tra la variabile risposta e `AtHome`. Tale grafico è un particolare tipo di diagramma a barre impilate che mostra la relazione che c'è tra due fattori. Il numero di colonne è uguale al numero livelli della variabile inserita sull'asse orizzontale. L'altezza delle barre in verticale, invece, è proporzionale al numero di osservazioni della variabile inserita sull'asse verticale per ciascun livello della variabile nell'asse orizzontale. In sostanza, il mosaicplot è una rappresentazione grafica di una tabella di contingenza che permette un confronto visivo tra gruppi. Nella Figura 3.3 c'è una leggera variazione dei risultati tra la squadra che gioca in casa e l'avversaria, infatti per le squadre che giocano in casa, c'è una maggior presenza di vittorie e di minor sconfitte. Naturalmente non c'è alcuna variazione per il pareggio dato che entrambe le squadre lo ottengono.

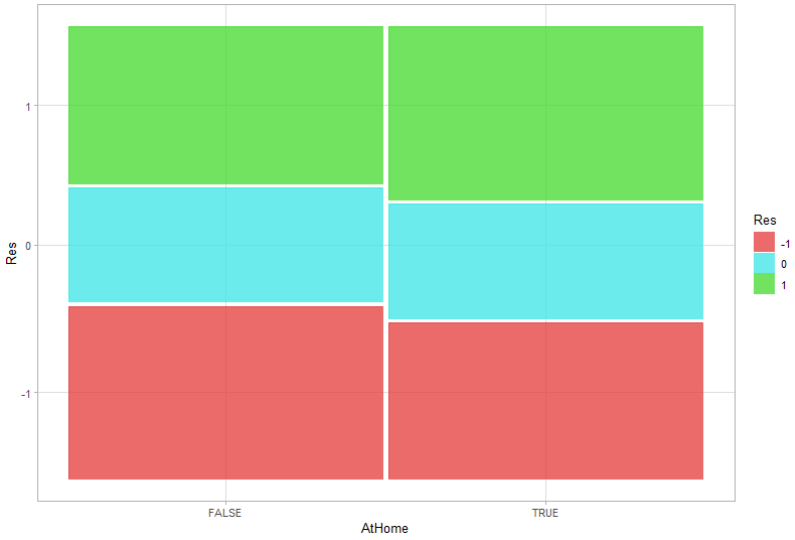


Figura 3.3: Mosaicplot che mostra la distribuzione degli esiti rispetto alle partite giocate in casa e fuori casa

Nella Figura 3.4 viene riportato il boxplot della distribuzione della variabile `Poss` rispetto ai valori della variabile risposta `Res`. Il boxplot è un grafico che consente di visualizzare il centro e la distribuzione dei dati. Inoltre, può essere un strumento visivo per la verifica della normalità o per l'identificazione di possibili outlier. Dal grafico si nota che `Poss` sembra essere significativa per l'esito. Infatti i valori crescono dal boxplot della sconfitta al boxplot della vittoria. C'è una buona distribuzione dei dati perché la lunghezza dei baffi per ogni boxplot è simmetrica. Si segnala che la mediana della sconfitta è più vicina al 3° quantile mentre la mediana della vittoria è più vicina al 1° quantile. Non sono presenti outliers.

Nella Figura 3.5 viene riportato il boxplot della distribuzione della variabile `SoT` rispetto ai valori della variabile risposta `Res`. Valori più alti sono presenti nella vittoria mentre valori molto più bassi sono presenti nella sconfitta. C'è una buona distribuzione dei valori nella vittoria dato che i baffi sono simmetrici, viceversa per le altri due boxplot

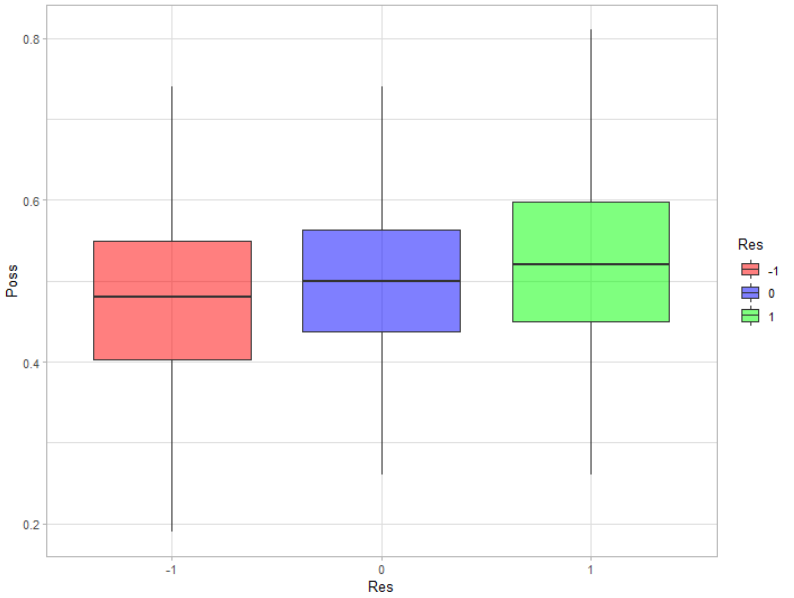


Figura 3.4: Boxplot della distribuzione della variabile `Poss` rispetto ai valori della variabile risposta `Res`

non c'è simmetria infatti, il baffo inferiore è molto più corto rispetto al baffo superiore, segno che la maggior parte dei valori sono bassi e simili tra loro. Inoltre alcuni outliers si discostano dalla distribuzione di tutti e tre i boxplot, questo perché ci sono state squadre che hanno tirato molte volte in porta. Le mediane dei boxplot pareggio e vittoria non sono equidistanti dai quantili ma più vicine al 1° quantile. Il boxplot della sconfitta ha una bassa varianza. In conclusione, avere un valore alto di tiri in porta sembra essere utile ai fini della vittoria.

Per la relazione tra la variabile risposta e la variabile `Sh`, si ha un boxplot molto simile al boxplot mostrato nella Figura 3.4. Il grafico di `Sh` rispetto al grafico di `Poss`, ha degli outliers e la mediana della sconfitta non è equidistante dai quantili ma più vicina al 1° quantile.

Nella Figura 3.6 viene riportato il boxplot della distribuzione della variabile `G/Sh` rispetto ai valori della variabile risposta `Res`. Si nota che ci sono valori molto bassi ma leggermente più alti per la vittoria. La distribuzione non è buona perché i baffi sono asimmetrici infatti, tutti i valori sono concentrati in basso e pochi verso il baffo superiore, segno che la maggior parte dei valori sono bassi e simili tra loro. C'è una bassa varianza tra i valori. C'è la presenza di outliers perché alcune squadre sono riuscite a ottenere il massimo da ogni tiro. I risultati mostrati, nonostante la distribuzione, sono comunque coerenti dato che non ci si aspetta dal rapporto tiri-gol un numero alto ma comunque una tendenza che favorisca la vittoria.

Nella Figura 3.7 viene riportato il boxplot della distribuzione della variabile `Saves` rispetto ai valori della variabile risposta `Res`. Come si può notare sembra che `Saves` sia poco significativa ai fini del risultato. Infatti c'è poca variazione tra un boxplot e l'altro

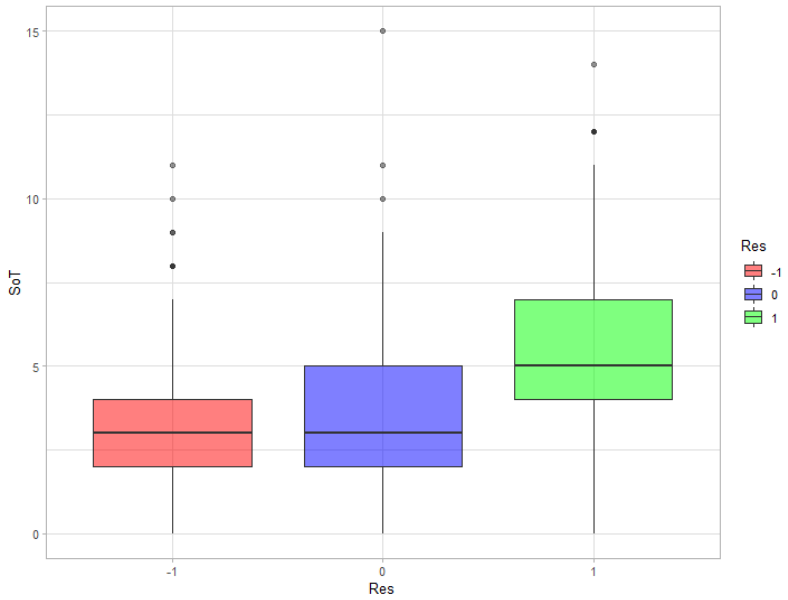


Figura 3.5: Boxplot della distribuzione della variabile `SoT` rispetto ai valori della variabile risposta `Res`

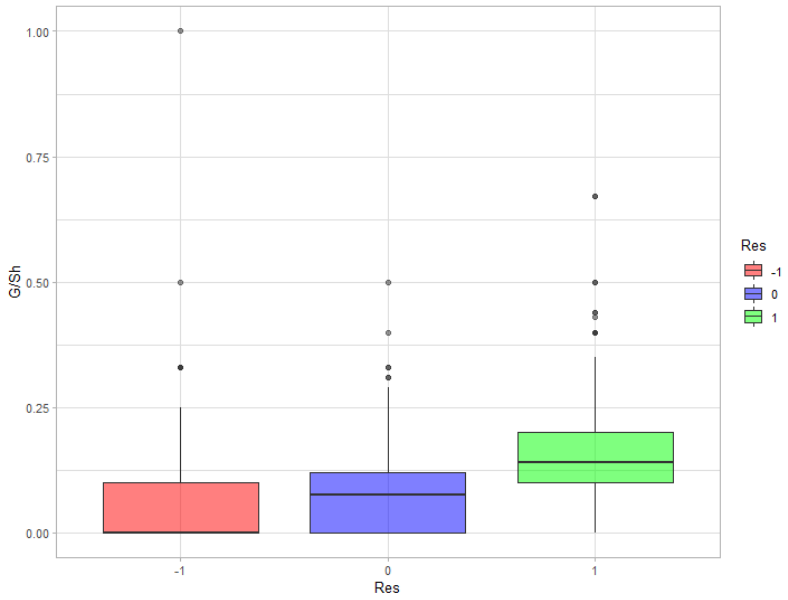


Figura 3.6: Boxplot della distribuzione della variabile `G/Sh` rispetto ai valori della variabile risposta `Res`

perché sembra che avere un alto numero di parate non è determinante a fini del risultato.

