UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA

Foundations of Probability and Statistics

COLLEGE BASKETBALL: NCAA MARCH MADNESS TOURNAMENT

Analisi svolta da:

Claudio Maffi N° matricola: 875789

Christian Internò N° matricola: 876238

Anno Accademico 2020/2021

INDICE

1. ABSTRACT

Nel presente articolo si sono svolte un'analisi esplorativa ed inferenziale sui microdati della stagione 2018-2019 del NCAA March Madness tournament.
Il Dataset Basketball College riassume le 353 squadre partecipanti al campionato nell'anno 2018-2019 e dispone di 8 variabili numeriche e 3 variabili qualitative.
Il Dataset è stato ottenuto dal sito ufficiale dell'associazione NCAA championships division, ossia l'associazione che gestisce e regola il campionato March Madness tra le squadre dei diversi college: https://www.ncaa.com/march-madness
L'obiettivo dello studio è stato approfondire le relazioni tra le principali variabili offensive e difensive del gioco del Basket, ricercando i pesi delle singole variabili e i possibili fattori esterni condizionanti.
Si è cominciato osservando le distribuzioni e la correlazione tra le variabili ADJOE (una stima dell'efficienza offensiva, punti segnati per 100 possessi) e ADJDE (una stima dell'efficienza difensiva, punti consentiti per 100 possessi).
Dopodiché si è cercato di adattare un modello di regressione lineare multiplo ponendo come risposta aleatoria il win-rate (rapporto tra vittorie e partite giocate).
Infine si è effettuata un'ANOVA test per ricercare significative differenze tra i gruppi riguardanti le variabili CONFERENCE (confederazione di appartenenza) e FTR (tiri liberi ottenuti).

2. DESCRIZIONE DELLE VARIABILI

TEAM: Nome del college team basketball school.

CONF: La CONFERENCE atletica nella quale la scuola partecipa.

(A10 = Atlantic 10, ACC = Atlantic Coast Conference, AE = America East, Amer = American, ASun = ASUN, B10 = Big Ten, B12 = Big 12, BE = Big East, BSky = Big Sky, BSth = Big South, BW = Big West, CAA = Colonial Athletic Association, CUSA = Conference USA, Horz = Horizon League, Ivy = Ivy League, MAAC = Metro Atlantic Athletic Conference, MAC = Mid-American Conference, MEAC = Mid-Eastern Athletic Conference, MVC = Missouri Valley Conference, MWC = Mountain West, NEC = Northeast Conference, OVC = Ohio Valley Conference, P12 = Pac-12, Pat = Patriot League, SB = Sun Belt, SC = Southern Conference, SEC = South Eastern Conference, Slnd = Southland Conference, Sum = Summit League, SWAC = Southwestern Athletic Conference, WAC = Western Athletic Conference, WCC = West Coast Conference)

G: Numero di partite giocate.

W: Numero di partite vinte.

ADJOE: Efficienza offensiva (una stima dell'efficienza offensiva, punti segnati per 100 possessi) che una squadra avrebbe contro la difesa media della divisione.

ADJDE: Efficienza difensiva (una stima dell'efficienza difensiva, punti consentiti per 100 possessi) che una squadra avrebbe contro l'attacco medio della divisione.

EFG_O: Percentuale effettiva di tiro sul campo.

EFG_D: Percentuale effettiva di tiro concessa sul campo.

FTR : Tiri liberi ottenuti (la frequenza con cui la squadra specifica effettua tiri liberi).

FTRD: Tiri liberi concessi.

POSTSEASON: Turno in cui la squadra in questione è stata eliminata o in cui si è conclusa la stagione (R68 = First Four, R64 = Round di 64, R32 = Round di 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Campione = vincitore del torneo NCAA March Madness).

3.ANALISI STATISTICA

Frequenze percentuali conference

```
CONF
      A10       ACC        AE      Amer      ASun       B10       B12        BE      BSky      BSth
3.966006 4.249292 2.549575 3.399433 2.266289 3.966006 2.832861 2.832861 3.399433 3.399433
       BW       CAA      CUSA      Horz       Ivy      MAAC       MAC      MEAC       MVC       MWC
2.549575 2.832861 3.966006 2.832861 2.266289 3.116147 3.399433 3.399433 2.832861 3.116147
      NEC       OVC       P12       Pat        SB        SC       SEC      SInd       Sum      SWAC
2.832861 3.399433 3.399433 2.832861 3.399433 2.832861 3.966006 3.682720 2.266289 2.832861
      WAC       WCC
2.549575 2.832861
```

