Nicolò Samuelli 866735
Carlotta Giacchetta 868779

NBA

OBIETTIVO DEL LAVORO:

Analizzare lo stipendio percepito da un giocatore del NBA in funzione delle statistiche realizzate nell'anno precedente al fine di comprendere quanto l'effettivo valore in campo del giocatore influisca sul salario percepito e quanto invece non sia spiegabile dai soli dati raccolti.

COSTRUZIONE DEL DATASET E SUA DESCRIZIONE:

Per costruire il dataset abbiamo unito una serie di dati presenti nel sito Basketball Statistics & History of Every Team & NBA and WNBA Players | Basketball-Reference.com. Ogni osservazione (2046) rappresenta un giocatore in un determinato anno (abbiamo considerato le ultime 5 stagioni)
In particolare abbiamo deciso di studiare:

- caratteristiche specifiche del giocatore quali: ruolo (Pos), età (Age), squadra (Tm);
- statistiche riferite a quanto un giocatore abbia giocato: numero di partite giocate (G), numero di partite giocate da titolare (GS), minuti giocati durante l'anno (MP);
- statistiche di tiro: tiri dal campo segnati (FG), tiri dal campo tentati (FGA), percentuale di tiri dal campo riusciti (FG.), tiri da tre punti segnati (X3P), tiri da tre punti tentati (X3PA), percentuale di tiri da tre punti riusciti (X3P.), tiri da due punti segnati (X2P), tiri da due punti tentati (X2PA), percentuale di tiri da due punti riusciti (X2P.), media pesata delle percentuali di tiro da tre punti e da due punti (con pesi rispettivamente 3 e 2) (eFG.), tiri liberi segnati (FT), tiri liberi tentati (FTA), percentuale di tiri liberi riusciti (FT.), punti totali in stagione (PTS);
- altre statistiche di gioco: rimbalzi totali (TRB), rimbalzi offensivi (ORB), rimbalzi difensivi (DRB), assist (AST), palle recuperate (STL), palle perse (TOV), stoppate (BLK), falli personali (PF).

Vediamo come si distribuisce la variabile target:

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
  19186  1836090  4200000  8501989 12138345 48070014        1
```

Il campo di variazione è molto elevato, dalla distribuzione si può notare che la maggior parte dei giocatori prende uno stipendio inferiore alla media e solo pochi prendono uno stipendio molto superiore alla media.
si nota la presenza di un dato mancante che verrà poi rimosso.

QUALCHE OSSERVAZIONE SU ALCUNE VARIABILI IN PARTICOLARE:

➔ POS

```
  C     C-PF       PF     PF-C    PF-SF       PG    PG-SG       SF    SF-SG       SG    SG-PG SG-PG-SF    SG-SF    SF-PF     SF-C    SG-PF
401        5      415        5        4      368        4      333        8      488        5        0        3        5        1        1
```

Abbiamo notato che per alcuni giocatori ci sono dei ruoli doppi → teniamo in considerazione solo del primario

