INTRODUCTION

The dataset I'm going to study contains statistics that NBA players produced during the 2021-2022 season, furthermore we can find anagraphical data and, most important, their salaries. I personally created the dataset by gathering information from the official NBA.com website for statistics and from Hoophype.com for salaries. The result includes 415 observations, equivalent to the number of players, and 16 variables. As they are game statistics, there are 4 integer variables and 12 numerical variables. In particular, we can observe: Age = Age of the player; Gp = Game played; W = Wins; TS.PCT = True shooting percentage, this is a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws; USG.PCT = Usage percentage, this is an estimate of the percentage of team plays used by a player while he was on the floor; POSS = Possession, which mean how many posseses does the player plays; X3PTS.PCT = 3 points percentage; FG.PCT = Field goal percentage; PTS = Points ; REB = Rebounds; AST = Assists; TOV = Tournovers; STL = Steal; BLK = Blocks; EFF = Efficiency; Salary (million of USD).

GOAL

The main objective of this project is to study and analyze how the statistics produced by players and their anagraphical information, affect their salary. However, it is important to remember that the analysis I am about to make does not concern an exact science, there is no direct cause and effect relationship between salary and performance, as there are dozens and dozens of additional variables to take into consideration. An example is the player's popularity, this is a statistic that cannot be estimated and will not be taken into account.

USE SUITABLE REPRESENTATION TO GRAPH THE DATA AND COMMMENT THE RELATION AMONG ALL VARIABLES

```
ggplot(NBAstats, aes(x = SALARY)) +
  geom_histogram(aes(fill = after_stat(count)), binwidth = 0.5) +
  scale_fill_gradient(low = "orangered", high = "orange") +
  theme_minimal() +
  labs( x = "Salary", y = "Frequency")
```

in figure 1 we can observe the frequency distribution of salary as we can see, there is a strong concentration of data at the beginning of the graph. this can be justified by the fact that the NBA has many more players under contract than those who actually have a significant impact on games. Suffice it to say that only 7/8 players are actually used for each team in the playoff games (final stage of the season), which leads to a total of about 200/240 players in the entire league actually useful in order to win the championship. This means that around 50% of the athletes are supporting actors or extras, who are contracted at the minimum wage (1.52 million in the 2021-2022 season) for purely bureaucratic reasons.

```
summary_boxplots <- data.frame(
  Minimum = c(1.52,-2,0,0),
  Maximum = c(45.78,38.8,64,30.3),
  Median = c(4.74,9.9,27,8.2),
  "First quartile" = c(2.265,6,17,4.8),
  "third Quartile" = c(12.015,14.5,38.5,13.05),
  row.names = c("Salary", "Efficency", "Wins", "Points")
```

