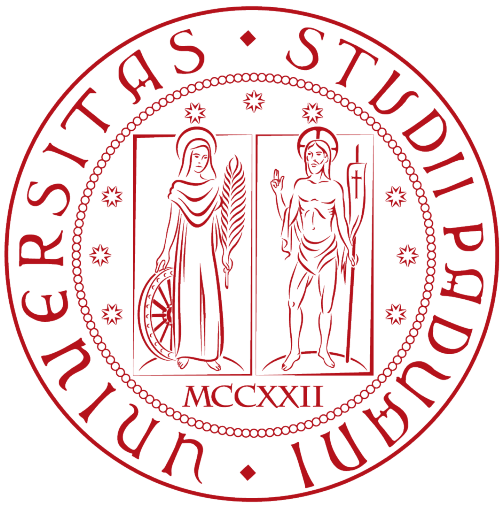
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea Magistrale in Informatica



Modellazione di un modello Bradley-Terry per l'individuazione delle variabili significative per l'esito di una partita di calcio nella Seria A italiana

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.
WORK IN PROGRESS + 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 descritto in dettaglio la raccolta dati effettuata per costruire il dataset riguardante le partite di calcio della Serie A italiana della stagione 2021/2022 e di come tale dataset è strutturato descrivendone le variabili e i dati al suo interno, utilizzati per l'analisi descritta precedentemente.

2.1 Serie A 2021/2022

L'analisi che è stata 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 ad esempio, Milan e Inter sono due squadre di 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 che in base al sorteggio della creazione del calendario delle partite decide quale delle due partite sarà giocata in casa oppure fuori casa (in casa dell'avversario).
Tale torneo nella 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 dove ogni turno è 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 vinta la squadra vincente guadagna 3 punti, per ogni pareggio le due squadre avversarie guadagnano entrambe un punto, mentre per ogni sconfitta la squadra perdente non guadagna punti. 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, dove il posto delle tre squadre retrocesse verrà presso da tre squadre della Serie B che hanno guadagnato la promozione alla Serie A.
La classifica della stagione 2021/2022 è mostrata 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 o il numero di tiri in porta prodotto dalla squadra in una determinata partita. L'obiettivo principale di questo lavoro è determinare l'influenza di queste variabili specifiche della partita.
Per creare il dataset per tale scopo, sono state raccolte un gran numero di variabili

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.

che a primo avviso possono essere significative, tali dati sono stati offerti dal sito web FBref.
FBref è un sito web dedicato al tracciamento delle statistiche relative ai calciatori e alle squadre di calcio 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, in più per rendere più facile l'esportazione, tali tabelle possono essere convertite in formato di CSV per poter essere poi trasportate in un file Excel.

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

Quindi per ogni squadra che ha partecipato alla stagione 2021/2022 di Serie A si è esportato per ogni partita giocata alcune variabili che ci interessavano, selezionando per prima cosa la macro aree dove si trovavano le variabili d'interesse e poi, modificando le tabelle per ottenere solo i dati di tali variabili. Ogni tabella generata veniva poi riconvertita in CSV per essere poi unita con tutte le altre in un file Excel che una volta completato, divenne il dataset per le nostre analisi. Per rendere più leggibile il file Excel, dato che le stringhe in CSV separavano i dati con il carattere separatore virgola, si è utilizzata la funzione di Excel "trasforma testo in colonne" per inserire tutti i dati in modo ordinato nelle celle del foglio Excel.

2.2.1 Struttura dataset

Il dataset risultate 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 perciò contiene informazioni riguardati solo la squadra indicata in Team fatta eccezioni per la data della partita (Date), il turno (Round), e gli spettatori (Spec). Quindi per ogni partita esistono due righe, una per ognuna delle due squadre coinvolte. Perciò ogni squadra appare nella colonna Team 38 e dato che si hanno 20 squadre si hanno perciò 760 righe totali. Per quanto riguarda le colonne se ne discuterà nella prossima sotto sezione.
La Tabella 2.2 mostra un breve estratto dei dati riguardanti le prime tre partite della stagione.

2.2.2 Covariate

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 saranno le possibili candidate a essere le covariate che costituiranno il modello. Ovviamente non è detto che tutte queste variabili saranno inserite nel modello perché prima di costruire un modello, ci sarà un analisi per verificare se sia sensato o no l'utilizzo di ognuna delle variabili verificando attraverso grafici (analisi grafica) e individuando possibili problemi di multicollinearità o di bassa significatività delle variabili.

Le possibili covariate sono le seguenti:

* AtHome: Tale variabile indica se la squadra indicata sulla variabile Team gioca nel suo 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, che 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

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.

calcio rispetto a altre sei squadre che hanno avuto l'effetto opposto, mentre le rimanti tre hanno avuto un effetto nullo. Alla luce di questo è stato deciso di inserire tale variabile per via del suo effetto nell'esito di una partita in generale.

* Res: Tale variabile indica se la squadra indicata sulla variabile Team ha vinto o ha pareggiato o ha perso. Per indicare se ha vinto viene inserito il valore 1, se ha pareggiato 0, altrimenti se ha perso -1. Chiaramente questa variabile sarà la nostra Y cioè la variabile risposta che il modello deve riuscire a prevedere.

* GF: Tale variabile indica il numero di gol fatti dalla squadra indicata sulla variabile Team.

  Questa variabile è stata inserita perché può permettere di valutare la qualità della fase offensiva della squadra e quindi essere significativa ai fine dell'analisi.

* GA: Tale variabile indica il numero di gol subiti dalla squadra indicata sulla variabile Team e quindi fatti dalla squadra indicata nella variabile Vs.

  Questa variabile è significativa perché subire pochi gol incide positivamente nell'esito della partita, infatti non espone la squadra a doversi sbilanciare in attacco per poter recuperare lo svantaggio e quindi non rischiare di subire altri gol dai avversari. Inoltre è un fatto riconosciuto che aver la miglior difesa del campionato porta con molta probabilità a vincere il campionato

* Team: Tale variabile indica il nome della squadra a cui i dati della riga fanno riferimento. É necessaria per il funzionamento del modello, nel prossimo paragrafo verrà approfondito il suo utilizzo nel modello.

* Vs: Tale variabile indica il nome della squadra avversaria. É necessaria per il funzionamento del modello, nel prossimo paragrafo verrà approfondito il suo utilizzo nel modello.

* Poss: Tale variabile indica in percentuale, la quantità di tempo di possesso della palla durante una partita di calcio della squadra indicata sulla 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. In poche parole 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 cioè ad esempio: rimesse laterali, calci di punizione ecc.. I tempi vengono poi trasformati in percentuali. Per una registrazione più sofisticata, si può utilizzare 22 cronometri, uno per ogni giocatore, in modo da registrare anche il possesso palla di ogni singolo giocatore per avere una registrazione più precisa.

  Tale variabile è stata inserita perché, la supremazia nel possesso palla è solitamente desiderabile e utile infatti si possono avere 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 la squadra A sta vincendo con un gol di scarto e per non rischiare attacchi dalla squadra B, "congela" il risultato mantenendo il possesso della palla.

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

* Sh: Tale variabile indica il numero di tiri totali fatti dalla squadra indicata sulla 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 pero capire quanto è precisa una squadra nel centrare la porta.

* SoT: Tale variabile indica il numero di tiri in porta totali fatti dalla squadra indicata sulla variabile Team.

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

* G/Sh: Tale variabile indica la proporzione tra gol e tiri fatti dalla squadra indicata sulla variabile Team.

  Tale variabile perciò permette di capire quanto la produzioni di tiri della squadra è efficace o meno. Con `Sh` e `SoT` si riesce a valutare quanto è offensiva la squadra cioè, se essa gioca costantemente in attacco o utilizza la tattica difesa e contropiede. Inoltre permette di capire quanto la squadra è precisa nel effettuare i tiri in porta.

* Saves: Tale variabile indica il numero di parate fatte del portiere della squadra indicata sulla variabile Team.

  La variabile è stata inserita perché permette di valutare se la squadra subisce tanti tiri dai avversari e la qualità del portiere nel salvare la squadra da un possibile gol subito.

* PAtt: Tale variabile indica il numero di tutti i passaggi tentati dai giocatori della squadra indicata sulla variabile Team.

  Utile a capire quanto la squadra sia incline a tentare i passaggi. Si studierà nell'analisi se tale variabile è significativa ma sicuramente ha un maggior significato se messa a confronto con la percentuale di passaggi riusciti `PCmp%`.

* PCmp%: Tale variabile indica la percentuale di passaggi riusciti ai giocatori della squadra indicata sulla variabile Team.

  Questa variabile 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.

* SPAtt: Tale variabile indica il numero di passaggi corti tentati dai giocatori della squadra indicata sulla variabile Team. Per passaggi corti si intendono quelli effettuati all'interno di una lunghezza tra i tre e 14 metri.

  Questa variabile è stata inserita per capire se un alto numero di passaggi corti possono essere determinanti ai fini dell'esito della partita. Ovviamente analogamente a `PAtt` occorre fare un confronto con la sua percentuale di passaggi corti riusciti `SPCmp%`.

* SPCmp%: Tale variabile indica la percentuale di passaggi corti riusciti ai giocatori della squadra indicata sulla variabile Team.

  Questa variabile 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.