```
  C  PF  PG  SF  SG
406 424 372 347 497
```

Osservando il box-plot non sembra esserci influenza sullo stipendio percepito da un giocatore in base al ruolo che ricopre.

➔ Age

I giocatori alle prime esperienze e a fine carriera percepiscono salari inferiori rispetto ai giocatori all'apice della carriera.

➔ G

All'aumentare delle partite giocate vi è un incremento dello stipendio percepito

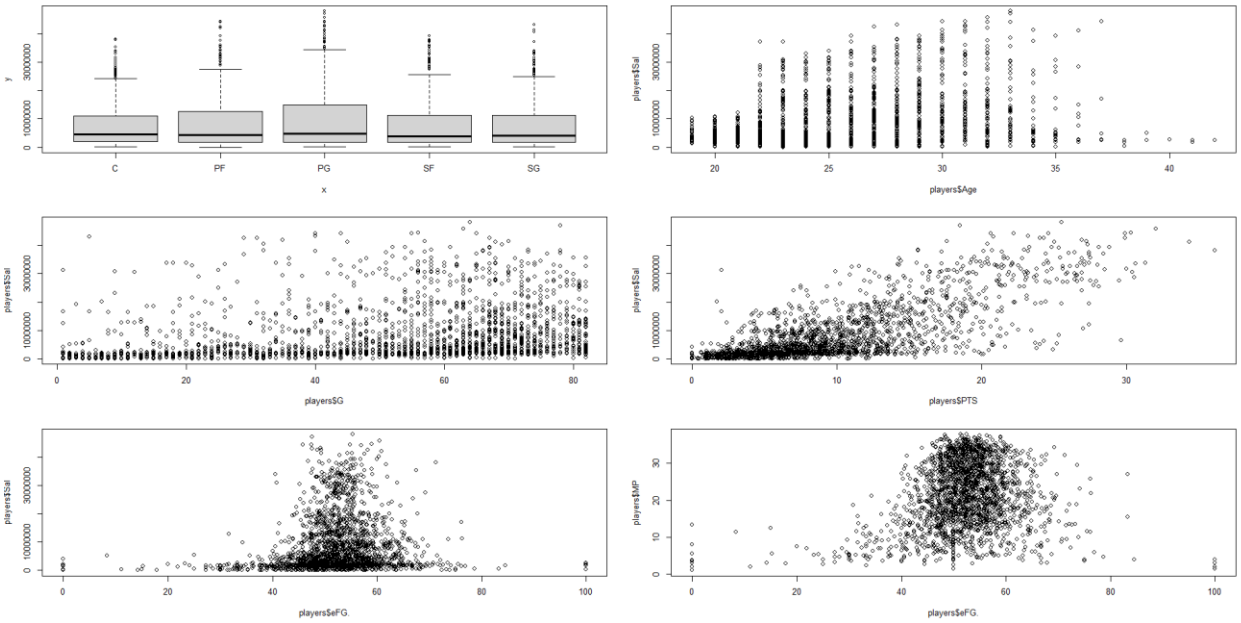
➔ PTS

Ci sono alcuni giocatori che hanno totalizzato pochi punti in stagione, la maggior parte di questi percepiscono stipendi bassi tuttavia vi sono dei valori inaspettati su cui indagheremo. Ad esclusione di chi totalizza pochi punti si nota un andamento crescente de salario al crescere dei punti totalizzati

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

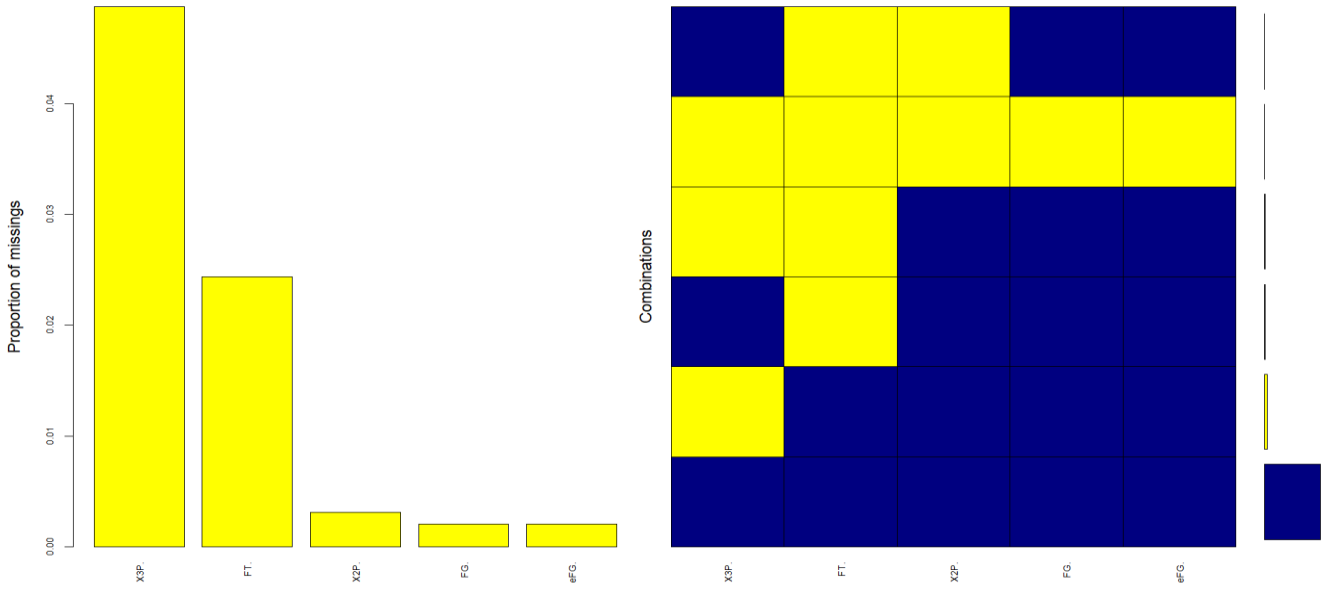
➔ eFG.

Questa distribuzione ci lascia un po' perplessi in quanto la maggior parte dei giocatori ha una percentuale di tiri realizzati compresa tra il 40%-65% e in questo range non sembra esserci una particolare relazione con lo stipendio. I giocatori con percentuali al di fuori di questo range percepiscono stipendi bassi. (perché nonostante percentuali >80% prendono uno stipendio basso?) → Proviamo ad osservare la relazione tra eFG. e minuti giocati



MISSING:

Per studiare i valori mancanti abbiamo rappresentato tramite grafico le combinazioni dei missing nelle sole variabili che presentano almeno un valore mancante, tali variabili risultano essere tutte legate alle percentuali di tiro segno quindi che non siano effettivamente mancanti ma rappresentative di giocatori che in stagione non hanno messo a segno nessun tiro. Abbiamo quindi deciso di sostituirli con zero.



Nicolò Samuelli 866735
Carlotta Giacchetta 868779

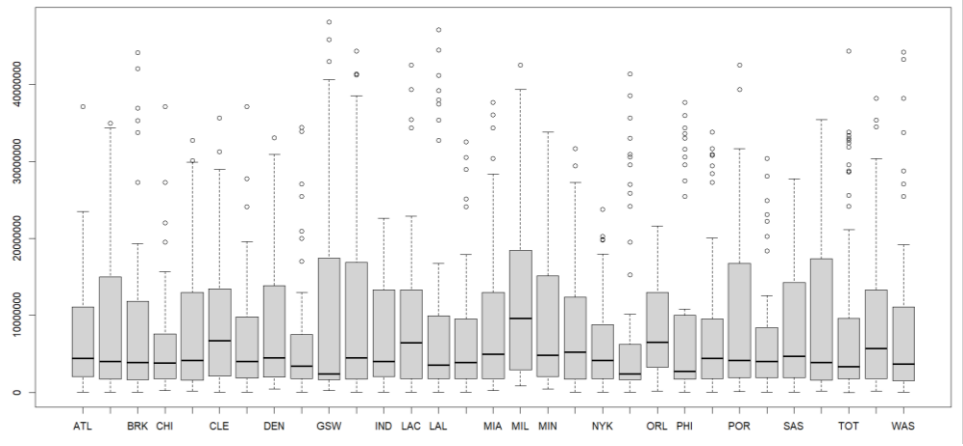
Dopo aver risolto i problemi con i missing possiamo creare delle variabili che meglio sintetizzino le variabili di partenza:

- *PER (Player Efficency Rate)* misura l'efficienza di un giocatore ed ottenuta dalla somma delle statistiche positive - la somma delle statistiche negative diviso per il numero delle partite totali moltiplicato per il numero di partite giocate. (In questo modo andiamo a penalizzare i giocatori che hanno giocato meno ma che potrebbero avere alti valori al numeratore).
  PER = (PTS-(FGA-FG)-(FTA-FT)+1.5*ORB+DRB+AST+STL+BLK-TOV-PF)*(G/82)
- *TRBW (Total Rebounds Weighted)* E' la combinazione lineare tra i rimbalzi offensivi e difensivi (diamo peso maggiore a quelli offensivi in quanto più difficili, rispettivamente 1,5 e 1)
- *PAS* numero di assist + 1 fatti sul totale di palle perse + 1 (abbiamo aggiunto + 1 per non avere denominatore nullo)

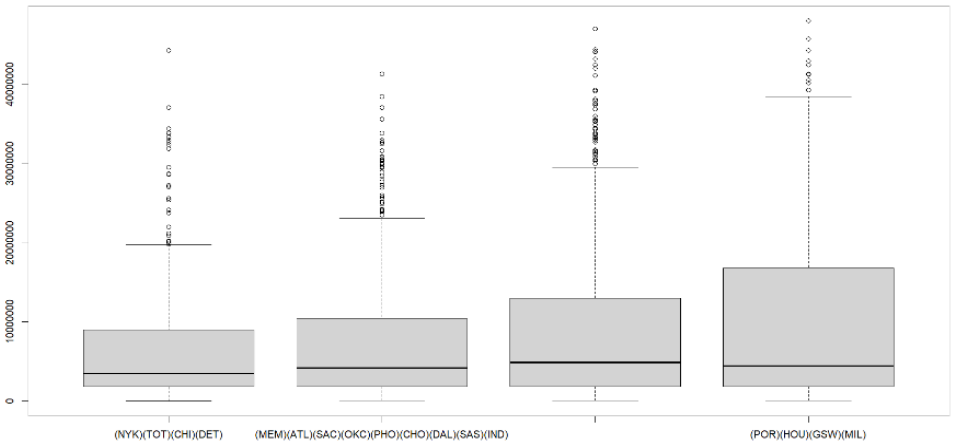
```
> summary(df1$PER)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1195  1.3390  4.5084  7.3902 33.9317
> summary(df1$TRBW)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.450   3.900   4.407   5.650  18.550
> summary(df1$PAS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4643  1.0625  1.2857  1.3756  1.6000  3.4706
```

eliminiamo "ORB", "TRB", "DRB", "AST", "TOV"

➔ Tm



Abbiamo troppi livelli e poco significativi → facciamo optimal grouping



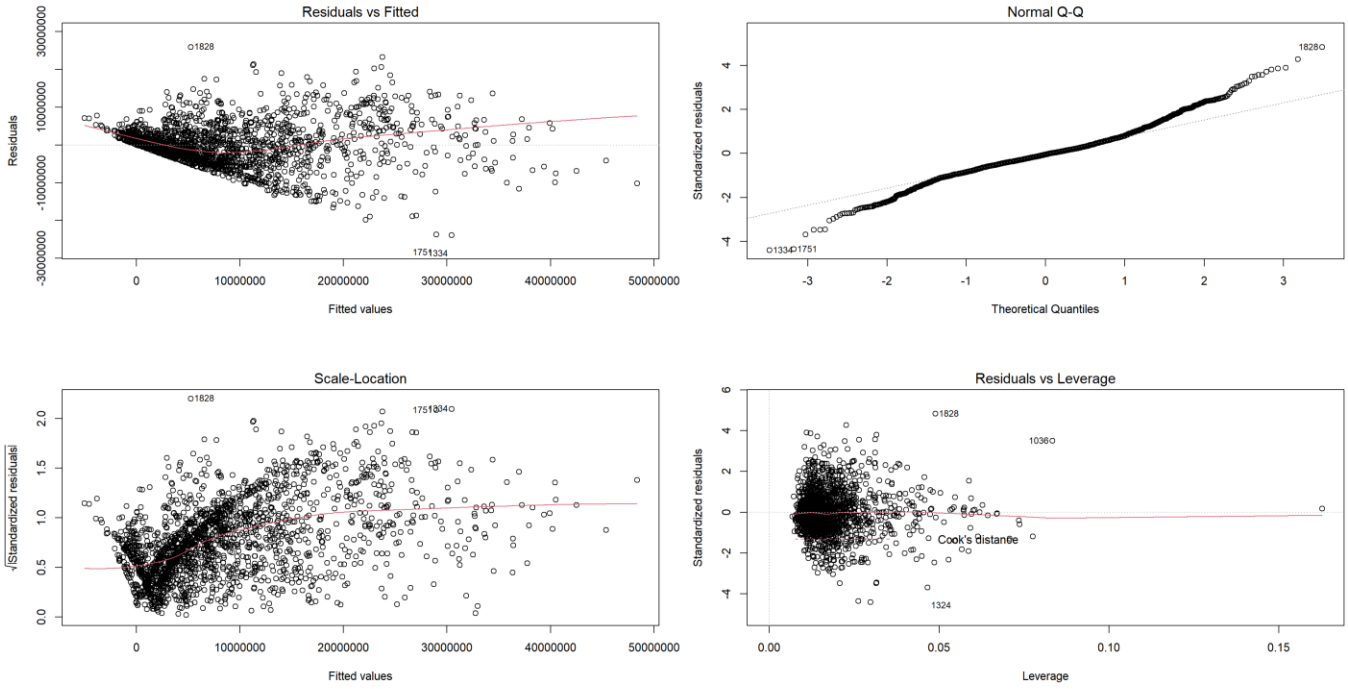
Notiamo che le mediane dei vari livelli sono praticamente uguali ma cambia il valore del terzo quartile

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

FIT DELPRIMO MODELLO: modello completo

Variable		N	Estimate	p
Pos	C	406	Reference	
	PF	423	367103.83 (-556970.99, 1291178.65)	0.436
	PG	372	-750360.44 (-2017097.97, 516377.10)	0.245
	SF	347	260358.33 (-861581.52, 1382298.17)	0.649
	SG	497	-627649.53 (-1793695.85, 538396.79)	0.291
Age		2045	479097.45 (414965.81, 543229.08)	<0.001
G		2045	-27814.11 (-42835.79, -12792.43)	<0.001
GS		2045	-76473.12 (-110727.56, -42218.68)	<0.001
MP		2045	21313.86 (-92904.24, 135531.96)	0.714
FG		2045	-1609452.23 (-9150360.70, 5931456.23)	0.676
FGA		2045	6329220.86 (1278768.71, 11379673.01)	0.014
FG.		2045	229887.21 (81614.06, 378160.35)	0.002
X3P		2045	-8026751.39 (-14848975.57, -1204527.21)	0.021
X3PA		2045	-4942992.00 (-9998472.45, 112488.44)	0.055
X3P.		2045	-11381.15 (-39257.15, 16494.85)	0.423
X2P		2045	-4822262.72 (-10169648.80, 525123.36)	0.077
X2PA		2045	-5911080.25 (-10996287.80, -825872.71)	0.023
X2P.		2045	817.50 (-46734.11, 48369.11)	0.973
eFG.		2045	-204433.66 (-346303.97, -62563.36)	0.005
FT		2045	-2010367.40 (-5482388.79, 1461653.98)	0.256
FTA		2045	-171670.02 (-1272570.59, 929230.54)	0.760
FT.		2045	-9679.76 (-27999.67, 8640.14)	0.300
STL		2045	702163.47 (-246900.82, 1651227.75)	0.147
BLK		2045	1142854.06 (252313.04, 2033395.07)	0.012
PF		2045	-881342.27 (-1446645.67, -316038.86)	0.002
PTS		2045	3151714.51 (-126141.93, 6429570.95)	0.059
Year	2017-18	377	Reference	
	2018-19	333	-36692.34 (-866735.61, 793350.93)	0.931
	2019-20	393	9282.91 (-799050.05, 817615.86)	0.982
	2020-21	411	110173.53 (-697098.81, 917445.87)	0.789
	2021-22	531	551053.99 (-219047.86, 1321155.83)	0.161
PER		2045	825286.18 (647350.98, 1003221.39)	<0.001
TRBW		2045	84170.74 (-142711.68, 311053.17)	0.467
PAS		2045	1408891.01 (501792.83, 2315989.18)	0.002
OG	(NYK)(TOT)(CHI)(DET)	410	Reference	
	(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)	533	388855.32 (-330009.21, 1107719.84)	0.289
	(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)	870	1265583.39 (610311.53, 1920855.25)	<0.001
	(POR)(HOU)(GSW)(MIL)	232	1293809.90 (383902.41, 2203717.40)	0.005