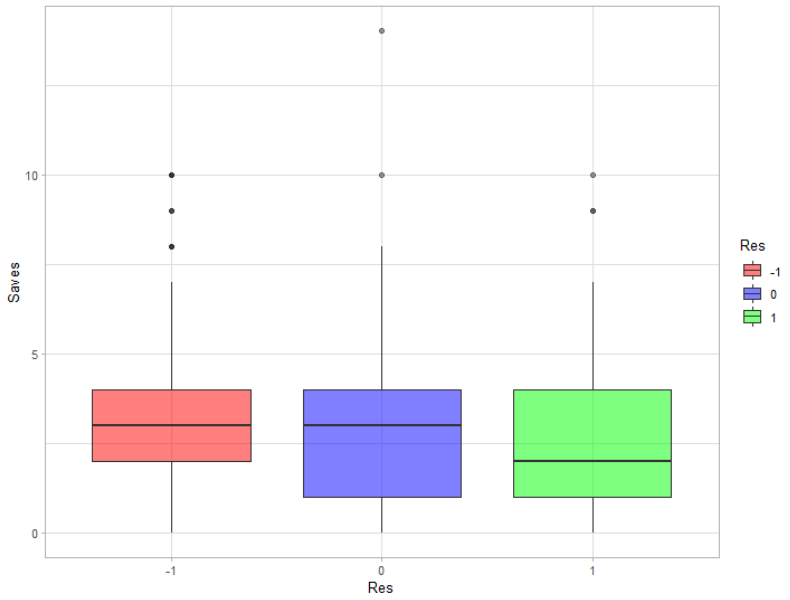


Figura 3.7: Boxplot della distribuzione della variabile `Saves` rispetto ai valori della variabile risposta `Res`

La Figura 3.8 viene riportato a sinistra il boxplot della variabile numerica `PAtt` rispetto ai valori della variabile risposta `Res` e a destra il boxplot della variabile numerica `PCmp%` rispetto ai valori della variabile risposta `Res`. Per entrambi sembra significativo l'elevato numero di passaggi tentati ma soprattutto quelli completati ai fini della vittoria. Nel grafico a sinistra, nel secondo e terzo boxplot il baffo superiore è più lungo rispetto al baffo inferiore, segno che molti valori sono bassi e simili tra loro, viceversa il primo boxplot ha una buona distribuzione perché i baffi sono simmetrici. Il boxplot della vittoria ha una maggiore varianza rispetto agli altri due e in più ha valori più alti; sia la mediana del boxplot della vittoria e sia quello del pareggio sono più vicine al 1° quantile, viceversa quella della sconfitta. I dati nel primo boxplot sembrano essere coerenti con l'esito della partita perché, maggior numero di passaggi si prova ad effettuare, maggiori sono le possibilità di vittoria. Occorre però sapere quanto è precisa la squadra e questo lo si può scoprire con la variabile `PCmp%`

Nel grafico a destra, si notano valori alti e molti outliers con valori bassi dovuti al fatto che ci sono state partite dove alcune squadre sono state poco precise nei passaggi. I baffi superiori di tutti e tre i boxplot sono molto meno lunghi rispetto ai baffi inferiori segno che molti valori sono alti e simili tra loro, inoltre, le varianze dei box sembrano essere uguali tra di loro. Sorprendentemente l'andamento invece di essere sempre crescente, prima scende da sconfitta a pareggio e poi sale da pareggio a vittoria.

Per la relazione tra la variabile risposta e la variabile `SPAtt`, si ha un grafico molto simile al grafico a sinistra della Figura 3.8. Il grafico di `SPAtt` rispetto al grafico di `PAtt`, ha un maggior numero di outliers soprattutto per la sconfitta rispetto al grafico

PAtt inoltre, c'è una minor varianza per tutti i tre boxplot oltre a valori più bassi in generale, questo è naturale perché PAtt contiene tutti i passaggi tentati e non solo quelli corti.

Per la relazione tra la variabile risposta e la variabile SPCmp%, si ha un grafico molto simile al grafico a destra della Figura 3.8. Il grafico di SPCmp% rispetto al grafico di PCmp%, il boxplot della sconfitta ha una maggior varianza, viceversa per la vittoria, che ha una minor varianza.

Per la relazione tra la variabile risposta e la variabile MPAtt, si ha un grafico molto simile al grafico a sinistra della Figura 3.8. Il grafico di MPAtt rispetto al grafico di PAtt, il boxplot della sconfitta ha una maggior varianza. In generale i valori sono più bassi rispetto al grafico di PAtt ma questo è naturale perché PAtt contiene tutti i passaggi tentati e non solo quelli medi.

Per la relazione tra la variabile risposta e la variabile MPCmp%, si ha un grafico molto simile al grafico a destra della Figura 3.8. Il grafico di MPCmp% rispetto al grafico di PCmp%, ha valori più alti e molti più outliers, inoltre i baffi inferiore dei boxplot della sconfitta e della vittoria sono più corti.

Per la relazione tra la variabile risposta e la variabile LPAtt, si ha un grafico molto simile al grafico a sinistra della Figura 3.8. Il grafico di LPAtt rispetto al grafico di PAtt, ha per il boxplot della sconfitta valori più bassi rispetto agli boxplot del pareggio e della vittoria inoltre, il boxplot del pareggio ha una maggior varianza valori mentre il boxplot della vittoria ha una minor varianza.
In generale i valori sono più bassi rispetto al grafico di PAtt ma questo è naturale perché PAtt contiene tutti i passaggi tentati e non solo quelli lunghi.

Per la relazione tra la variabile risposta e la variabile LPCmp%, si ha un grafico molto simile al grafico a destra della Figura 3.8. Il grafico di LPCmp% rispetto al grafico di PCmp%, ha valori più bassi, la distribuzione dei valori per il boxplot della sconfitta è ben equilibrata perché i baffi sono della stessa lunghezza e in più la mediana è equidistante dai due quantili, analogamente anche il boxplot del pareggio ha una distribuzione equilibrata ma con più varianza e una mediana equidistante dai quantili.

Nella Figura 3.9 viene riportato il boxplot della distribuzione della variabile ToDefPen rispetto ai valori della variabile risposta Res. Si nota che non c'è nessuna variazione dei tre boxplot, oltre ad avere la stessa varianza. L'esito può essere giustificato dal fatto che le squadre cercano di rimane fuori il più possibile dalla propria area di rigore per non portare troppo vicino alla porta l'avversario. Da ciò si può ipotizzare che ToDefPen non è significativa per la variabile risposta. Prima di escluderla si andrà ad analizzare se c'è qualche interazione con altre variabili che la fanno diventare significativa.

Nella Figura 3.10 viene riportato il boxplot della distribuzione della variabile ToAttPen rispetto ai valori della variabile risposta Res. Contrariamente quanto visto con la Figura 3.9 qui si nota una certa variazione tra i boxplot infatti, i valori crescono dal boxplot della sconfitta fino al boxplot della vittoria. C'è una maggior varianza per il boxplot della vittoria rispetto agli altri due boxplot. Per tutti e tre i boxplot i baffi inferiori sono leggermente meno lunghi rispetto ai baffi superiori, segno che i valori sono bassi e simili tra loro infatti, ci sono alcuni outliers sopra al baffo superiore, segno

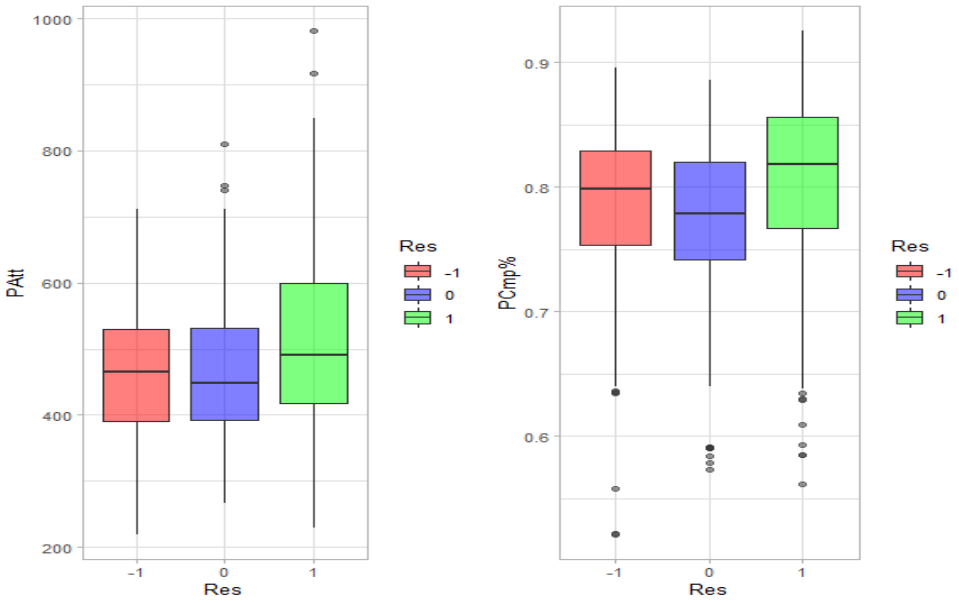


Figura 3.8: A sinistra il boxplot della variabile numerica `PAtt` rispetto ai valori della variabile risposta `Res` e a destra il boxplot della variabile numerica `PCmp%` rispetto ai valori della variabile risposta `Res`

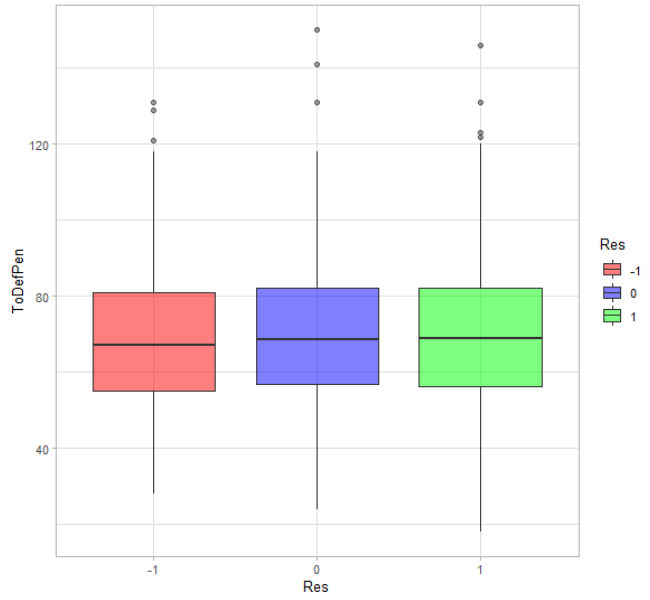


Figura 3.9: Boxplot della distribuzione della variabile `ToDefPen` rispetto ai valori della variabile risposta `Res`

che alcune squadre in qualche partita, si sono particolarmente rese note nel produrre un quantitativo di tocchi maggiore rispetto alla distribuzione, ciò però non sembra influenzare l'esito. Le mediane sono equidistanti.

Per la relazione tra la variabile risposta e la variabile `ToDef3rd`, si ha un grafico molto simile a quello mostrato nella Figura 3.10. Il grafico di `ToDef3rd` rispetto al grafico di `ToAttPen`, ha un minore numero di outliers soprattutto per il boxplot del pareggio, tale boxplot ha inoltre una varianza simile al boxplot della sconfitta. Il boxplot della vittoria invece, ha una distribuzione ben equilibrata.

Per la relazione tra la variabile risposta e la variabile `ToMid3rd`, si ha un grafico molto simile a quello mostrato nella Figura 3.10. Il grafico di `ToMid3rd` rispetto al grafico di `ToAttPen`, ha un minore numero di outliers e la varianza del boxplot della sconfitta è molto simile alla mediana del boxplot del pareggio ma con la mediana più vicina al 3° quantile.

Per la relazione tra la variabile risposta e la variabile `ToAtt3rd`, si ha un grafico molto simile a quello mostrato nella Figura 3.10. Il grafico di `ToAtt3rd` rispetto al grafico di `ToAttPen`, ha una minor varianza in generale per tutti e tre i boxplot e una distribuzione sbilanciata verso valori più bassi dato che tutti i baffi inferiori sono più corti rispetto ai baffi superiori. L'andamento però rimane lo stesso presente nella Figura 3.10.

