Corso di Laurea Magistrale in
Artificial Intelligence and Data Engineering

Prima Relazione del Corso di Statistica II

# Prevedere il numero di vittorie di una squadra NBA

*Edoardo Ruffoli*

21 Novembre 2021

## Indice

# 1 Introduzione

## 1.1 Scopo dell'analisi

L'obiettivo dell'analisi è quello di costruire un modello di regressione lineare in grado di prevedere il numero di vittorie che una squadra NBA riuscirà a totalizzare al termine di una stagione date le attuali statistiche per partita della squadra stessa.

Un modello simile trova applicazione, principalmente, nei seguenti contesti: i bookmakers e gli analisti sportivi possono farne uso per il calcolo delle quote sportive; può risultare estremamente utile anche agli allenatori delle squadre stesse per migliorare gli aspetti dello stile di gioco che influenzano maggiormente il numero di vittorie. Si può ipotizzare, quindi, che l'analisi sia stata commissionata da una società di scommesse sportive oppure dalla dirigenza di una franchigia NBA.

# 2 Dataset

L'analisi è stata svolta facendo uso delle tabelle *Teams General Traditional* e *Teams General Advanced* reperibili sul sito ufficiale NBA che contengono le statistiche stagionali di ogni squadra:

https://www.nba.com/stats/teams/traditional
https://www.nba.com/stats/teams/advanced

Dal momento che non vi è un link diretto per scaricare le sopracitate tabelle, è stato utilizzato uno script Python scaricabile al seguente link il cui funzionamento sarà analizzato nel dettaglio nell'appendice di questa relazione.

Per raggiungere una numerosità adeguata al metodo di analisi sono stati utilizzati i dati delle ultime 21 stagioni NBA: la scelta di utilizzare osservazioni riguardanti le medesime squadre, registrate in anni differenti, non comporta il rischio di avere una ripetizione degli stessi andamenti. Da una stagione all'altra sono solite variare, talvolta anche radicalmente, la composizione ma soprattutto le prestazioni delle singole squadre.

## 2.1 Contenuto del Dataset

La tabella contiene 42 colonne e 626 osservazioni. Ai fini dell'analisi sono state utilizzate le seguenti 12 colonne:

- **W**: numero di vittorie (intero positivo).
- **FGA**: numero di tiri tentati a partita (decimale positivo).
- **FG3A**: numero di tiri da tre punti tentati a partita (decimale positivo).
- **FTA**: numero di tiri liberi tentati a partita (decimale positivo).
- **REB**: numero di rimbalzi catturati a partita (decimale positivo).
- **AST**: numero di assist a partita (decimale positivo).
- **TOV**: numero di palle perse per partita (decimale positivo).
- **STL**: numero di palle recuperate per partita (decimale positivo).
- **BLK**: numero di stoppate a partita (decimale positivo).

- **PTS**: numero di punti segnati a partita (decimale positivo).
- **OFF_RATING**: numero di punti segnati da una squadra per 100 possessi di gioco (decimale positivo).
- **DEF_RATING**: numero di punti subiti da una squadra per 100 possessi di gioco (decimale positivo).

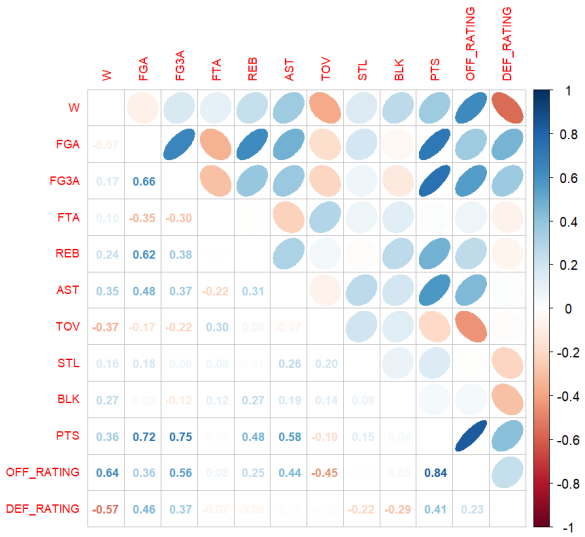
NB: l'uso di valori espressi nella forma "per partita" è una scelta obbligata per lo scopo di questa analisi. Se fossero stati utilizzati i valori totali, non sarebbe stato possibile avere una previsione delle vittorie a fine della stagione date le statistiche di gioco attuali di una squadra.

### 2.2 Data Preprocessing

Il numero di partite giocate non è costante durante le stagioni considerate: in particolare le stagioni 2011-12, 2019-20 e 2020-21 hanno un numero di partite inferiore alle regolari 82 per squadra. Un numero di partite non costante tra le varie osservazioni può compromettere l'attendibilità dell'analisi in quanto il numero di vittorie andrebbe a dipendere anche dal numero di partite giocate e non solo dalle statistiche di gioco. Piuttosto che andare a valutare la percentuale di vittorie o una "standardizzazione" su 82 partite, è stato preferito andare a eliminare le osservazioni in questione soprattutto perché relative a stagioni fortemente influenzate da fattori esterni (ad esempio la pandemia di Covid 19). Il dataset passa quindi da 626 a 534 osservazioni.