* MPAtt: Tale variabile indica il numero di passaggi medi tentati dai giocatori della squadra indicata sulla variabile Team. Per passaggi medi si intendono quelli effettuati all'interno di una lunghezza tra i 13 e 27 metri. Questi passaggi possono essere considerati come passaggi filtranti cioè un tipo di passaggio non diretto direttamente 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 intercettino la palla. Nella Figura 2.2 viene mostrato l'esecuzione di un passaggio filtrante.

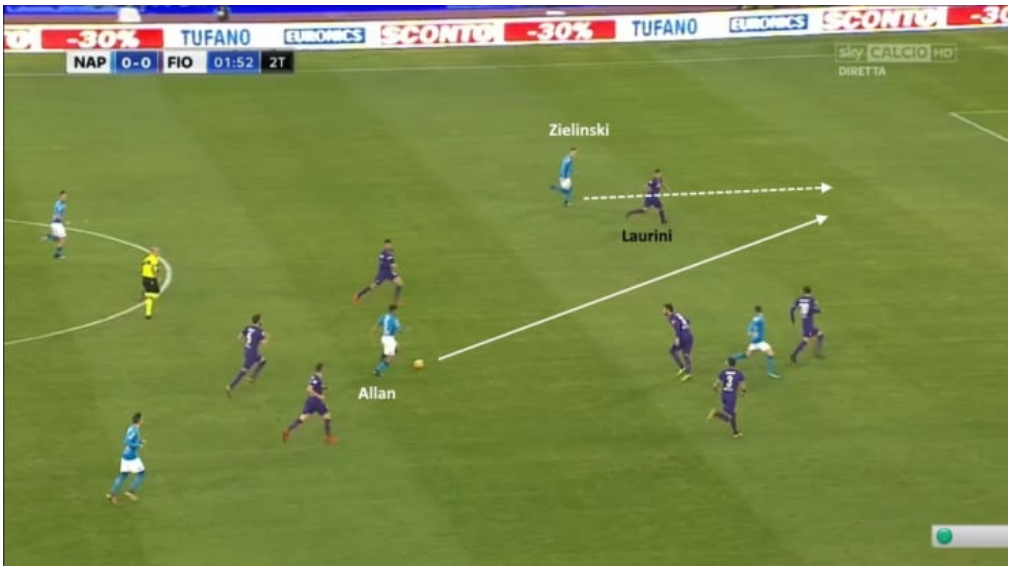


Figura 2.2: Esecuzione di un passaggio filtrante

Questa variabile è stata inserita per capire se un alto numero di passaggi medi possono essere determinanti ai fini dell'esito della partita. Ovviamente analogamente a `PAtt` occorre fare un confronto con la sua percentuale di passaggi corti riusciti `MPCmp%`.

* MPCmp%: Tale variabile indica la percentuale di passaggi medi riusciti ai giocatori della squadra indicata sulla variabile Team. Questa variabile 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.

* LPAtt: Tale variabile indica il numero di passaggi lunghi tentati dai giocatori della squadra indicata sulla variabile Team. Per passaggi corti si intendono quelli effettuati all'interno di una lunghezza superiore ai 27 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 2.3.

  Questa variabile è stata inserita per capire se un alto numero di passaggi lunghi possono essere determinanti ai fini dell'esito della partita. Ovviamente analogamente a `PAtt` occorre fare un confronto con la sua percentuale di passaggi corti riusciti `LPCmp%`.

* LPCmp%: Tale variabile indica la percentuale di passaggi lunghi riusciti ai giocatori della squadra indicata sulla variabile Team.

  Questa variabile 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.

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

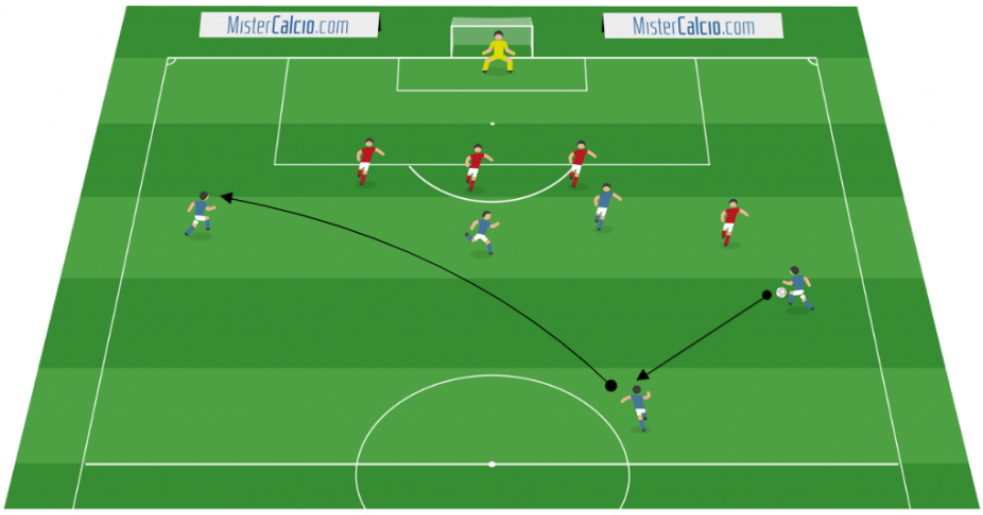


Figura 2.3: Esecuzione di un cambio di gioco

Questa variabile è stata inserita perché può essere utile per capire come il possesso della palla viene gestito, cioè 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 ToDef3rd, ToMid3rd, ToAtt3rd e ToAttPen permette di capire se il possesso della palla fatto della squadra è utile e porta benefici ai fini del risultato oppure è sterile. Inoltre si vuole capire attraverso l'analisi in che misura può influenzare il risultato della partita con un alto o un basso valore di numero di tocchi nella propria area di rigore la cui area nel campo da calcio è indicata nella Figura 2.4.

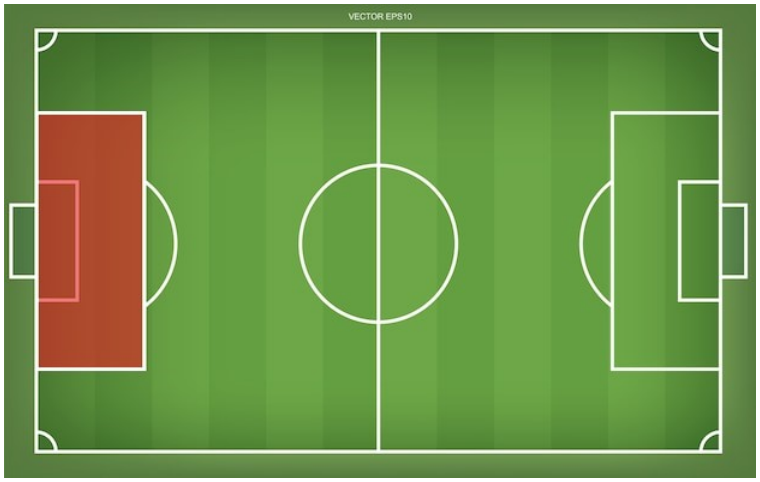


Figura 2.4: In rosso l'area di rigore in un campo da calcio.

* ToDef3rd: Tale variabile indica il numero di tocchi fatti dai giocatori della squadra

indicata sulla variabile Team nella propria mediana o trequarti difensiva.

Questa variabile è stata inserita perché può essere utile per capire come il possesso della palla viene gestito, cioè 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 ToDef3rd, ToMid3rd, ToAtt3rd e ToAttPen permette di capire se il possesso della palla fatto della squadra è utile e porta benefici ai fini del risultato oppure è sterile. Inoltre si vuole capire attraverso l'analisi in che misura può influenzare il risultato della partita con un alto o un basso valore di numero di tocchi nella propria mediana la cui area nel campo da calcio è indicata nella Figura 2.5.

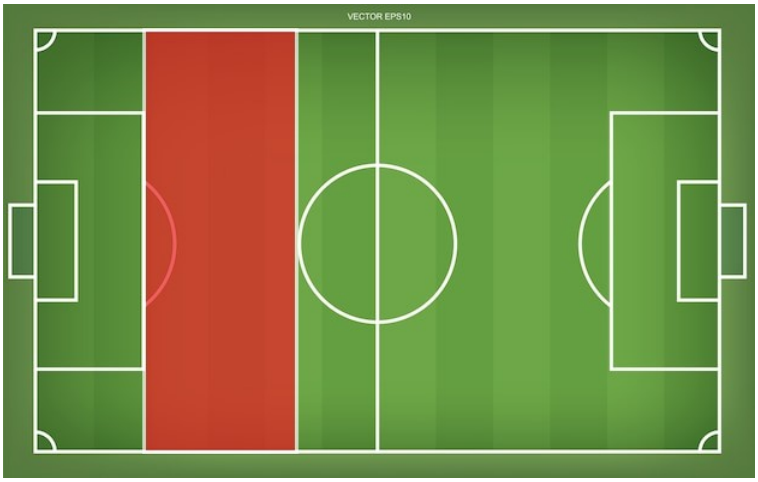


Figura 2.5: In rosso la mediana nel campo da calcio.

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

  Questa variabile è stata inserita perché può essere utile per capire come il possesso palla viene gestito, cioè 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 ToDef3rd, ToMid3rd, ToAtt3rd e ToAttPen permette di capire se il possesso della palla fatto dalla squadra è utile e porta benefici ai fini del risultato oppure è sterile. Inoltre si vuole capire attraverso l'analisi in che misura può influenzare il risultato della partita con un alto o un basso valore di numero di tocchi a centrocampo la cui area nel campo da calcio è indicata nella Figura 2.6.