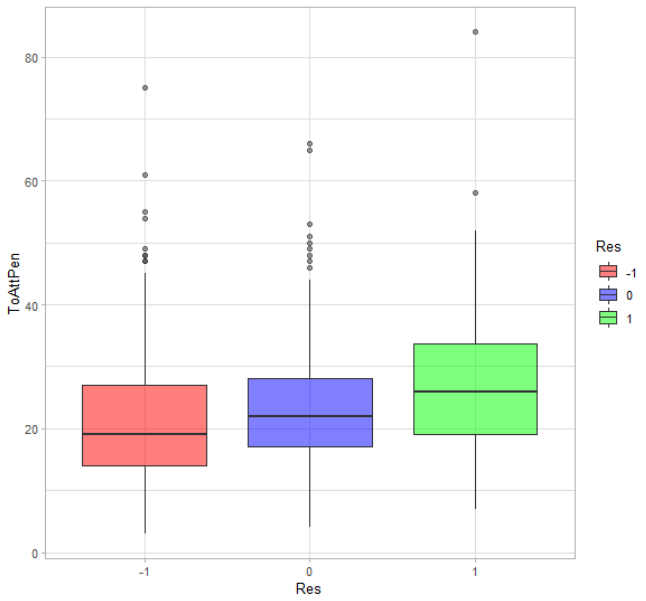


Figura 3.10: Boxplot della distribuzione della variabile `ToAttPen` rispetto ai valori della variabile risposta `Res`

Nella Figura 3.11 vengono riportati a sinistra il boxplot della variabile numerica `Fls` rispetto ai valori della variabile risposta `Res` e a destra il boxplot della variabile numerica `Fld` rispetto ai valori della variabile risposta `Res`. Nel boxplot a sinistra si può notare che i valori più alti sono nel boxplot del pareggio e della vittoria ma nel boxplot

del pareggio ci sono più valori alti. Ciò fa ipotizzare che subire molti falli può impedire la vittoria alla squadra che li subisce. Per quanto riguarda la distribuzione sembra essere buona; c'è una minor varianza per quanto riguarda il boxplot della sconfitta.

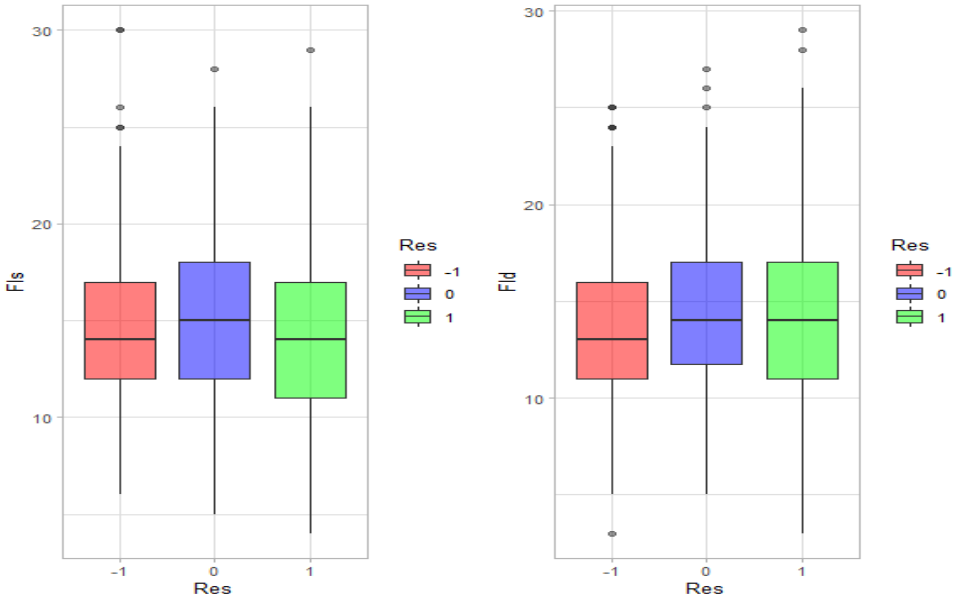


Figura 3.11: A sinistra il boxplot della variabile numerica `Fls` rispetto ai valori della variabile risposta `Res` e a destra il boxplot della variabile numerica `Fld` rispetto ai valori della variabile risposta `Res`

Nel secondo boxplot si hanno valori valori più alti nel boxplot della vittoria e una maggior varianza rispetto al boxplot della sconfitta. Sembra perciò che dal grafico si può intuire che se la squadra non commette dei falli allora sarà più soggetta a perdere.

Per la relazione tra la variabile risposta e la variabile `Off`, si ha un grafico molto simile a quello mostrato nella Figura 3.7. Il grafico di `Off` rispetto al grafico di `Saves`, ha un numero minore di valori per il boxplot della sconfitta rispetto agli altri due boxplot inoltre, le mediane del boxplot della sconfitta e del pareggio sono attaccate al 1° quantile.

Per la relazione tra la variabile risposta e la variabile `Crs`, si ha un grafico molto simile a quello mostrato nella Figura 3.12. Il grafico di `Crs` rispetto al grafico di `Saves`, ha per il boxplot della sconfitta maggior varianza e il baffo inferiore dei boxplot della sconfitta e della vittoria sono più corti rispetto ai baffi superiori.

Nella Figura 3.12 viene riportato il boxplot della distribuzione della variabile `Int` rispetto ai valori della variabile risposta `Res`. Sorprendentemente valori più alti sono registrati nel boxplot della sconfitta, anche se la mediana risulta essere più vicina al 1° quantile sottolineando che c'è un maggior numero di valori bassi piuttosto che alti. Le mediane dei restanti boxplot invece, sono ben equilibrate ma il boxplot del pareggio risulta avere meno varianza. Sembra perciò che effettuare troppi intercettazioni dei passaggi avversari contrariamente da quanto si pensi sia controproducente per la

vittoria. Si segnala inoltre la presenza di alcuni outliers con valori alti di intercettazioni, che si discostano dalle distribuzioni.

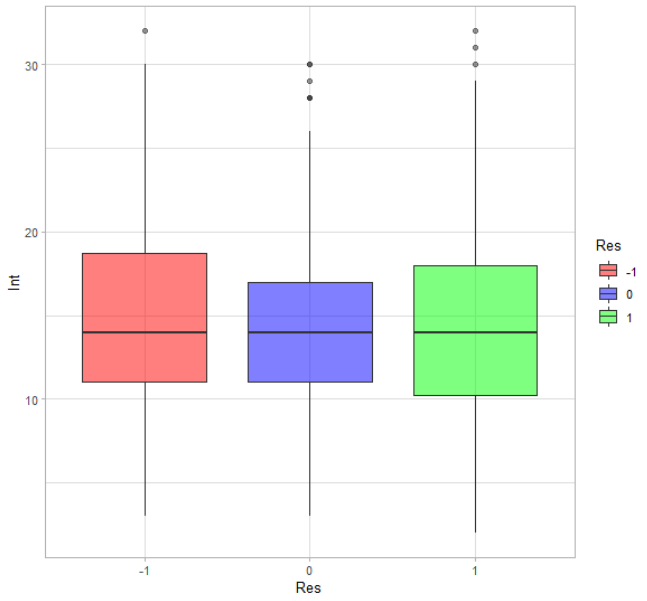


Figura 3.12: Boxplot della distribuzione della variabile `Int` rispetto ai valori della variabile risposta `Res`

Nella Figura 3.13 viene riportato il boxplot della distribuzione della variabile `TklWin` rispetto ai valori della variabile risposta `Res`. Come si può notare, vincere più contrasti possibili evita di subire una sconfitta. Infatti ci sono valori più alti nei boxplot del pareggio e della vittoria rispetto al boxplot della sconfitta. Nello specifico però si nota che: nella distribuzione ci sono maggior valori alti nella vittoria rispetto al pareggio, graficamente lo si vede dalla mediana che nel boxplot del pareggio è più vicina al 1° quindi ha valori più bassi e lo si nota anche dal baffo inferiore che è meno lungo rispetto a quello superiore viceversa, la mediana del boxplot della vittoria risulta più vicina al 3° oltre ad avere il baffo superiore più corto rispetto a quello inferiore. C'è inoltre qualche outliers con valori più alti di contrasti vinti ma sembrano non influenzare la classificazione.

Infine nella Figura 3.14 viene riportato il Boxplot della distribuzione della variabile `Recov` rispetto ai valori della variabile risposta `Res`. Per entrambi i boxplot la distribuzione sembra più sbilanciata verso valori bassi quindi ad una loro maggior presenza, infatti entrambe i baffi inferiori sono più corti rispetto a quelli superiori. Per quanto riguarda la mediana sembra equidistante dai quantili per entrambi i tre boxplot. Si nota che il boxplot del pareggio presenta minor varianza rispetto agli altri due boxplot ma valori più alti soprattutto nei confronti del boxplot della vittoria. Sembra perciò che un eccessivo numero di recuperi non porti alla vittoria. Si nota inoltre che ci sono numerosi outliers.

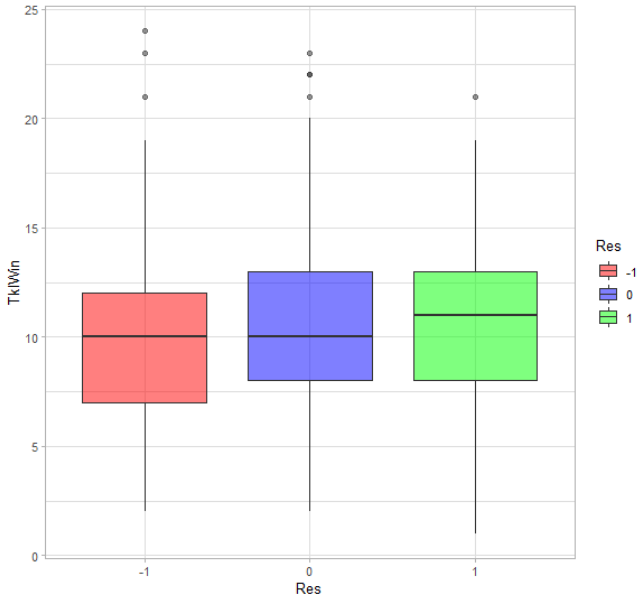


Figura 3.13: Boxplot della distribuzione della variabile `TklWin` rispetto ai valori della variabile risposta `Res`

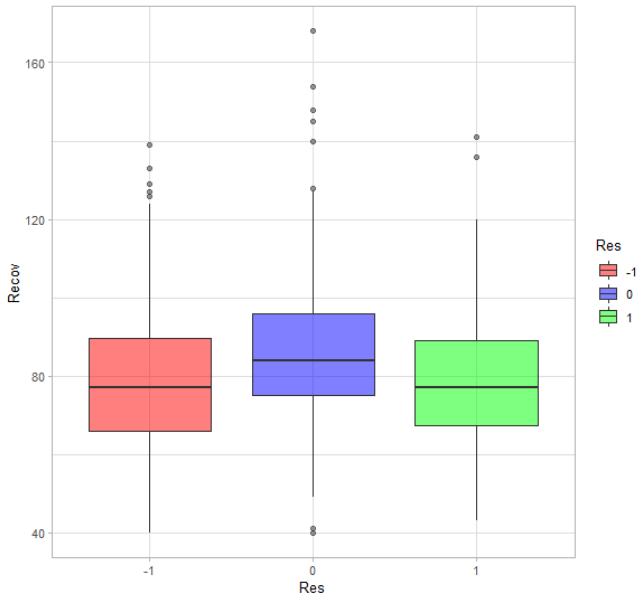


Figura 3.14: Boxplot della distribuzione della variabile `Recov` rispetto ai valori della variabile risposta `Res`

3.2.2 Analisi possibili interazioni

Per concludere l'attività di preprossening, non resta che analizzare le relazioni tra le covariate per individuare possibili interazioni tra di loro che possono influenzare la variabile risposta. Chiaramente dato che ci sono più di trenta variabili e dunque, un grandissimo numero di combinazioni, non si sono esaminate tutte le relazioni ma sono state selezionate solo alcune per l'analisi, basandosi su teorie calcistiche esaminate durante la fase di studio del problema.

