```
suppressMessages(library(corrplot))
suppressMessages(library(car))
suppressMessages(library(FactoMineR))
suppressMessages(library(factoextra))
suppressMessages(library(MASS))
```

Caricamento dei dati con omissione degli NA e summary

```
dati <- read.csv("data/dati_meteo_O3.csv")
dati <- na.omit(dati)
rownames(dati) <- c(1:168)
summary(dati)
```

```
##      DATA               campag        wd_vett          ws_vett
##  Length:168         Min.   :1.00   Min.   :  1.483   Min.   :0.05839
##  Class :character   1st Qu.:2.00   1st Qu.:192.590   1st Qu.:0.65502
##  Mode  :character   Median :4.00   Median :236.324   Median :1.13442
##                     Mean   :4.47   Mean   :213.707   Mean   :1.18978
##                     3rd Qu.:7.00   3rd Qu.:267.032   3rd Qu.:1.62999
##                     Max.   :8.00   Max.   :348.723   Max.   :4.17368
##     ws_scal           umidit          temp_med        radiazione
##  Min.   :0.8083   Min.   : 22.50   Min.   : 1.942   Min.   : 12.76
##  1st Qu.:1.7375   1st Qu.: 48.25   1st Qu.: 7.418   1st Qu.: 67.67
##  Median :2.2854   Median : 67.23   Median :12.652   Median :107.89
##  Mean   :2.2413   Mean   : 65.18   Mean   :14.263   Mean   :147.74
##  3rd Qu.:2.6708   3rd Qu.: 80.95   3rd Qu.:19.729   3rd Qu.:254.42
##  Max.   :4.7500   Max.   :100.71   Max.   :31.946   Max.   :327.90
##    pressione      precipitazione        AH_day            O3
##  Min.   : 979.4   Min.   :0.00000   Min.   :1.891   Min.   :  1.435
##  1st Qu.:1001.6   1st Qu.:0.00000   1st Qu.:3.668   1st Qu.: 11.467
##  Median :1004.8   Median :0.00000   Median :5.155   Median : 25.457
##  Mean   :1005.0   Mean   :0.07842   Mean   :5.335   Mean   : 38.693
##  3rd Qu.:1009.4   3rd Qu.:0.02500   3rd Qu.:6.952   3rd Qu.: 61.315
##  Max.   :1021.2   Max.   :1.20833   Max.   :9.865   Max.   :130.565
```

ws Nome della colonna che rappresenta la velocita' del vento (wind speed).

wd Nome della colonna che rappresenta la direzione del vento (wind direction).

AH_day Nome della colonna che rappresenta l'umidita' assoluta (Absolute humidity).

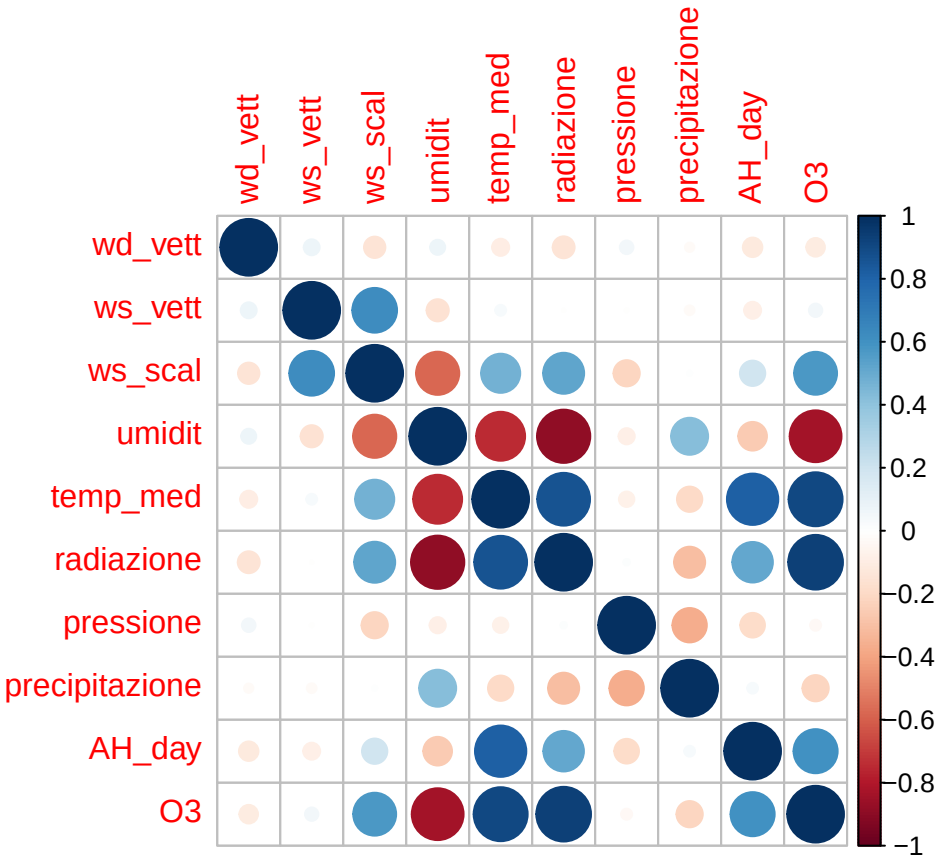
La prima variabile non e' numerica e la seconda è categoriale, il resto sono numeriche