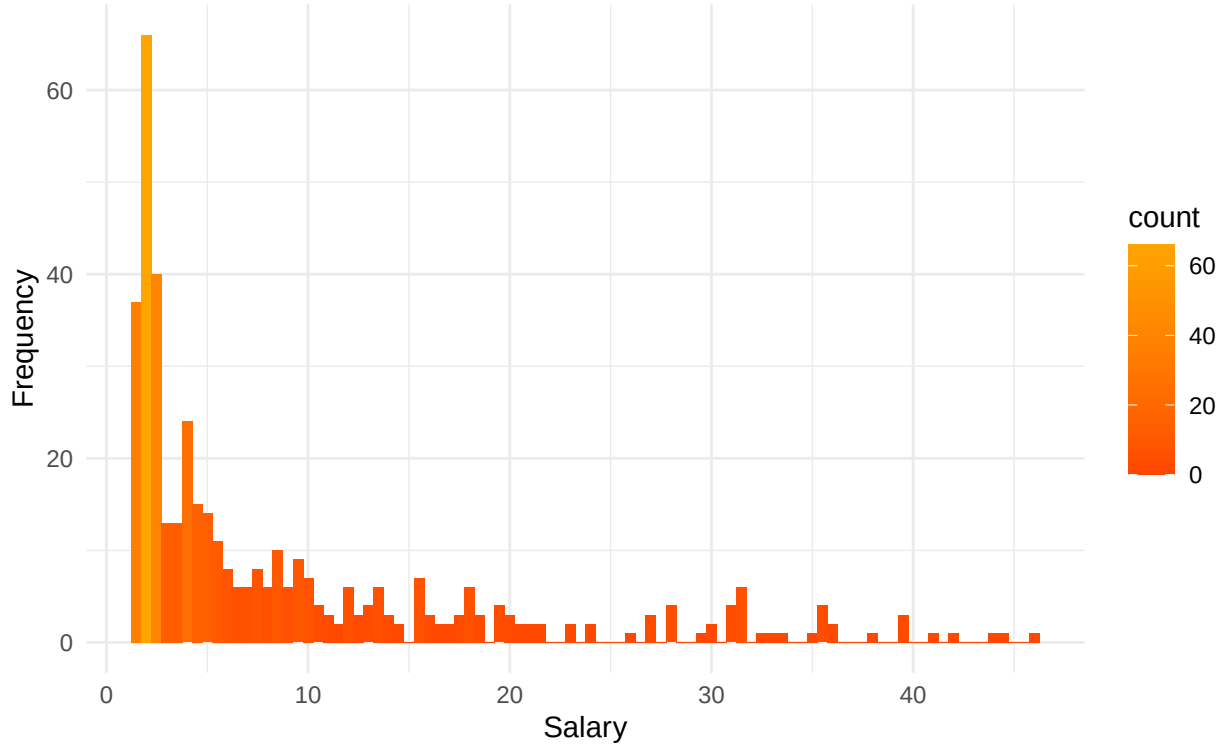


Figure 1: Histogram based on salary frequency

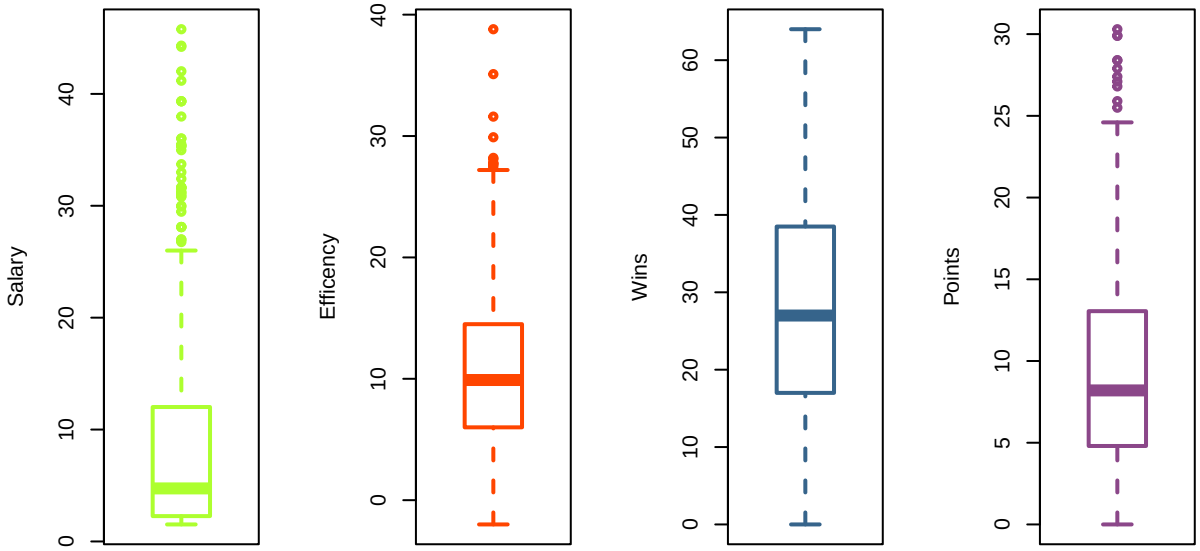


Figure 2: Boxplots of Salary, efficency, wins and points

```
)
kable(summary_boxplots)
```

	Minimum	Maximum	Median	First.quartile	third.Quartile
Salary	1.52	45.78	4.74	2.265	12.015
Efficency	-2.00	38.80	9.90	6.000	14.500
Wins	0.00	64.00	27.00	17.000	38.500
Points	0.00	30.30	8.20	4.800	13.050

The boxplots presented here allow us to observe the individual variables and their distribution among the players in the best way possible. In particular, thanks to them, it is possible to observe the median, the first and the third quartile, the minimum and maximum values (shown in the table above). We can say, for example, that the salary one is no symmetric since the median is positioned down in the boxplot and that both Salary, Efficency and Points present many extreme values.

FITTING THE MODEL

```
all<- lm(SALARY ~ ., data=NBAstats)
```

the ols presented above shows the following multiple linear model $\hat{y_i} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$ for every i that goes from 1 to p. β_0 = is the intercept, that describe the value of $\hat{y_i}$ when every predictor is equal to 0, β_i = describes how the $\hat{y_i}$ varies for an increase of 1 unit of the x while keeping every other β constant. Residual standard error: Indicates the standard deviation of the residuals of the model, or how far the observed values are from the values estimated by the model. R^2: Represents the percentage of variance in the dependent variable that can be explained by the independent variables in the model. $adjR^2$: Indicates the R-squared corrected for the number of independent variables in the model. F-statistic: Indicates the F-value for the model, which is a measure of the overall significance of the model. A high F-value (and a low p-value) indicates that at least one of the independent variables is significant in the model.

```
ols<-regsubsets(SALARY ~ ., data=NBAstats,nvmax = 15)
summ = summary(ols)
```

the regsubset function compute the best subset selection that consist in choosing the best model among the $\binom{p}{k}$ possible models. Here best is defined as having the smallest RSS or largest R^2.

CRITERIA

While the bic, the Cp mallow and the adjusted R^2 can be easily found in the summary, we need to compute manually the aic, we can do this by extrapolating it from the bic formula.

```
#aic
p=15
n=nrow(NBAstats)
aic=matrix(NA,p,1)
for(j in 1:p){
  aic[j]= summ$bic[j] - (j+2)*log(n)+2*(j+2)
}
```

we need the model that maximizes the adjR^2 and minimizes all the other criteria, moreover, by computing the cross validation (using the k folder system) we can observe which model has the lowest MSE.

```
par(pty="s",mfrow=c(1,5),mar=c(2,1,2,1))
# BIC
plot(summ$bic, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Drop in BIC")
abline (v=which.min(summ$bic),col = 2, lty=2)
# Cp
plot(summ$cp, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Mallow' Cp")
abline (v=which.min(summ$cp),col = 2, lty=2)
#R2
plot(summ$adjr2, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Adjusted R^2")
abline (v=which.max(summ$adjr2),col = 2, lty=2)
#Aic
plot(aic, type="b", pch=19, xlab="Number of predictors", ylab="", main="Drop in AIC")
abline (v=which.min(aic),col = 2, lty=2)
#Cv
p <- 15
k <- 10
set.seed (1)
folds <- sample (1:k,nrow(NBAstats),replace =TRUE)
cv.errors <- matrix (NA ,k, p, dimnames =list(NULL , paste (1:p) ))
for(j in 1:k){
  best.fit =regsubsets (SALARY ~ ., data=NBAstats[folds!=j,],nvmax = 15)
  for(i in 1:p) {
    mat <- model.matrix(as.formula(best.fit$call[[2]]), NBAstats[folds==j,])
    coefi <- coef(best.fit ,id = i)
    xvars <- names(coefi )
    pred <- mat[,xvars ]%*% coefi
    cv.errors[j,i] <- mean((NBAstats$SALARY[folds==j] - pred)^2)
  }
}
cv.mean <- colMeans(cv.errors)
plot(cv.mean ,type="b",pch=19,xlab="Number of predictors",ylab="CV error", main = "Cv")
abline(v=which.min(cv.mean), col=2, lty=2)
```

Thanks to figure 3, we can notice how the best model is the one that has 7 covariates. In fact, in the case of Cp mallow, AIC, and $adjR^2$, we can see that after the seventh predictor there is a plateau. It's important to take the first value before the plateau because, according to the occam's razor principle, using multiple predictors can lead to various problems as well as a high consumption of time and money. The cross-validation also confirms this hypothesis.