Per l'analisi delle interazioni si sono utilizzati i grafici di dispersione. Un grafico di dispersione mostra la relazione tra due variabili continue. A tali grafici si è inserito una terza variabile, la variabile risposta `Res`, dove ogni punto è colorato in tre possibili colori che rappresentano una delle tre categorie di `Res`. Di conseguenza il grafico permette di visualizzare se le categorie sono ben separati e quindi se un'interazione può spiegare l'andamento dei punti della variabile risposta.
Inoltre è stato utilizzato l'indice di correlazione, che indica la forza dell'associazione lineare espressa in valori compresi tra -1 e 1. Tale misura permette di escludere da subito alcune relazioni tra variabili se l'indice è troppo alto o basso, infatti, le relazioni troppo forti vanno escluse perché può presentarsi il fenomeno della collinearità. La collinearità è quel fenomeno che va a nasconde il legame tra le variabili e la variabile risposta, a causa di un legame troppo forte tra le covariate.
Nella Figura 3.15 viene mostrato il valore della correlazione per ogni possibile relazione tra variabili numeriche. Si nota che ci sono molte relazioni che hanno un valore di correlazione molto vicino a 1, in basso a sinistra del grafico. Ad esempio notiamo che la variabile `SPCmp%` ha una relazione molto forte con la variabile `PCmp%` (correlazione = 0.82), ciò è coerente perché, la variabile `SPCmp%` contiene solo i passaggi corti completati mentre `PCmp%` contiene tutti i tipi di passaggi completati, ne consegue che la ridondanza dei dati causa questa alta correlazione. Analogamente la stessa motivazione la si può applicare tra la variabile `PAtt` e la variabile `SPAtt` (correlazione = 0.91). Perciò tale motivazione è applicabile a tutte le variabili relative ai passaggi completati o relative ai passaggi tentati.
Di seguito si riporteranno le interazioni che sono state individuate come significative.

Sono state individuate le seguenti tre interazioni con la variabile `Sh`:

* Interazione tra la variabile `Sh` e la variabile `ToAttPen`. È ragionevole ipotizzare che il numero di tocchi fatti nell'area di rigore avversaria possano creare azioni da tiro. È quindi possibile che tra le due variabili possa esserci una relazione. La Figura 3.16 mostra una relazione positiva tra le due variabili infatti, quando aumenta la variabile `Sh` aumenta anche la variabile `ToAttPen` e viceversa. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta, inoltre la correlazione tra le due variabili non è troppo alta (0.72). Ne consegue che un'interazione tra la variabile `Sh` e la variabile `ToAttPen`, sembra essere significativa rispetto alla variabile risposta.

* Interazione tra la variabile `Sh` e la variabile `G/Sh`. È ragionevole ipotizzare che ci sia un legame naturale tra tiri fatti e rapporto tiri-gol. La Figura 3.17 mostra una relazione negativa tra le due variabili infatti, quando aumenta la variabile `Sh` diminuisce anche la variabile `G/Sh` e viceversa. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta infatti, i punti della categoria vittoria sono più in alto mentre i punti delle categorie pareggio

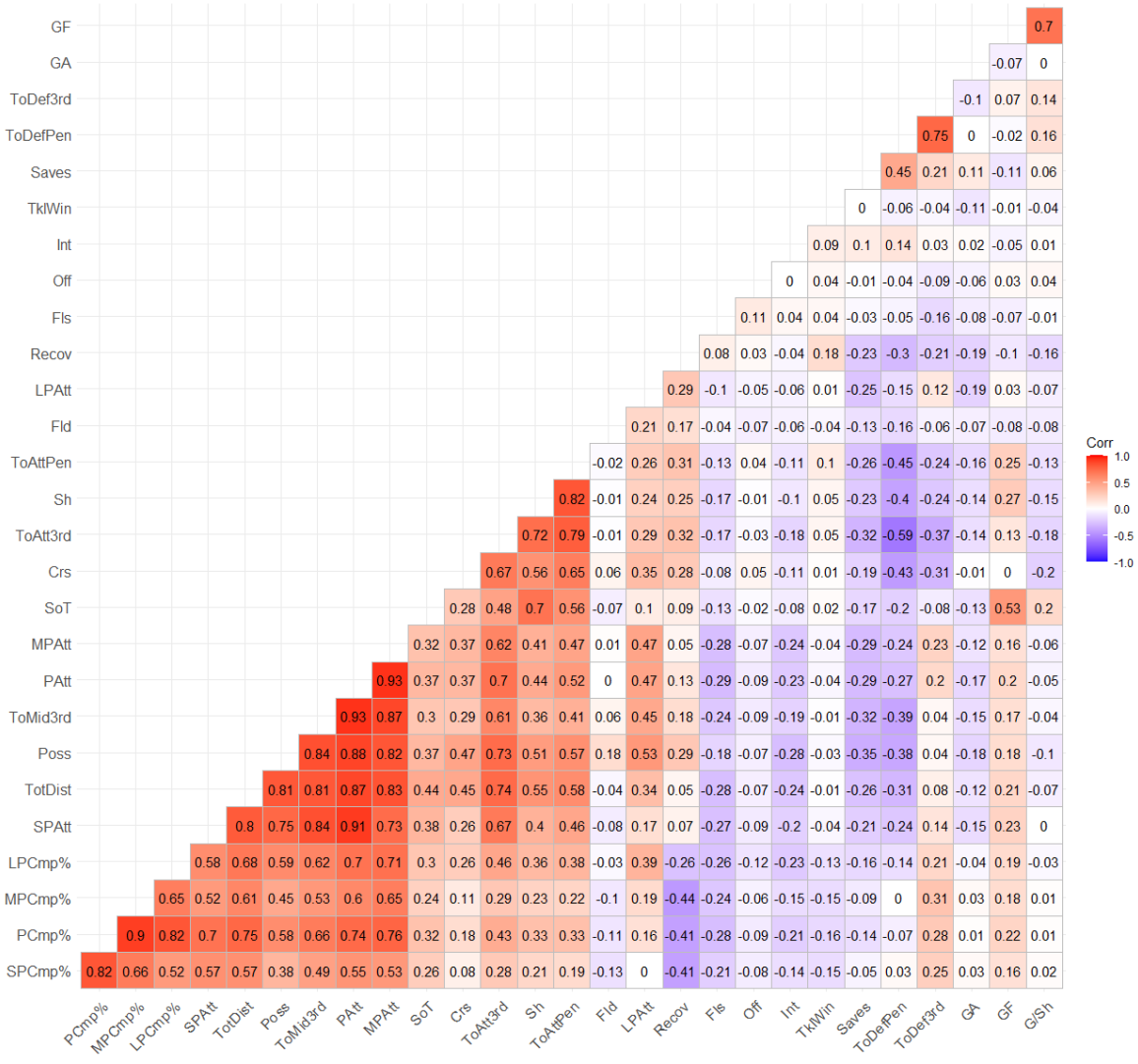


Figura 3.15: Grafico delle correlazioni di ogni coppia di variabili

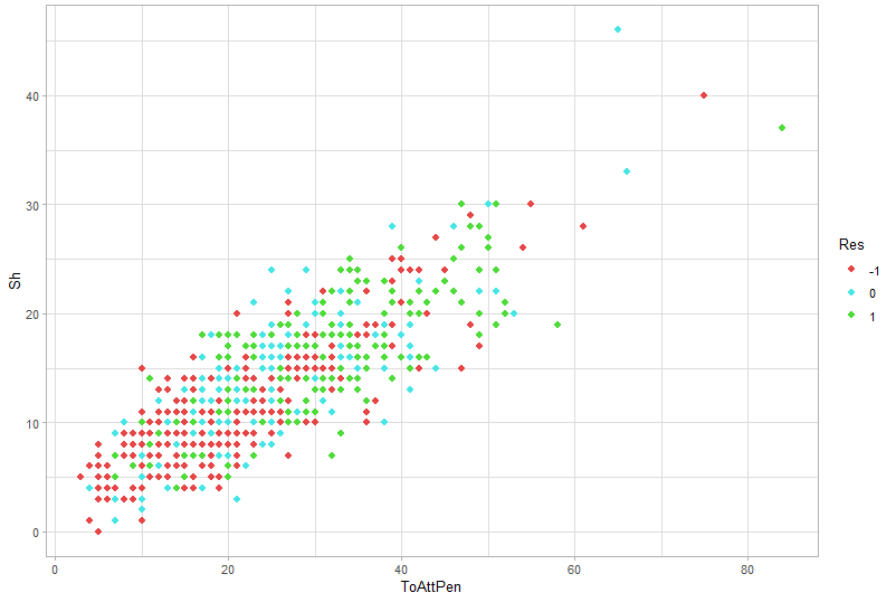


Figura 3.16: Scatterplot della distribuzione della variabile `Sh` rispetto ai valori della variabile `ToAttPen`

e sconfitta più in basso. Inoltre la correlazione tra le due variabili non è bassa (-0.15). Ne consegue che un'interazione tra la variabile `Sh` e la variabile `G/Sh`, sembra essere significativa rispetto alla variabile risposta.

* Interazione tra la variabile `Sh` e la variabile `Poss`. Generalmente è possibili ipotizzare che il possesso della palla possa favorire nel effettuare i tiri. Infatti, la Figura 3.18 mostra una relazione positiva tra le due variabili, quando aumenta la variabile `Sh` aumenta anche la variabile `Poss` e viceversa. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta, inizialmente i vari punti sono mescolati tra di loro ma, con l'avanzamento emergono le direzioni di ogni categoria infatti, i punti della categoria vittoria vanno più verso destra mentre i punti delle categoria sconfitta si spostano verso l'alto senza tendere verso destra, i punti della categoria pareggio pareggio invece, si muovono in mezzo ai punti delle altre due categorie. La correlazione tra le due variabili non è alta (0.51). Ne consegue che un'interazione tra la variabile `Sh` e la variabile `Poss`, sembra essere significativa rispetto alla variabile risposta.

Sono state individuate le seguenti tre interazioni con la variabile `ToMid3rd`:

* Interazione tra la variabile `ToMid3rd` e la variabile `LPAtt`. Si suppone che tra le due variabili ci sia una relazione perché molti lanci lunghi per le punte partono proprio del centrocampo. La Figura 3.19 mostra un andamento un po' a "nuvola" ma comunque, è possibile individuare una relazione positiva tra le due variabili infatti, quando aumenta la variabile `ToMid3rd` aumenta anche la variabile `LPAtt` e viceversa. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta, inizialmente i vari punti sono mescolati tra di loro ma, successivamente i punti della categoria vittoria vanno molto in alto

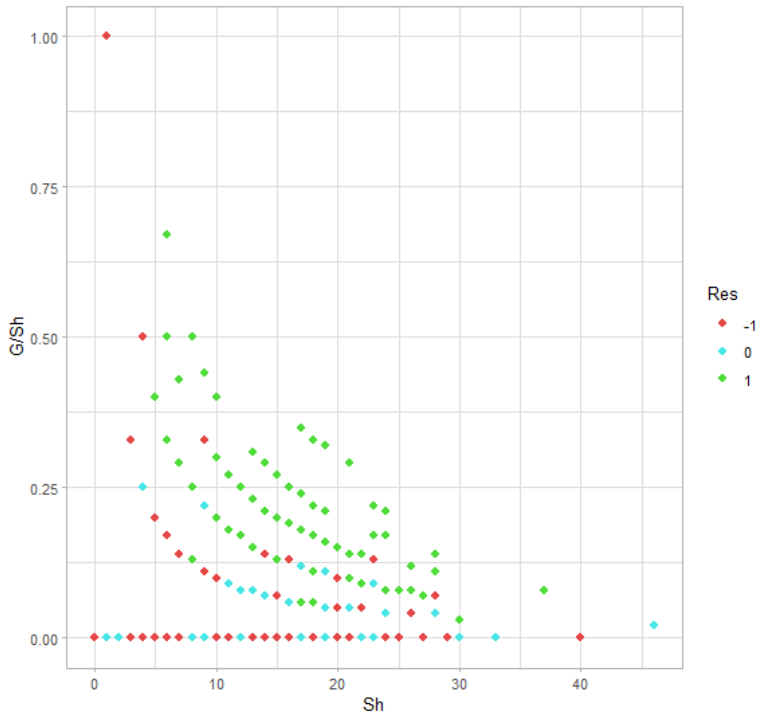


Figura 3.17: Scatterplot della distribuzione della variabile `Sh` rispetto ai valori della variabile `G/Sh`