La selezione dei fattori di ingresso è stata effettuata in modo da evitare la presenza di fattori contenenti informazioni già presenti in altre colonne ma espresse in forma diversa. Inoltre non sono stati considerati tutti i fattori di ingresso espressi in valori percentuali o nominali.

### 2.3 Considerazioni preliminari




Come è possibile osservare dal grafico riassuntivo delle correlazioni, le vittorie risultano essere nettamente correlate ($|0.6|$) con i due indicatori avanzati OFF_RATING e DEF_RATING; correlato positivamente con l'OFF_RATING e negativamente con il DEF_RATING. Ciò è dovuto alla natura opposta delle due statistiche: vincerà un numero maggiore di partite la squadra che totalizzerà un grande numero di punti, OFF_RATING elevato, e ne subirà il meno possibile, DEF_RATING ridotto. Le palle perse, gli assist e i punti hanno invece delle correlazioni più lievi (circa $|0.35|$). Non vi sono ulteriori correlazioni significative.

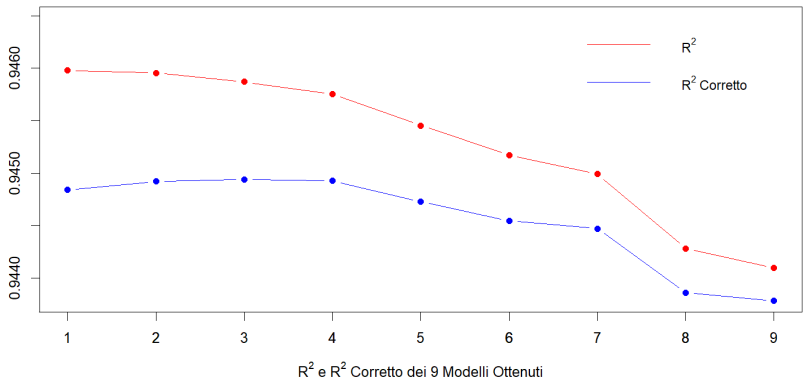
## 3 Analisi

Per effettuare l'analisi sono stati confrontati un modello di regressione lineare e un modello di regressione non lineare, logaritmico. Dato il consistente numero di fattori di ingresso, per entrambi i modelli, è stata operata una riduzione valutando di volta in volta, i valori di *varianza spiegata*, *varianza spiegata corretta* e i p-value dei fattori di ingresso.

## 3.1 Modello di Regressione Lineare

Il primo dei due modelli, è stato ottenuto tramite una riduzione in 9 passaggi totali. Il settimo modello è il primo a presentare dei p-value significativamente bassi e anche un buon valore di $R^2_{Adj}$ confrontato con gli altri modelli. Rappresenta un buon compromesso tra proporzione di varianza spiegata e numero di attributi e, pertanto, è stato selezionato come modello di regressione lineare.

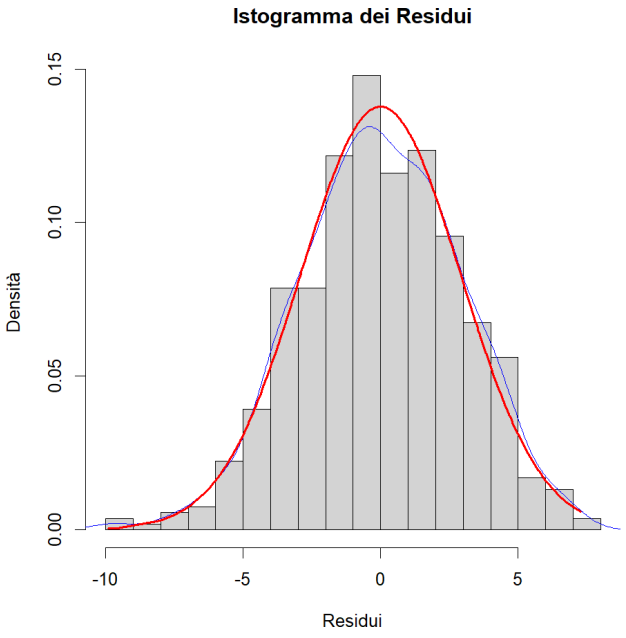
```
> summary(lm.7)

Call:
lm(formula = W ~ .-FG3A-BLK-PTS-FTA-STL-REB, data = data)

Residuals:
Min       1Q   Median       3Q      Max
-9.8940  -1.9632  -0.0257   2.0399   7.2707

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.87061    5.69700   8.578  < 2e-16 ***
FGA         -0.10798    0.04829  -2.236  0.02576 *
AST          0.15593    0.08206   1.900  0.05796 .
TOV         -0.32933    0.12602  -2.613  0.00922 **
OFF_RATING   2.53288    0.04284  59.119  < 2e-16 ***
DEF_RATING  -2.51066    0.04209 -59.647  < 2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 ' ' 1

Residual standard error: 2.91 on 528 degrees of freedom
Multiple R-squared:  0.945, Adjusted R-squared:  0.9445
F-statistic:  1814 on 5 and 528 DF,  p-value: < 2.2e-16
```