AGE	GP	W	TS.PCT	USG.PCT	POSS	X3PTS.PCT	FG.PCT	PTS	REB	AST	TOV	STL	BLK	EFF
*	*	*		*	*	*		*						

According to the best subset selection, the best model with 7 predictors is the following: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$ * AGE + $\hat{\beta}_2$ * W + $\hat{\beta}_3$ * GP + $\hat{\beta}_4$ * USAGE PERCENTAGE + $\hat{\beta}_5$ * 3 POINT PERCENTAGE + $\hat{\beta}_6$ * POSSESSIONS + $\hat{\beta}_7$ * POINTS + ε

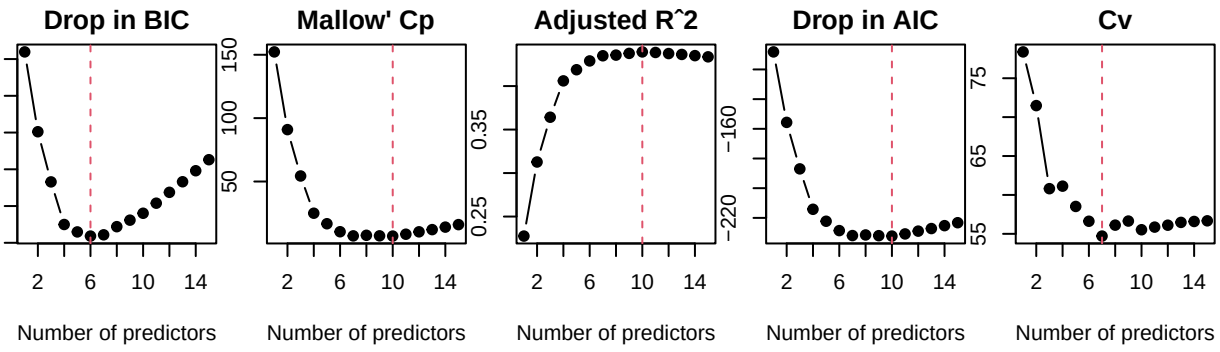


Figure 3: variable selection 1

```
model = lm(SALARY ~ AGE + W + GP + USG.PCT + X3PTS.PCT + POSS + PTS, data = NBAstats)
summ1=summary(model)
```

CHECKING FOR COLLINEARITY

```
correlation_matrix <- cor(NBAstats[,c(1,2,3,5,6,7,9,16)])
corrplot(correlation_matrix, method = 'ellipse')
```

```
correlazione = data.frame(GP = c(0.80566341,0.84734630),row.names = c("W","POSS")
)
kable(correlazione)
```

	GP
W	0.8056634
POSS	0.8473463

```
vif(model)
```

```
##       AGE         W        GP   USG.PCT X3PTS.PCT      POSS       PTS
##  1.055655  2.956888  6.024797  1.503694  1.147940  5.185181  1.580242
```

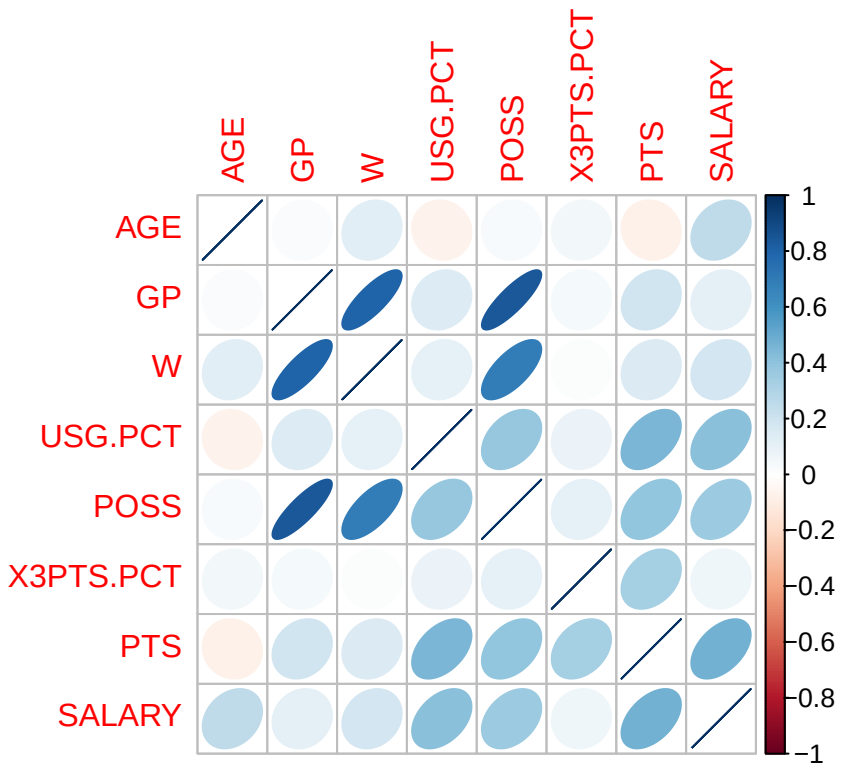


Figure 4: collinearity plot

```
VIF = vif(model)
plot(VIF, pch = 16, ylim = c(0,10),
ylab = "Vif values", main = "Variance Inflation plot")
abline( h = 2.5, col = 'grey', lty = 2, lwd = 2)
abline( h = 5, col = 'blue', lty = 2, lwd = 2)
abline( h = 10, col = 'red', lty = 2, lwd = 2)
text(x = VIF, labels = names(VIF), cex = 0.8, pos = 3)
legend(x = "topright", legend = c("2.5", "5", "10"),
lty = 2, lwd = 2, col = c('grey','blue','red') )
```

The correlation table shows all possible parameter pairs and returns how related they are. in my case we can see how there is a strong correlation between games played and victories and between games played and possessions. to analyze multicollinearity in a linear regression system, it is necessary to analyze the vif. in the case of vif, values above the threshold of 5 or 10 could lead to problems, to solve there are two possible solutions: either directly eliminate the value in question, or carry out an interaction between values, in my case however, as we can see in figure 5, the variables possessions and games played slightly exceed the threshold of 5. However both by carrying out an interaction and by eliminating one of the two variables, the resulting model is no better than the original one, so I will keep it unchanged.