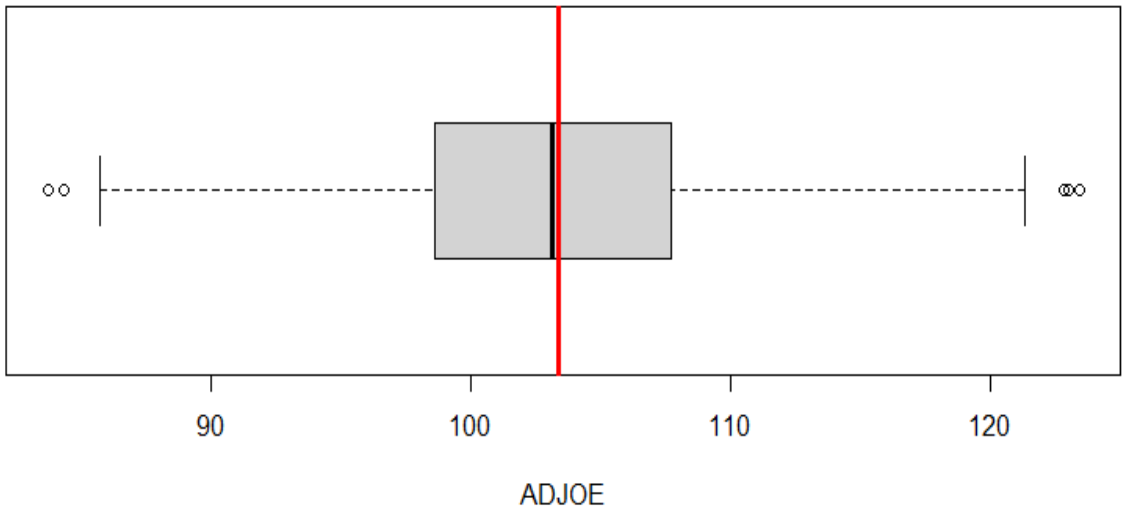
Analizzando le frequenze percentuali delle conference troviamo un'alta eterogeneità di appartenenza. La conference con il maggior numero di squadre presente è ACC (Atlantic Coast Conference) col 4.25%, mentre le conference con il minor numero di squadre in rapporto al totale sono Ivy e Sum col 2.27% ciascuna.

Media, mediana e boxplot ADJOE e ADJDE

Si è calcolato media e mediana dell'Efficienza offensiva modificata (una stima dell'efficienza offensiva, punti segnati per 100 possessi) che una squadra avrebbe contro la difesa media della divisione I (ADJOE):

Mean=103.336 Median=103.1 Min=83.7 Max=123.4

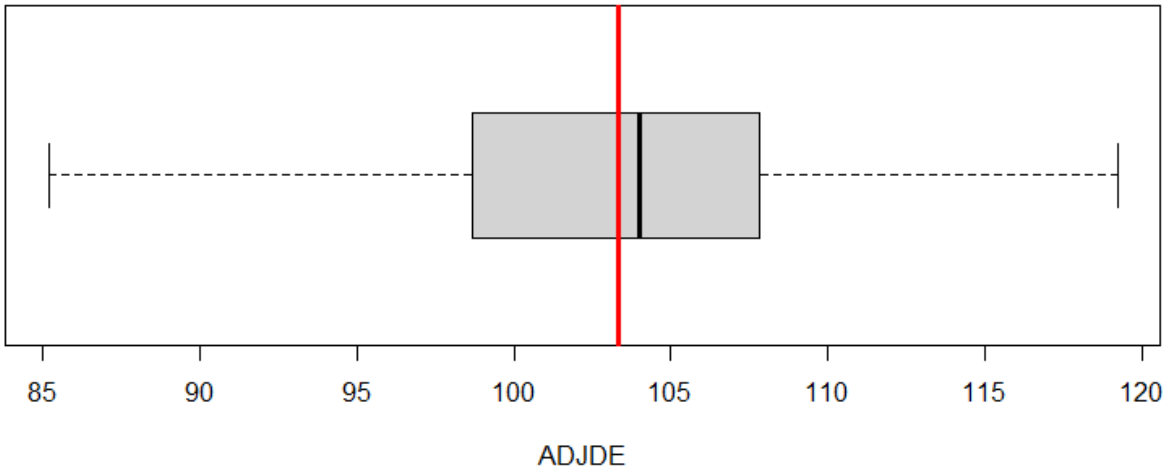
1st Qu.=98.6 3rd Qu.=107.7



Il boxplot della variabile ADJOE mostra la presenza di outliers sia a destra che a sinistra della distribuzione. Nonostante questo, la media (rappresentata dalla linea rossa) e la mediana sono quasi identiche.

Si è calcolato media e mediana dell'efficienza difensiva modificata (una stima dell'efficienza difensiva, punti concessi per 100 possessi) che una squadra avrebbe contro l'attacco medio della divisione I (ADJDE):