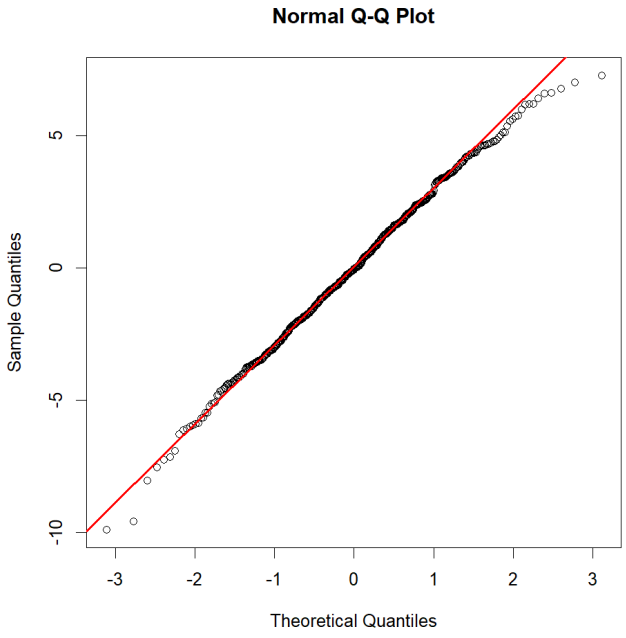



La regressione lineare multivariata basata sul modello risulta avere una proporzione di *varianza spiegata aggiustata* superiore al 94% e un p-value generale molto basso.

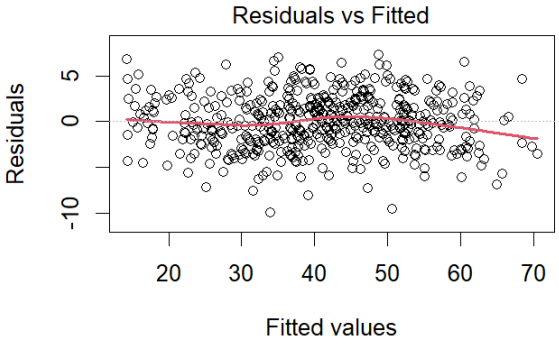
### 3.1.1 Analisi dei Residui

Osservando l'istogramma dei residui, essi appaiono distribuiti normalmente: la distribuzione della densità si avvicina molto a quella di una distribuzione normale, rappresentata nel grafico dalla curva rossa. Anche il Q-Q plot ci suggerisce la normalità dei residui, data l'aderenza della quasi totalità dei punti alla bisettrice a meno di qualche outlier agli estremi.



```
> skewness=mean(((lm.r-mean(lm.r))/sd(lm.r))^3)
> skewness
[1] -0.1370172
> kurtosi=mean(((lm.r-mean(lm.r))/sd(lm.r))^4)-3
> kurtosi
[1] -0.07409105
> shapiro.test(lm.r)

   Shapiro-Wilk normality test

data:  lm.r
W = 0.99671, p-value = 0.3498
```




Osservando il grafico di dispersione dei valori residui e dei valori stimati, si nota come i residui si disperdano in modo simile su tutto il grafico, senza seguire nessun pattern e quindi verosimilmente con varianza omogenea. Anche in questo caso possiamo notare la presenza di qualche piccolo outlier, posizionati prevalentemente nella parte centrale-bassa del grafico. La linea rossa, rappresentante la tendenza dei residui, è abbastanza sovrapponibile alla linea tratteggiata, corrispondente alla media zero, allora l'ipotesi di linearità è verificata, ovvero i residui si distribuiscono in modo casuale intorno allo 0.

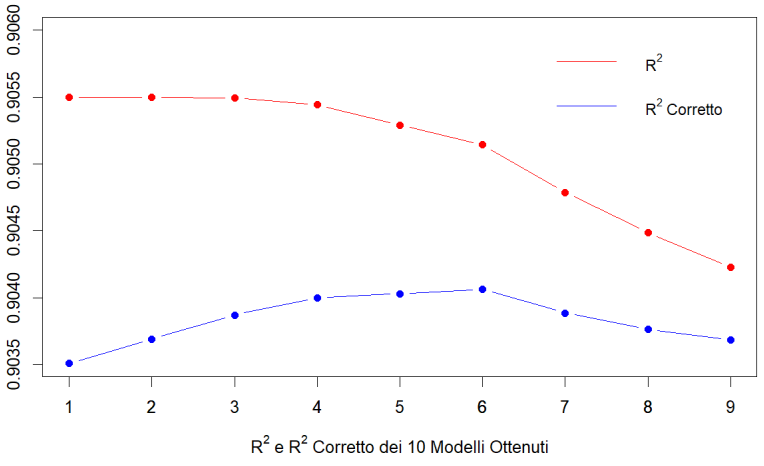
Andando a verificare la normalità dei residui, il test di Shapiro Wilk restituisce un p-value decisamente superiore ai livelli di significatività a cui di solito si fa riferimento (alpha=0.05): ciò impedisce di rigettare l'ipotesi nulla ovvero la normalità della distribuzione.