REGRESSION DIAGNOSTICS

```
par(mfrow = c(1,2))
plot(fitted(model), residuals(model),
```

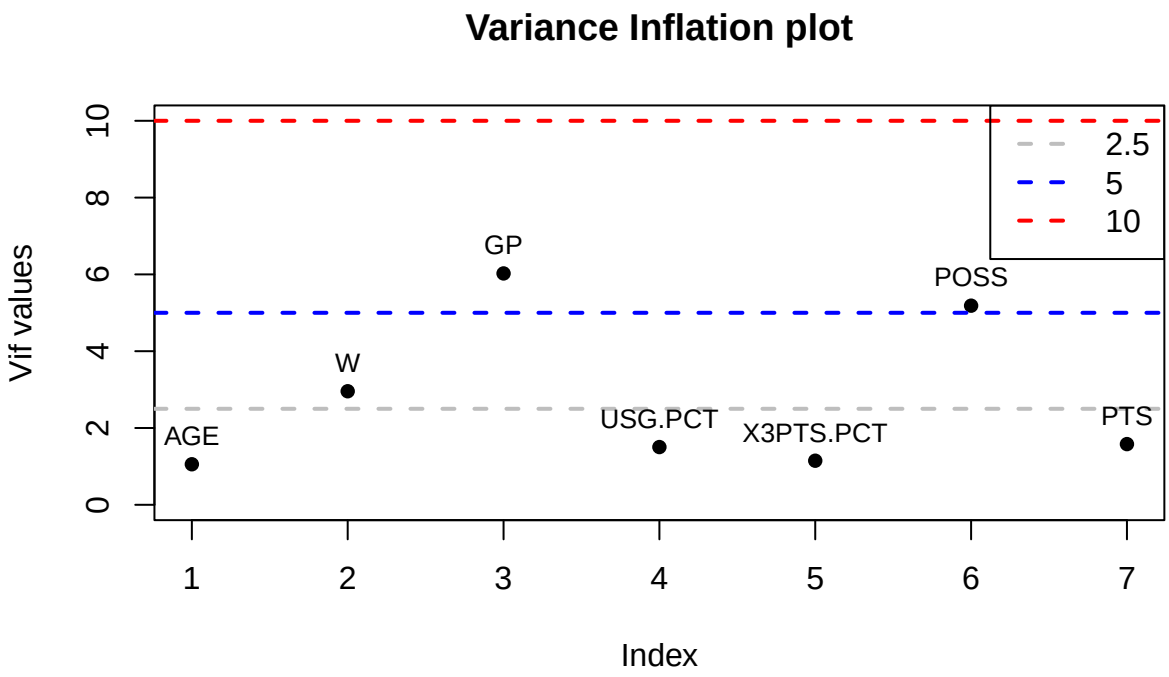


Figure 5: vif plot

```
xlab="Fitted values", ylab="Residuals",main = "residual plot", pch=19, cex=0.8,col = "indianred")
abline(h = 0, col="darkorchid4", lwd = 2)
qqnorm (residuals (model), ylab="Residuals", col = "indianred")
qqline (residuals (model), col = "darkorchid4", lwd = 2)
```

figure 6 represents the residual plot and the normal plot. Residuals are the differences between the observed values and the values predicted by the model. The residuals are plotted on the y-axis and the predicted values by the model are plotted on the x-axis. Each point represents an observation from the dataset, and the overall shape of the points can indicate the presence or absence of heteroscedasticity (i.e. if the variance of the residuals is constant throughout the range of predicted values) or non-linear relationships between the variables. The horizontal line represents the value of 0 and indicates the ideal situation where the residuals are symmetrically distributed around zero. The purple line represents the reference line for residuals equal to zero. If the points are uniformly distributed around the purple line and do not show any clear pattern, then it can be presumed that the model is adequate for explaining the data. Otherwise, there may be issues. in my case, we can see that there is a cluster in the initial part of the graph, this is a clear example of non costant variance, in particular an example of right opening megaphone. it will therefore be necessary to adjust the response variable. Talking about the second graph in figure 6, we can say that one of the properties of the residuals is that they follow a normal distribution with mean 0 and variance sigma squared, therefore looking at the graph we should be able to confirm this thesis. in this case however we can observe how the errors are not aligned on the line of the normal, we can instead observe a long tailed distribution, therefore an intervention will be necessary to solve this problem.

```
shapiro.test(residuals(model))
```

```
##
```