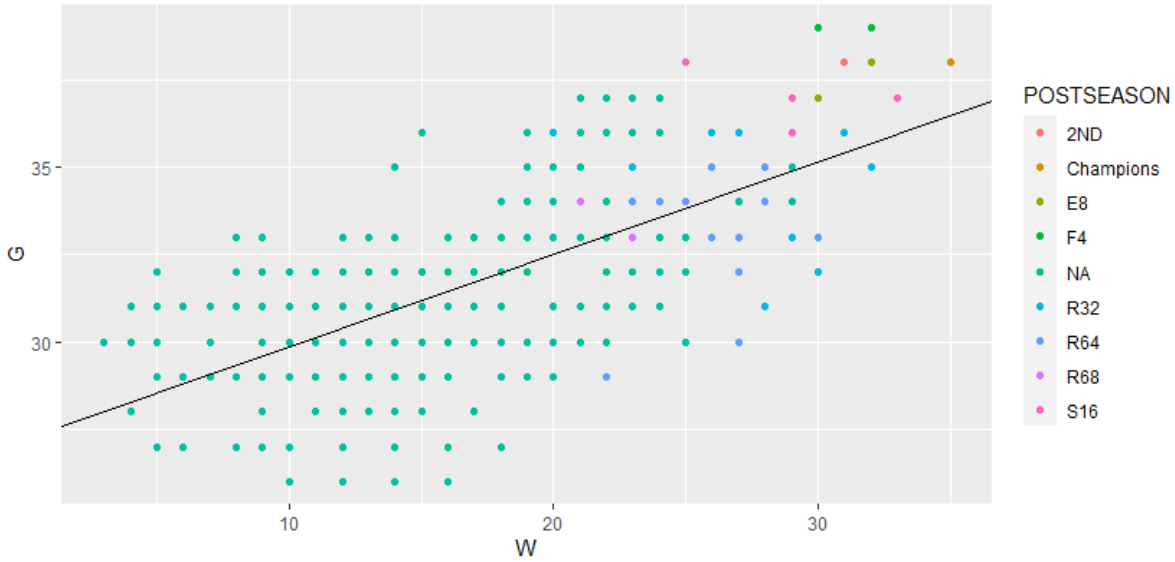
Mean=103.3363 Median=104 Min=85.2 Max=119.2
1st Qu.=98.7 3rd Qu.=107.8



Il boxplot della variabile ADJDE non mostra la presenza di outliers nella distribuzione. Nonostante questo, la media (rappresentata dalla linea rossa) e la mediana sono leggermente diverse, portando ad avere una distribuzione leggermente spostata a sinistra.

Rappresentazione win rate

Si è calcolato il win rate come rapporto tra vittorie e partite giocate (W/G). Il grafico sottostante rappresenta le vittorie e le partite giocate da ogni squadra, differenziate per turno eliminatorio raggiunto (POSTSEASON).