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

  Questa variabile è stata inserita perché può essere utile per capire come il possesso della palla viene gestito, cioè 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 ToDef3rd, ToMid3rd, ToAtt3rd e ToAttPen permette di capire

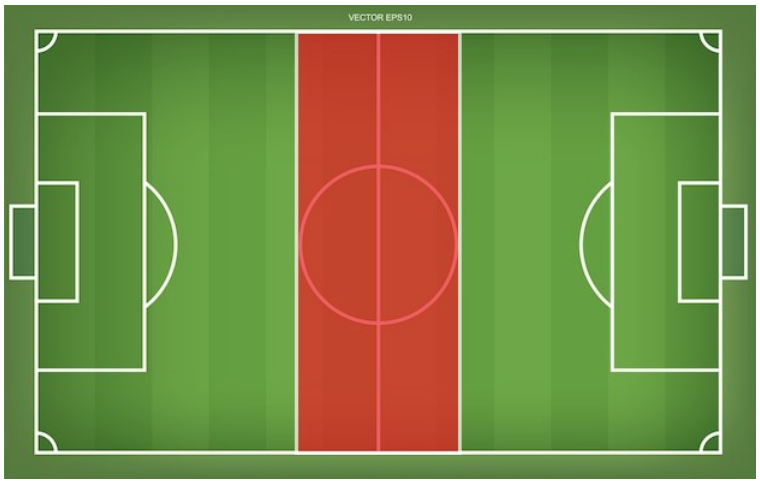


Figura 2.6: In rosso il centrocampo nel campo da calcio.

se il possesso della palla fatto della squadra è utile e porta benefici ai fini del risultato oppure è sterile. Inoltre si vuole capire attraverso l'analisi in che misura può influenzare il risultato della partita con un alto o un basso valore di numero di tocchi nella trequarti dell'avversario la cui area nel campo da calcio è indicata nella Figura 2.7.

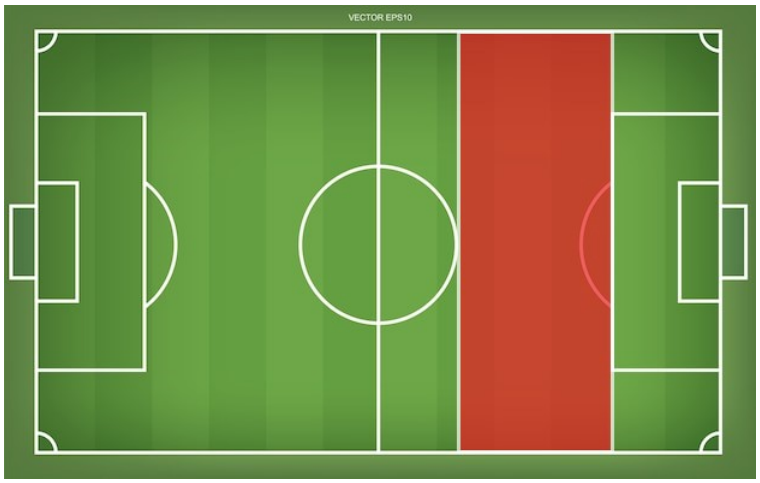


Figura 2.7: In rosso la trequarti dell'avversario nel campo da calcio.

∗ ToAttPen: Tale variabile indica il numero di tocchi fatti dai giocatori della squadra indicata sulla variabile Team a nell'area di rigore dell'avversario.

Questa variabile è stata inserita perché può essere utile per capire come il possesso della palla viene gestito, cioè 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 ToDef3rd, ToMid3rd, ToAtt3rd e ToAttPen permette di capire se il possesso della palla fatto della squadra è utile e porta benefici ai fini del risultato oppure è sterile. Inoltre si vuole capire attraverso l'analisi in che misura può influenzare il risultato della partita con un alto o un basso valore di numero di tocchi nell'area di rigore dell'avversario.

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

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

* Fls: Tale variabile indica il numero di falli dai giocatori della squadra indicata sulla 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 può andare ad influire sull'esito della partita, ricordando però che una che commette molti falli è più soggetta a ricevere cartellini gialli o rossi che condizionano la prestazione dei giocatori.

* Fld: Tale variabile indica il numero di falli subiti dai giocatori della squadra indicata sulla variabile Team da parte della squadra avversaria indicata sulla 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. Si vuole perciò capire come questa variabile può andare ad influire sull'esito della partita.

* Off: Tale variabile indica il numero di volte che la squadra indicata sulla 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 perché non possono essere usate per controllare il pallone; si trova nella nella metà campo avversaria ed è più vicina alla linea di porta avversaria sia rispetto al pallone e sia rispetto al penultimo giocatore difendente avversario; tale penultimo avversario può essere anche il portiere, se un compagno di questi è più vicino di lui alla linea di porta. Una rappresentazione grafica del fuorigioco è mostrata nell'immagine 2.8.

  Tale variabile è stata inserita perché, se una squadra viene colta molte volte in fuorigioco allora il suo gioco sarà interrotto e darà un vantaggio alla squadra avversaria che farà ripartire la sua azione a suo favore.

* Crs: Tale variabile indica il numero di cross effettuati dalla squadra indicata sulla 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 se eseguito permette al compagno di squadra posizionato vicino alla porta avversaria, di colpire la palla al volo di testa

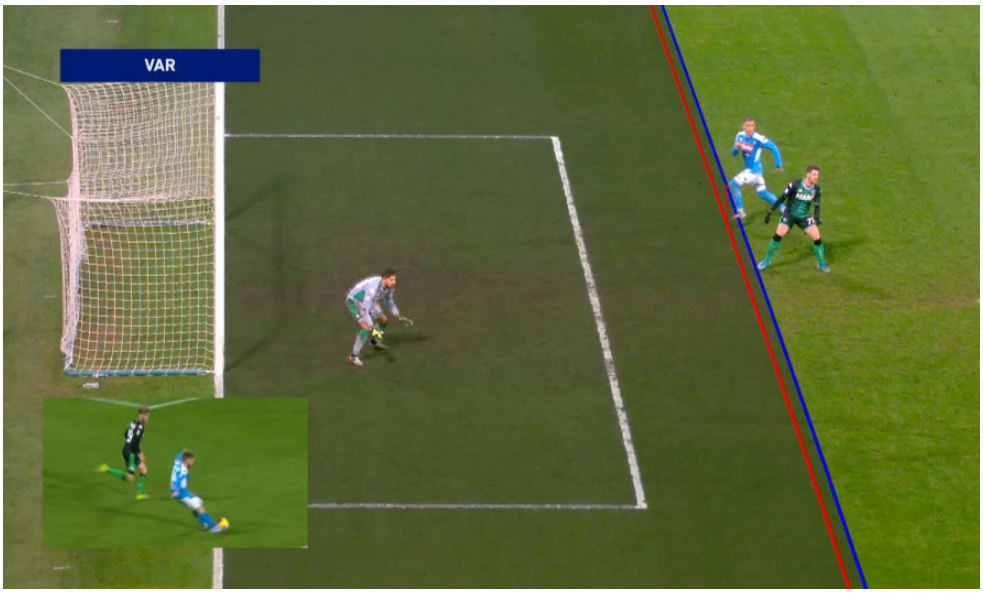


Figura 2.8: Rappresentazione del fuorigioco

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.

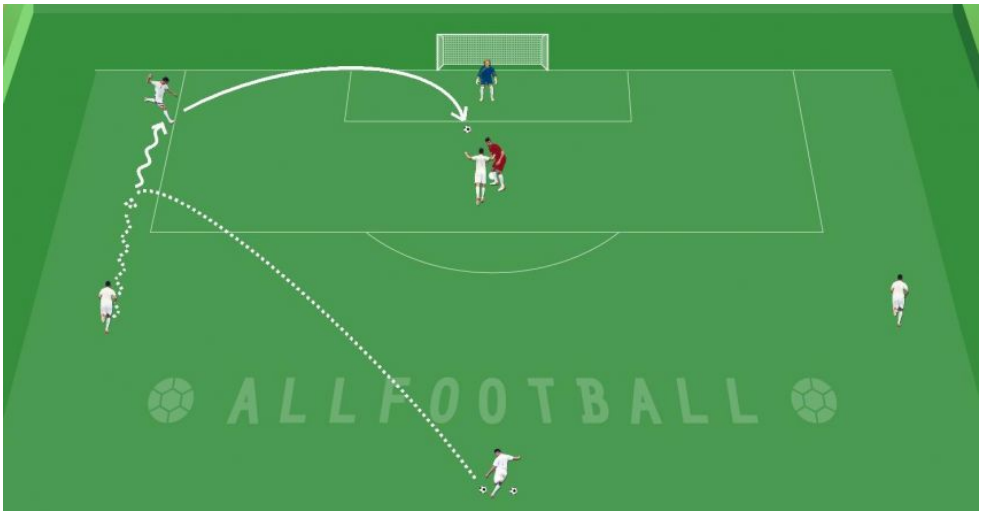


Figura 2.9: Rappresentazione di un cross

Per tale motivo si è deciso di inserire una variabile specifica per questo tipo di passaggio nell'analisi. Una rappresentazione di un cross è mostrata nella Figura 2.9.