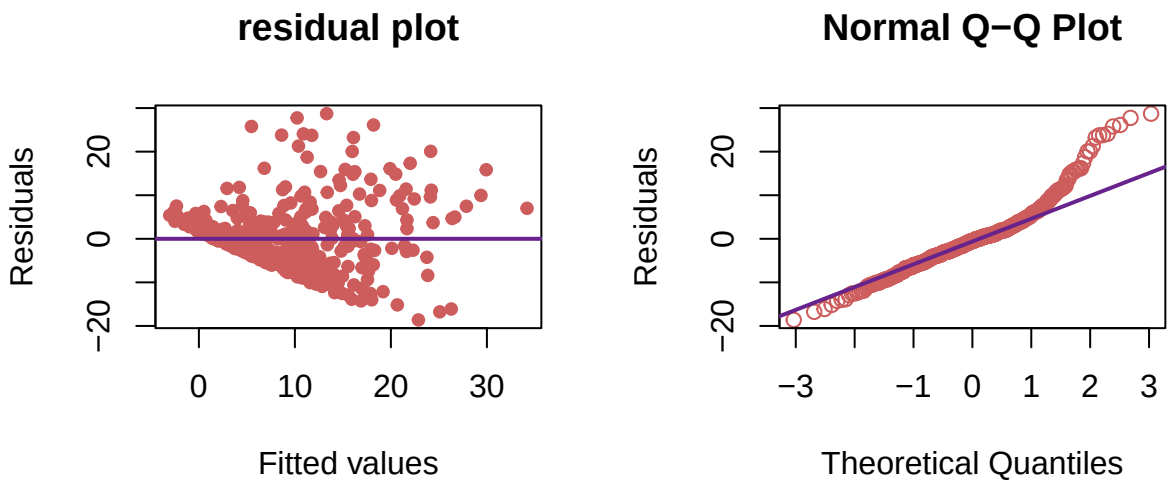


Figure 6: residual plots

```
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.94048, p-value = 7.712e-12
```

the shapiro test perform the following test hypothesis:

$$y_i = \begin{cases} H_0 & \text{residuals are normal distribuited} \\ H_1 & \text{residuals are not} \end{cases}$$

Since the p-value is very small, we reject the null hypotesis. To solve that we try to use the log of the response variable

```
model1 = lm(log(SALARY) ~ AGE + W + GP + USG.PCT + X3PTS.PCT + POSS + PTS, data = NBAstats)
summ2 = summary(model1)
par(mfrow = c(1,2))
plot(fitted(model1), residuals(model1), xlab="Fitted values", ylab="Residuals", main = "residual plot",
abline(h = 0, col="darkorchid4", lwd = 2)
qqnorm (residuals (model1), ylab="Residuals", col = "indianred")
qqline (residuals (model1), col ="darkorchid4", lwd = 2)
```

```
shapiro.test(residuals(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.99587, p-value = 0.3498
```

by running again the shapiro test, we can see how the normality problem is solved, indeed we no longer have the long tailed distribution. as far as it concerns the heteroschedasticity problem, we got a new graph

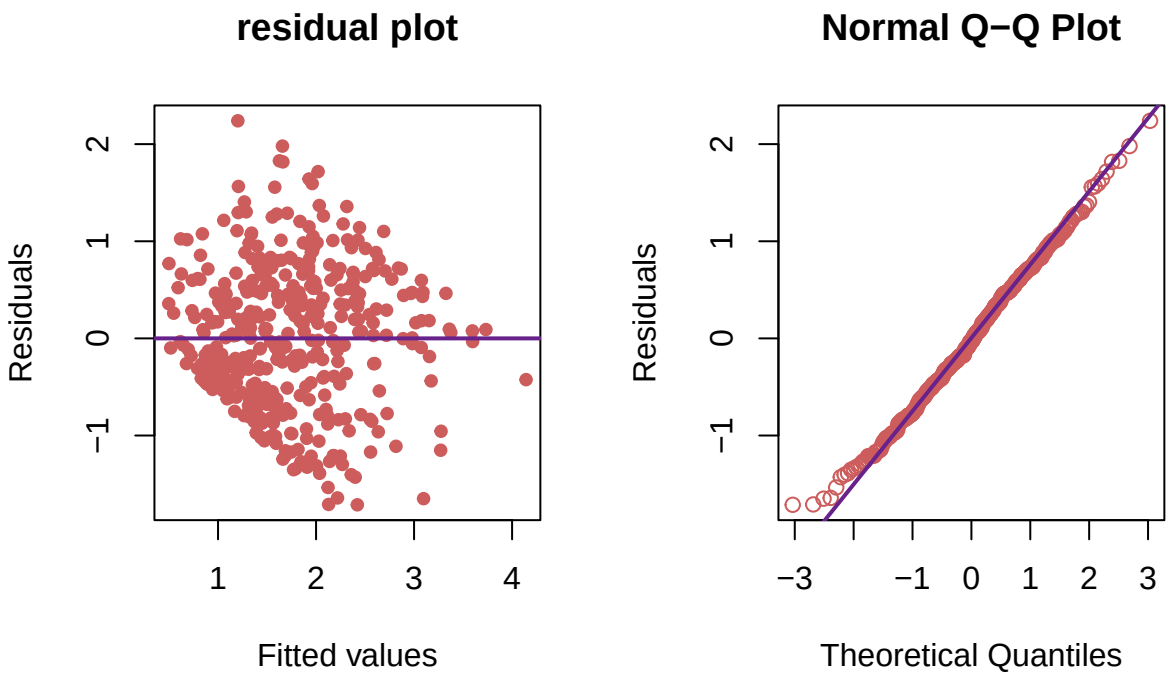


Figure 7: residual plots

that shows a double otward bows. It's possible to solve this this problem in different ways: using another transformation of the response variable, using the wls or using alternative regression models. I tried to use different transformations of the response variable, however, I got a worsening both from the point of view of heteroskedasticity and of the errors graph, therefore considering the tools I currently have available, I will use the above model. Let's now compare the two models by lookinng at the R^2

```
kable(data.frame(R_squared1=summ1$r.squared,
R_squared2=summ2$r.squared))
```

R_squared1	R_squared2
0.444393	0.4501274

```
ols_log<-regsubsets(I(log(SALARY)) ~ ., data=NBAstats,nvmax = 15)
summ_log = summary(ols_log)
```

as we can see, the second model present an higher R^2, but since we changed the response variable, we need to study once again the subset selection