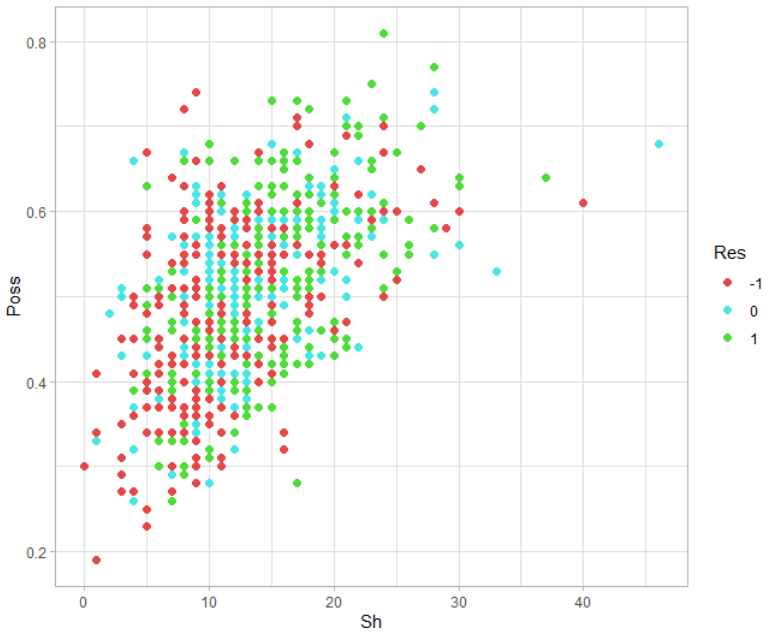


Figura 3.18: Scatterplot della distribuzione della variabile `Sh` rispetto ai valori della variabile `Poss`

mentre i punti della categoria sconfitta rimangono molto più bassi muovendosi verso destra, invece i punti della categoria pareggio anche essi vanno verso destra ma rimanendo più alti rispetto ai punti della categoria sconfitta. La correlazione tra le due variabili non è alta (0.45). Ne consegue che un'interazione tra la variabile `ToMid3rd` e la variabile `LPAtt`, sembra essere significativa rispetto alla variabile risposta.

* Interazione tra la variabile `ToMid3rd` e la variabile `PCmp%`. Per le stesse ragioni illustrate nel punto precedente si ipotizza una relazione tra le variabili. La Figura 3.20 mostra una relazione positiva tra le due variabili infatti, quando aumenta la variabile `ToMid3rd` aumenta anche la variabile `PCmp%`, con un'andamento simile ad una funzione esponenziale. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta, dove i punti più in alto sono della categoria del pareggio, leggermente più sotto ci sono i punti della vittoria che pero verso la fine del grafico raggiungono i valori più alti, e infine i punti della sconfitta. La correlazione tra le due variabili non è alta (0.66). Ne consegue che un'interazione tra la variabile `ToMid3rd` e la variabile `PCmp%`, sembra essere significativa rispetto alla variabile risposta.

Infine sono state individuate le seguenti interazioni:

* Interazione tra la variabile `TotDist` e la variabile `PCmp%`. Naturalmente per effettuare i passaggi e completarli è possibile farlo solo se ci si muove con la palla. La Figura 3.21 mostra una relazione positiva tra le due variabili infatti, quando aumenta la variabile `TotDist` aumenta anche la variabile `PCmp%`, con un'andamento simile ad una funzione esponenziale. Sono distinguibili tre differenti

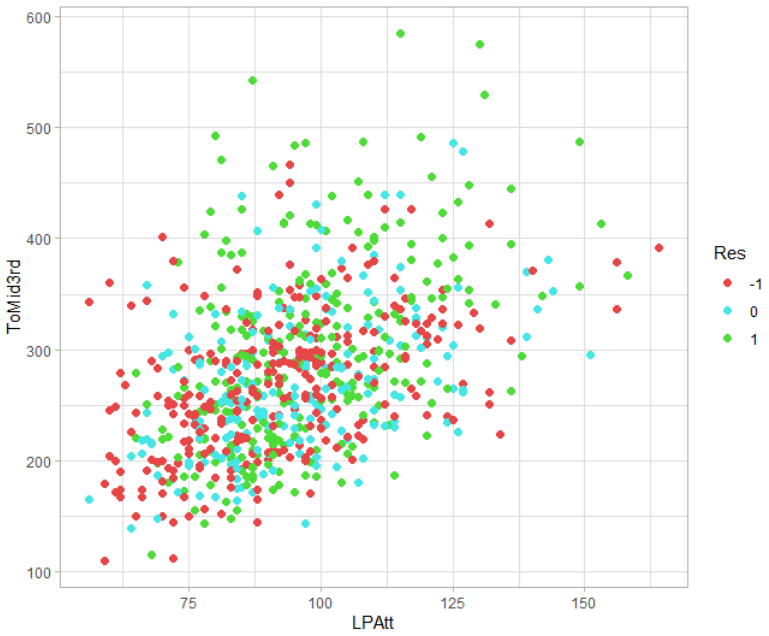


Figura 3.19: Scatterplot della distribuzione della variabile `ToMid3rd` rispetto ai valori della variabile `LPAtt`

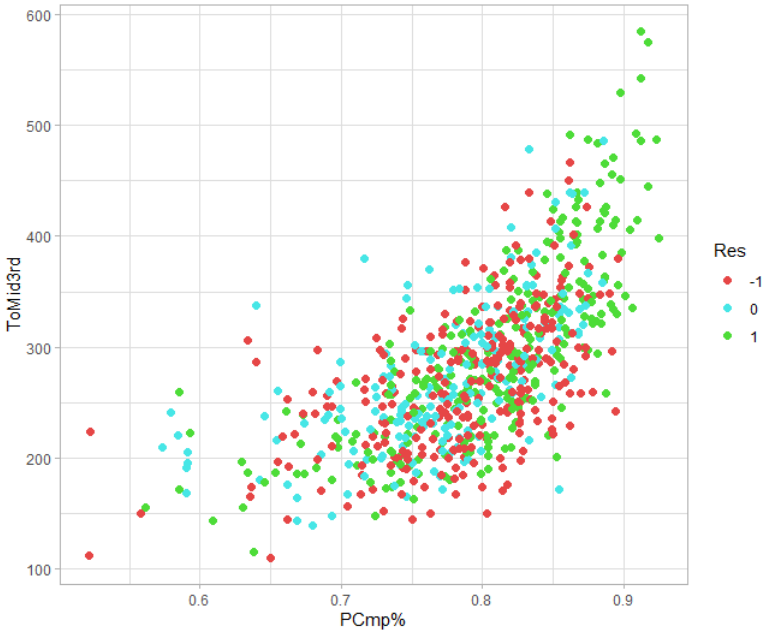


Figura 3.20: Scatterplot della distribuzione della variabile `ToMid3rd` rispetto ai valori della variabile `PCmp%`

gruppi che rappresentano le tre categorie della variabile risposta, dove i punti più in alto sono della categoria del pareggio, leggermente più sotto ci sono i punti della vittoria e infine i punti della sconfitta. La correlazione tra le due variabili non è troppo alta (0.75). Ne consegue che un'interazione tra la variabile `TotDist` e la variabile `PCmp%`, sembra essere significativa rispetto alla variabile risposta.

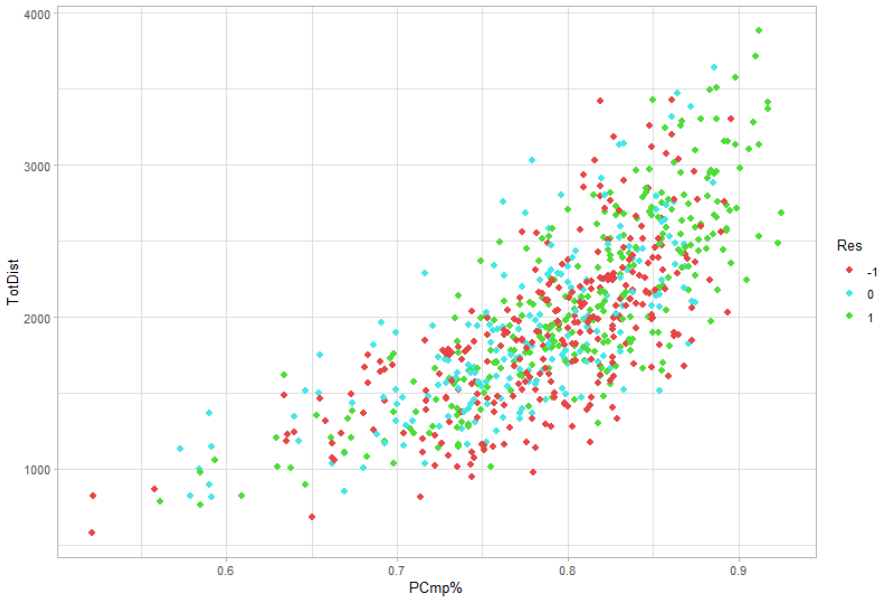


Figura 3.21: Scatterplot della distribuzione della variabile `TotDist` rispetto ai valori della variabile`PCmp%`

* Interazione tra la variabile `PAtt` e la variabile `PCmp%`. Data la loro naturale correlazione si ipotizza che ci sia un'interazione. La Figura 3.22 mostra una relazione positiva tra le due variabili infatti, quando aumenta la variabile `PAtt` aumenta anche la variabile `PCmp%`, con un'andamento simile ad una funzione esponenziale. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta, dove i punti più in alto sono della categoria del pareggio, leggermente più sotto ci sono i punti della vittoria e infine i punti della sconfitta. La correlazione tra le due variabili non è troppo alta (0.74). Ne consegue che un'interazione tra la variabile `PAtt` e la variabile `PCmp%`, sembra essere significativa rispetto alla variabile risposta.

* Interazione tra la variabile `ToDefPen` e la variabile `ToAttPen`. Come ci si può aspettare la Figura 3.23 mostra una relazione negativa tra le due variabili, quando aumenta la variabile `ToDefPen` diminuisce anche la variabile `ToAttPen` e viceversa. Sono distinguibili tre differenti gruppi che rappresentano le tre categorie della variabile risposta infatti, i punti della categoria vittoria sono quelli più distanti dallo zero mentre i punti delle categorie pareggio e sconfitta sono più vicini allo zero. Inoltre la correlazione tra le due variabili non è bassa (-0.45). Ne consegue che un'interazione tra la variabile `ToDefPen` e la variabile `ToAttPen`, sembra essere significativa rispetto alla variabile risposta.