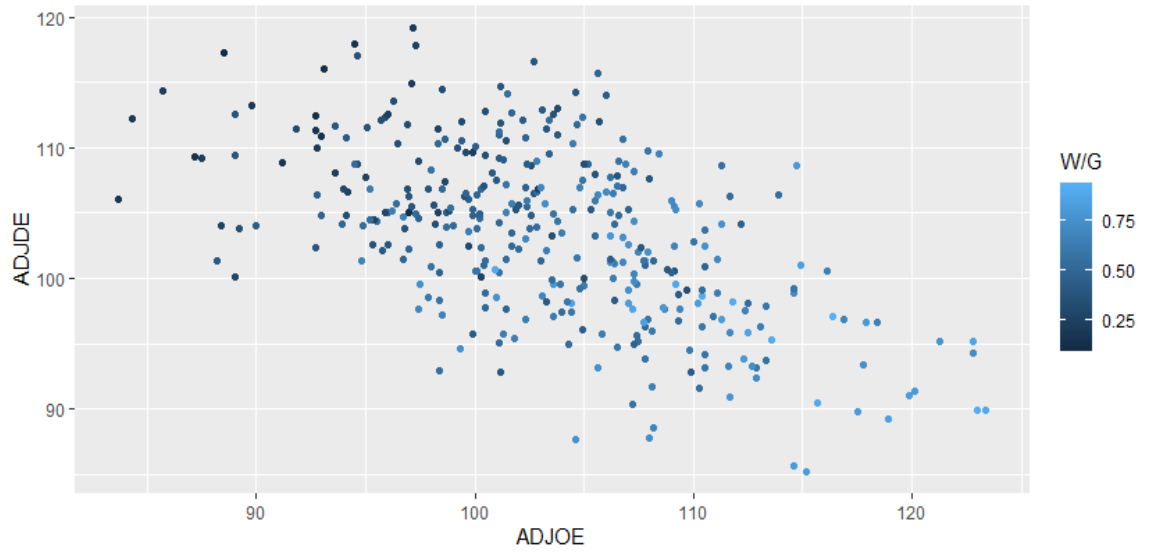


L'equazione della retta di regressione è pari a: Y=0.26376*X+27.23125

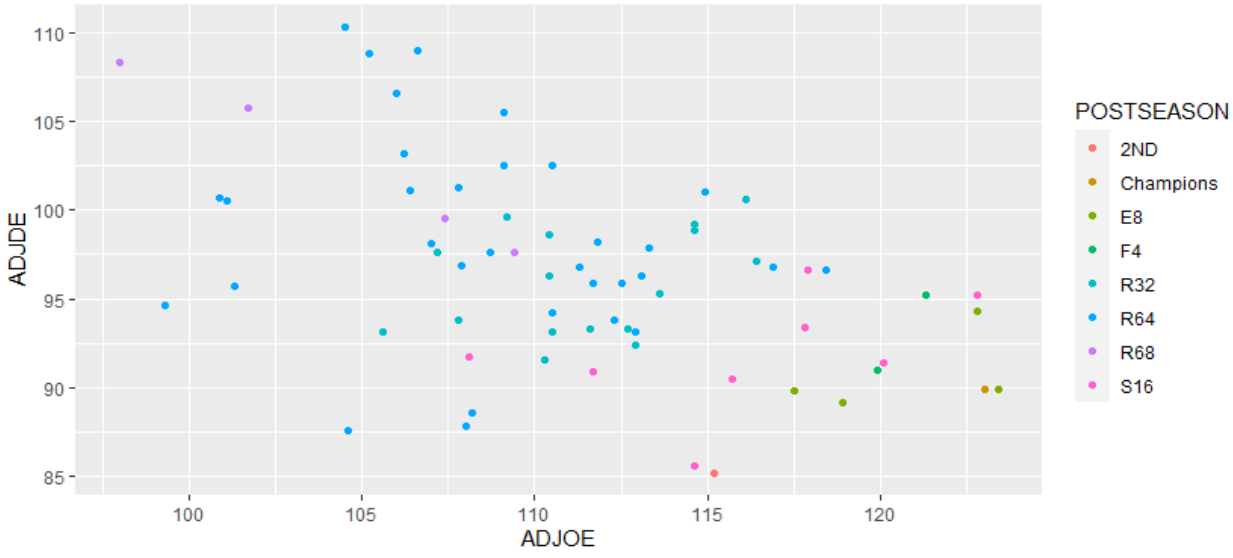
Il modello di regressione lineare tra la variabile indipendente G e la variabile dipendente W ha restituito un coefficiente di determinazione, o R quadrato, pari a 0.4473: questo rappresenta la bontà del modello, che riesce a spiegare il 44.73% dei dati campionari.

Correlazione tra ADJOE e ADJDE

Il grafico sottostante rappresenta il tasso di punti offensivo (ADJOE) e il tasso di punti difensivo (ADJDE) di ogni squadra, colorate in base al loro win rate (W/G).



Dalla rappresentazione delle variabili ADJOE e ADJDE in relazione al win rate si può notare che le squadre che hanno un win rate elevato sono quelle che hanno un elevato tasso offensivo e un basso tasso difensivo.
Si è calcolato il coefficiente di correlazione lineare di Pearson risultando pari a -0.5687.
Questo conferma la direzione inversa della relazione tra le variabili ADJOE e ADJDE.

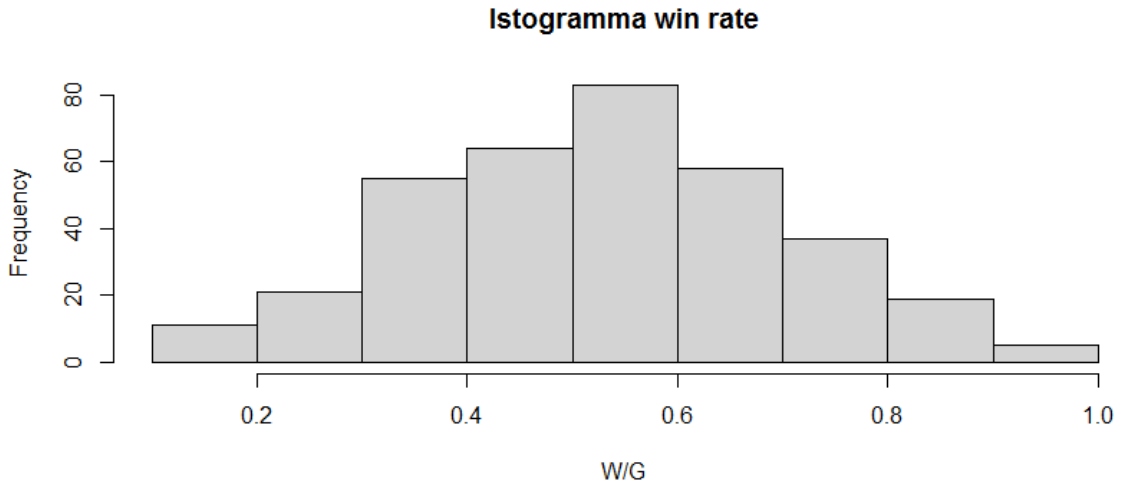


Il precedente grafico rappresenta le prestazioni offensive e difensive delle squadre che hanno superato la fase di qualificazione e sono divise in base al turno eliminatorio raggiunto. L'indice offensivo ADJOE risulta più determinante nella fase eliminatoria del torneo rispetto all'indice difensivo ADJDE. Come si può osservare a parità di indice difensivo le squadre con un indice offensivo più alto hanno raggiunto le ultime fasi eliminatorie del torneo.