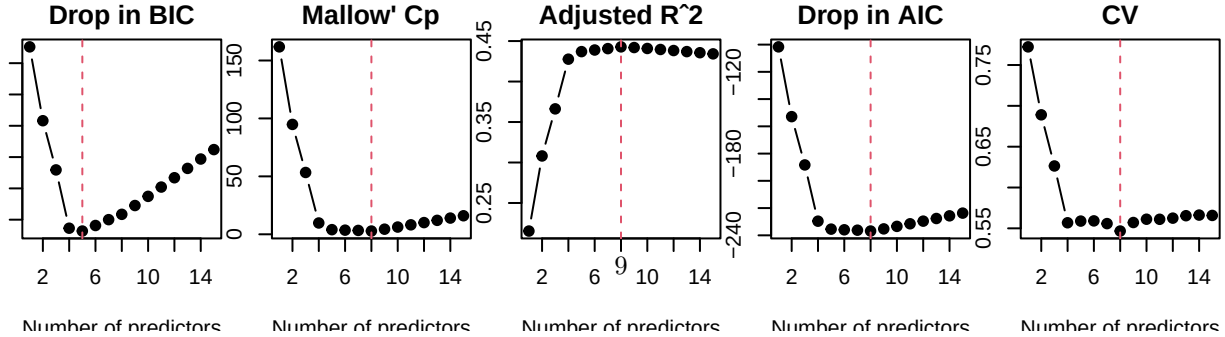


Figure 8: variable

```
## Analysis of Variance Table
##
## Model 1: log(SALARY) ~ AGE + W + GP + USG.PCT + X3PTS.PCT + POSS + PTS
## Model 2: log(SALARY) ~ AGE + GP + USG.PCT + POSS + EFF
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    407 214.59
## 2    409 217.07 -2   -2.4822 2.3539 0.09629 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova produces the following test of hypothesis: $y_i = \begin{cases} H_0 & \text{models can be considered equal} \\ H_1 & \text{models can't be considered equal} \end{cases}$

As we can see, the 2 models are considered to be equal by the anova test, however, for the Occam's razor principle, a model with less predictors is preferible. Let's now look for the high leverage points, outliers and influential points

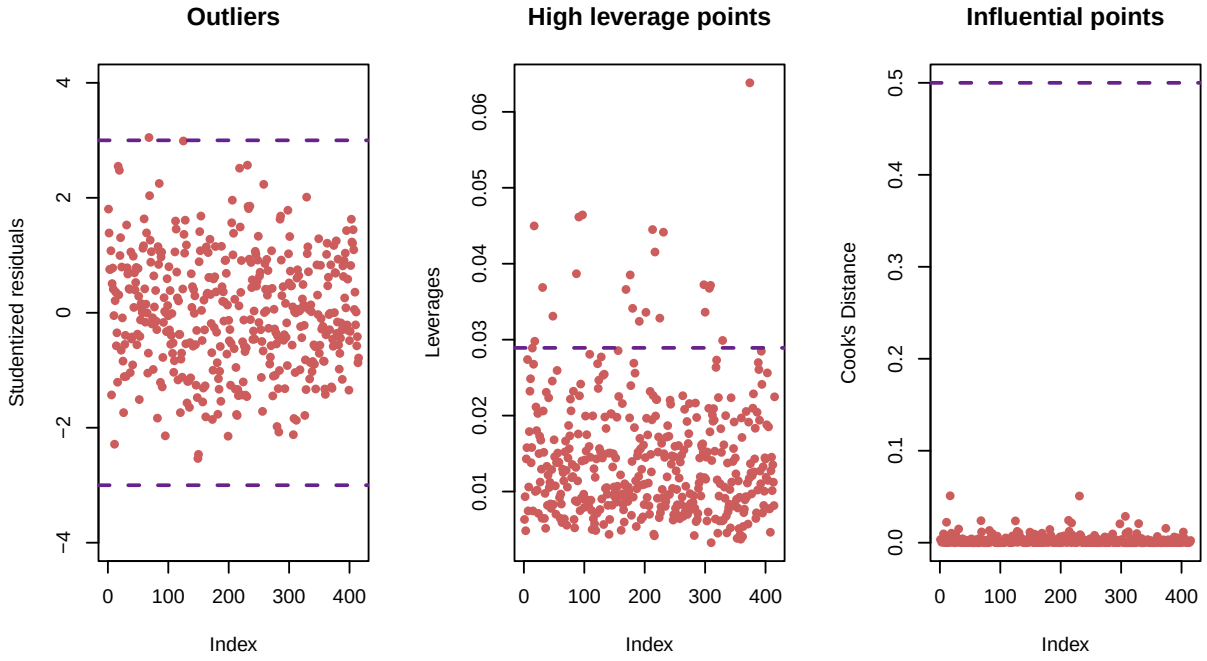


Figure 9: high leverage points, outliers and influential points

outliers are values that deviate significantly from the distribution of observations: $r_i = \frac{e_i}{\hat{\sigma}\sqrt{i+h_{ii}}}$ the rule of thumb is that if $r_i > |3|$ than the i-th value is a possible outlier. In my case, as we can see in figure 9, we got only one value slightly over the 3 marks, which is:

```
outliers <- rsta[rsta>3]
outliers
```

```
## Kevin Love
##   3.048775
```

as we can see, there is only one outliers, which value is approximable at 3, in any case single outliers do not cause problems, clustsers of outliers instead requires investigation.

high leverage points are data points that have a high influence on the estimated parameters of a regression model. The leverage value is so calculated: $h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ rule of thumb: if $h_{ii} > \frac{2(p+1)}{n}$ the data point is an high leverage point. In my case we got several high leverage points