L'analisi dei residui è stata decisamente positiva, anche perché effettuata su un numero sufficientemente elevato di osservazioni. Abbiamo la conferma che il modello abbia catturato l'essenza del problema dato che gli errori sembrano essere dovuti esclusivamente al caso.

### 3.2 Modello di Regressione Non Lineare

Il modello di regressione lineare ha dato ottimi risultati, tuttavia è stato ritenuto necessario andare a verificare se un modello di regressione logaritmico potesse fornire previsioni migliori, magari riuscendo a catturare componenti non lineari del problema. Dei nove modelli ottenuti tramite riduzione, questa volta è stato selezionato il nono poiché il primo a presentare p-value sufficientemente bassi.

```
> summary(lm.log)

Call:
lm(formula = W ~ .-FG3A-FTA-PTS-BLK-STL-REB-AST-FGA, data = l

Residuals:
     Min       1Q   Median       3Q      Max
-0.61935 -0.05781  0.00720  0.07747  0.25661

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.10873    0.87219   4.711 3.16e-06 ***
TOV          -0.16678    0.06454  -2.584     0.01 *
OFF_RATING    7.26847    0.14497  50.137  < 2e-16 ***
DEF_RATING   -7.26919    0.13770 -52.790  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 ' ' 1

Residual standard error: 0.1065 on 530 degrees of freedom
Multiple R-squared:  0.9042,    Adjusted R-squared:  0.9037
F-statistic:  1668 on 3 and 530 DF,  p-value: < 2.2e-16
```

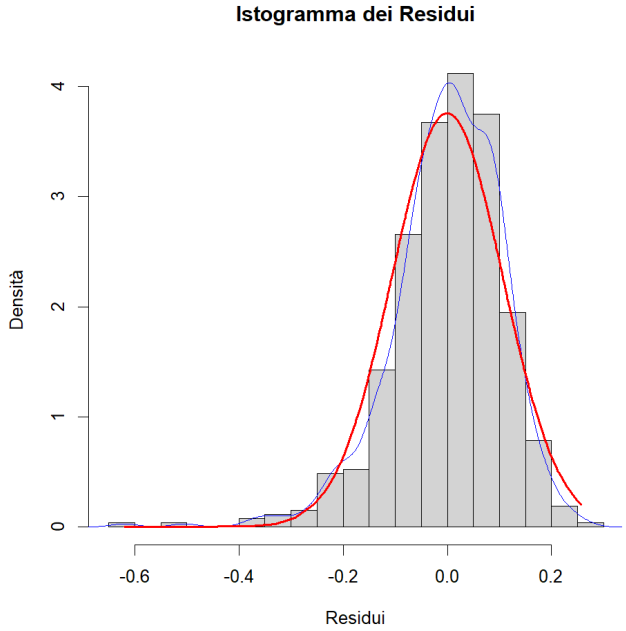


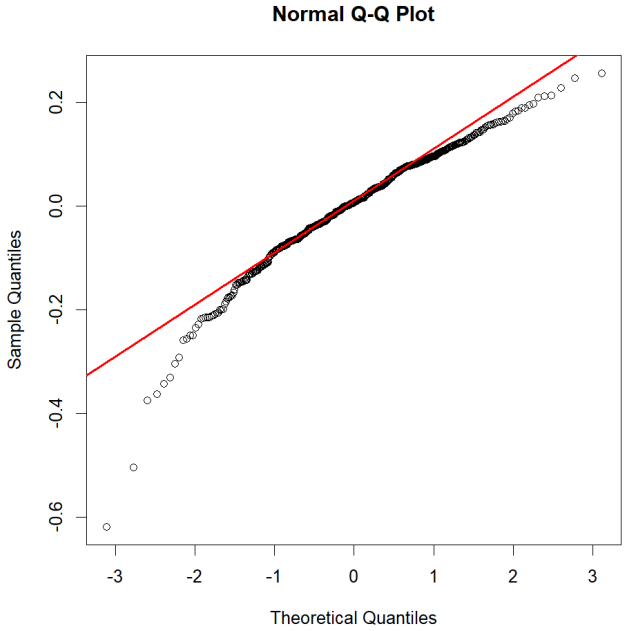

Rispetto al modello lineare, la proporzione di varianza spiegata e varianza spiegata aggiustata calano di circa 4.5% e, inoltre, il modello logaritmico ridotto possiede un numero

inferiore di fattori di ingresso considerati. Anche in questo modello gli indicatori avanzati di OFF_RATING e DEF_RATING hanno i p-value più bassi, come mostrato nelle considerazioni preliminari, sono i fattori di ingresso più correlati al numero di vittorie.