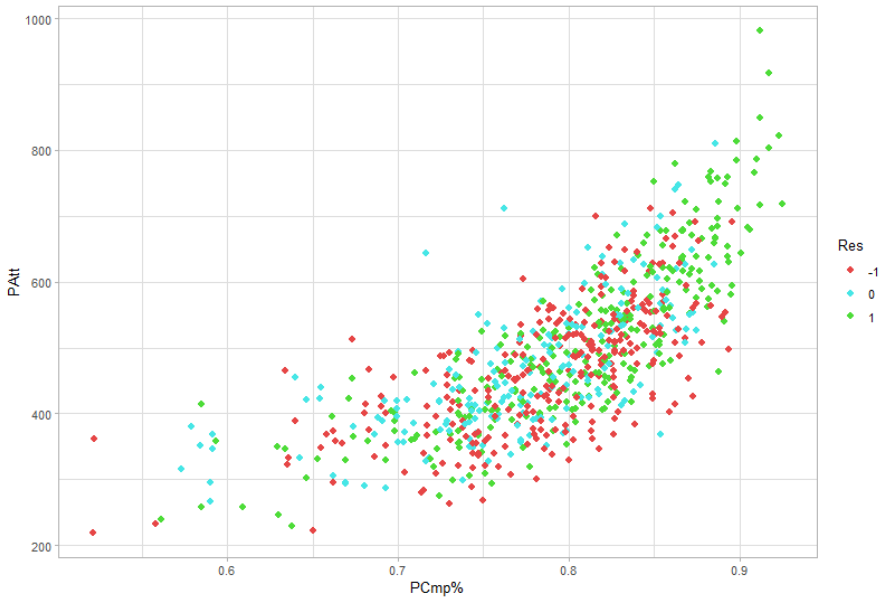


Figura 3.22: Scatterplot della distribuzione della variabile `PAtt` rispetto ai valori della variabile `PCmp%`

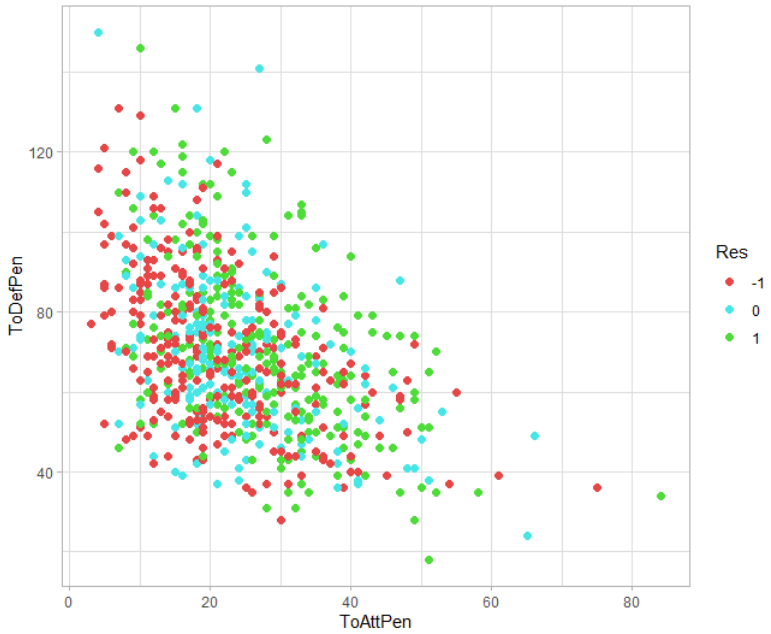


Figura 3.23: Scatterplot della distribuzione della variabile `ToDefPen` rispetto ai valori della variabile `ToAttPen`

4 | Il modello Bradley-Terry

Nel seguente capitolo verranno introdotti differenti modelli per il confronto a coppie, iniziando con il modello Bradley-Terry versione base fino a presentare tutte le sue estensioni usate per l'analisi trattata. Infine verrà illustrata la penalizzazione applicata.

4.1 Modello Bradley-Terry base

Il modello Bradley-Terry (Bradley e Terry, 1952) è un modello probabilistico che permette di predire il risultato di un confronto a coppie, dove per confronto a coppie si intende un processo di comparazione tra due oggetti α_i e α_j con i,j $\in$ {1,...,n}, appartenenti a un set di oggetti $\{\alpha_1, ..., \alpha_n\}$, dove gli oggetti sono l'entità che vengono confrontate tra di loro.
Formalmente, dato un set di oggetti $\{\alpha_1, ..., \alpha_n\}$, un set di parametri $\{\gamma_1, ..., \gamma_n\}$ che rappresentano ciascuno l'abilità/forza del i-esimo oggetto e la variabile casuale associata al risultato del confronto a coppie $Y_{i,j}$ con i<j $\in$ {1,...,n}, la probabilità che il risultato sia $\alpha_i \succ \alpha_j$ è

$$P(\alpha_i \succ \alpha_j) = P(Y_{i,j} = 1) = \frac{exp(\gamma_i - \gamma_j)}{1 + exp(\gamma_i - \gamma_j)} \tag{4.1}$$

Il risultato $\alpha_i \succ \alpha_j$ può essere letto come "l'oggetto α_i è preferito all'oggetto α_j","α_i batte l'oggetto α_j" oppure "α_i è migliore dell'oggetto α_j". La variabile casuale è di tipo binario cioè $Y_{i,j} = 1$ se l'oggetto α_i è preferito all'oggetto α_j e $Y_{i,j} = 0$ se l'oggetto α_j è preferito all'oggetto α_i. I parametri γ_i sono stimati dal modello attraverso la massima verosimiglianza. È necessario imporre un vincolo per identificare gli oggetti. Tali vincoli possono essere, il vincolo di somma $\sum_{i=1}^{n} \gamma_i = 0$ oppure il vincolo dell'oggetto di riferimento. Per il vincolo dell'oggetto di riferimento si intende che viene fissato $\gamma_i = 0$ per un oggetto $\alpha_i \in$ {1, ..., n}, mentre il valore dei parametri γ_j degli altri oggetti α_j sarà la differenza rispetto all'oggetto di riferimento α_i.
Il modello precedentemente descritto è chiamato modello non strutturato, inoltre il modello base non considera covariate e, in generale, non presta alcuna attenzione all'eterogeneità causata dai soggetti dei confronti.

Il modello può essere alternativamente espresso in forma di logit lineare:

$$logit(\alpha_i \succ \alpha_j) = log\left(\frac{P(\alpha_i \succ \alpha_j)}{P(\alpha_j \succ \alpha_i)}\right) = log\left(\frac{exp(\gamma_i)}{exp(\gamma_j)}\right) = \gamma_i - \gamma_j \tag{4.2}$$

4.2 Modello Bradley-Terry con categorie di risposta ordinate

In molti contesti di comparazione tra oggetti, è possibile che sia richiesto di dare una scala di preferenza tra un oggetto e un altro, ossia la variabile casuale deve avere K possibili categorie di risposta con K > 2. Inoltre tali scelte devono avere un ordine di preferenza, dal risultato meno gradevole al più gradevole per l'i-esimo oggetto, ad esempio si preferisce il pareggio piuttosto che perdere. Perciò il modello 4.2 che ha una variabile casuale binaria non è adeguato.
Avere K categorie di risposta ordinate con K > 2 è di interesse per le comparazioni calcistiche dato che non è sufficiente stimare la probabilità di vittoria o sconfitta ma deve essere obbligantemente preso in considerazione anche il pareggio come risultato.

Modelli che consentono un numero generale di categorie K, sono stati proposti da (Bradley e Terry, 1952) a (Tutz, 1986) . In particolare (Tutz, 1986) mostrò come due modelli per l'analisi di dati ordinati possono essere adattati per i confronti a coppie.

Il primo modello presentato è detto a collegamento cumulativo e sfrutta la rappresentazione nelle variabili latenti. In generale, data la variabile continua casuale latente $Z_{i,j}$ sia K il numero di gradi della scala di preferenza e siano $\theta_1 < \theta_2 < .... < \theta_{K-1}$ le soglie tale che $Y_{i,j} = k$ quando $\theta_{k-1} < Z_{i,j} < \theta_k$. Allora:

$$P(Y_{i,j} \leq k) = \frac{exp(\theta_k + \gamma_i - \gamma_j)}{1 + exp(\theta_k + \gamma_i - \gamma_j)} \tag{4.3}$$

con k ∈ {1,....,K} che indica le possibili categorie di risposta. I parametri θ_k rappresentano le cosiddette soglie per le singole categorie di risposta, che determinano la preferenza per le specifiche categorie. In particolare, $Y_{i,j} = 1$ rappresenta la massima preferenza per un oggetto i rispetto a un oggetto j.
In generale vi è imposta una simmetria del modello in modo che valga: $P(Y_{i,j} = k) = P(Y_{j,i} = K - k + 1)$. È quindi necessario che le soglie siano ristrette a θ_k = -θ_{K-k} e se, K è dispari, $\theta_{K/2} = 0$; per garantire che le probabilità siano simmetriche cioè il risultato opposto abbia la stessa probabilità di verificarsi. Per garantire che le probabilità siano non negative per le singole categorie di risposta vi è imposta la seguente limitazione: $-\infty = \theta_0 < \theta_1 < ... < \theta_{K-1} < \theta_K = \infty$. Dato che la soglia per l'ultima categoria è fissata a $\theta_K = \infty$ allora $P(Y_{i,j} \leq K) = 1$. Si sottolinea che le soglie sono parametri che vanno stimate dai dati. Inoltre, la probabilità di una singola categorie di risposta può essere derivata dalla differenza tra categorie adiacenti cioè:

$$P(Y_{i,j} = k) = P(Y_{i,j} \leq k) \text{ - } P(Y_{i,j} \leq k - 1).$$

Il modello ha anche una rappresentazione logit lineare ed è la seguente:

$$logit(Y_{i,j} \leq k) = \theta_k + \gamma_i - \gamma_j \tag{4.4}$$

Il secondo modello invece proposto da (Agresti, 1992) è detto modello a categorie adiacenti. In questo caso il collegamento è applicato alle probabilità di risposte adiacenti piuttosto che alle probabilità cumulative, riducendosi così al modello Bradley-Terry quando sono consentite solo due categorie mentre quando sono consentite solo tre categorie, si riduce al modello proposto da (Davidson, 1970) che verrà presentato di seguito al prossimo paragrafo.

Il modello a categorie adiacenti è più semplice da interpretare rispetto ai modelli a collegamenti cumulativi poiché la probabilità si riferisce a un determinato risultato anziché a raggruppamenti di risultati.
Perciò dal modello proposto da (Davidson, 1970), sia θ il parametro stimato dai dati che indica quanto è auspicabile la non preferenza, allora:

$$P(Y_{i,j} = 2|Y_{i,j} \neq 0) = \frac{exp(\gamma_i - \gamma_j)}{1 + exp(\gamma_i - \gamma_j)}, \tag{4.5}$$

$$P(Y_{i,j} = 1) = \frac{\theta\sqrt{exp(\gamma_i) * exp(\gamma_j)}}{exp(\gamma_i) + exp(\gamma_j) + \theta\sqrt{exp(\gamma_i) * exp(\gamma_j)}}, \tag{4.6}$$

$$P(Y_{i,j} = 0|Y_{i,j} \neq 1) = \frac{exp(\gamma_j - \gamma_i)}{1 + exp(\gamma_j - \gamma_i)} \tag{4.7}$$

Si è riportato la modellazione di tutti e tre i possibili risultati, con γ_n che rappresenta la forza degli oggetti in comparazione. La probabilità che l'oggetto α_i batta l'oggetto α_j è rappresentata da (4.5), mentre la probabilità che l'oggetto α_j batta l'oggetto α_i è rappresentata da (4.7). Sia (4.5) e sia (4.7) rimangono uguali alla probabilità (4.2) descritta precedentemente. Invece, per la probabilità che l'oggetto α_i pareggi con l'oggetto α_j (4.6), viene aggiunto il parametro θ. Il parametro θ rappresenta quanto è auspicabile il pareggio.