```
Residual standard error: 5505000 on 2009 degrees of freedom
Multiple R-squared:  0.6831,    Adjusted R-squared:  0.6776
F-statistic: 123.7 on 35 and 2009 DF,  p-value: < 0.00000000000000022
```



1) Grafico dei residui VS interpolati: non è una retta → non c'è linearità nei residui → non c'è linearità del modello → PROBABILE TRASFORMAZIONE DELLA Y
   PROBLEMI: stime dei coefficienti distorte (stimatori distorti)
2) Grafico QQPLOT dei res standardizzati: i residui non si distribuiscono come una normale. In particolare problemi sulla coda detra → residui alti si discostano molto dalla normale

PROBLEMI: nessuno (OLS affidabili, corretti ed eefficienti) → non cambia molto dalla distribuzione "corretta" degli stimatori

3) Residui standardizzati VS interpolati: non è una retta → no omoschedasticità e correlazione tra i residui e i valori interpolati
   PROBLEMI: std non corretti e inferenza sui parametri non corretta.
4) Residual VS leverage: pare che non abbiamo valori influenti sul modello (con distanza di Cook troppo alta → verifica poi)

ANALISI DELLA COLLINEARITÀ

Per variabili quantitative:

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
Age	1.1493	0.8701	13.1166	13.7196	0.9328	-0.0685	0	0.0099	0.1623
G	1.8759	0.5331	76.9610	80.4991	0.7301	-0.1118	0	0.0061	0.5835
GS	14.0958	0.0709	1150.7234	1203.6243	0.2664	-0.8400	1	0.0008	1.1610
MP	16.1489	0.0619	1331.1274	1392.3217	0.2488	-0.9624	1	0.0007	1.1723
FG	5058.8210	0.0002	444428.5347	464859.7327	0.0141	-301.4690	1	0.0000	1.2494
FGA	10059.2138	0.0001	883810.8765	924441.2897	0.0100	-599.4561	1	0.0000	1.2495
FG.	33.9678	0.0294	2896.8672	3030.0415	0.1716	-2.0242	1	0.0003	1.2129
X3P	619.0680	0.0016	54309.3636	56806.0651	0.0402	-36.8920	1	0.0000	1.2476
X3PA	2277.6082	0.0004	200044.5734	209240.9908	0.0210	-135.7289	1	0.0000	1.2491
X3P.	2.3978	0.4170	122.8272	128.4738	0.6458	-0.1429	0	0.0047	0.7285
X2P	1715.2416	0.0006	150629.6613	157554.3842	0.0241	-102.2159	1	0.0000	1.2489
X2PA	5307.8013	0.0002	466306.3203	487743.2804	0.0137	-316.3064	1	0.0000	1.2494
X2P.	4.5364	0.2204	310.7414	325.0267	0.4695	-0.2703	1	0.0025	0.9742
eFG.	28.0832	0.0356	2379.7858	2489.1889	0.1887	-1.6736	1	0.0004	1.2052
FT	424.6269	0.0024	37223.9152	38935.1671	0.0485	-25.3047	1	0.0000	1.2467
FTA	63.5618	0.0157	5497.2802	5750.0003	0.1254	-3.7878	1	0.0002	1.2300
FT.	1.7170	0.5824	63.0026	65.8989	0.7632	-0.1023	0	0.0066	0.5218
STL	2.4919	0.4013	131.0937	137.1203	0.6335	-0.1485	0	0.0046	0.7482
BLK	2.1516	0.4648	101.1918	105.8438	0.6817	-0.1282	0	0.0053	0.6689
PF	2.9028	0.3445	167.1993	174.8857	0.5869	-0.1730	0	0.0039	0.8192
PTS	7456.9956	0.0001	655155.0947	685273.7806	0.0116	-444.3828	1	0.0000	1.2495
PER	19.9037	0.0502	1661.0613	1737.4233	0.2241	-1.1861	1	0.0006	1.1869
TRBW	5.6311	0.1776	406.9350	425.6426	0.4214	-0.3356	1	0.0020	1.0277
PAS	1.9456	0.5140	83.0910	86.9109	0.7169	-0.1159	0	0.0058	0.6074

passaggi intermedi:

1) Eliminiamo FG, X3P, X2P, FT ovvero tutte le statistiche di tiri riusciti, tenendo solo i tentanti e la percentuale
2) Eliminiamo FGA, X3PA, X2PA, FTA, eFG., ovvero tutte le statistiche sui tiri tentati e eFG. che è costruita come c.l. delle altre percentuali. Teniamo tutte le altre percentuali
3) Eliminiamo MP in quanto in relazione sia con le partite giocate che con le statistiche di gioco
4) Eliminiamo PTS in quanto già spiegata da PER

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
Age	1.0853	0.9214	14.4503	15.7717	0.9599	-0.4432	0	0.0054	0.1477
G	1.8217	0.5489	139.1360	151.8595	0.7409	-0.7438	0	0.0032	0.8471
GS	7.3577	0.1359	1076.5762	1175.0248	0.3687	-3.0044	1	0.0008	1.6228
FG.	3.5982	0.2779	439.9538	480.1858	0.5272	-1.4692	1	0.0016	1.3561
X3P.	1.2823	0.7798	47.8043	52.1758	0.8831	-0.5236	0	0.0046	0.4135
X2P.	3.1502	0.3174	364.0981	397.3934	0.5634	-1.2863	1	0.0019	1.2819
FT.	1.3317	0.7509	56.1657	61.3019	0.8666	-0.5438	0	0.0044	0.4678
STL	2.1011	0.4759	186.4530	203.5033	0.6899	-0.8579	0	0.0028	0.9842
BLK	2.0639	0.4845	180.1556	196.6300	0.6961	-0.8428	0	0.0029	0.9681
PF	2.5215	0.3966	257.6353	281.1950	0.6298	-1.0296	0	0.0023	1.1332
PAS	1.6273	0.6145	106.2143	115.9272	0.7839	-0.6645	0	0.0036	0.7239
PER	8.1073	0.1233	1203.5034	1313.5589	0.3512	-3.3105	1	0.0007	1.6464
TRBW	3.9887	0.2507	506.0935	552.3737	0.5007	-1.6287	1	0.0015	1.4072

Sussistono ancora dei lievi problemi ma decidiamo di conservare le variabili per non perdere troppe informazioni

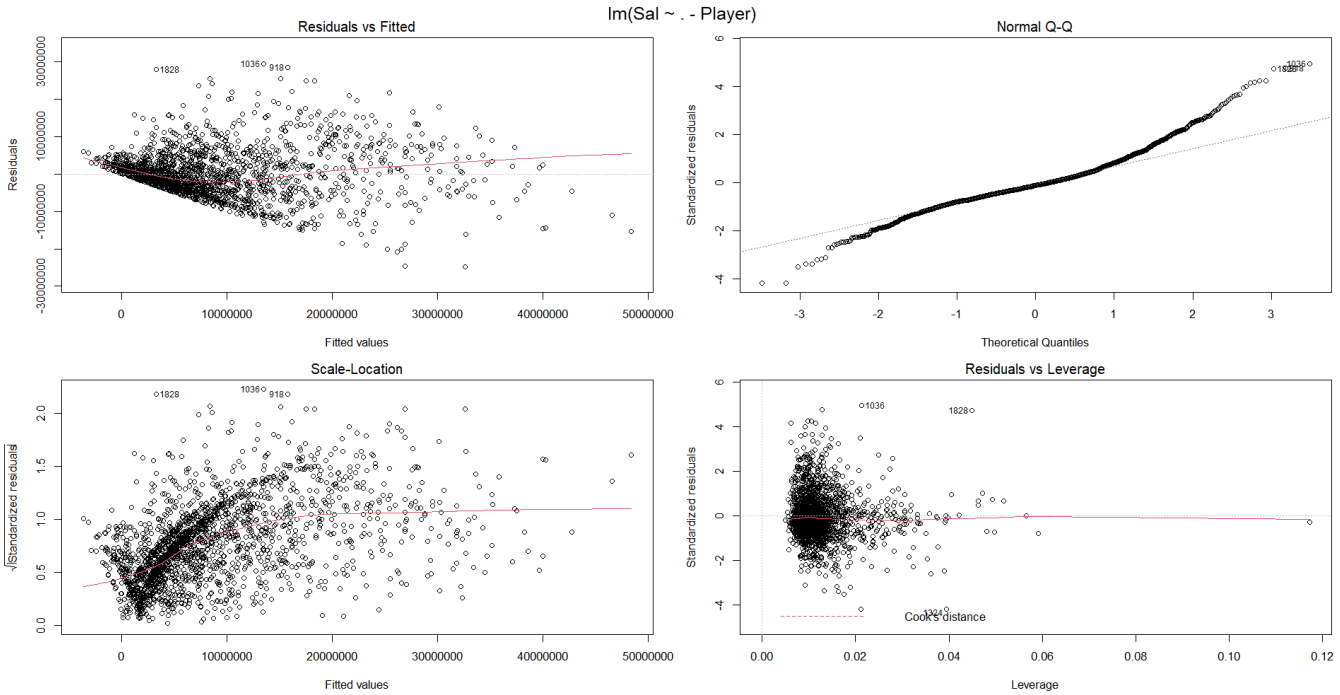
Nicolò Samuelli 866735
Carlotta Giacchetta 868779

Per variabili qualitative:

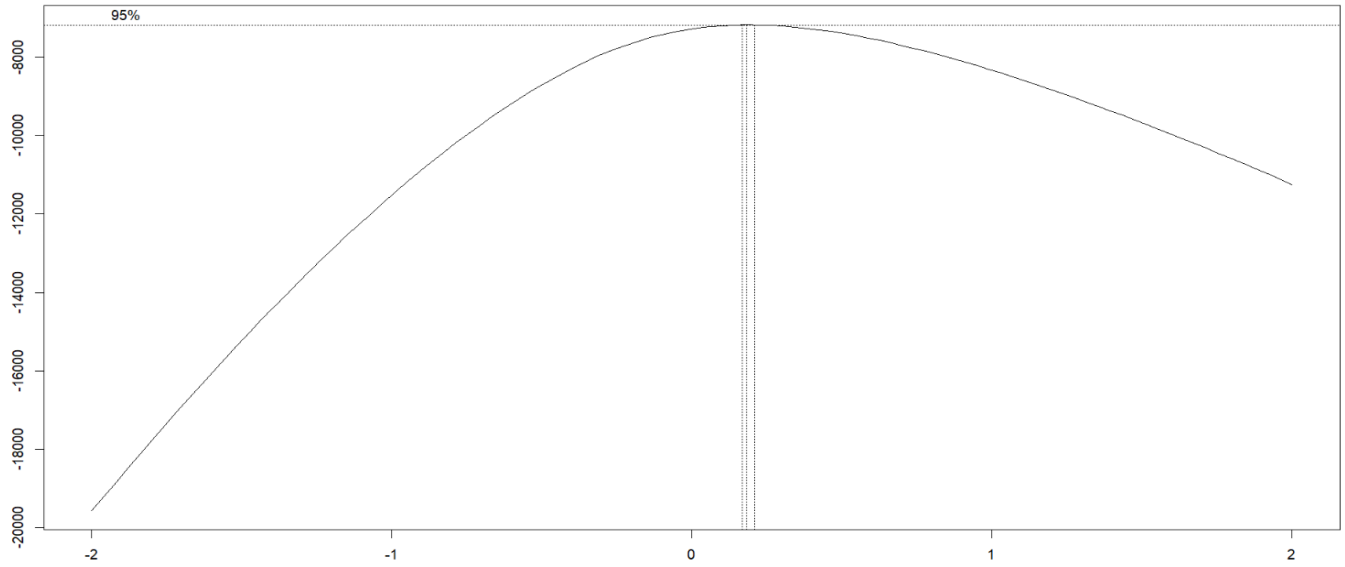
	X1	Row	Column	Chi.Square	df	p.value	n	u1	u2	nMinu1u2	Chi.Square.norm
1	1	Pos	Year	7.150	16	0.970	2045	4	4	8180	0.0008741174
2	2	Pos	OG	5.541	12	0.937	2045	4	3	6135	0.0009031682
3	3	Year	OG	5.670	12	0.932	2045	4	3	6135	0.0009242617

Non si evidenzia nessun problema, come ci aspettavamo.

Notiamo miglioramenti nel plot:



BOX-COX



Lambda = 0.22 → proviamo a trasformare il salario utilizzando il logaritmo

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

Prima della trasformazione:

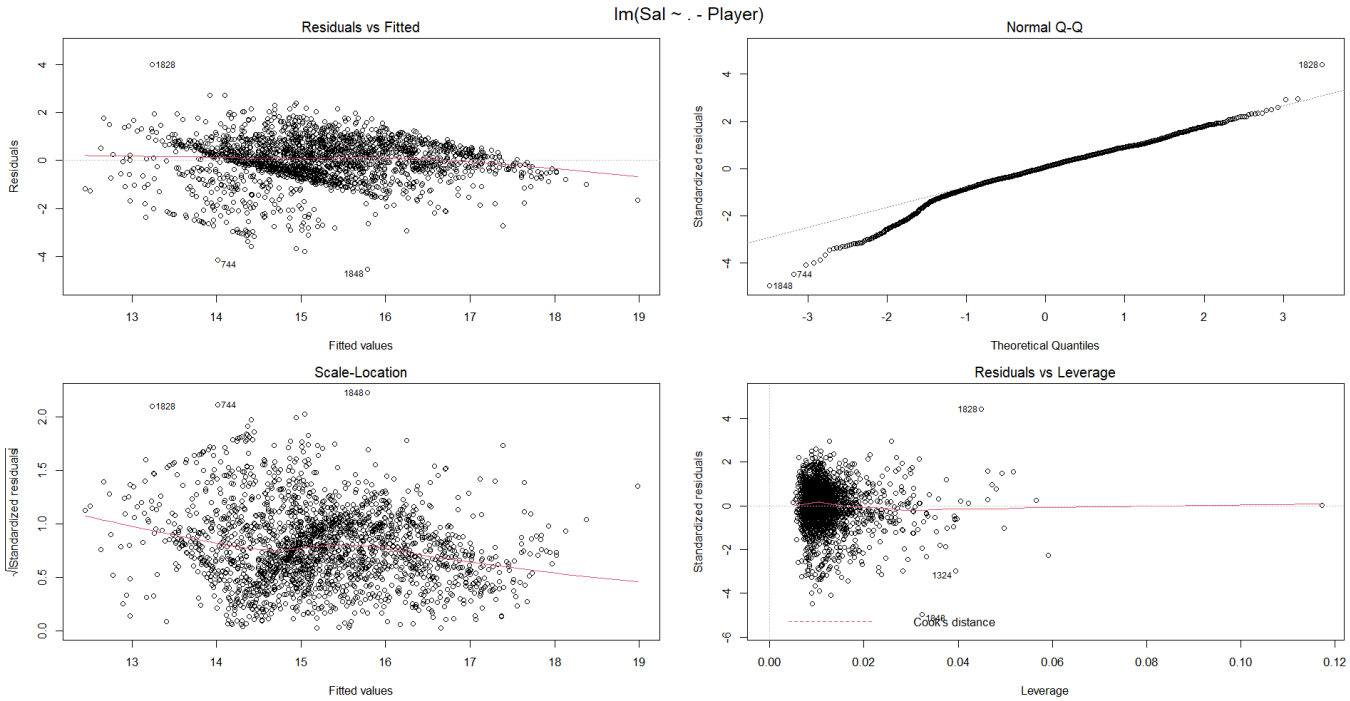
```
> summary(df2$Sal)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   19186  1836090  4200000  8501989 12138345 48070014
```

Dopo la trasformazione:

```
> summary(df2_log$Sal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.862  14.423  15.251  15.248  16.312  17.688
```

Notiamo che il range si restringe moltissimo e i valori sono molto più contenuti → abbiamo delle stime dei coefficienti molto più basse e degli standard error molto più piccoli

Dall'analisi dei grafici di diagnostica sul modello:

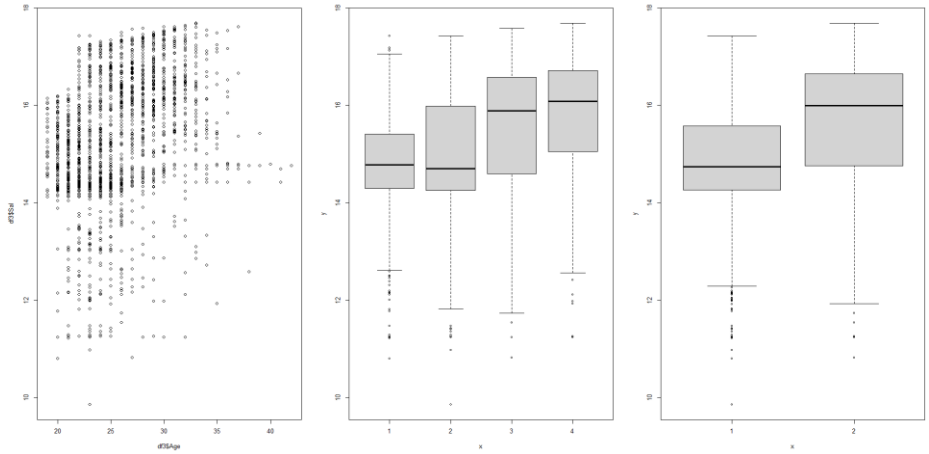


BINNING

Abbiamo deciso di rendere fattoriale la variabile Age in quanto la sua distribuzione evidenzia la presenza di categorie ben definite: in primo luogo abbiamo deciso di raggruppare in base ai quartili e poi di raggrupparli in due ulteriori gruppi in quanto aventi distribuzioni simili, come si può notare dal box-plot centrale.