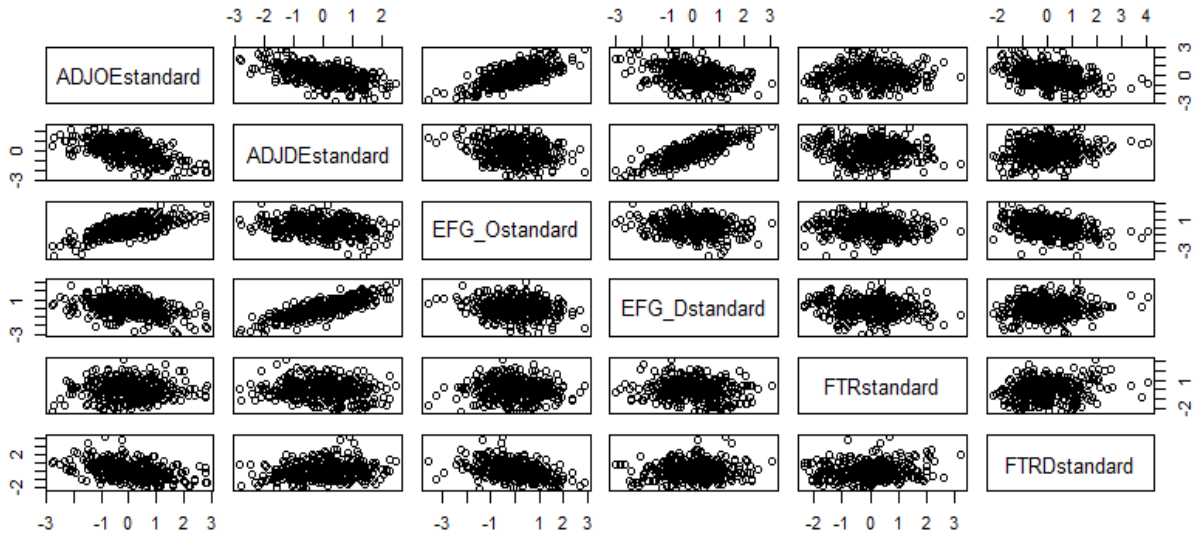
Modello di regressione multipla

Si è ipotizzato di porre il win rate come riposta aleatoria a un modello di regressione multipla utilizzando tutte le variabili offensive e difensive come predittori.
Si è svolta un'analisi di normalità attraverso Shapiro test:

Winrate p-value: 0.2156



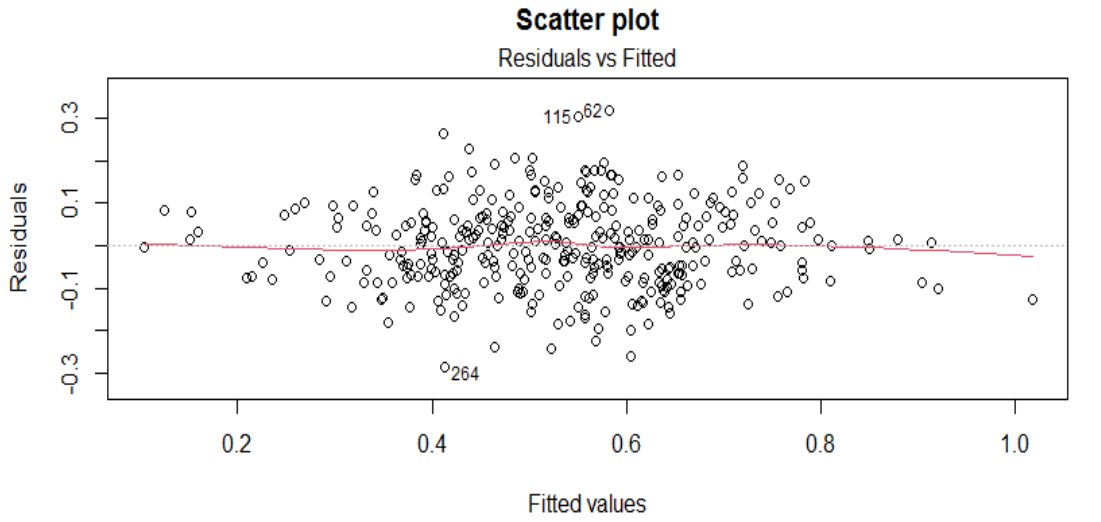
Dopodichè si sono analizzate le distribuzioni delle variabili offensive e difensive.
Un'osservazione preliminare è che le variabili potrebbero essere eccessivamente correlate.