4.3 Bradley–Terry Model con effetti dell'ordine

Nel modello descritto nella sezione 4.2, è necessario imporre la simmetria tra le categorie di risposta. Purtroppo la simmetria imposta risulta essere non adeguata in alcuni contesti. Tra questi vi è anche il calcio poiché l'ordine dei oggetti (le squadre) conta. Infatti in una partita di calcio, la prima squadra che viene indicata tra le due squadre, è quella che gioca in casa, dove teoricamente dovrebbe avere un vantaggio sull'avversario. Perciò, il presupposto che le categorie di risposta siano simmetriche non vale più.
Un possibile modello riadattato al problema esposto è il seguente:

$$P(\alpha_i \succ \alpha_j) = P(Y_{i,j} = 1) = \frac{exp(\delta + \gamma_i - \gamma_j)}{1 + exp(\delta + \gamma_i - \gamma_j)}. \tag{4.8}$$

L'effetto dell'ordine (il vantaggio di giocare in casa in ambito calcistico) viene trattato come un parametro δ. Se $\delta > 0$ allora viene attribuito un vantaggio all'oggetto α_i; aumentando la probabilità che vinca il confronto o nel caso di categorie di risposta ordinate, di avere un risultato superiore rispetto all'oggetto α_j. Chiaramente il peso di δ deve essere stimato dai dati.

Invece un modello con categorie di risposta ordinate riadatto è il seguente:

$$P(Y_{i,j} \leq h) = \frac{exp(\delta + \theta_h + \gamma_i - \gamma_j)}{1 + exp(\delta + \theta_h + \gamma_i - \gamma_j)} \tag{4.9}$$

Il modello (4.8) e il modello (4.9) , hanno anche una rappresentazione logit lineare e sono le seguenti:

Per (3.8)

$$logit(\alpha_i \succ \alpha_j) = \delta + \gamma_i - \gamma_j \tag{4.10}$$

Per (3.9)

$$logit(\alpha_i \succ \alpha_j) = \delta + \theta_h + \gamma_i - \gamma_j \tag{4.11}$$

4.4 Bradley–Terry Model con variabili esplicative

È stato presentato un modello che valutasse il grado di preferenza per un oggetto α_i rispetto a un oggetto α_j, senza considerare nessuna covariata. Tale modello risulta essere inutile, dato che siamo interessati a capire quali covariate possono influenzare il risultato della comparazione. Prima di esporre il modello con covariate, è necessario fare una distinzione tra soggetti e oggetti e successivamente distinguere i tre tipi di covariate di un confronto a coppie, ovvero: le covariate specifiche al soggetto x_p, le covariate specifiche all'oggetto z_i e infine le covariate specifiche al soggetto e all'oggetto $z_p i$ per i soggetti p, p = 1,.....,m e gli oggetti α_i, i = 1,....,n.
Gli oggetti sono le entità che vengono confrontate in un confronto a coppie. I soggetti invece, sono le unità che stabiliscono la preferenza tra gli oggetti in un confronto a coppie. Nel calcio gli oggetti sono le squadre di calcio, mentre i soggetti sono le partite di calcio dove avviene la comparazione tra le squadre.

Di seguito vengono illustrate le tre tipologie di covariate in un confronto a coppie:

* `specifiche al soggetto`: Caratterizzano i soggetti che eseguono i confronti tra oggetti, e quindi queste covariate variano solo tra soggetti. Ad esempio nel calcio, covariate come il numero spettatori o il meteo sono specifiche al soggetto. Perciò, sia x_p un vettore di covariate specifiche al soggetto, β_i il peso stimato delle covariate per ogni oggetto α_i e β_{i0} l'intercetta, allora l'abilità γ_{pi} dell'oggetto α_i nel soggetto p sarà

  $$\gamma_{pi} = \beta_{i0} + x_p^T \beta_i.$$

  Con l'inclusione di covariate specifiche al soggetto, il modello è in grado di spiegare l'eterogeneità sui soggetti. Le covariate specifiche al soggetto nei confronti a coppie sono state considerate da (Francis, Dittrich e Hatzinger, 2010) a (Turner e Firth, 2012).

* `specifiche all'oggetto`: Caratterizzano gli oggetti che vengono confrontati ma, non variano tra i soggetti ma tra gli oggetti. Nel calcio una covariata specifica all'oggetto può essere il valore di mercato della rosa della squadra di calcio. Un loro utilizzo lo si può trovare in (Schauberger e Tutz, 2017). Perciò, sia z_i un vettore di covariate specifiche all'oggetto, τ il peso uguale per tutti gli oggetti e β_{i0} l'intercetta, allora l'abilità γ_i dell'oggetto α_i sarà

$$\gamma_{pi} = \gamma_i = \beta_{i0} + z_i^T \tau.$$

Il peso τ è un parametro globale, che insieme a z_i rappresenta l'abilità spiegata delle covariate mentre β_{i0} rappresenta la parte dell'abilità non spiegata dalle covariate.

* `specifiche al soggetto e all'oggetto`: Questi tipi di covariate possono variare sia per oggetti e sia per i soggetti, ad esempio nel calcio il possesso palla è una covariata che varia per ogni singola squadra e per ogni singola partita. Tali variabili vengono approfondite da (Thurner e Eymann, 2000) a (Mauerer et al., 2015). Perciò, sia $z_p i$ un vettore di covariate specifiche al soggetto e all'oggetto, η_i il peso stimato delle covariate per ogni oggetto, β_{i0} l'intercetta, allora l'abilità γ_{pi} dell'oggetto α_i nel soggetto p sarà

$$\gamma_{pi} = \beta_{i0} + z_{pi}^T \eta_i.$$

Contrariamente alle coviariate specifiche al soggetto, le covariate specifiche al soggetto e all'oggetto posso essere modellate con un effetto globale, quindi γ_{pi} sarà

$$\gamma_{pi} = \beta_{i0} + z_{pi}^T \tau$$

dove τ rappresenta il peso stimato delle covariate. Come si può notare il parametro τ non ha alcun indice, questo perché l'effetto della covariate è uguale su tutti gli oggetti.

Nei vari punti presentati precedentemente, veniva aggiunto il parametro β_{i0}. Tale parametro è l'intercetta che è un parametro specifico all'oggetto. Tale parametro spiegata la maggior parte della forza dell'oggetto, infatti le covariate possono essere viste come estensioni contenenti effetti aggiuntivi dell'abilità dell'oggetto che non sono spiegati dall'intercetta. In tal senso, gli effetti della covariata possono aiutare a spiegare i risultati (imprevisti) di un soggetto che non possono essere completamente spiegati esclusivamente dall'intercetta.
Nella Sezione 4.3, viene presentato l'effetto dell'ordine degli oggetti in competizione. Invece dell'effetto d'ordine globale δ , che è uguale per tutti gli oggetti, è possibile specificare l'effetto d'ordine specifico per ogni oggetto α_i, quindi δ_i.
Nella Tabella 4.1 vengono riassunti tutti i tipi di covariate e tutte le possibili parametrizzazioni che possono essere applicate.
Quindi, il parametro abilità γ_{pi} di un oggetto α_i con i = 1,....,n su un soggetto p, p = 1,.....,m non è altro che una combinazione lineare dei parametri precedentemente spiegati. Da ciò si ottiene il modello capace di utilizzare le covariate. Tale modello viene chiamato modello strutturato e fa parte dei *generalized linear models* (GLMs). Riprendendo il modello 4.9 può essere riadatto nella seguente forma

$$P(Y_{p(i,j)} \leq h) = \frac{exp(\delta + \theta_h + \beta_{i0} - \beta_{j0} + x_{pi}^T \eta_i - x_{pj}^T \eta_j)}{1 + exp(\delta + \theta_h + \beta_{i0} - \beta_{j0} + x_{pi}^T \eta_i - x_{pj}^T \eta_j)} \tag{4.12}$$

4.5 Stima e penalizzazione

È importante considerare che con l'inserimento di un elevato numero di covariate si ha un aumento di complessità del modello. Dato che si utilizza un modello lineare, un

Tipo di covariate	**Tipo di effetto**	$\gamma_{pi} =$	$\gamma_{pj} =$	$\gamma_{p(ij)} = \gamma_{pi} - \gamma_{pj}$
Intercetta	Spec. all'oggetto	β_{i0}	β_{j0}	$\beta_{i0} - \beta_{j0}$
Effetto dell'ordine	Globale	$+ \delta$		$+ \delta$
Effetto dell'ordine	Spec. all'oggetto	$+ \delta_i$		$+ \delta_i$
Spec. al soggetto x_p	Spec. all'oggetto	$+ x_p^T \beta_i$	$+ x_p^T \beta_j$	$+ x_p^T (\beta_i - \beta_j)$
Spec.all'oggetto z_i	Globale	$+ z_i^T \tau$	$+ z_{si}^T \tau$	$+ (z_i - z_j)^T \tau$
Spec. al soggetto e all'oggetto $z_p i$	Globale	$+ z_{pi}^T \tau$	$+ z_{pj}^T \tau$	$+ (z_{pi} - z_{pj})^T \tau$
Spec. al soggetto e all'oggetto $z_p i$	Spec. all'oggetto	$+ x_{pi}^T \eta_i$	$+ x_{pj}^T \eta_i$	$+ x_{pi}^T \eta_i - x_{pj}^T \eta_j$

Tabella 4.1: La Tabella riassuntiva di tutti i tipi di covariate e di tutte le possibili parametrizzazioni applicabili.

eccessivo livello di complessità può portare a problemi di identificabilità ed efficienza. Infatti includendo soltanto una covariata specifica al soggetto e all'oggetto, questa ha un peso pari a n covariate dove n sono il numero di oggetti in considerazione. Oltretutto per ogni oggetto c'è la sua intercetta, perciò è necessario limitare il più possibile la complessità del modello. La soluzione è utilizzare metodi di *shrinkage* che includono termini di penalizzazione nelle procedure di stima. L'obiettivo è quello di ottenere un modello con una moderata complessità utilizzando solo i parametri realmente necessari.
Con l'inclusione della penalizzazione dei termini il modello potrebbe migliorare o leggermente peggiorare, ma la variabilità associata alle stime sarà minore. C'è perciò un trade-off di cui occuparsi, infatti più è forte la penalità inserita, più sarà elevata la varianza perché molte informazioni sulle variabili vengono perse. Non si massimizzerà la verosimiglianza ma la verosimiglianza penalizzata

$$l(\varepsilon)_p = l(\varepsilon) - \lambda J(\varepsilon)$$

dove $l(\varepsilon)$ è la log verosimiglianza con ε che rappresenta il vettore contenente tutti i parametri del modello. $J(\varepsilon)$ è un termine di penalizzazione. Il parametro λ è il parametro di Turing che stabilisce quanto forte deve essere la penalizzazione sui parametri. Per eseguire la penalizzazione è necessario trasformare in scale comparabili tutte le covariate.
Sono state utilizzate solo alcune modalità di penalizzazione tra quelle disponibili, quindi verranno esposte solo quelle effettivamente utilizzate. In (Schauberger e Tutz, 2019) vi è una trattazione completa di tutte le penalizzazioni applicabili.
Come metodo di penalizzazione verrà applicato l'*Adaptive Lasso* proposto da (Zou, 2006). Il metodo riduce i coefficienti ed esegue una selezione delle covariate applicando penalità di tipo L_1 per le differenze di coefficienti, di seguito verrà illustrato come è stato applicato.