### 3.2.1 Analisi dei Residui

L'istogramma dei residui è simile a quello nel caso lineare mentre il Q-Q plot mostra le code della gaussiana discostarsi, anche notevolmente, dalla bisettrice.

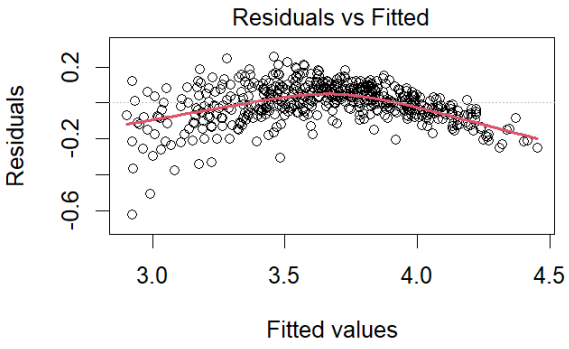



Il grafico di dispersione dei valori residui e dei valori stimati evidenzia una relazione particolare tra i residui: la presenza di un pattern evidente nella distribuzione dei residui suggerisce che il modello logaritmico possa non essere il più adeguato all'analisi del fenomeno.

Nonostante la forte similitudine con il grafico di una distribuzione normale, il test di Shapiro Wilk ha un p-value molto ridotto che costringe a rifiutare l'ipotesi nulla di normalità dei residui.

```
> skewness=mean(((lm.log.r-mean(lm.log.r))/sd(lm.log.r))^3)
> skewness
[1] -1.009664
> kurtosi=mean(((lm.log.r-mean(lm.log.r))/sd(lm.log.r))^4)-3
> kurtosi
[1] 3.05411
> shapiro.test(lm.log.r)

    Shapiro-Wilk normality test

data: lm.log.r
W = 0.95504, p-value = 1.101e-11
```



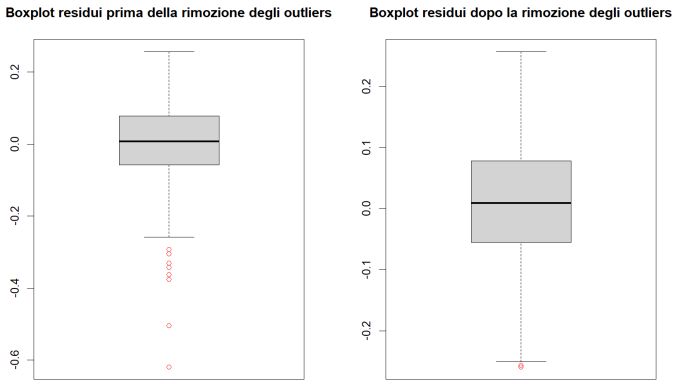

Inizialmente, è stata fatta l'ipotesi che la non normalità potesse essere dovuta al processo di riduzione ma, come mostrato in figura, il p-value ottenuto eseguendo il test sul modello non ridotto restituisce risultati circa uguali a quelli del test effettuato sul modello ridotto.

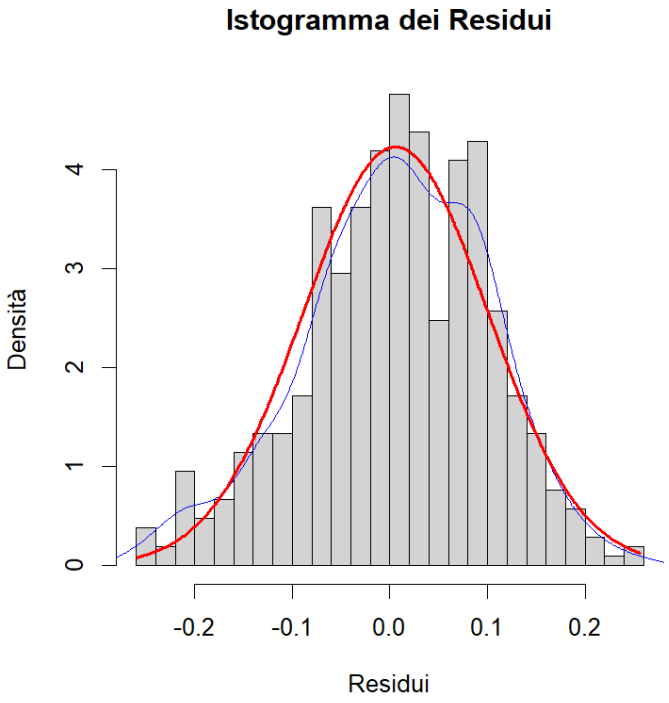
```
> # valutazione dei residui del modello di regressione non ridotto
> lm.log.1.r = residuals(lm.log.1)
> shapiro.test(lm.log.1.r)

   Shapiro-Wilk normality test

data:  lm.log.1.r
W = 0.95491, p-value = 1.051e-11
```