Per realizzare il correlogramma rimuovo le prime due colonne

```
dati_cor <- dati[,c(3:12)]
```

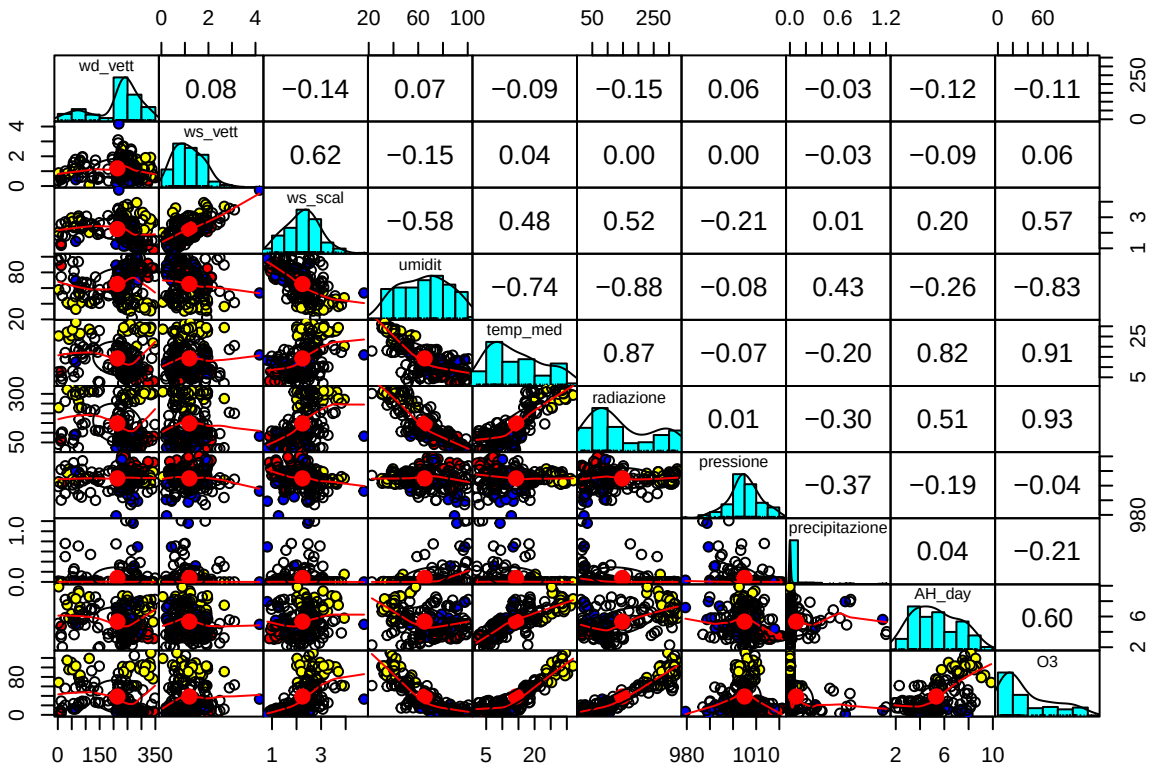
Creo il correlogramma

```
R <- cor(dati_cor)
corrplot(R)
```



Creo anche un grafico coi valori delle correlazioni, gli scatter plot e le loro distribuzioni

```
suppressMessages(library(psych))
pairs.panels(dati_cor,
             gap = 0,
             bg = c("red", "yellow", "blue")[dati$campag],
             pch=21)
```



Sotto la diagonale vengono mostrati gli scatter plot, sopra la diagonale vengono mostrati le correlazioni e lungo la diagonale le distribuzioni. Le distribuzioni sono utili per l'analisi preliminare delle variabili

Condizioni metereologiche: temperatura, pressione, umidita', precipitazioni e il vento colonne: "campag" "wd_vett" "ws_vett" "ws_scal" "umidit" "temp_med" "pressione" "precipitazione" "AH_day"
Caratteristiche delle giornate: "radiazione" "O3"