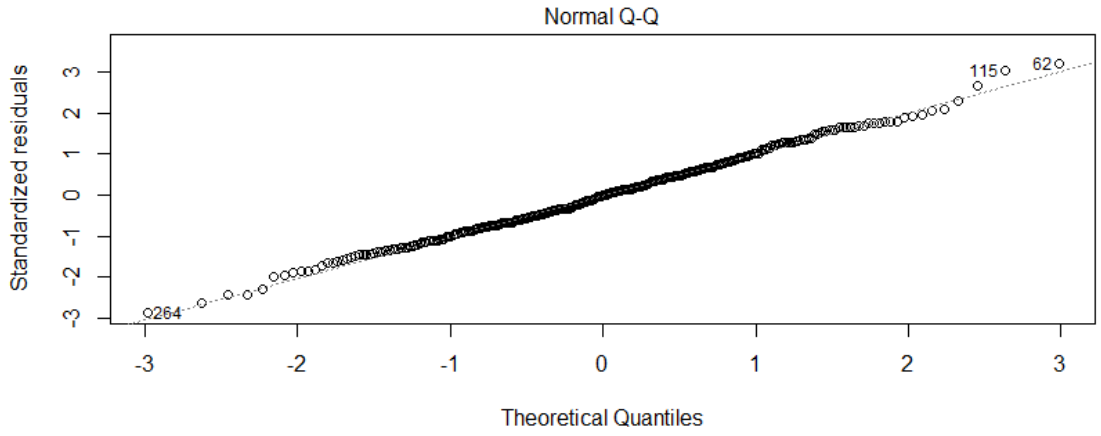


```
Residuals:
     Min       1Q   Median       3Q      Max
-0.28472 -0.06945 -0.00105  0.06577  0.31744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.068020   0.202104   0.337   0.7367
ADJOE        0.007026   0.001362   5.157 4.23e-07 ***
ADJDE       -0.001956   0.001723  -1.136   0.2568
EFG_O        0.020335   0.002747   7.402 1.03e-12 ***
EFG_D       -0.024066   0.003519  -6.839 3.63e-11 ***
FTR          0.001972   0.001175   1.677   0.0944 .
FTRD         0.002053   0.001190   1.725   0.0854 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1001 on 346 degrees of freedom
Multiple R-squared:  0.6743,    Adjusted R-squared:  0.6687
F-statistic: 119.4 on 6 and 346 DF,  p-value: < 2.2e-16
```

Il modello ha un buon livello di adattamento e un coefficiente di determinazione pari a 0.6473, ossia spiega il 64.73% dei dati campionari.

Per testare la bontà del modello si è proceduto all’analisi dei residui, ottenendo i seguenti risultati:





L’ipotesi di normalità sui residui risulta rispettata e gli indicatori di bontà complessiva sono positivi. Tuttavia, si è approfondita l’analisi di correlazione tra predittori attraverso il VIF.

```
   ADJOE    ADJDE    EFG_O    EFG_D      FTR     FTRD
3.212628 4.338819 2.289095 3.296231 1.075888 1.283497
```

Si può notare che alcune variabili si trovano in presenza di leggera multicollinearità.