* Int: Tale variabile indica il numero di intercettazioni fatte dai giocatori della squadra indicata sulla variabile Team. Per intercettazione della palla si intende

interrompere un passaggio della squadra avversaria entrando in possesso del pallone che era stato lanciato per un passaggio ma che una volta intercettato non è andato a buon fine cioè non è arrivato al compagno del giocatore avversario che ha effettuato il passaggio.

Quindi si è deciso di inserire questa variabile perché indica quante volte si è tolto il possesso della palla all'avversario interrompendone il suo gioco.

* TklWin: Tale variabile indica il numero di contrasti vinti dai giocatori della squadra indicata sulla 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. Si sottolinea che i contrasti vengono anche effettuati per allontanare dalle zone pericolose l'avversario. La Figura 2.10 mostra un contrasto di gioco.

Figura 2.10: Rappresentazione di un contrasto in scivolata

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

* Recov: Tale variabile indica il numero di palle vaganti recuperate dalla squadra indicata sulla 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. Si ha che nessuno ha in possesso il pallone e quindi si ha una palla vagante.

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

2.3 Preprocessing dei dati

Nella sezione precedente si è descritto come si è costruito il dataset che verrà utilizzato per l'analisi e come esso è stato strutturato. Tale struttura ha il vantaggio di essere di facile interpretazione per un essere umano ma vi sono alcune criticità che non lo permettono di essere utilizzato correttamente all'interno del modello messo a disposizione dal pacchetto `BradleyTerry2`.
Dopo aver importato il dataset sul software R, sono state perciò necessarie apportare alcune modifiche attraverso la scrittura di codice che andasse a modificare la struttura del dataset per poi essere correttamente utilizzabile nel modello.

Innanzitutto il modello richiede per il suo funzionamento che le due variabili che indicano quali delle due squadra hanno partecipato alla partita in esame; devono essere o di tipo fattore oppure un data.frame. Una variabile fattore è un variabile non numerica, espressa in termini verbali ad esempio una categoria. Un data.frame è una lista di vettori, che devono avere tutti la stessa lunghezza, ma possono essere di tipo diverso: variabili nominali cioè fattori, variabili cardinali cioè vettori numerici; un data.frame può essere visto come una matrice ma con il tipo dei valori che può essere diverso.
Dato che le due variabili in questione Team e Vs erano solo di tipo `character` e si voleva inserire un collegamento che faccia capire al modello, quali valori sono legati alla squadra indicata in Team e quali in Vs nella stessa partita; si è deciso perciò di trasformare Team e Vs in `data.frame` inserendo al loro interno tutte le covariate descritte nella sezione precedente, ad esempio Poss, Int ecc..

Si sottolinea inoltre che sono state necessarie ulteriore modifiche per quanto riguarda la variabile AtHome; dato che al momento dell'importazione del dataset, i valori venivano interpretati come stringhe, è stato necessario trasformarli in valori logici con il commando as.logical(soccern$AtHome). Ciononostante pero il valore logico non era accettato dal modello ma era accettato un valore numerico per indicare se la squadra giocava in casa o no; si è quindi convertito il valore `TRUE` in 1 mentre FALSE in 0.

2.3.1 Codice per l'adattamento del dataset

Di seguito viene mostrato il codice applicato per adeguare il dataset con le modifiche scritte precedentemente.

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

```
ToMid3rdVs <- c()
ToAtt3rdVs <- c()
ToAttPenVs <- c()
ToDistVs <- c()
FlsVs <- c()
FldVs <- c()
OffVs <- 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]
                 SavesVs[k] <- soccern$Saves[j]
                 PAttVs[k] <- soccern$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$LPCmp.[j]
                 ToDefPenVs[k] <- soccern$ToDefPen[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$Fld[j]
                 OffVs[k] <- soccern$Off[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
   }
}
```

Con il codice precedente si 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` vi 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`.

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

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

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

```
> soccern3$Team <- data.frame(team = soccern3$Team, GF = soccern
    3$GF, GA = soccern3$GA,  at.home = 1, Poss = soccern3$Poss, Sh
     = soccern3$Sh, SoT = soccern3$SoT, G.Sh = soccern3$G.Sh,
    Saves = soccern3$Saves, PAtt = soccern3$PAtt, PCmp. = soccern
    3$PCmp., SPAtt = soccern3$SPAtt, SPCmp. = soccern3$SPCmp.,
    MPAtt = soccern3$MPAtt, MPCmp. = soccern3$MPCmp., LPAtt =
    soccern3$LPAtt, LPCmp. = soccern3$LPCmp., ToDefPen = soccern3$
    ToDefPen, ToDef3rd = soccern3$ToDef3rd, ToAtt3rd = soccern3$
    ToAtt3rd, ToAttPen = soccern3$ToAttPen, TotDist = soccern3$
    TotDist, Fls = soccern3$Fls, Fld = soccern3$Fld, Off = soccern
    3$Off, Crs = soccern3$Crs, Int = soccern3$Int, TklWin =
    soccern3$TklWin, Recov = soccern3$Recov)
```

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

Con il precedente commando 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 vengo riportati attraverso l'inserimento dei vettori costruiti e riempiti precedentemente.

Si segnala inoltre che il dataset non contiene valori mancati dato che, in quei rari casi in cui venivano individui valori mancanti durante la raccolta, veniva ricercato il dato da altre fonti attendibili.

2.4 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 cercherà di: individuare possibili outliers o anomalie, quali distribuzioni hanno i dati ma soprattutto valutare le relazione tra covariate e variabile di risposta e tra due covariate, con lo scopo di individuare quali covariate possono essere significative per la variabile risposta e quali interazioni tra covariate emergono dall'analisi grafica.

Come primo passo dell'analisi, viene valutata la distribuzione delle classi della variabile risposta `Res` all'interno delle osservazioni disponibili. Tale distribuzione è mostrata nella Figura 2.11.

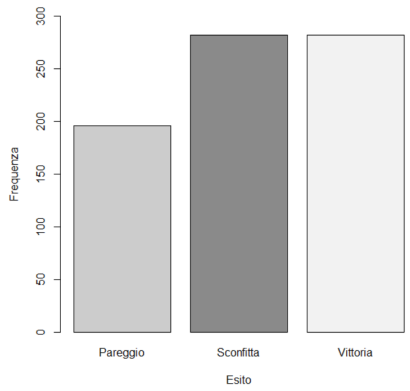


Figura 2.11: Barplot della distribuzione della variabile di risposta `Res`

Come si può notare le classi sembrano ben distribuite dato che abbiamo 196 pareggi e 282 vittorie e altrettante sconfitte. Si ha quindi un campione abbastanza ampio e distribuito e corretto per le nostre analisi.

Aumentando il livello di dettaglio è di interessante analizzare come queste classificazione sono distribuite tra le varie squadre.

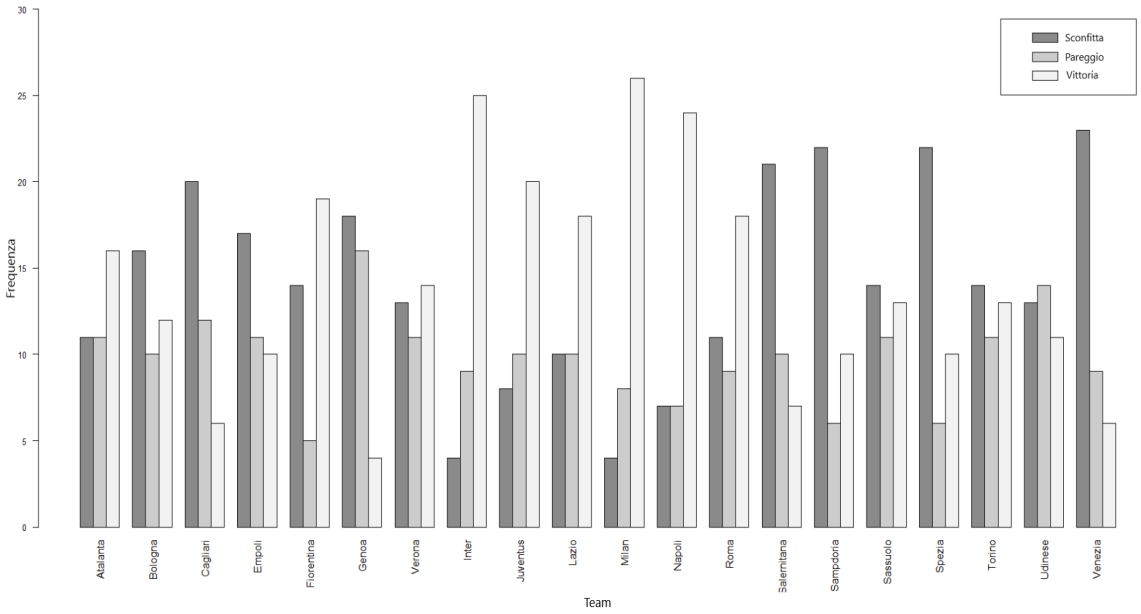


Figura 2.12: Barplot della distribuzione della variabile di risposta per squadra`Res`

Nella Figura 2.12 si può notare come la distribuzione di vittorie, pareggi e sconfitte non è omogenea tra le squadre. Ovviamente è un risultato che ci si aspettava ma che sottolinea prima di tutto la correttezza dei dati ma soprattutto che vi è qualche elemento nascosto che ha determinato tale distribuzione.

Come secondo step si analizzerà le relazione tra variabile di risposta con alcune covariate.