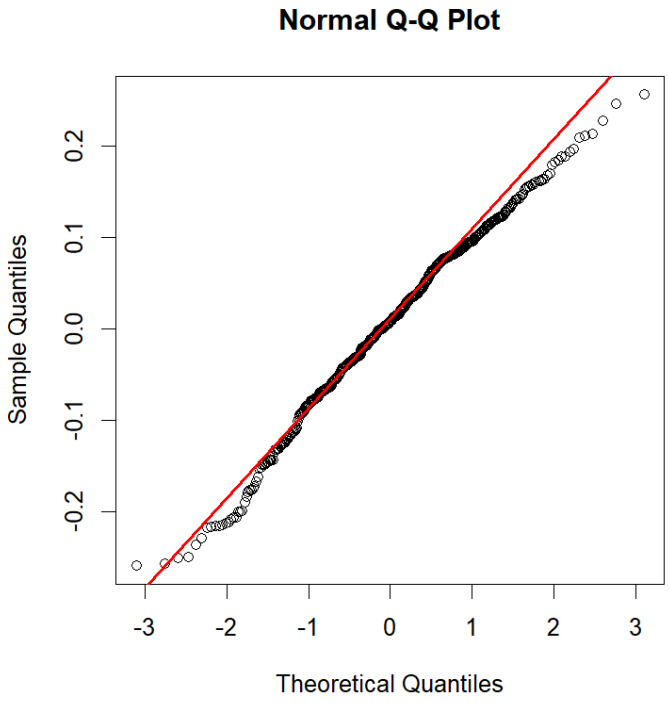
La seconda ipotesi è stata che potesse essere ricondotta alla presenza di outliers, che come visibile dall'analisi del Q-Q plot, sembrano avere un'incidenza negativa decisamente più amplificata rispetto al modello lineare. Andando a rimuovere gli outliers, si possono ottenere risultati decisamente più soddisfacenti.

```
# rimozione dei residui outliers
r_outliers <- boxplot(lm.log.r, plot=FALSE)$out
r_outliers <- rev(sort(r_outliers))
r_outliers
lm.log.r<-lm.log.r[-which(lm.log.r %in% r_outliers[1:length(r_outliers)])]
```

```
> # indicatori post rimozione outliers
> skewness=mean(((lm.log.r-mean(lm.log.r))/sd(lm.log.r))^3)
> skewness
[1] -0.3017812
> kurtosi=mean(((lm.log.r-mean(lm.log.r))/sd(lm.log.r))^4)-3
> kurtosi
[1] -0.05749205
> shapiro.test(lm.log.r)

   Shapiro-Wilk normality test

data:  lm.log.r
W = 0.99107, p-value = 0.002883
```




Dopo la rimozione di 8 outliers, il Q-Q plot mostra i residui molto più aderenti alla bisettrice, il valore di skewness e soprattutto di kurtosi sono diminuiti; il test di Shapiro