Si è proseguito rimuovendo alcuni predittori che generavano eccessiva correlazione. I predittori rimossi sono ADJOE e ADJDE, perché essendo indici generali sull’efficienza risultavano troppo correlati agli indicatori più specifici.

I risultati del VIF sul nuovo modello sono:

```
    EFG_O    EFG_D      FTR     FTRD
1.242347 1.069474 1.057118 1.247995
```

I VIF risultano bassi e c’è scarsa probabilità di multicollinearità.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.29946 -0.07687  0.00297  0.06192  0.31699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5016851  0.1856393   2.702  0.00722 **
EFG_O        0.0307569  0.0021485  14.315  < 2e-16 ***
EFG_D       -0.0323374  0.0021279 -15.197  < 2e-16 ***
FTR          0.0027907  0.0012369   2.256  0.02467 *
FTRD         0.0007245  0.0012456   0.582  0.56117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1063 on 348 degrees of freedom
Multiple R-squared:  0.6309,    Adjusted R-squared:  0.6266
F-statistic: 148.7 on 4 and 348 DF,  p-value: < 2.2e-16
```

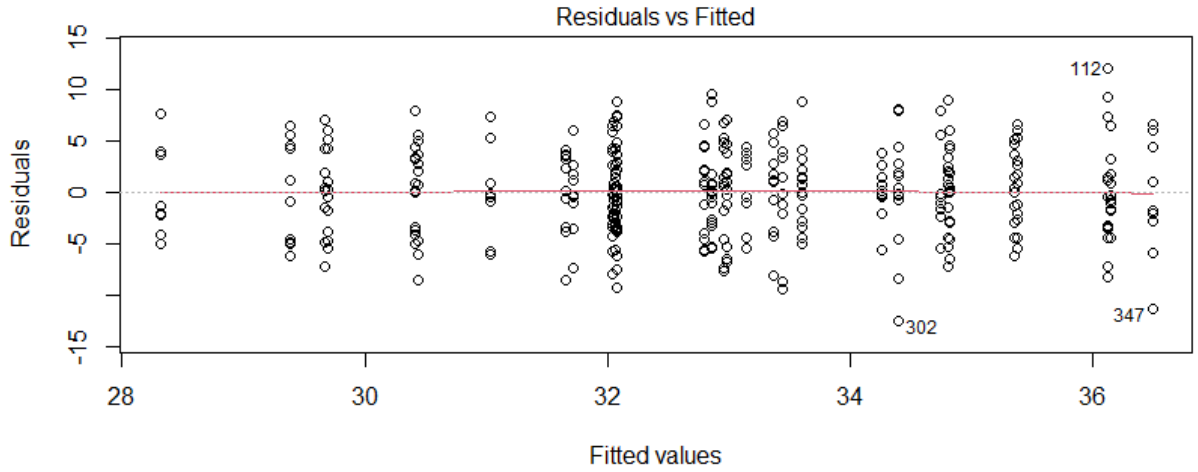
Il nuovo modello mantiene quasi inalterato il coefficiente di determinazione e quindi il livello di bontà di adattamento, e migliora i p-value associati all’ipotesi di nullità dei singoli predittori tranne per la variabile FTRD.

Analisi della varianza (ANOVA)

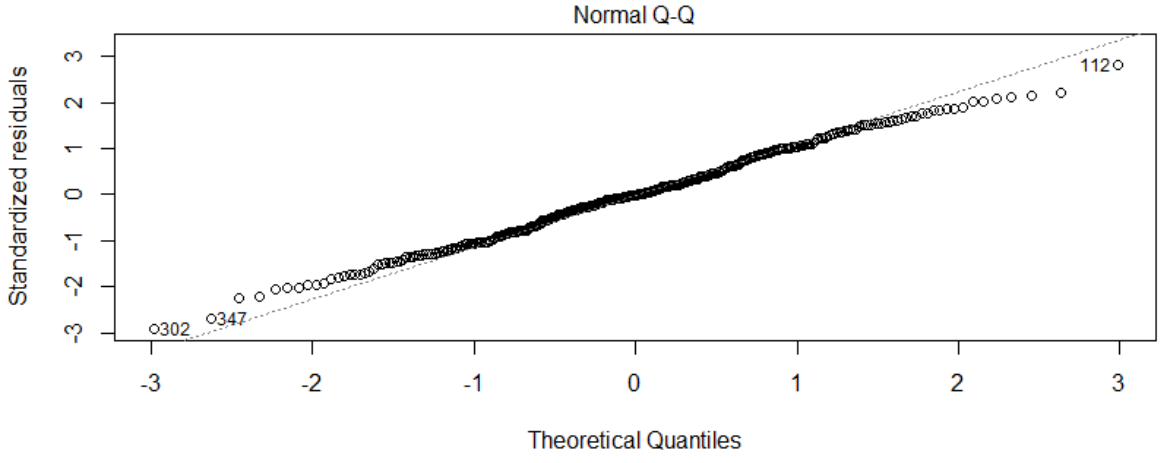
Vogliamo sapere se c'è qualche differenza significativa tra la conference di appartenenza (CONF) e i numeri medi di tiri liberi ottenuti (FTR):