La prima relazione che si analizza è quella con la variabile categorica `AtHome`. Nella Figura 2.13 si può vedere che c'è una leggera variazione dei risultati tra la squadra che gioca in casa oppure no. Infatti c'è una leggera tendenza a favorire la vittoria per la squadra che gioca in casa piuttosto che la vittoria per la squadra fuori casa. Naturalmente non deve esserci alcuna variazione per quanto riguarda il pareggio dato che entrambe le squadre lo ottengono. Risulta perciò significativa la variabile `AtHome`.

Analizzando invece la relazione tra variabile di risposta e `Poss`, dalla Figura 2.14 si nota che tale variabile sembra essere significativa per l'esito. Infatti vi è un relazione positiva dove valori più alti di possesso palla sono registrati nel box della vittoria e ciò può portare a una maggiore probabilità di vittoria. Vi è una buona distribuzione dei dati, infatti le code sono simmetriche mentre vi è una variabilità quasi identica; si segnala solo che la mediana della sconfitta è più vicina al 3° quantile mentre quella della vittoria è più vicina al 1° quantile. Inoltre non vi sono presenti outliers.

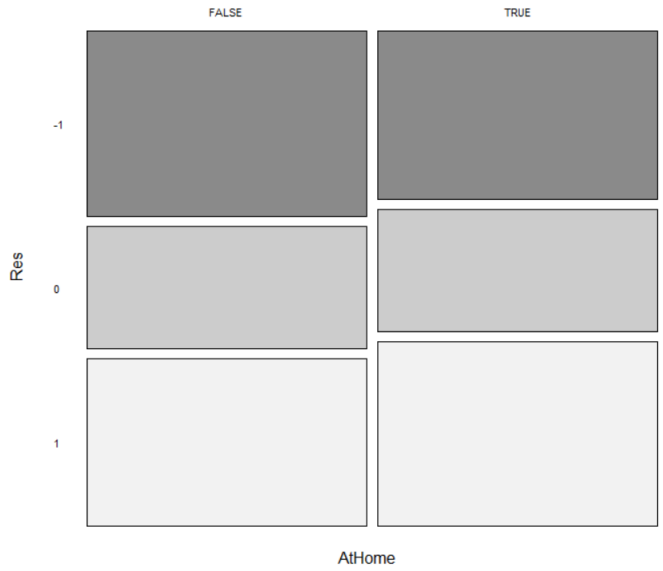


Figura 2.13: Mosaicplot che mostra la distribuzione degli esiti rispetto alle partite giocate in casa e fuori casa

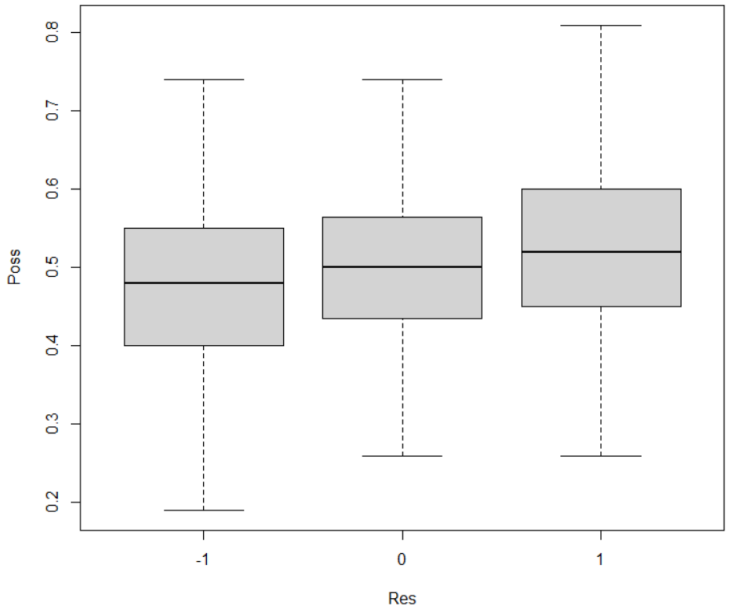


Figura 2.14: Boxplot della variabile risposta e della variabile numerica `Poss`

La Figura 2.15 mostra come si comporta la relazione con `SoT`. Come ci si aspetta si hanno valori più alti nella vittoria e valori molto più bassi nella sconfitta, si ha una buona distribuzione dei valori nella vittoria dato che le code sono simmetriche, per le altre due classi non c'è simmetria dato che ci sono valori più bassi rispetto a valori più alti. Vi sono inoltre alcuni outliners che si discostano dalla distribuzione di tutte e tre le classi dovuti al fatto che ci sono state squadre che hanno tirato molto in porta. Le mediane dei box pareggio e vittoria non sono equidistanti dai quantili ma più vicine al 1° quantile. Il box della sconfitta ha una bassa varianza. In conclusione avere un valore alto di tiri in porta sembra essere significativo ai fini della vittoria. Si segnala inoltre che si ottengono i stessi risultati anche con `Sh` solo con valori meno alti per la vittoria rispetto a `SoT`.

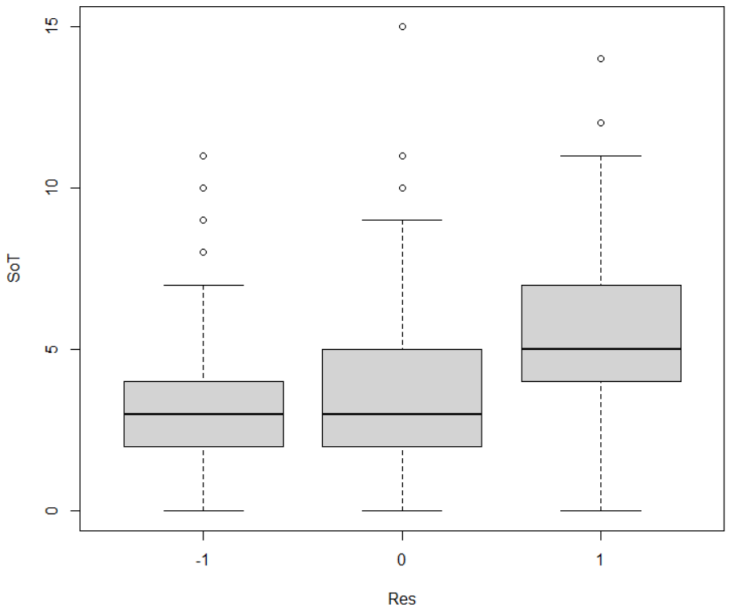


Figura 2.15: Boxplot della variabile risposta e della variabile numerica `SoT`

La Figura 2.16 mostra come si comporta la relazione con `G/Sh`. Si nota che vi sono valori molto bassi ma leggermente più alti per la vittoria. La distribuzione non è buona perché le code sono asimmetriche infatti tutti i valori sono concentrati in basso e pochi verso la coda in alto, per di più c'è una bassa varianza tra i valori. Vi è la presenza di outliners dovuti a partite dove le squadre sono riuscite a ottenere il massimo da ogni tiro. I risultati mostrati nonostante la pessima 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.

La Figura 2.17 mostra come si comporta la relazione con `Saves`. Come si può notare sembra che tale variabile sia poco significativa ai fini del risultato. Infatti c'è poca variazione tra una classe e l'altra dato che avere un alto numero di parate non è determinante a fini del risultato.