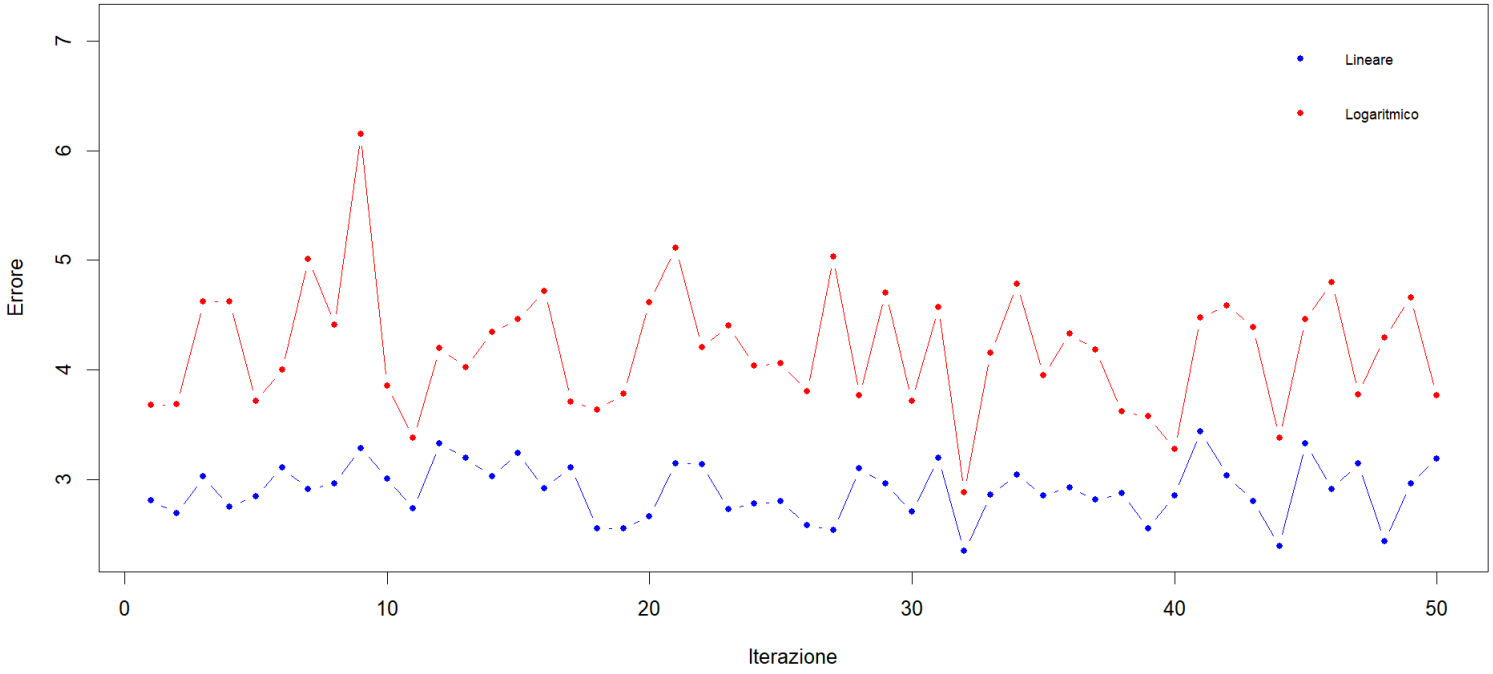
Wilk restituisce un p-value decisamente migliorato però non permette ancora di rigettare l'ipotesi nulla di normalità.

L'analisi dei residui sul modello non lineare, dopo la rimozione degli outliers, ha dato risultati accettabili, siamo comunque lontani dalla normalità dei residui trovata con il modello lineare.

# 4 Autovalutazione e Previsione

Avendo due modelli con proporzione di varianza spiegata elevata, rispettivamente di 94% e 90%, l'analisi dei residui potrebbe non essere sufficiente a distinguere chiaramente quale dei due sia un modello migliore.

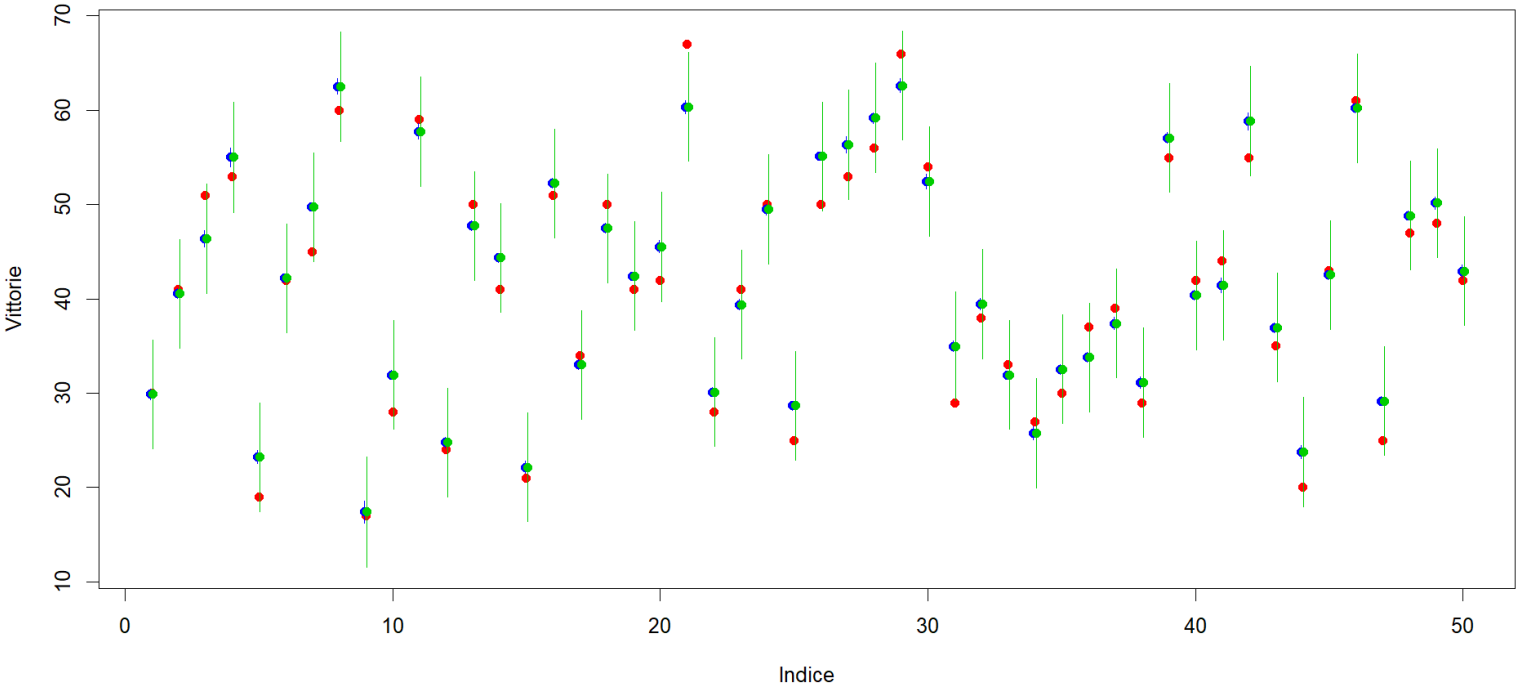
Per esprimere un giudizio definitivo sui due modelli analizzati, sono stati messi a confronto: le 534 osservazioni sono state divise in un sottoinsieme di 50, utilizzate come *test set*, e le restanti 484, utilizzate come *training set*. Questo procedimento è stato ripetuto per un totale di 50 iterazioni in modo da garantire una certa rilevanza statistica. I risultati delle 50 iterazioni sono riportati nel seguente grafico.