```
> summary(df3$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   22.00   25.00   25.72   29.00   42.00
```

Distribuzione di Age prima del binning



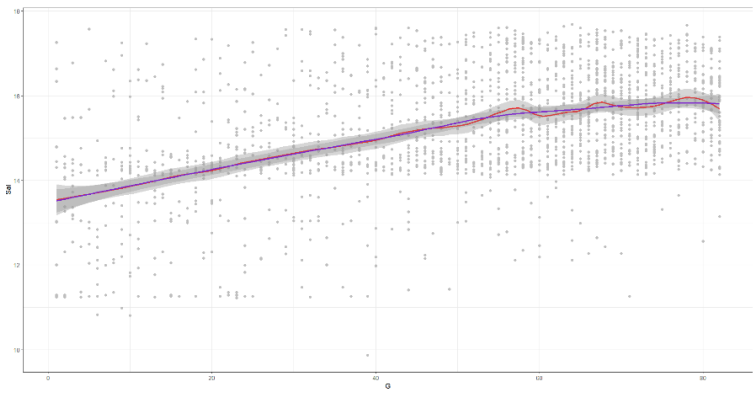
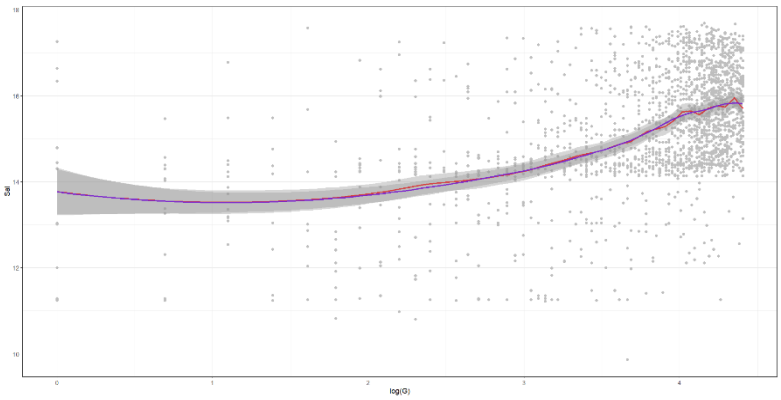
Nicolò Samuelli 866735
Carlotta Giacchetta 868779

GAM

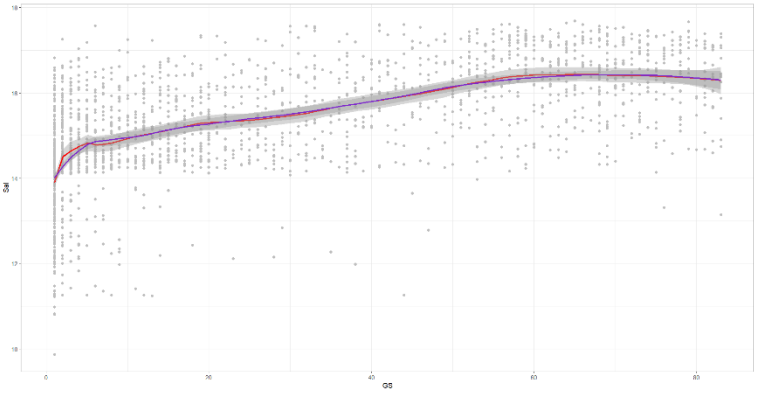
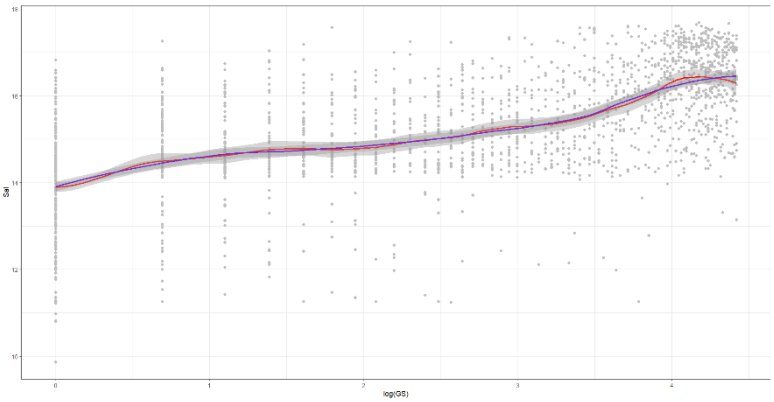
```
           edf Ref.df      F               p-value
s(G)     1.618  2.021 15.394          0.000000395 ***
s(GS)    1.001  1.001 71.416 < 0.0000000000000002 ***
s(FG.)   2.825  3.675  0.576               0.55260
s(X3P.)  1.001  1.001  1.584               0.20844
s(X2P.)  3.286  4.149  1.494               0.19380
s(FT.)   2.547  3.162  3.851               0.00754 **
s(STL)   2.872  3.647  7.459          0.000020310 ***
s(BLK)   3.568  4.448  1.342               0.21429
s(PF)    1.004  1.008  0.072               0.79135
s(PER)   6.522  7.682 21.266 < 0.0000000000000002 ***
s(TRBW)  3.188  4.068  1.947               0.09579 .
s(PAS)   2.460  3.163  1.466               0.21755
```

Stampiamo i singoli grafici delle variabili che la GAM ritiene significative:

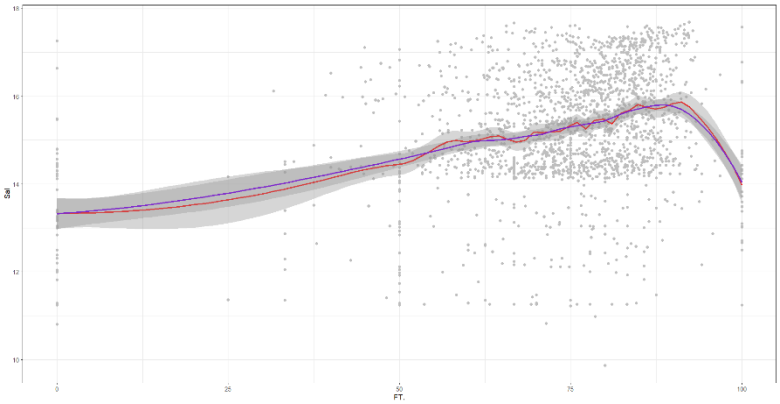
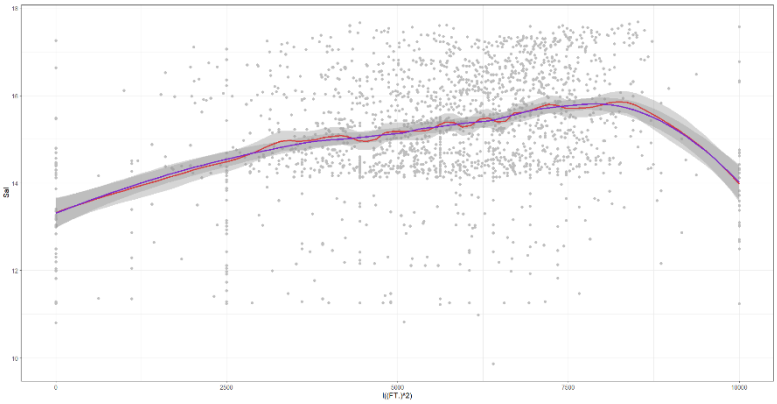
- G (Le variabili a cui è stato applicato il logaritmo sono state traslate di 1)

- GS

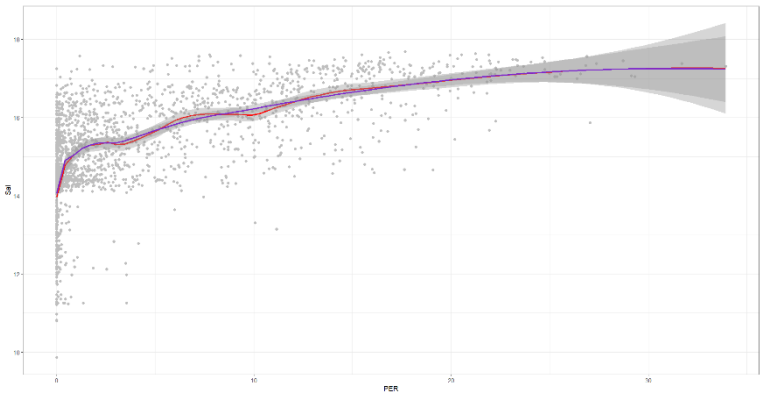



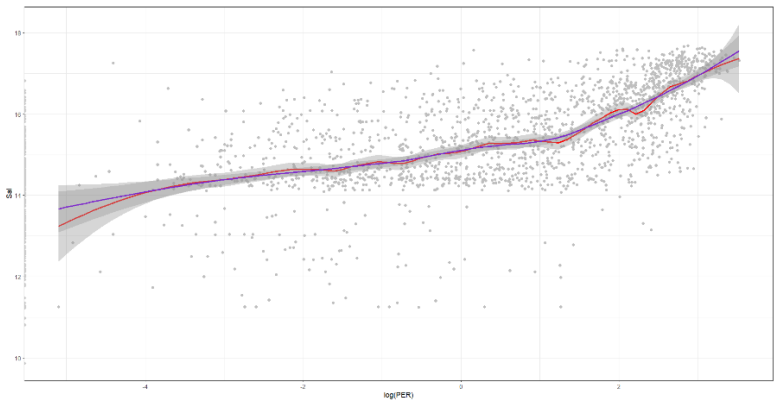
- FT.

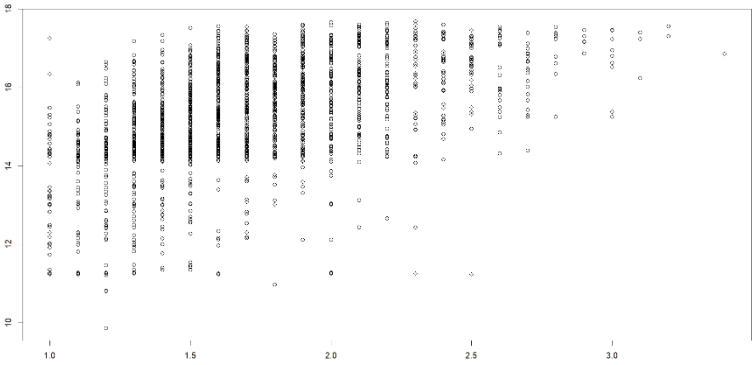
Nicolò Samuelli 866735
Carlotta Giacchetta 868779

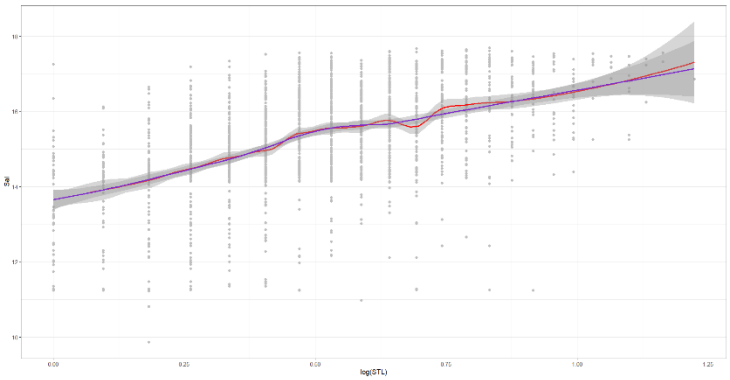
➔ PER





➔ STL





Confronto modello pre vs post gam

```
Model 1: Sal ~ (Player + Pos + G + GS + FG. + X3P. + X2P. + FT. + STL +
    BLK + PF + Year + PER + TRBW + PAS + OG + Age_cl_og) - Player
Model 2: Sal ~ log(G) + log(GS) + FG. + X3P. + X2P. + I((FT.)^2) + FT. +
    log(STL) + BLK + PF + log(PER) + TRBW + PAS + Pos + Year +
    OG + Age_cl_og
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2020 1716.2
2   2019 1662.6  1    53.631 65.127 1.193e-15 ***
```

```
Model:
Sal ~ log(G) + log(GS) + FG. + X3P. + X2P. + I((FT.)^2) + FT. +
    log(STL) + BLK + PF + log(PER) + TRBW + PAS + Pos + Year +
    OG + Age_cl_og
            Df Sum of Sq    RSS     AIC  F value                Pr(>F)
<none>                  1667.7 -365.08
log(G)       1    15.637 1683.3 -348.00  18.9305       0.0000142331267 ***
log(GS)      1     1.640 1669.3 -365.07   1.9857             0.1589483
FG.          1     0.190 1667.9 -366.85   0.2295             0.6319423
X3P.         1     2.935 1670.6 -363.48   3.5536             0.0595585 .
X2P.         1     0.029 1667.7 -367.04   0.0356             0.8503773
I((FT.)^2)   1     2.539 1670.2 -363.97   3.0733             0.0797405 .
FT.          1     0.056 1667.8 -367.01   0.0672             0.7954306
log(STL)     1    34.847 1702.5 -324.79  42.1878       0.0000000001041 ***
BLK          1     0.037 1667.7 -367.03   0.0445             0.8329658
PF           1     0.000 1667.7 -367.08   0.0001             0.9908705
log(PER)     1    80.783 1748.5 -270.35  97.7999 < 0.00000000000000022 ***
TRBW         1    10.539 1678.2 -354.20  12.7593             0.0003625 ***
PAS          1     2.132 1669.8 -364.47   2.5812             0.1082992
Pos          4     8.789 1676.5 -362.33   2.6600             0.0312192 *
Year         4    35.974 1703.7 -329.44  10.8881       0.0000000098154 ***
OG           3    11.085 1678.8 -357.53   4.4735             0.0038841 **
Age_cl_og    1   125.997 1793.7 -218.14 152.5384 < 0.00000000000000022 ***
```

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

AIC VS SBC

```
Step:  AIC=-373.79
Sal ~ log(G) + log(GS) + X3P. + I((FT.)^2) + log(STL) + log(PER) +
    TRBW + PAS + Pos + Year + OG + Age_cl_og

             Df Sum of Sq    RSS     AIC
<none>                   1668.8 -373.79
- log(GS)     1     1.870 1670.6 -373.50
- PAS         1     2.048 1670.8 -373.28
+ FG.         1     0.932 1667.8 -372.93
+ X2P.        1     0.783 1668.0 -372.74
- X3P.        1     3.106 1671.9 -371.98
+ BLK         1     0.072 1668.7 -371.87
+ FT.         1     0.039 1668.7 -371.83
+ PF          1     0.007 1668.8 -371.79
- Pos         4     9.413 1678.2 -370.28
- OG          3    11.283 1680.0 -366.00
- TRBW        1    14.653 1683.4 -357.91
- log(G)      1    19.410 1688.2 -352.14
- I((FT.)^2)  1    19.871 1688.6 -351.58
- Year        4    37.004 1705.8 -336.93
- log(STL)    1    39.136 1707.9 -328.38
- log(PER)    1    88.857 1757.6 -269.69
- Age_cl_og   1   125.951 1794.7 -226.98
```

```
Step:  AIC=-299.82
Sal ~ log(G) + I((FT.)^2) + log(STL) + log(PER) + TRBW + Year +
    Age_cl_og