Nel modello si è inserito l'effetto partita in casa come parametro con effetto specifico all'oggetto δ_i, la penalizzazione risultate è data dalle differenze assolute tra tutti i confronti.

$$P(\delta_1, ...\delta_m)_\delta = \sum_{i<j} |\delta_i - \delta_j|$$

È importante sottolineare che se ci sono molte differenze pari a zero, si ottengono gruppi di oggetti (nel nostro caso squadre) con un effetto identico della covariata penalizzata e che quindi la covariata deve avere un effetto globale piuttosto che specifico all'oggetto. Quindi con la penalizzazione è possibile capire quale tipo di effetto è più opportuno applicare.
Dato che non vi sono dubbi che l'effetto casa sia determinante per l'esito di una partita di calcio (Lago-Peñas et al., 2016), non verrà applicata nessun altra penalizzazione.
La penalizzazione per tutte le altre covariate (specifiche al soggetto e all'oggetto) è la seguente

$$P_\eta(\eta_1, ...\eta_m) = \sum_{p=1}^{m} \sum_{i<j} |\eta_{ip} - \eta_{jp}| + \sum_{p=1}^{m} \sum_{i<j}^{n} |\eta_{ip}|.$$

Rispetto alla penalizzazione precedente è stata aggiunta una penalizzazione al valore assoluto delle covariate. Questo perché non sappiamo in anticipo se una variabile è influente oppure no.

Le penalizzazione illustrate precedentemente se combinate permettono di ottenere il parametro $J(.) = P_\delta(.) + P_\eta(.)$.

4.5.1 Scelta del parametro di Turing

Un punto cruciale per le tecniche di *shrinkage* è la determinazione del parametro di Turing ottimo λ, cioè il grado di penalizzazione che ci da il miglior trade-off. Per farlo ci si affiderà alla *K-Fold Cross-Validation* (con k = 10), che sceglierà la miglior λ rispetto alla metrica *ranked probability score* (RPS).
Il RPS (Gneiting e Raftery, 2007) per categorie di risposte ordinate $y \in \{1, ..., K\}$ misura quanto siano buone le previsioni espresse come distribuzioni di probabilità rispetto ai valori osservati. Il RPS può essere cosi espresso

$$RPS(y, \pi(k)) = \sum_{k=1}^{K} (\pi(k) - \mathbb{1}(y \leq k))^2$$

dove $\pi(k)$ rappresenta la probabilità cumulativa $\pi(k) = P(y \leq k)$. A differenza delle altre possibili misure dell'errore, ad esempio la devianza, il RPS tiene conto dell'ordine di preferenza.

5 | Conclusioni

MEMO Riassunto del lavoro/risultati ottenuti, possibili estensione e migliorie che possono essere apportate. Sottolineare che alcune variabili possono avere un peso differente a seconda della lega in cui si svolge la partita, (ad esempio Premier league è un campionato più fisico con alti ritmi rispetto alla Serie A che è più "tattica") TO DO

6 | Appendice A

TO REWRITE

6.1 Codice di adattamento dataset per il trasferimento dati

Nella Figura 6.1 viene mostrato il codice applicato per adeguare il dataset con le modifiche scritte precedentemente.
Tale codice ha l'obbiettivo di prendere le due righe di ogni partita e di unirle insieme formando un unica riga per ogni partita. Successivamente si elimineranno le righe delle partite giocate fuori casa (AtHome = FALSE) dalle squadre indicate in Team mentre le righe delle partite giocate in casa (AtHome = TRUE) dalle squadre indicate in Team conteranno il risultato della fusione.
Perciò si è creato un vettore vuoto per ogni covariata presente nel dataset, ad eccezione di AtHome che verrà gestita in un modo diverso. Il vettore `del` è il vettore che tiene traccia di quali righe saranno da eliminare. `k` è l'indice usato per scorre il dataset per trovare i dati dell'avversario; `z` l'indice usato per inserire un nuovo elemento nel vettore `del`.
Il primo ciclo `for` scorre tutto il dataset alla ricerca delle righe con i dati delle partite giocate in casa dalla squadra indicata in `Team`, infatti al suo interno il primo costrutto `if` controlla se la partita è in casa per `Team` se sì, parte un secondo ciclo `for` che anche esso scorre tutto il dataset per cercare la riga con la partita giocata della squadra indicata in `Vs`; giocata ovviamente fuori casa. Perciò all'interno del secondo ciclo `for` c'è un costrutto `if` che controlla se la j-esima riga si riferisce alla stessa partita indicata nella i-esima riga, se sì allora si salvano tutti i dati nei vettori e si incrementa l'indice `k`. Se il primo `if` da esito negativo allora si andrà a inserire l'indice dell'i-esima riga nel vettore `del` perché contiene informazioni di una partita giocata fuori casa dalla squadra indicata in Team e viene incrementato l'indice di uno `z`.

,

```
PossVs <- c()
ShVs <- c()
ShTVs <- c()
G.ShVs <- c()
PAttVs <- c()
PCmp.Vs <- c()
SPAttVs <- c()
SPCmp.Vs <- c()
MPAttVs <- c()
MPCmp.Vs <- c()
LPAttVs <- c()
LPCmp.Vs <- c()
ToDef3rdVs <- c()
ToMid3rdVs <- c()
ToAtt3rdVs <- c()
```

```
ToAttPenVs <- c()
ToDistVs <- c()
FlsVs <- c()
FldVs <- c()
CrsVs <- c()
IntVs <- c()
TklWinVs <- c()
RecovVs <- c()
del <-c()
k <- 1
z <- 1
for(i in 1:nrow(soccern)){
  if(soccern$AtHome[i] == TRUE){
    for(j in 1:nrow(soccern)){
      if((soccern$Team[j] == soccern$Vs[i]) && (soccern$Team[i] ==
    soccern$Vs[j]) && (soccern$AtHome[j] == FALSE)){
        PossVs[k] <- soccern$Poss[j]
        ShVs[k] <- soccern$Sh[j]
        ShTVs[k] <- soccern$SoT[j]
        G.ShVs[k] <- soccern$G.Sh[j]
        PAttVs[k] <- soccern$PAtt[j]
        PCmp.Vs[k] <- soccern$PCmp.[j]
        SPAttVs[k] <- soccern$SPAtt[j]
        SPCmp.Vs[k] <- soccern$SPCmp.[j]
        MPAttVs[k] <- soccern$MPAtt[j]
        MPCmp.Vs[k] <- soccern$MPCmp.[j]
        LPAttVs[k] <- soccern$LPAtt[j]
        LPCmp.Vs[k] <- soccern$LPCmp.[j]
        ToDef3rdVs[k] <- soccern$ToDef3rd[j]
        ToMid3rdVs[k] <- soccern$ToMid3rd[j]
        ToAtt3rdVs[k] <- soccern$ToAtt3rd[j]
        ToAttPenVs[k] <- soccern$ToAttPen[j]
        ToDistVs[k] <- soccern$TotDist[j]
        FlsVs[k] <- soccern$Fls[j]
        FldVs[k] <- soccern$Fld[j]
        CrsVs[k] <- soccern$Crs[j]
        IntVs[k] <- soccern$Int[j]
        TklWinVs[k] <- soccern$TklWin[j]
        RecovVs[k] <- soccern$Recov[j]
        k <- k + 1
      }
    }
  }
  else{
    del[z] <- i
    z <- z + 1
  }
}
```

Di seguito vengono riportati i comandi fatti per applicare le modifiche al dataset.

```
  > soccern3 <- soccern2[-del,]
```

Con il precedente comando si va a creare un nuovo dataset con 380 righe, eliminando tutte quelle righe con valore `FALSE` su AtHome.

Con il comando mostrato nella Figura 6.1 si va a modificare Team rendendolo un `data.frame`, andando a inserire i dati della riga relativi alla squadra che gioca in casa. Si inserisce come chiave `team = soccern3$Team` e si indica che la partita è in casa per

la squadra di riferimento con `at.home = 1`.

```
> soccern3$Team <- data.frame(team = soccern3$Team, GF = soccern3$GF,
  GA = soccern3$GA,  at.home = 1, Poss = soccern3$Poss, Sh = soccern3$
  Sh, SoT = soccern3$SoT, G.Sh = soccern3$G.Sh, PAtt = soccern3$PAtt,
  PCmp. = soccern3$PCmp., SPAtt = soccern3$SPAtt, SPCmp. = soccern3$
  SPCmp., MPAtt = soccern3$MPAtt, MPCmp. = soccern3$MPCmp., LPAtt =
  soccern3$LPAtt, LPCmp. = soccern3$LPCmp., ToDef3rd = soccern3$
  ToDef3rd, ToAtt3rd = soccern3$ToAtt3rd, ToAttPen = soccern3$ToAttPen,
   TotDist = soccern3$TotDist, Fls = soccern3$Fls, Fld = soccern3$Fld,
  Crs = soccern3$Crs, Int = soccern3$Int, TklWin = soccern3$TklWin,
  Recov = soccern3$Recov)
```

Listing 6.1: Codice per la creazione del data.frame Team

Con il comando mostrato nella Figura 6.2 si va a modificare `Vs` rendendolo un `data.frame`, andando a inserire i dati della riga relativi alla squadra che gioca fuori casa. Si inserisce come chiave `team = soccern3$Vs` e si indica che la partita è fuori casa per la squadra `Vs` con `at.home = 0`.
Per quanto riguarda il resto dei dati, vengono riportati attraverso l'inserimento dei vettori costruiti e riempiti precedentemente.

```
> soccern3$Vs <- data.frame(team = soccern3$Vs, GF = GFVs, GA = GAVs,
  at.home = 0, Poss = PossVs, Sh = ShVs, SoT = ShTVs, G.Sh = G.ShVs,
  PAtt = PAttVs, PCmp. = PCmp.Vs, SPAtt = SPAttVs, SPCmp. = SPCmp.Vs,
  MPAtt = MPAttVs, MPCmp. = MPCmp.Vs, LPAtt = LPAttVs, LPCmp. = LPCmp.
  Vs, ToDef3rd = ToDef3rdVs, ToAtt3rd = ToAtt3rdVs, ToAttPen =
  ToAttPenVs, TotDist = ToDistVs, Fls = FlsVs, Fld = FldVs, Crs = CrsVs
  , Int = IntVs, TklWin = TklWinVs, Recov = RecovVs)
```

Listing 6.2: Codice per la creazione del data.frame Vs

Bibliografia

Riferimenti bibliografici

Agresti, Alan (1992). Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **41** (2), 287–297.

Bradley, Ralph Allan e Milton E Terry (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39** (3/4), 324–345.

Davidson, Roger R (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65** (329), 317–328.

Francis, Brian, Regina Dittrich e Reinhold Hatzinger (2010). Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do Europeans get their scientific knowledge? *The annals of applied statistics*, 2181–2202.

Gneiting, Tilmann e Adrian E Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* **102** (477), 359–378.

Lago-Peñas, Carlos et al. (2016). Home advantage in football: Examining the effect of scoring first on match outcome in the five major European leagues. *International Journal of Performance Analysis in Sport* **16** (2), 411–421.

Mauerer, Ingrid et al. (2015). Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the lasso approach. *Journal of choice modelling* **16**, 23–42.

R Core Team (2022). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. URL: `http://www.R-project.org/`.

Schauberger, Gunther e Gerhard Tutz (2017). Subject-specific modelling of paired comparison data: A lasso-type penalty approach. *Statistical Modelling* **17** (3), 223–243.

— (2019). BTLLasso: a common framework and software package for the inclusion and selection of covariates in Bradley-Terry models. *Journal of Statistical Software* **88**, 1–29.

Springall, A (1973). Response Surface Fitting Using a Generalization of the Bradley-Terry Paired Comparison Model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **22** (1), 59–68.

Thurner, Paul W e Angelika Eymann (2000). Policy-specific alienation and indifference in the calculus of voting: A simultaneous model of party choice and abstention. *Public Choice* **102** (1), 49–75.

Turner, Heather e David Firth (2012). Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software* **48**, 1–21.

Tutz, Gerhard (1986). Bradley-Terry-Luce models with an ordered response. *Journal of mathematical psychology* **30** (3), 306–316.

Zou, Hui (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101** (476), 1418–1429.

Sitografia

Riferimenti bibliografici

Library bradleyterry2. Libreria per la modellazione delle paired comparisons. URL: https://cran.r-project.org/package=BradleyTerry2.