```
             Df Sum Sq Mean Sq F value   Pr(>F)
CONF         31   1408   45.42   2.279 0.000205 ***
Residuals   321   6397   19.93
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Poiché il valore p-value è inferiore al livello di significatività di 0.05, possiamo concludere che ci sono differenze significative tra i gruppi evidenziati nel riepilogo del modello.
Il test ANOVA presuppone che i dati siano normalmente distribuiti e la varianza tra i gruppi sia omogenea. Nel grafico sottostante, non ci sono relazioni evidenti tra residui e valori stimati, il che è positivo. Quindi, possiamo assumere l'omogeneità delle varianze.

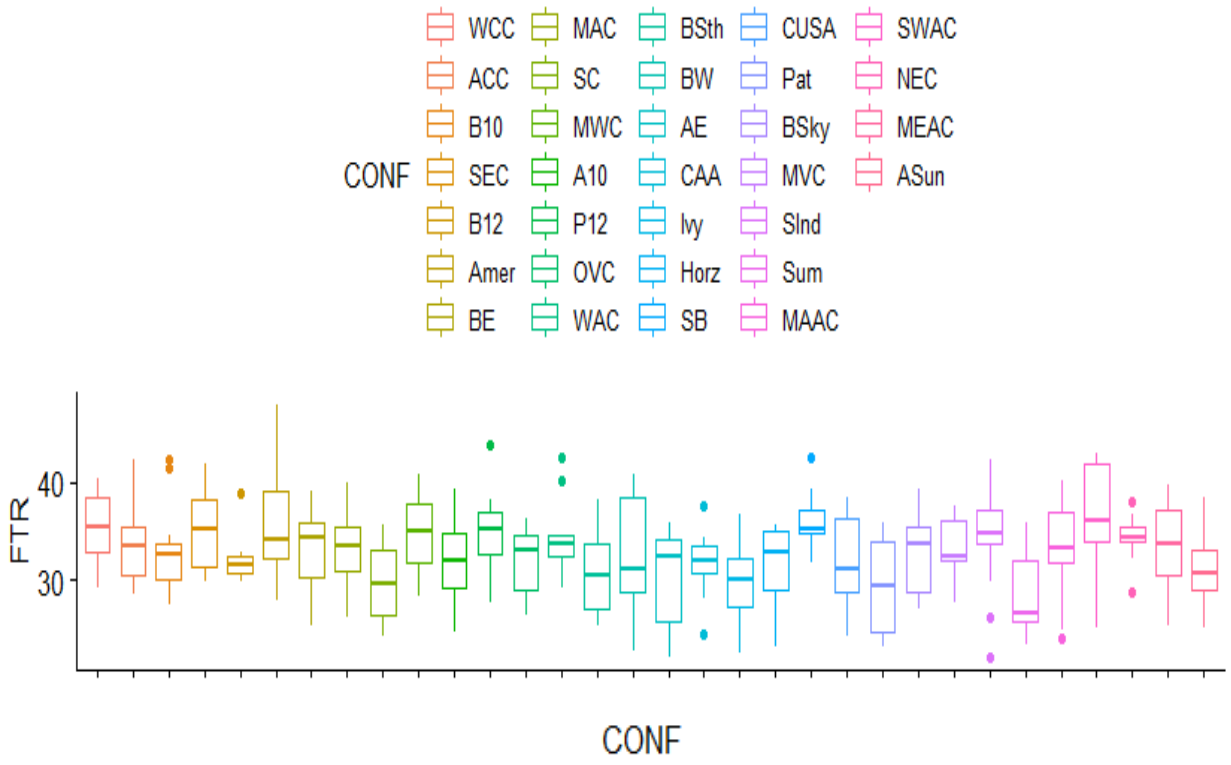


Nel grafico di normalità dei residui sottostante, i quantili dei residui vengono tracciati rispetto ai quantili della distribuzione normale. Il grafico di probabilità normale dei residui viene utilizzato per verificare l'ipotesi che i residui siano distribuiti normalmente.



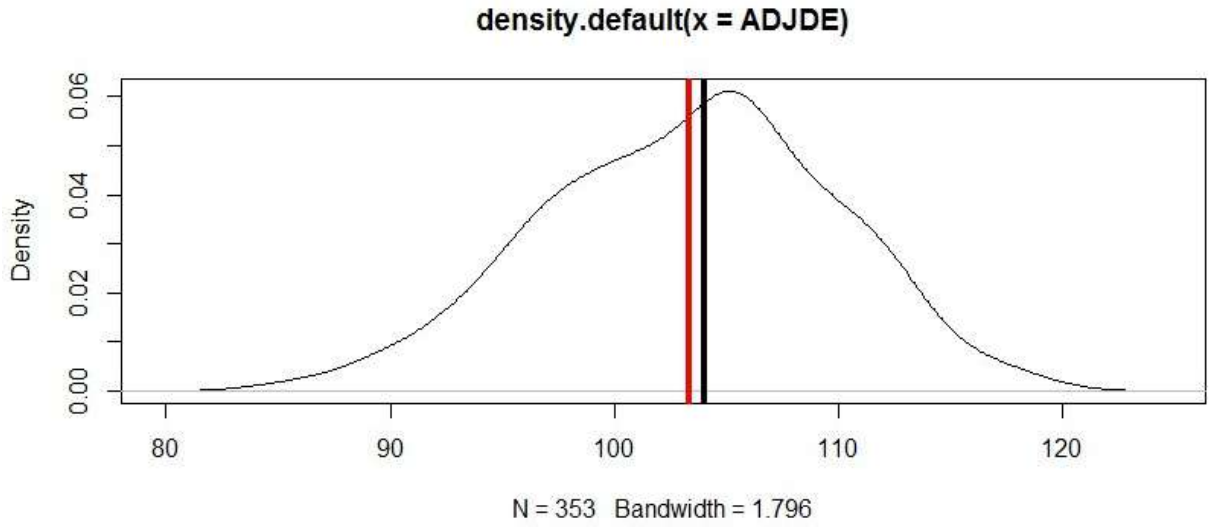
Poiché tutti i punti cadono approssimativamente lungo questa linea di riferimento, possiamo assumere la normalità. Si può assumere quindi la validità dell'ANOVA test.

L'analisi delle varianze conferma la significativa differenza tra i gruppi, come dimostrato dai boxplot delle conference.



4. CONCLUSIONI

Si osservata un'alta eterogeneità delle CONFERENCE di appartenenza. Si sono analizzate le distribuzioni di ADJOE e ADJDE, nel primo caso nonostante la presenza di vari outliers media e mediana coincidono, presupponendo la simmetria della distribuzione. Nel seondo caso media e mediana non coincidono, anche se non si è rilevata la presenza di outliers. Si conferma una asimmetria della distribuzione verso sinistra.



Partendo dall'analisi grafica del win-rate si è cercata la correlazione tra le variabili ADJOE e ADJDE. Si può concludere che vi è una relazione inversa tra le due variabili. Un'ulteriore osservazione risultante dall' analisi diagnostica dimostra che l'indice ADJOE è particolarmente determinante per l'arrivo alla vittoria del campionato.

Il modello di regressione multipla con il win-rate come risposta aleatoria ha restituito un coefficiente di determinazione pari a 0.6309. L'analisi dei VIF ci conferma la bassa probabilità di problemi di multicollinearità. I risultati ottenuti derivano dalla rimozione di due predittori che creavano eccessiva correlazione.

Nell'analisi della varianza è risultata una significativa differenza tra i gruppi.
Osservando i box-plot delle CONFERENCE si possono notare le differenze di tiri liberi ottenuti dalle squadre appartenenti alle diverse CONFERENCE.
Concludendo, si può ipotizzare che FTR sia condizionata da fattori esterni (quali fattori di campo come i diversi arbitraggi) portando significative differenze tra le CONFERENCE.