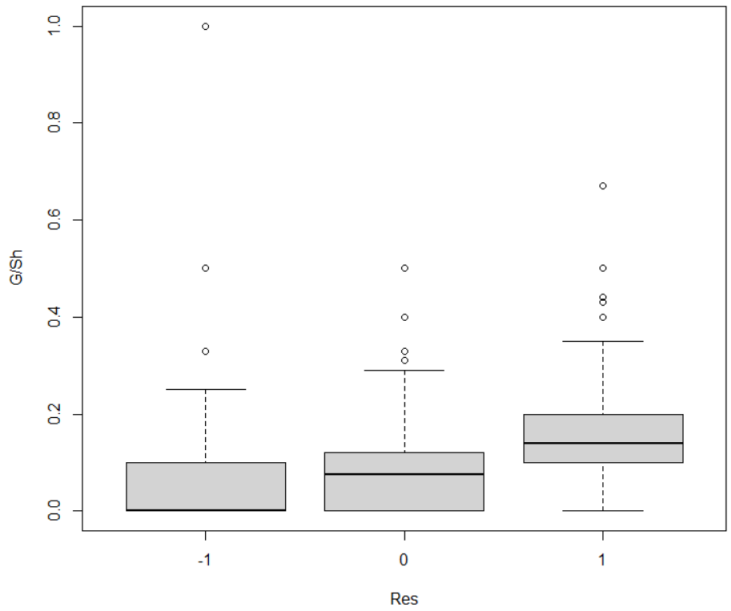


Figura 2.16: Boxplot della variabile risposta e della variabile numerica `G/Sh`

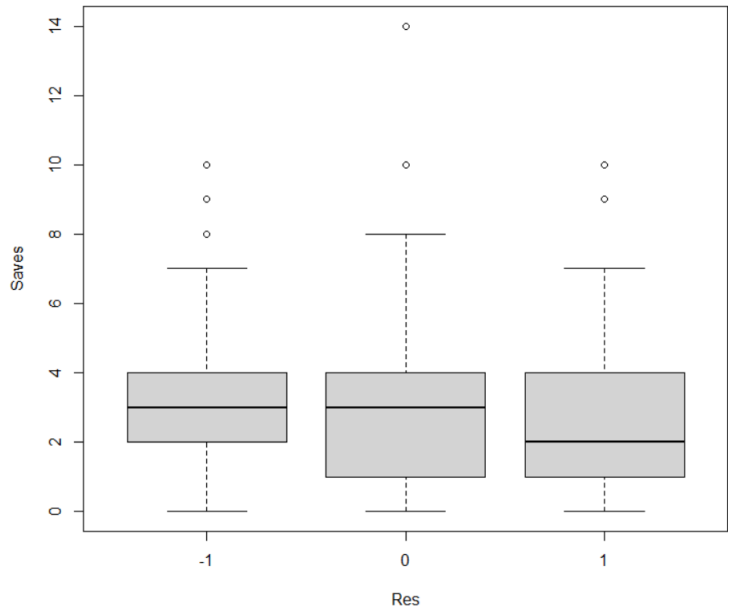


Figura 2.17: Boxplot della variabile risposta e della variabile numerica `Saves`

La Figura 2.18 mostra come si comporta la relazione con `PAtt` e con `PCmp%`. Per entrambi sembra significativo l'alto numero di passaggi tentati ma soprattutto quelli completati ai fini della vittoria. Nel primo boxplot la coda più in alto è più lunga rispetto alla coda in basso, quindi abbiamo valori più concentrati verso il basso che verso l'alto. Sempre nel primo boxplot il box della vittoria ha una maggiore variabilità rispetto ai altri due è varia di più avendo valori più alti; sia la mediana del box 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 dato che maggior numero di passaggi si prova ad effettuare maggiori sono le possibilità di vittoria, occorre pero sapere quanto è precisa la squadra.
Nel secondo boxplot si notano valori alti e molti outliners con valori bassi dovuti al fatto che ci sono state partite dove alcune squadre sono state poco precise nei passaggi. A differenza del primo boxplot il secondo boxplot ha molti valori alti, infatti la coda in alto e molto meno lunga rispetto alla coda in basso e le variabilità dei box sembrano essere uguali tra di loro; anche qui le code non sono simmetriche e quindi non c'è una buona distribuzione dei dati. Sorprendentemente sembra che avere un buona precisione pero non da la sicurezza di una vittoria, inoltre l'andamento prima scende da sconfitta a pareggio e poi sale da pareggio a vittoria.
Per quanto riguarda le variabili delle altre tipologie di passaggi si discostano di poco dai boxplot in Figura 2.18

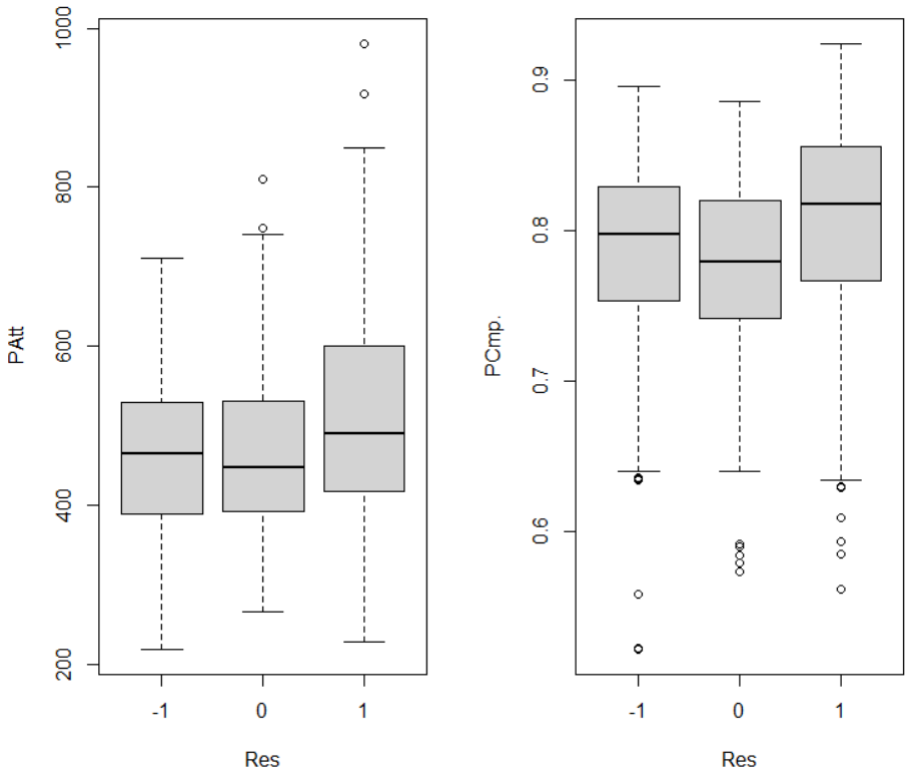


Figura 2.18: Boxplot della variabile risposta e della variabile numerica `PAtt` e `PCmp%`

La Figura 2.19 mostra come si comporta la relazione con `ToDefPen`. Come si può notare questa non è per nulla significativa per la variabile risposta, infatti non c'è una minima variazione e i box hanno tutti la stessa varianza. Tale 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 quindi tale variabile non sarà inserita nel modello.

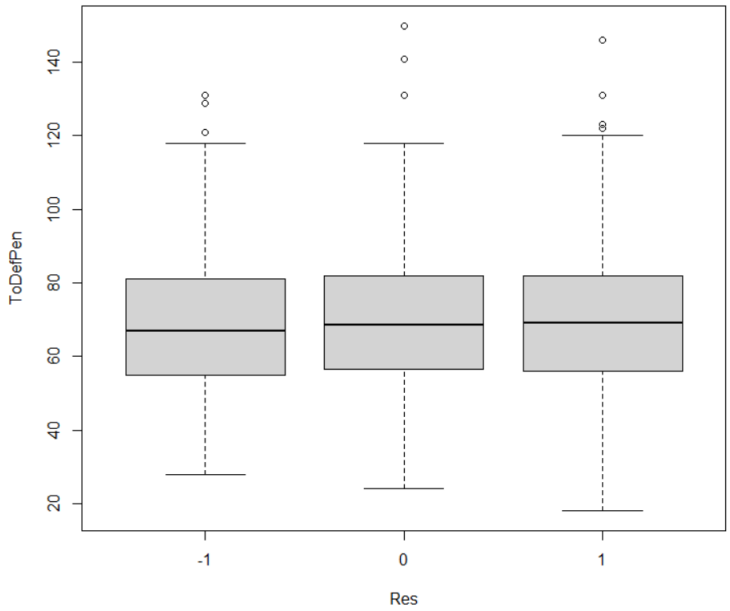


Figura 2.19: Boxplot della variabile risposta e della variabile numerica `ToDefPen`

La Figura 2.19 mostra come si comporta la relazione con `ToAttPen`. Contrariamente quanto visto con la Figura 2.19 qui si nota una certa variazione da una un box e l'altro, infatti vi è una tendenza positiva che porta ad aver valori più alti in caso di vittoria. Si ha una maggior varianza per quanto riguarda la vittoria rispetto ai altri due esiti e la distribuzione di tutti e tre è abbastanza bilanciata se non che la coda più bassa è leggermente meno lunga rispetto all'altra coda; la mediana invece è equilibrata. Si nota inoltre che vi sono alcuni outliners segno che alcune squadre in qualche partite, si sono particolarmente rese note nel produrre un quantitativo di tocchi maggiore rispetto alla distribuzione, ciò pero non sembra influenzare l'esito.

Per quanto riguarda `ToDef3rd`, `ToMid3rd` e `ToAtt3rd`, esse si comportano come `ToAttPen`. Perciò è stato omesso il loro grafico.

Nella Figura 2.21 vengono mostrati gli andamenti delle variabile dei falli, `Fls` e `Fld`. Nel boxplot a sinistra si può notare che i valori più alti sono nel box del pareggio mentre sono presenti valori più bassi nel box vittoria. Ciò fa pensare che subire molti falli può impedire la vittoria alla squadra che li subisce. Per quanto riguarda la distribuzione sembra essere buona; c'è minor varianza per quanto riguarda la sconfitta.

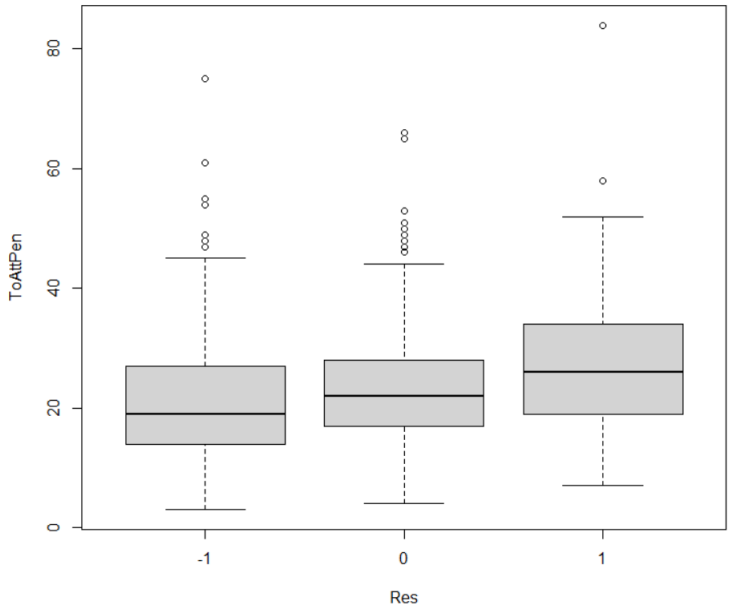


Figura 2.20: Boxplot della variabile risposta e della variabile numerica `ToAttPen`