             Df Sum of Sq    RSS      AIC
<none>                   1695.2 -299.820
+ X3P.        1     2.812 1692.4 -295.592
+ log(GS)     1     2.050 1693.1 -294.672
+ FG.         1     1.287 1693.9 -293.751
+ X2P.        1     1.209 1694.0 -293.656
+ OG          3    12.882 1682.3 -292.551
+ BLK         1     0.273 1694.9 -292.527
+ PAS         1     0.173 1695.0 -292.405
+ FT.         1     0.029 1695.1 -292.232
+ PF          1     0.015 1695.2 -292.215
- Year        4    40.873 1736.0 -281.589
- log(G)      1    23.316 1718.5 -279.507
- I((FT.)^2)  1    25.180 1720.3 -277.291
+ Pos         4     6.486 1688.7 -277.167
- TRBW        1    33.105 1728.3 -267.892
- log(STL)    1    47.964 1743.1 -250.384
- Age_cl_og   1   141.640 1836.8 -143.337
- log(PER)    1   196.987 1892.2  -82.627
```

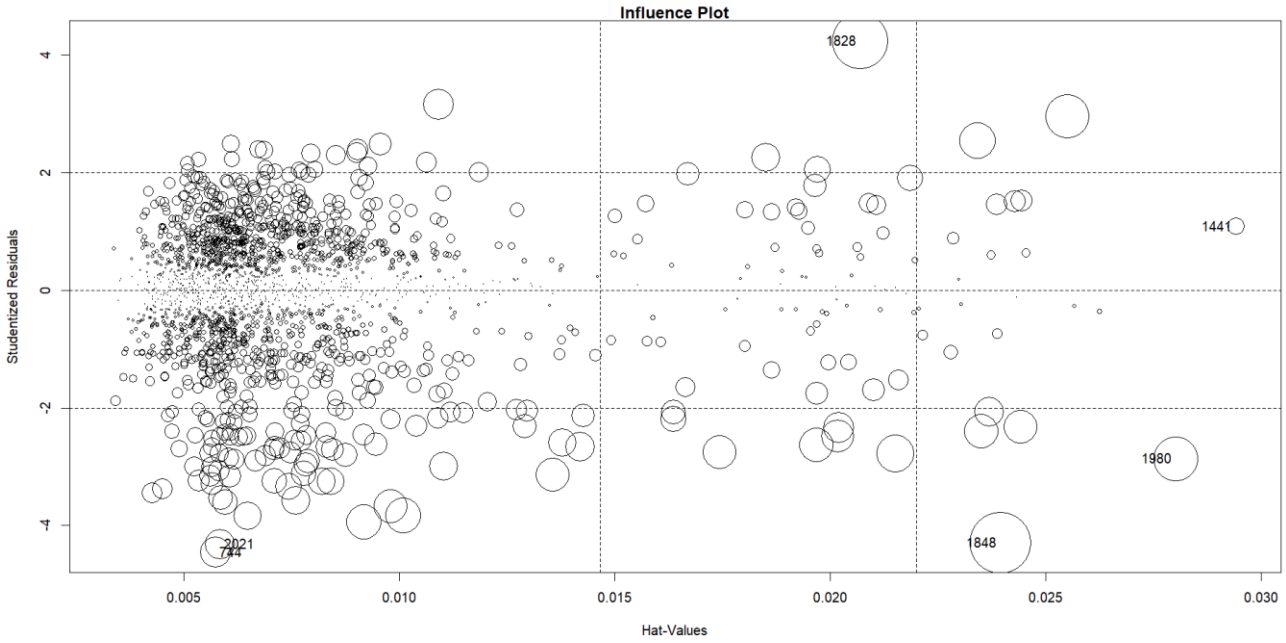
Per selezionare le variabili abbiamo deciso di tenere in considerazione il parere di entrambi i criteri di model selection (AIC e SBC) conservando nel modello le variabili mantenute da entrambi e selezionando fra quelle conservate solamente dall'AIC (meno restrittivo rispetto all' SBC) quelle più significative. Il modello ottenuto è di seguito riportato tramite il suo summary:

```
Call:
lm(formula = Sal ~ log(G) + log(GS) + I(FT.^2) + FT. + log(STL) +
    log(PER) + TRBW + Year + Age_cl_og + OG, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1036 -0.4559  0.0968  0.5692  3.8978

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     1.248e+01  1.444e-01  86.470  < 2e-16 ***
log(G)                                                          2.178e-01  4.112e-02   5.297 1.30e-07 ***
log(GS)                                                        -1.202e-01  3.902e-02  -3.080 0.002098 **
I(FT.^2)                                                        6.277e-05  3.971e-05   1.581 0.114109
FT.                                                             8.036e-05  4.626e-03   0.017 0.986141
log(STL)                                                        8.467e-01  1.129e-01   7.497 9.68e-14 ***
log(PER)                                                        6.341e-01  6.002e-02  10.564  < 2e-16 ***
TRBW                                                            5.843e-02  1.068e-02   5.474 4.96e-08 ***
Year2018-19                                                    -4.459e-02  6.857e-02  -0.650 0.515533
Year2019-20                                                     1.140e-01  6.581e-02   1.733 0.083282 .
Year2020-21                                                     1.976e-02  6.515e-02   0.303 0.761722
Year2021-22                                                     3.299e-01  6.156e-02   5.359 9.32e-08 ***
Age_cl_og2                                                      5.447e-01  4.185e-02  13.015  < 2e-16 ***
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)                 7.385e-02  6.021e-02   1.226 0.220170
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)  1.968e-01  5.487e-02   3.586 0.000344 ***
OG(POR)(HOU)(GSW)(MIL)                                          1.646e-01  7.500e-02   2.195 0.028302 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9086 on 2029 degrees of freedom
Multiple R-squared:  0.5468,    Adjusted R-squared:  0.5435
F-statistic: 163.2 on 15 and 2029 DF,  p-value: < 2.2e-16
```

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

OUTILER

Dal test di Breusch Pagan rifiutiamo la presenza di omoschedasticità con un p-value< 2.2e-16.

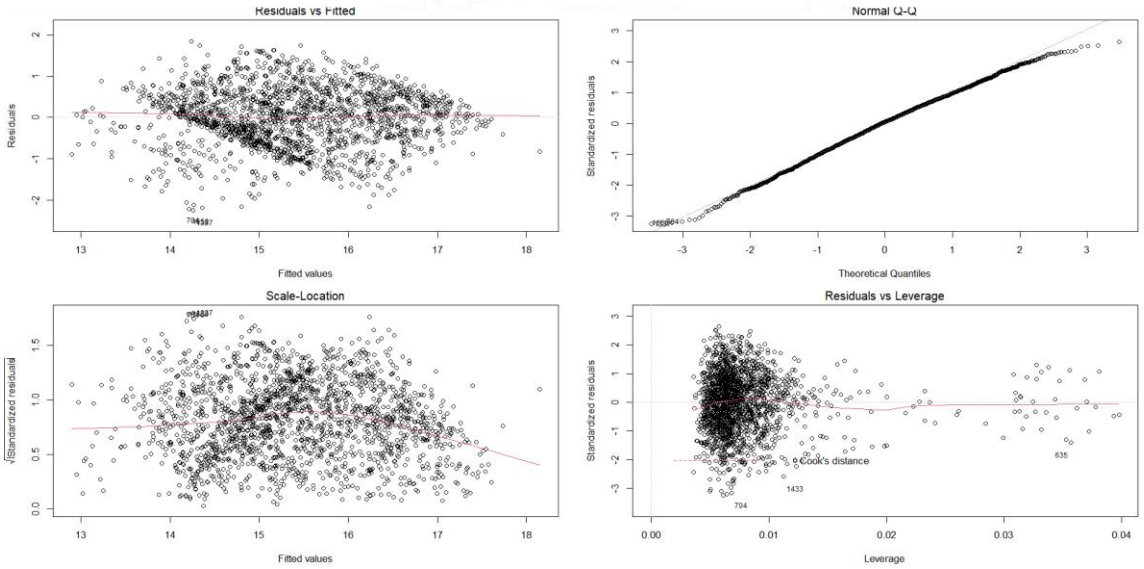


Andiamo a studiare i valori influenti sul modello:

```
> df3[1827,-17]
            Player Pos G GS FG. X3P. X2P. FT. STL BLK PF      Sal    Year      PER TRBW PAS Age_cl_og
1828 Gordon Hayward  SF 2  2  50    0  100   0   1   0  1 17.25639 2017-18 2.012195    1   1         2
> df3[1847, -17]
           Player Pos G GS  FG. X3P. X2P. FT. STL BLK  PF      Sal    Year PER TRBW PAS Age_cl_og
1848 Andre Ingram  SG 3  1 47.1 55.6 37.5 100 2.5 1.5 1.5 11.24159 2017-18   2 3.25 1.8         2
> df3[1979, -17]
            Player Pos G GS  FG. X3P. X2P. FT. STL BLK PF     Sal    Year PER TRBW PAS Age_cl_og
1980 Devin Robinson  SF 2  1 33.3    0 33.3   0   2   0  2 11.2548 2017-18   2    6   1         1
```

```
> muvec
       G       GS       FG.      X3P.      X2P.       FT.       STL       BLK       PF       Sal       PER      TRBW       PAS
51.043521 26.498289 45.766015 30.530122 51.594181 73.127237  1.685966  0.444890  1.804597 15.248035  6.508378  4.406797  1.375554
```

Modello senza outliers:



Nicolò Samuelli 866735
Carlotta Giacchetta 868779

ETEROSCHEDASTICITÀ

```
        studentized Breusch-Pagan test
```

Presenta lieve eteroschedasticità

```
data:  lm_OUT
BP = 31.373, df = 15, p-value = 0.007827
```

Summary del modello con stime degli standard error non corretti:

```
Call:
lm(formula = Sal ~ log(G) + log(GS) + I(FT.^2) + FT. + log(STL) +
    log(PER) + TRBW + Year + Age_cl_og + OG, data = filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30024 -0.44216  0.02576  0.46745  1.79746

Coefficients:
                                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                        1.263e+01  1.382e-01  91.392  < 2e-16 ***
log(G)                                                             1.795e-01  3.519e-02   5.100 3.74e-07 ***
log(GS)                                                           -1.844e-01  3.082e-02  -5.982 2.63e-09 ***
I(FT.^2)                                                           4.328e-06  3.468e-05   0.125 0.900673
FT.                                                                8.628e-03  4.215e-03   2.047 0.040820 *
log(STL)                                                           7.922e-01  8.945e-02   8.856  < 2e-16 ***
log(PER)                                                           6.913e-01  4.718e-02  14.652  < 2e-16 ***
TRBW                                                               4.847e-02  8.363e-03   5.796 7.96e-09 ***
Year2018-19                                                        1.268e-02  5.450e-02   0.233 0.816074
Year2019-20                                                        6.437e-02  5.240e-02   1.228 0.219434
Year2020-21                                                        1.764e-02  5.170e-02   0.341 0.732983
Year2021-22                                                        2.309e-01  4.883e-02   4.729 2.42e-06 ***
Age_cl_og2                                                         5.440e-01  3.270e-02  16.637  < 2e-16 ***
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)                    6.705e-02  4.715e-02   1.422 0.155179
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)  1.510e-01  4.306e-02   3.507 0.000463 ***
OG(POR)(HOU)(GSW)(MIL)                                             1.399e-01  5.900e-02   2.371 0.017860 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.683 on 1884 degrees of freedom
Multiple R-squared:  0.6434,    Adjusted R-squared:  0.6406
F-statistic: 226.6 on 15 and 1884 DF,  p-value: < 2.2e-16
```

Coefftest → stime degli standard error e inferenza corretti

```
t test of coefficients:

                                                                   Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                                                      1.2634e+01  1.2261e-01 103.0405 < 2.2e-16 ***
log(G)                                                           1.7948e-01  3.7670e-02   4.7645 2.038e-06 ***
log(GS)                                                         -1.8439e-01  3.1822e-02  -5.7943 8.025e-09 ***
I(FT.^2)                                                         4.3284e-06  3.3326e-05   0.1299  0.896676
FT.                                                              8.6279e-03  3.8914e-03   2.2171  0.026733 *
log(STL)                                                         7.9217e-01  8.6745e-02   9.1321 < 2.2e-16 ***
log(PER)                                                         6.9128e-01  4.7293e-02  14.6169 < 2.2e-16 ***
TRBW                                                             4.8472e-02  8.0567e-03   6.0164 2.138e-09 ***
Year2018-19                                                      1.2677e-02  5.4384e-02   0.2331  0.815704
Year2019-20                                                      6.4374e-02  5.0753e-02   1.2684  0.204820
Year2020-21                                                      1.7640e-02  5.2214e-02   0.3378  0.735521
Year2021-22                                                      2.3090e-01  4.7357e-02   4.8758 1.175e-06 ***
Age_cl_og2                                                       5.4396e-01  3.3602e-02  16.1883 < 2.2e-16 ***
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)                  6.7053e-02  4.7140e-02   1.4224  0.155064
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)  1.5101e-01  4.4255e-02   3.4122  0.000658 ***
OG(POR)(HOU)(GSW)(MIL)                                           1.3987e-01  5.7944e-02   2.4138  0.015882 *
```

Osserviamo che FT: non è molto significativa, quindi proviamo a toglierla e osserviamo cosa cambia nel modello:

1) Anova

```
Model 1: Sal ~ log(G) + log(GS) + I(FT.^2) + FT. + log(STL) + log(PER) +
    TRBW + Year + Age_cl_og + OG
Model 2: Sal ~ log(G) + log(GS) + log(STL) + log(PER) + TRBW + Year +
    Age_cl_og + OG
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1884 878.99
2   1886 906.24 -2   -27.253 29.207 3.224e-13 ***
```

2) Bptest

```
        studentized Breusch-Pagan test

data:  lm_OUT_FT
BP = 26.325, df = 13, p-value = 0.01537
```

Nonostante il test Anova affermi che vi siano differenze significative tra i due modelli e che il modello con FT. abbia devianza residua minore, abbiamo deciso di eliminarla in quanto, al netto di una differenza nelle devianze residue non eccessivamente alta, guadagniamo in riduzione dell’eteroschedasticità e semplicità del modello.

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

Coeftest per il modello senza FT.

```
t test of coefficients:

                                                    Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                                       12.9631866  0.1218395 106.3956 < 2.2e-16 ***
log(G)                                             0.2668606  0.0344014   7.7573 1.410e-14 ***
log(GS)                                           -0.2032427  0.0318245  -6.3864 2.134e-10 ***
log(STL)                                           0.8742098  0.0873032  10.0135 < 2.2e-16 ***
log(PER)                                           0.7279310  0.0459023  15.8583 < 2.2e-16 ***
TRBW                                               0.0350190  0.0075805   4.6196 4.104e-06 ***
Year2018-19                                        0.0171177  0.0545707   0.3137  0.753799
Year2019-20                                        0.0797432  0.0507249   1.5721  0.116102
Year2020-21                                        0.0424719  0.0525122   0.8088  0.418732
Year2021-22                                        0.2543031  0.0475744   5.3454 1.012e-07 ***
Age_cl_og2                                         0.5738539  0.0337238  17.0163 < 2.2e-16 ***
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)    0.0638763  0.0480747   1.3287  0.184112
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)  0.1419649  0.0450035   3.1545  0.001633 **
OG(POR)(HOU)(GSW)(MIL)                             0.1385272  0.0585993   2.3640  0.018181 *
```

Tutte le variabili sono significative tranne le differenze tra i vari anni e l'anno 2017-2018, proviamo a togliere anno e vediamo come cambia il modello:

1) Anova

```
Model 1: Sal ~ log(G) + log(GS) + log(STL) + log(PER) + TRBW + Age_cl_og +
    OG
Model 2: Sal ~ log(G) + log(GS) + log(STL) + log(PER) + TRBW + Year +
    Age_cl_og + OG
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1890 924.88
2   1886 906.24  4    18.634 9.6951 9.234e-08 ***
```

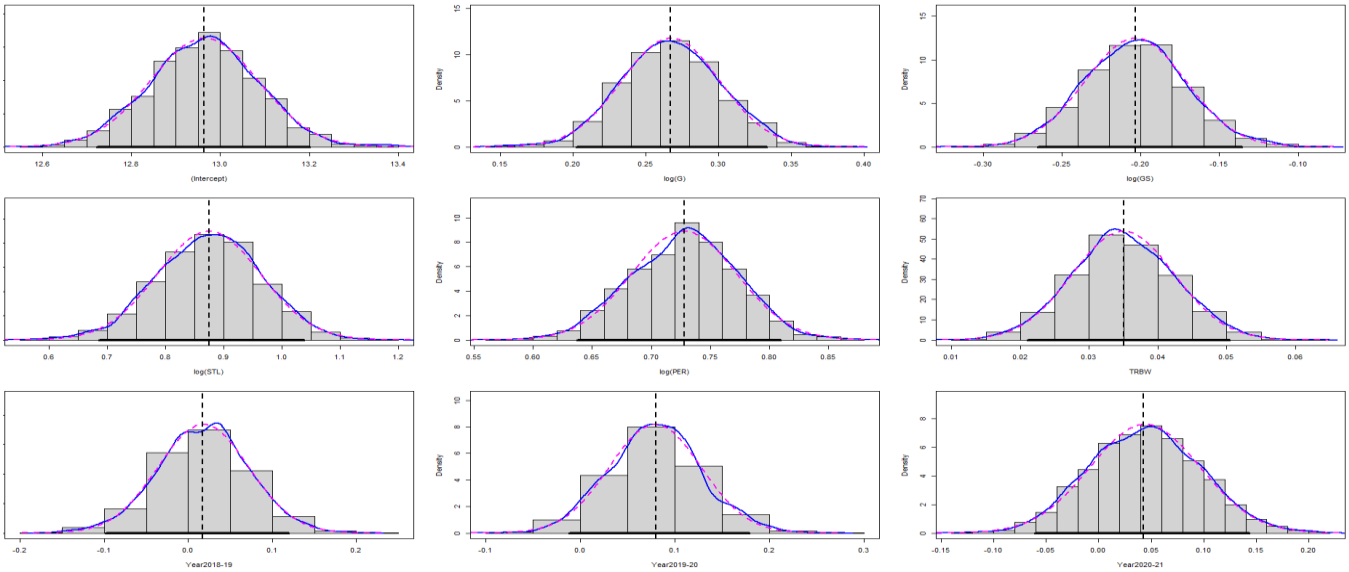
2) Bptest

```
        studentized Breusch-Pagan test

data:  lm_OUT_FT_YR
BP = 21.469, df = 9, p-value = 0.01072
```

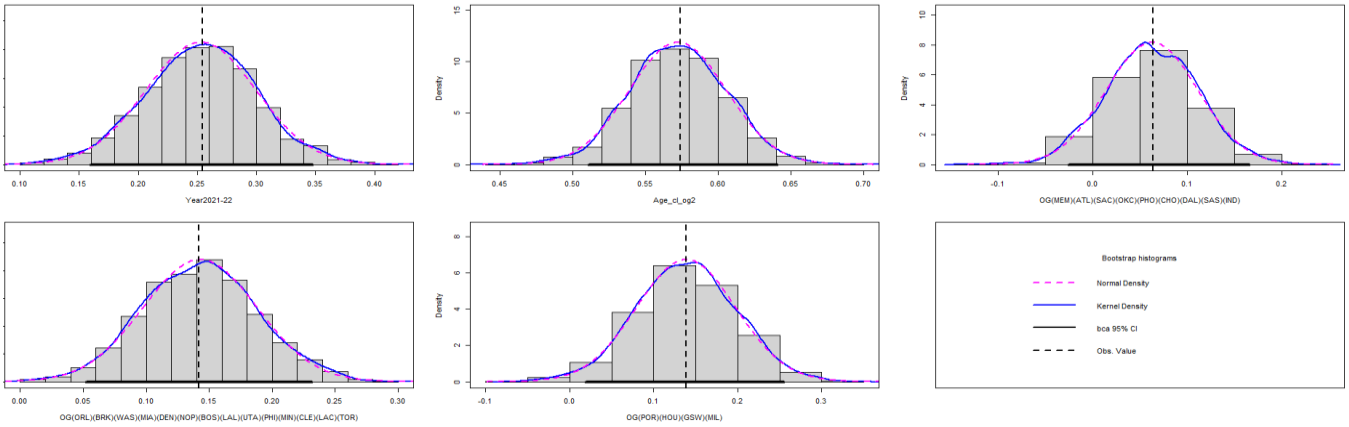
Deidiamo di tenere anno.