Come visibile, il modello lineare ha dato risultati migliori: ha una media di errore più bassa e una deviazione standard inferiore.

```
> # media errori
> mean(err_lin)
[1] 2.903416
> mean(err_log)
[1] 4.1868
> # deviazione standard errori
> sd(err_lin)
[1] 0.2566418
> sd(err_log)
[1] 0.5712424
```

Infine, sono state confrontate le previsioni ottenute con il modello lineare con gli intervalli di confidenza (in blu) e predizione (in verde) forniti da R. Non sono state analizzate le previsioni del modello non lineare in quanti gli intervalli di confidenza sono ottenuti sotto l'ipotesi di gaussianità dei residui, che è confermata solo per il modello lineare.




## 5 Conclusioni

In conclusione, il modello lineare è sicuramente quello più affidabile e soprattutto più adatto al problema: dai risultati ottenuti dall'analisi emerge che la relazione tra il numero di vittorie e i fattori di ingresso sia di tipo lineare.

Osservando i grafici riguardanti il modello di regressione non lineare, che sulla carta rimane comunque un buon modello ($R^2 = 90\%$) con anche un numero inferiore di fattori di ingresso, si può concludere che esso risulti meno adatto a catturare le legge che lega i parametri di ingresso all'output:

- il grafico di dispersione dei valori residui e dei valori stimati presenta un pattern ben definito che indica che l'errore nelle previsioni non sia esclusivamente dovuto al caso ma che sia presente anche una componente dovuta alla natura del modello stesso.
- nell'analisi dei residui effettuata, è emerso come gli outliers abbiano un'incidenza negativa decisamente più marcata rispetto al modello lineare. Le previsioni del modello non lineare riguardanti questi dati presentano un errore decisamente maggiore rispetto a quelle del modello lineare.
- i risultati ottenuti dalla validazione son sempre nettamente inferiori rispetto a quelli del modello lineare.

Inoltre, un'altra conclusione interessante può essere ricavata dalla composizione dei modelli ridotti: è infatti possibile riconoscere quali siano le statistiche che svolgono un ruolo più importante nel decidere quante partite vincerà una squadra. I parametri avanzati selezionati per l'analisi, ovvero OFF_RATING e DEF_RATING, ma anche le palle perse (TOV), sono presenti in entrambi i modelli ottenuti.

# A Appendice

Lo script utilizzato per ottenere le due tabelle utilizzate nell'analisi, fa uso delle *nba-api* ufficiali per Python. La funzione *get_data()*, per ogni stagione indicata nella lista "seasons", scarica i dati sotto forma di liste di *dictonaries* utilizzando la funzione *leaguedashteamstatsLeagueDashTeamStats* delle *nba-api*. Specificando il valore "Base" del parametro *measure_type_detailed_defense* sarà possibile scaricare le informazioni della tabella *Teams General Traditional*; specificando invece il valore "Advanced" le informazioni saranno quelle della tabella *Teams General Advanced*.

Una volta ottenute entrambe le liste di *dictionaries*, *get_data()* le unisce usando il codice seguente e le aggiunge al dataframe da restituire.

```
d = defaultdict(dict)
for l in (allTeamsTraditionalList, allTeamsAdvancedList):
    for elem in l:
        d[elem['TEAM_ID']].update(elem)
result = d.values()
```

Dal dataframe ottenuto tramite *get_data()*, vengono rimosse tutte le informazioni che non sono presenti nelle tabelle ai link riportati in precedenza. In particolare vengono rimosse le colonne contenenti i ranking relativi ai vari aspetti del gioco (PTS_RANK, AST_RANK...). Infine, il dataframe viene salvato in formato .csv con il nome "tabella.csv".