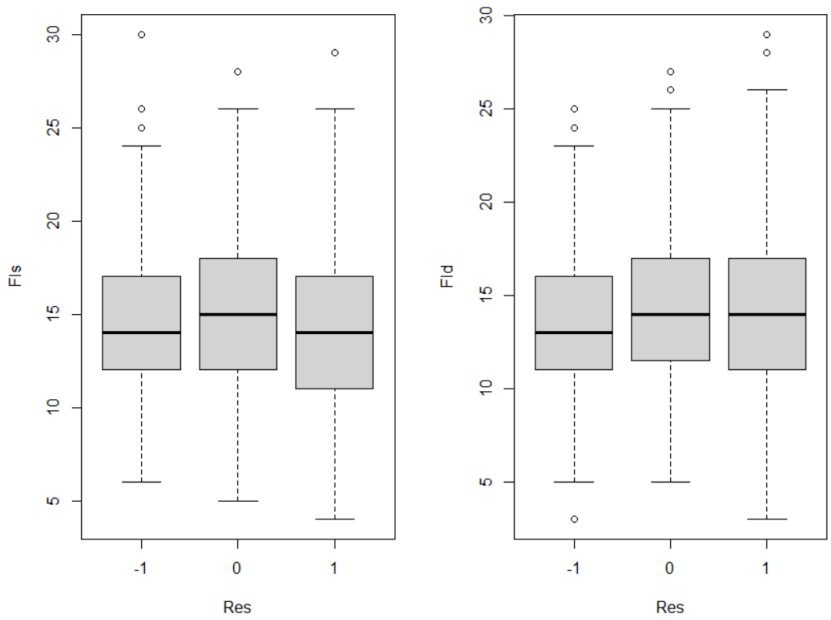


Figura 2.21: A sinistra il boxplot della variabile risposta e della variabile numerica `Fls` e a destra il boxplot della variabile risposta e della variabile numerica `Fld`

Nel secondo boxplot si può notare che i valori più alti sono presenti sia sul pareggio e sia sulla vittoria e sempre qui si ha una maggior distribuzione rispetto alla sconfitta. Sembra perciò che dal grafico si può intuire che se la squadra non commette dei falli allora sarà più soggetta a perdere.

La Figura 2.22 mostra come si comporta la relazione con `Int`. Sorprendentemente valori più alti sono registrati nella sconfitta, anche se la mediana risulta essere più vicina al 1 ° quantile sottolineando che vi è un maggior numero di valori bassi piuttosto che alti. La mediana dei restanti esiti invece e ben equilibrata ma il pareggio risulta avere meno variabilità. Sembra perciò che effettuare troppi intercettazioni dei passaggi avversari contrariamente da quanto si pensi sembra essere controproducente per la vittoria. Si segnala inoltre la presenza di alcuni outliners con valori alti di intercettazioni, che si discostano dalle distribuzioni.

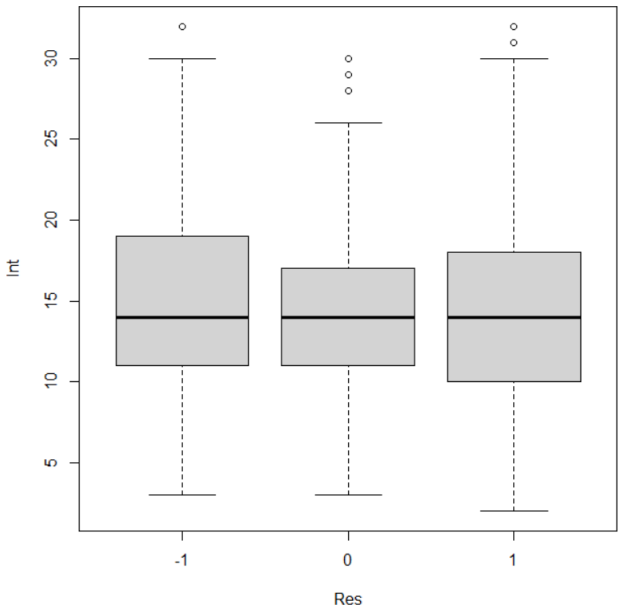


Figura 2.22: Boxplot della variabile risposta e della variabile numerica `Int`

La Figura 2.23 mostra come si comporta la relazione con `TklWin`. Come si può notare, vincere più contrasti possibili evita di subire una sconfitta. Infatti vi sono valori più alti in pareggio e vittoria oltre a una maggiore varianza rispetto alla sconfitta. Nello specifico pero si nota che, nella distribuzione dei valori vi sono maggior valori alti nella vittoria rispetto al pareggio, graficamente lo si vede dalla mediana che nel pareggio è più vicina al 1° quindi a valori più bassi e lo si nota anche dalla coda più bassa che è meno lunga rispetto a quella in alto; invece la mediana della vittoria risulta più vicina al 3° oltre ad avere la coda in alto più corta rispetto a quella in basso. Vi è inoltre qualche outliners con valori più alti di contrasti vinti ma sembrano non influenzare la classificazione.

Infine la Figura 2.24 mostra come si comporta la relazione con `Recov`. Per entrambe le classi la distribuzione sembra più sbilanciata verso valori bassi quindi ad una loro

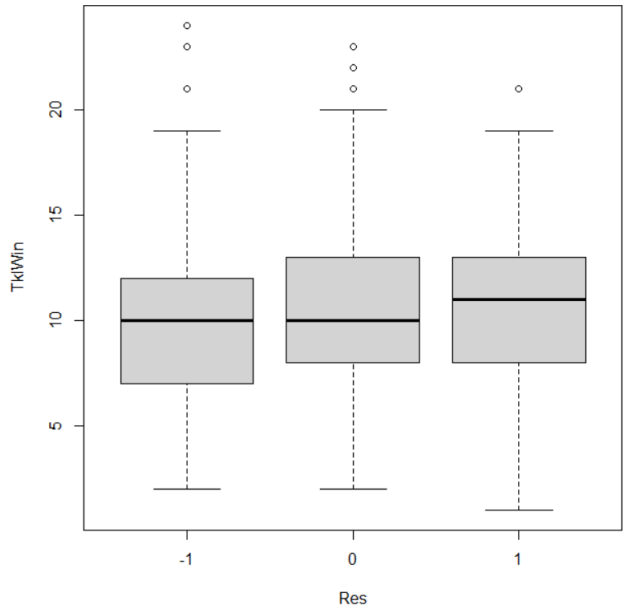


Figura 2.23: Boxplot della variabile risposta e della variabile numerica `TklWin`

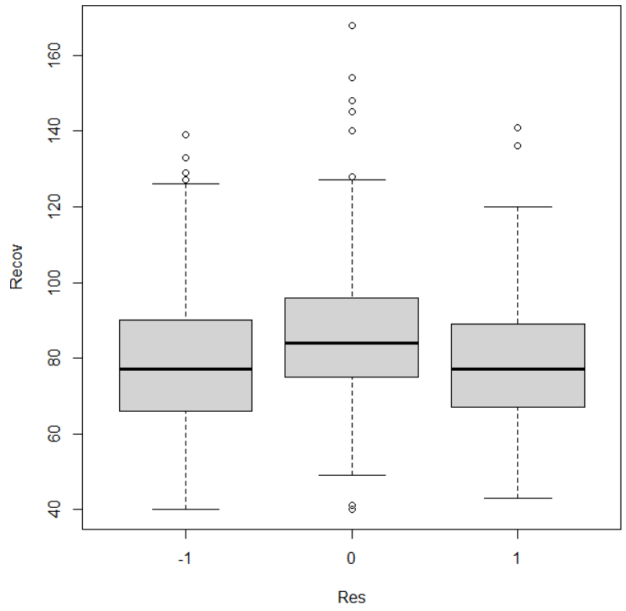


Figura 2.24: Boxplot della variabile risposta e della variabile numerica `Recov`

maggior presenza, infatti entrambe le code più in basso sono più corte rispetto a quelle più in alto che sono più lunghe. Per quanto riguarda la mediana sembra per entrambe le classi equidistante dai quantili. Si nota che il pareggio presenta minor varianza rispetto alle altre due classi ma valori più alti soprattutto nei confronti della vittoria oltre ad averne anche di più rispetto alle altre classi. Sembra perciò che un eccessivo numero di recuperi non porti alla vittoria. Si nota inoltre che vi sono numerosi outliners sopratutto per il pareggio.

3 | Modeling Paired Comparisons

Nel seguente capitolo verranno introdotti differenti modelli per la paired comparisons, iniziando con il Bradley-Terry model versione standard fino a presentare tutte le sue estensioni usate per l'analisi trattata. TO DO

3.1 Il Bradley-Terry Model

Il Bradley-Terry model (**bradley1952rank**) asserisce che in una competizione tra due qualsiasi giocatori, detti player i e player j (i,j $\in \{1,...,n\}$), la probabilità che i sia preferito a j è data dal rapporto tra α_i e α_j, dove α_i e α_j sono parametri che rappresentano la cosiddetta abilità dei due giocatori. Il modello standard non considera covariate e in generale, non presta nessuna attenzione all'eterogeneità causata dai soggetti dei confronti.