```
length(hat[which(hat>(12/nrow(NBAstats)))])
```

```
## [1] 22
```

```
head(hat[which(hat>(12/nrow(NBAstats)))])
```

```
##     Kevin Durant     Kyrie Irving     Jayson Tatum    Miles Bridges
##       0.04497660       0.02979086       0.03687206       0.03309158
##      Luka Doncic Boban Marjanovic
##       0.03868363       0.04615214
```

As we can see, there are a lot of high leverage points (22), in the section above we can look at 6 of them. Those type of data points are not a big issue since, as we will see below, there aren't any influential points. Influential points are data points that have a significant effect on the estimated parameters of a regression model and on the predictions made by the model. They are points that, if removed from the dataset, would cause a significant change in the estimated parameters and/or in the predicted values. To study if a data point is an influential points we need to perform the cook distance: $D_i = \frac{1}{n} r_i^2 \frac{h_{ii}}{1-h_{ii}}$ if D_i is greater that 0.5 the data point observed is an influential points. In my case we don't have influential points.

REPORT THE COEFFICIENTS OF THE BEST MODEL OBTAINED

```
summ3
```

```
##
## Call:
## lm(formula = log(SALARY) ~ AGE + GP + USG.PCT + POSS + EFF, data = NBAstats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83714 -0.48549 -0.00201  0.52049  2.20410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.297e-01  2.927e-01  -2.493   0.0131 *
## AGE          6.253e-02  8.121e-03   7.699 1.04e-13 ***
## GP          -2.161e-02  3.633e-03  -5.948 5.83e-09 ***
## USG.PCT      2.281e-02  7.901e-03   2.887   0.0041 **
## POSS         4.075e-04  5.294e-05   7.697 1.05e-13 ***
## EFF          3.896e-02  6.070e-03   6.418 3.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7285 on 409 degrees of freedom
## Multiple R-squared:  0.4438, Adjusted R-squared:  0.437
## F-statistic: 65.26 on 5 and 409 DF,  p-value: < 2.2e-16
```

The usual interpretation of an estimated coefficient is as a rate of change, so for example increasing POSS rate by 1 cent, with all the other regressors in the model held fixed, is associated with a change in log(salary) of about 4.075eˆ-4 millions of dollar. The standard deviation of the model is equal to 0.72, which is a very high value since the mean of the log salary is equal to 1.7138, also the standard errors of the individual coefficients are quite high which means that there is some uncertainty on the estimation of the effects of the predictors on the model.

```
coefplot(model2,intercept = F, ci = T)
```

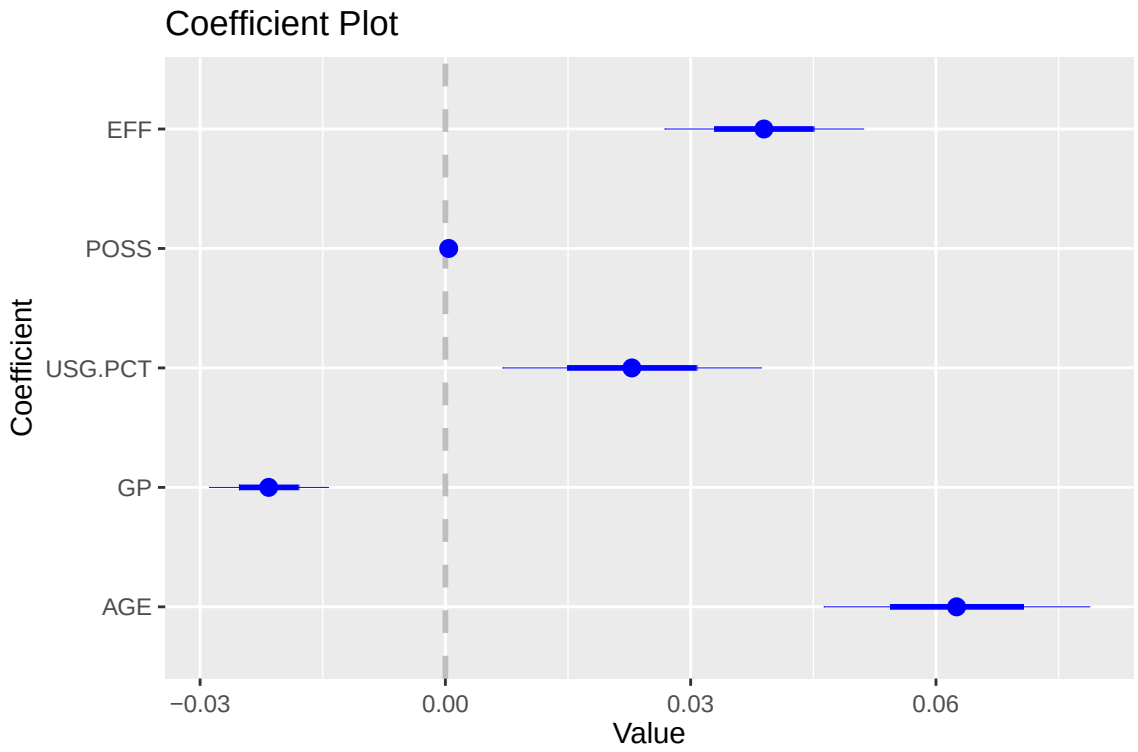


Figure 10: coef

```
kable(confint(model2,level = 0.95))
```

	2.5 %	97.5 %
(Intercept)	-1.3050542	-0.1543637
AGE	0.0465610	0.0784901
GP	-0.0287534	-0.0144689
USG.PCT	0.0072753	0.0383388
POSS	0.0003034	0.0005116
EFF	0.0270256	0.0508905

The above graph shows the value of each beta within the model except for the intercept, and the line below the points shows the corresponding 95% confidence interval. By looking at it we can notice that only the games played are inversely proportional to an increase in salary, this can be justified by the fact that the NBA teams during the regular season (remember that the dataset deals only with this span of games and

not with the playoffs) prefer rest their best players in order to avoid injuries and to allow them to get fit for the final tournament. All the other variables instead have a directly proportional relationship to the salary, which implies that the more a player produces statistics during the game the better his salary will be. We can also see that the covariate POSS has a confidence interval almost equal to 0, which can also be observed in the table that reports all the intervals.

TEST EACH BETA TO BE 0

the following table shows the value of the p-values for a two-tailed hypothesis test, performed for each beta.