```
                    Value   p-value                  Decision
Global Stat       18.9507 0.0008037 Assumptions NOT satisfied!
Skewness          12.2942 0.0004544 Assumptions NOT satisfied!
Kurtosis           0.5903 0.4423155    Assumptions acceptable.
Link Function      5.1873 0.0227524 Assumptions NOT satisfied!
Heteroscedasticity 0.8789 0.3485071    Assumptions acceptable.
```

BOOTSTRAP

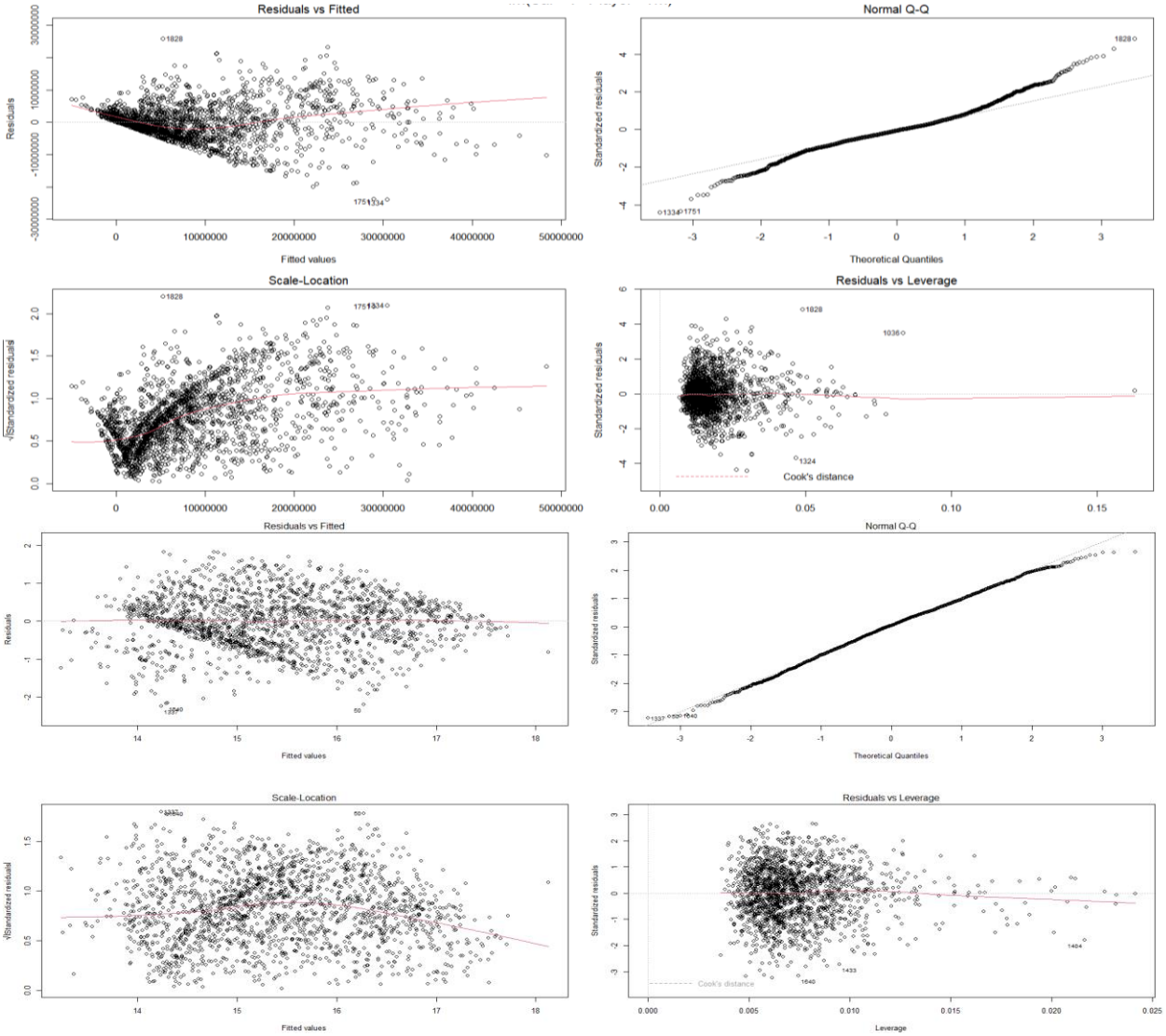


Nicolò Samuelli 866735
Carlotta Giacchetta 868779



Ad eccezione di un livello della variabile OG ed ai livelli delle variabili YEAR tutte le variabili sono significative in quanto il loro intervallo di confidenza non include il valore nullo.

CONFRONTO TRA MODELLO INIZIALE E MODELLO FINALE



Nicolò Samuelli 866735
Carlotta Giacchetta 868779

Come confermato anche dal test globale sulle assunzioni del modello

	Value	p-value	Decision
Global Stat	399.425	0.000e+00	Assumptions NOT satisfied!
Skewness	28.203	1.092e-07	Assumptions NOT satisfied!
Kurtosis	296.923	0.000e+00	Assumptions NOT satisfied!
Link Function	73.121	0.000e+00	Assumptions NOT satisfied!
Heteroscedasticity	1.177	2.779e-01	Assumptions acceptable.

	Value	p-value	Decision
Global Stat	18.9507	0.0008037	Assumptions NOT satisfied!
Skewness	12.2942	0.0004544	Assumptions NOT satisfied!
Kurtosis	0.5903	0.4423155	Assumptions acceptable.
Link Function	5.1873	0.0227524	Assumptions NOT satisfied!
Heteroscedasticity	0.8789	0.3485071	Assumptions acceptable.

Confronto stime parametri nel modello iniziale e nel modello finale

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.290e+07	1.476e+06	-8.742	< 2e-16	***
PosPF	3.671e+05	4.712e+05	0.779	0.436014	
PosPG	-7.504e+05	6.459e+05	-1.162	0.245496	
PosSF	2.604e+05	5.721e+05	0.455	0.649083	
PosSG	-6.276e+05	5.946e+05	-1.056	0.291265	
Age	4.791e+05	3.270e+04	14.651	< 2e-16	***
G	-2.781e+04	7.660e+03	-3.631	0.000289	***
GS	-7.647e+04	1.747e+04	-4.378	1.26e-05	***
MP	2.131e+04	5.824e+04	0.366	0.714431	
FG	-1.609e+06	3.845e+06	-0.419	0.675578	
FGA	6.329e+06	2.575e+06	2.458	0.014067	*
FG.	2.299e+05	7.561e+04	3.041	0.002391	**
X3P	-8.027e+06	3.479e+06	-2.307	0.021134	*
X3PA	-4.943e+06	2.578e+06	-1.918	0.055315	.
X3P.	-1.138e+04	1.421e+04	-0.801	0.423405	
X2P	-4.822e+06	2.727e+06	-1.769	0.077120	.
X2PA	-5.911e+06	2.593e+06	-2.280	0.022733	*
X2P.	8.175e+02	2.425e+04	0.034	0.973107	
eFG.	-2.044e+05	7.234e+04	-2.826	0.004760	**
FT	-2.010e+06	1.770e+06	-1.136	0.256284	
FTA	-1.717e+05	5.614e+05	-0.306	0.759779	
FT.	-9.680e+03	9.341e+03	-1.036	0.300225	
STL	7.022e+05	4.839e+05	1.451	0.146950	
BLK	1.143e+06	4.541e+05	2.517	0.011920	*
PF	-8.813e+05	2.883e+05	-3.058	0.002261	**
PTS	3.152e+06	1.671e+06	1.886	0.059483	.
Year2018-19	-3.669e+04	4.232e+05	-0.087	0.930924	
Year2019-20	9.283e+03	4.122e+05	0.023	0.982034	
Year2020-21	1.102e+05	4.116e+05	0.268	0.788996	
Year2021-22	5.511e+05	3.927e+05	1.403	0.160677	
PER	8.253e+05	9.073e+04	9.096	< 2e-16	***
TRBW	8.417e+04	1.157e+05	0.728	0.466966	
PAS	1.409e+06	4.625e+05	3.046	0.002349	**
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)	3.889e+05	3.666e+05	1.061	0.288889	
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)	1.266e+06	3.341e+05	3.788	0.000156	***
OG(POR)(HOU)(GSW)(MIL)	1.294e+06	4.640e+05	2.789	0.005344	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5505000 on 2009 degrees of freedom
Multiple R-squared: 0.6831, Adjusted R-squared: 0.6776

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.963187	0.120374	107.691	< 2e-16	***
log(G)	0.266861	0.032247	8.275	2.39e-16	***
log(GS)	-0.203243	0.030847	-6.589	5.74e-11	***
log(STL)	0.874210	0.090007	9.713	< 2e-16	***
log(PER)	0.727931	0.045864	15.872	< 2e-16	***
TRBW	0.035019	0.007867	4.451	9.04e-06	***
Year2018-19	0.017118	0.055295	0.310	0.75692	
Year2019-20	0.079743	0.053141	1.501	0.13362	
Year2020-21	0.042472	0.052360	0.811	0.41738	
Year2021-22	0.254303	0.049428	5.145	2.95e-07	***
Age_cl_og2	0.573854	0.032824	17.483	< 2e-16	***
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)	0.063876	0.047847	1.335	0.18203	
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)	0.141965	0.043677	3.250	0.00117	**
OG(POR)(HOU)(GSW)(MIL)	0.138527	0.059875	2.314	0.02080	*

Nicolò Samuelli 866735
Carlotta Giacchetta 868779

Interpretazione dei coefficienti delle variabili categoriali:

```
> (exp(lm_OUT_FT$coefficients[7:14])-1)*100
                                  Year2018-19                                    Year2019-20
                                     1.726504                                       8.300895
                                  Year2020-21                                    Year2021-22
                                     4.338671                                      28.956260
                                   Age_cl_og2        OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)
                                    77.509496                                       6.596053
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)     OG(POR)(HOU)(GSW)(MIL)
                                    15.253622                                      14.858088
```

MODELLO LOGISTICO

Abbiamo deciso di prendere in esame la variabile target rendendola fattoriale come segue:

- 0 = giocatori che guadagnano meno di 30000000 $
- 1 = giocatori che guadagnano almeno 30000000 $

```
glm(formula = Sal ~ log(G) + log(STL) + log(PER) + TRBW + Year +
    Age_cl + OG, family = "binomial", data = df4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7889  -0.1079  -0.0086  -0.0018   3.9714

Coefficients:
                                                                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                              -4.96215    0.88488  -5.608 2.05e-08 ***
log(G)                                                                   -3.34124    0.37631  -8.879  < 2e-16 ***
log(STL)                                                                  2.38989    0.67934   3.518 0.000435 ***
log(PER)                                                                  5.01253    0.55028   9.109  < 2e-16 ***
TRBW                                                                     -0.14595    0.04793  -3.045 0.002327 **
Year2018-19                                                               1.07745    0.49961   2.157 0.031038 *
Year2019-20                                                               1.43985    0.50423   2.856 0.004297 **
Year2020-21                                                               1.89004    0.48224   3.919 8.88e-05 ***
Year2021-22                                                               2.08965    0.45856   4.557 5.19e-06 ***
Age_cl2                                                                   2.00829    0.31167   6.444 1.17e-10 ***
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND)                          -0.19767    0.48172  -0.410 0.681563
OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)  0.70621    0.42402   1.665 0.095813 .
OG(POR)(HOU)(GSW)(MIL)                                                    0.94602    0.48064   1.968 0.049040 *
```

```
         predicted
observed    0    1
       0 1901   19
       1   62   63
> accuracy
[1] 0.9603912
```

Il modello logistico fittato come riportato nel summary presenta coefficienti significativamente diversi da zero per tutte le variabili tranne un livello di OG. L'accuracy del modello (casi classificati correttamente / osservazioni totali) risulta essere pari al 96%.

Per interpretare i coefficienti del modello logistico riportiamo il loro esponenziale che in indica l'odds ratio.

```
exp(lm_logit$coefficients)
                                    (Intercept)                                                                   log(G)
                                    0.006997881                                                              0.035392886
                                       log(STL)                                                                 log(PER)
                                   10.912321743                                                            150.284052090
                                           TRBW                                                              Year2018-19
                                    0.864201541                                                              2.937183091
                                    Year2019-20                                                              Year2020-21
                                    4.220077562                                                              6.619643626
                                    Year2021-22                                                                  Age_cl2
                                    8.082086052                                                              7.450542600
OG(MEM)(ATL)(SAC)(OKC)(PHO)(CHO)(DAL)(SAS)(IND) OG(ORL)(BRK)(WAS)(MIA)(DEN)(NOP)(BOS)(LAL)(UTA)(PHI)(MIN)(CLE)(LAC)(TOR)
                                    0.820643763                                                              2.026292825
                         OG(POR)(HOU)(GSW)(MIL)
                                    2.575444960
```