Formalmente, sia $Y_{i,j}$ la variabile casuale associata al risultato della *paired comparison* tra oggetti i e j, con $j > i \in \{1,...,n\}$, dove nella forma più semplice, il modello dato è il seguente:

$$P(i \succ j) = P(Y_{i,j} = 1) = \frac{exp(\alpha_i - \alpha_j)}{1 + exp(\alpha_i - \alpha_j)} \tag{3.1}$$

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

$$logit(i \succ j) = log(\frac{P(i \succ j)}{P(j \succ i))}) = log(\frac{exp(\alpha_i)}{exp(\alpha_j)}) = \alpha_i - \alpha_j \tag{3.2}$$

La risposta del modello rappresenta la probabilità che un certo oggetto i è preferito rispetto su un altro oggetto j, $i \succ j$. La variabile $Y_{i,j}$ essendo binaria può assumere solo due valori, $Y_{i,j} = 1$ se l'oggetto i è preferito sull'oggetto j e $Y_{i,j} = 0$ viceversa. I parametri α_n come scritto precedentemente rappresentano l'attrattiva o la forza del loro corrispondente oggetto. Chiaramente questi parametri di abilità devono essere stimati dal modello attraverso la massima verosimiglianza. Infine si noti che vi è necessario un vincolo per identificare i parametri, ad esempio: il vincolo di somma $\sum_{i=1}^{n} \alpha_i = 0$ oppure il vincolo dell'oggetto di riferimento, $\alpha_i = 0$ per un oggetto $i \in \{1, ..., n\}$. Se il vincolo dell'oggetto di riferimento è usato, allora il valore dei parametri abilità degli altri oggetti j sarà la differenza rispetto all'oggetto di riferimento i.

Si sottolinea inoltre che il modello precedentemente descritto è chiamato modello non strutturato e l'obbiettivo dell'analisi è di fare inferenza sul valore dei parametri abilità α_n per stilare una classifica finale di tutti gli oggetti.

3.2 Il Bradley-Terry Model con ordered response categories

In molti contesti di comparazione tra oggetti, è possibile che sia richiesto di dare una scala di preferenza tra un oggetto e un altro. Supponiamo che due oggetti i e j siano confrontati e che la preferenza ora non sia più espressa i termini di: preferisco i al posto di j o viceversa ma, attraverso una scala di preferenza ad esempio, dando una forte preferenza a i rispetto a j o una leggera preferenza a i rispetto a j o non dando nessuna preferenza o preferendo leggermente j rispetto a i oppure preferire fortemente j rispetto a i. Dal modello descritto nella precedente sezione si passa da due classi di preferenza a cinque classi di preferenza.
Ovviamente il caso descritto 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. Si necessità perciò di un estensione del classico Bradley-Terry model descritto precedentemente.

Modelli che consentono un numero generale di categorie K, sono stati proposti da (**tutz1986bradley**) e da (**bradley1952rank**), in particolare quest'ultimo mostrò come due modelli per l'analisi di dati ordinati possono essere adattati per le *ordinal paired comparisons*.

Il primo modello è il *cumulative link model* che sfrutta la rappresentazione della variabile casuale latente. In generale, sia H il numero di gradi della scala di preferenza e sia $Z_{i,j}$ una variabile continua casuale latente e siano $\theta_1 < \theta_2 < .... < \theta_{H-1}$ le soglie tale che $Y_{i,j} = h$ quando $\theta_{h-1} < Z_{i,j} < \theta_h$. Allora:

$$P(Y_{i,j} \leq h) = \frac{exp(\theta_h + \alpha_i - \alpha_j)}{1 + exp(\theta_h + \alpha_i - \alpha_j)} \tag{3.3}$$

con h $\in$ {1,....,H} che indica le possibili *response categories*. I parametri θ_h rappresentano le cosiddette soglie per le singole *response categories*, 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: $P(Y - i, j = h) = P(Y_{i,j} = H - h + 1)$. È quindi necessario che le soglie siano ristrette a θ_i = -θ_{H-h} e se, H è dispari, $\theta_{H/2} = 0$; per garantire che le probabilità siano simmetriche. Per garantire che le probabilità siano non negative per le singole *response categories* vi è imposta la seguente limitazione: $-\infty = \theta_0 < \theta_1 < ... < \theta_{H-1} < \theta_H = \infty$. Dato che la soglia per l'ultima categoria è fissata a $\theta_H = \infty$ allora vale che $P(Y_{i,j} \leq H) = 1$. Si sottolinea che le soglie sono parametri che vanno stimate dai dati; inoltre la probabilità di una singola *response category* può essere derivata dalla differenza tra categorie adiacenti cioè:

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

Il modello delle *adjacent categories model*, così come il modello Bradley-Terry, ha anche una rappresentazione logit lineare ed è il seguente:

$$logit(Y_{i,j} \leq h) = \theta_h + \alpha_i - \alpha_j \tag{3.4}$$

Il secondo modello invece proposto da (**agresti1992analysis**) è il *adjacent categories model*. 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 e al modello proposto da (**davidson1970extending**) quando sono consentite solo tre categorie.
Il modello proposto da (**davidson1970extending**) risulta essere adatto per l'analisi sulle partite di calcio.
Il *adjacent categories model* è più semplice da interpretare rispetto ai *cumulative link models* poiché l'odds ratio si riferisce a un determinato risultato anziché a raggruppamenti di risultati.
Perciò dal modello proposto da (**davidson1970extending**), sia θ il parametro stimato dai dati che indica quanto è auspicabile la non preferenza, nel nostro caso il pareggio, allora:

$$P(Y_{i,j} = 2|Y_{i,j} \neq 0) = \frac{exp(\alpha_i - \alpha_j)}{1 + exp(\alpha_i - \alpha_j)}, \tag{3.5}$$

$$P(Y_{i,j} = 1) = \frac{\theta\sqrt{exp(\alpha_i) * exp(\alpha_j)}}{exp(\alpha_i) + exp(\alpha_j) + \theta\sqrt{exp(\alpha_i) * exp(\alpha_j)}}, \tag{3.6}$$

$$P(Y_{i,j} = 0|Y_{i,j} \neq 1) = \frac{exp(\alpha_j - \alpha_i)}{1 + exp(\alpha_j - \alpha_i)} \tag{3.7}$$

Come si può vedere si è riportato la modellazione di tutti e tre i possibili risultati, con α_n che rappresenta la forza degli oggetti in comparazione da stimare dai dati. La modellazione vittoria e la sconfitta dell'oggetto i contro l'oggetto j rimane uguale alla modellazione (3.1) descritta precedentemente. Diversamente per il pareggio dove viene aggiunto il parametro θ.

3.3 Il Bradley–Terry Model con variabili esplicative

Fin ad ora è stato presentato un modello che valutasse il grado di preferenza per un oggetto i rispetto a un oggetto j, senza che considerasse nessuna variabile. Chiaramente tale modello risulta essere inutile per le nostre analisi, dato che siamo interessati a capire quali variabili possono influenzare il risultato della comparazione. Si necessita perciò di un modello che tenga conto anche di variabili esplicative inserite durante l'analisi.
Sia $x_i=(x_{i1}, ...x_{iK})$ il vettore di K variabili esplicative per un certo oggetto i e $\beta = (\beta_1, ...\beta_P)$ il vettore dei pesi stimati per ogni variabile presente in x_i, allora si ha che il parametro abilità α_i di un certo oggetto i è uguale a:

$$\alpha_i = \beta_1 x_{i1} + ... + \beta_P x_{iP} \text{ con i=1,....,n}$$

Si ha quindi che il parametro abilità α_i per un certo oggetto i è una combinazione lineare di variabili.
Il modello è stato presentato per la prima volta da (**springall1973response**); tale modello viene chiamato modello strutturato.

Grazie a questo modello se vi sono covariate che hanno un legame con la variabile risposta, tanto da influenzarne l'esito con quest'ultima allora, sarà possibile inserirle nel modello. Nel caso calcistico tali covariate possono essere il possesso della palla o il numero di falli fatti.

3.3.1 Il Bradley–Terry Model con effetto partite in casa

Nel modello descritto nella sezione 2.2, si era scritto che, era 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 conta. Infatti nel 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:

$$P(i \succ j) = P(Y_{i,j} = 1) = \frac{exp(\delta + \alpha_i - \alpha_j)}{1 + exp(\delta + \alpha_i - \alpha_j)} \tag{3.8}$$

Il qual'è il modello (3.1) riadatto e da cui possiamo derivare il modello (3.3) riadatto che è il seguente:

$$P(Y_{i,j} \leq h) = \frac{exp(\delta + \theta_h + \alpha_i - \alpha_j)}{1 + exp(\delta + \theta_h + \alpha_i - \alpha_j)} \tag{3.9}$$

Come si può vedere il vantaggio di giocare in casa, in generale l'effetto d'ordine; viene trattato come una variabile esplicativa. Infatti se $\delta > 0$ allora viene attribuito un vantaggio all'oggetto i, nel contesto calcistico significa che gioca in casa; aumentando la probabilità che vinca il confronto o nel caso di *ordered response categories*, di avere un risultato superiore rispetto all'oggetto j. Chiaramente il peso di δ deve essere stimato dai dati.

Il modello (3.8) cosi come il modello (3.9) , hanno anche una rappresentazione logit lineare e sono le seguenti:

Per (3.8)

$$logit(i \succ j) = \delta + \alpha_i - \alpha_j \tag{3.10}$$

Per (3.9)

$$logit(Y_{i,j} \leq h) = \delta + \theta_h + \alpha_i - \alpha_j \tag{3.11}$$

4 | 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