```
kable(summ3$coefficients[,4],col.names = "p-value")
```

	p-value
(Intercept)	0.0130543
AGE	0.0000000
GP	0.0000000
USG.PCT	0.0041011
POSS	0.0000000
EFF	0.0000000

the hypothesis test carried out is the following: $y_i = \begin{cases} H_0 & \beta_j = 0 \\ H_1 & \beta_j \neq 0 \end{cases}$ As we can see, every predictor has a p values lower than 0.05 which mean that we never reject the null hypothesis.

TEST A GROUP OF REGRESSORS

```
model3 = lm (log(SALARY) ~ AGE + GP + POSS + EFF, data = NBAstats)
anova(model2,model3)
```

```
## Analysis of Variance Table
##
## Model 1: log(SALARY) ~ AGE + GP + USG.PCT + POSS + EFF
## Model 2: log(SALARY) ~ AGE + GP + POSS + EFF
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    409 217.07
## 2    410 221.49 -1   -4.4222 8.3323 0.004101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$y_i = \begin{cases} H_0 & \text{models can be considered equal} \\ H_1 & \text{models can't be considered equal} \end{cases}$

As we can note from the summary the best model, all the variables have an extremely low p value, the only one that has a p value above the average is the usage percentage, therefore I create an alternative model without the variable in question and perform the anova. the p value less than 0.05 indicates that we reject the null hypothesis that the two models are equal and therefore I continue to use the best model.

GOODNESS OF FIT

```
summ3$r.squared
```

```
## [1] 0.4437668
```

The coefficient of determination R squared in the linear model is 0.4438, which means that the model explains 44.38% of the variation in the data. This rather low value may depend on various factors:

-Data variability: If the data has a large variance, the R squared may be low. This means that the model fails to explain the variation in the data.

-Missing variables: If there are many missing variables or if the data sample is very small, the R squared may be low. In this case, the model may not have enough information to explain the variation in the data. I explained at the beginning how the cases analyzed by me do not represent an exact science, therefore there are several variables that it is impossible for me to take into consideration.

-the model may also not fit the data well and, therefore, cause a lowering of the R squared

PREDICTION OF A NEW OBSERVATION

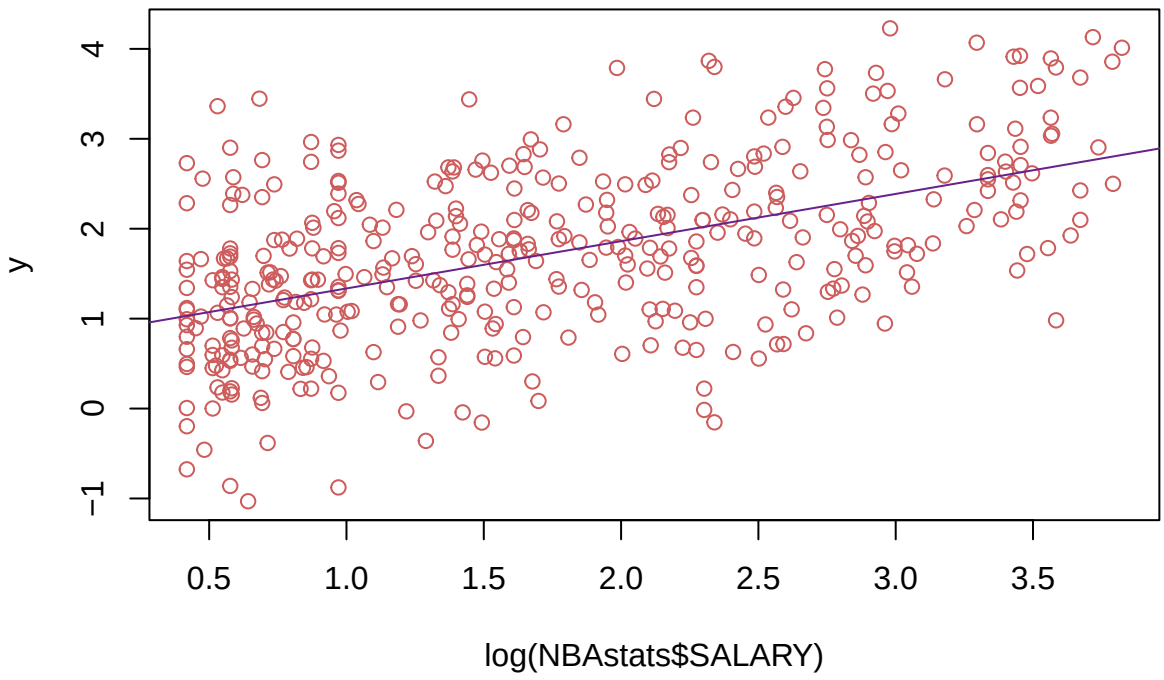
```
data_new = data.frame( "AGE" = 32, "GP" = 82, "USG.PCT" = 32.2, "POSS" = 4529, "EFF"= 31.4)
y_pred = predict(model2, newdata = data_new)
salary_new = exp(y_pred)
salary_new
```

```
##        1
## 27.17585
```

suppose there is a new player in the league, aged 32, who plays all 82 available games, with an efficiency of 31.4, 32.2 usage percentage and 4529 possessions played. according to my model, the player in question would make about $27 million

SIMULATE N DATA POINTS

```
set.seed(32)
beta=coefficients(model2)
X=model.matrix(model2)
y=X%*%beta+rnorm(n, 0, sigma(model2))
regressione = lm(log(NBAstats$SALARY) ~ y )
plot(log(NBAstats$SALARY),y, col = "indianred")
abline(regressione, col = "darkorchid4")
```



the vector y contains 415 data points, representing the estimate of the wage logarithm according to my best model (model2). By plotting the predicted observation against the actual observation, we can see that there is still a big problem of heteroschedasticity. This was totally expected because, as we can see in the diagnostics paragraph, I didn't solve the non costant variance problem (go back to that paragraph for the explaination). However we can see some kind of trend going on.