Studio quindi le correlazioni tra O3 e radiazione con il resto delle variabili

Per l'ozono noto correlazione positiva elevata con la temperatura media (0.90547118), sempre correlata positivamente ma in maniera minore con la velocita' del vento scalare (0.572159470) e con l'umidita' assoluta (0.60408933). E' presente inoltre correlazione negativa elevata con l'umidita' (-0.83439596)

Per le radazioni noto correlazione positiva elevata con la temperatura media (0.86603771), sempre correlata positivamente ma in maniera minore con la velocita' del vento scalare (0.522133215) e con l'umidita' assoluta (0.51049568). E' presente inoltre correlazione negativa elevata con l'umidita' (-0.88191581)

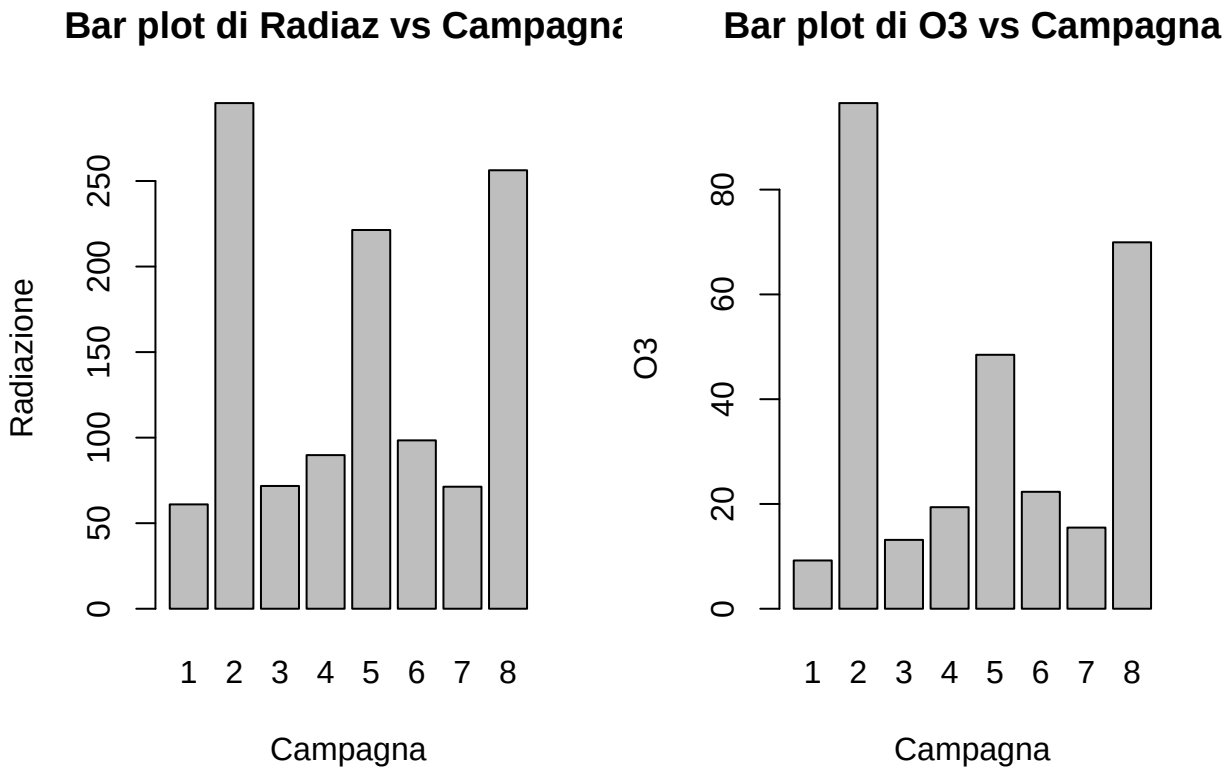
Possiamo concludere che le due variabili studiate sono molto correlate (0.93026623) tra loro e hanno un comportamento simile rispetto alle variabili metereologiche, ovvero valore alto durante le giornate con temperatura, velocita' e umidita' assolute alte e umidita' bassa.

Sono presenti correlazioni molto elevate, quindi sono possibili problemi di collinearita' in caso di analisi fattoriale o regressioni lineari con questi dati.

Realizzo grafico a barre per studiare comportamento della presenza di ozono e delle radiazioni in funzione delle campagne

```
mean_radizioni <- tapply(dati$radiazione,dati$campag , mean)
mean_O3 <- tapply(dati$O3,dati$campag , mean)
par(mfrow=c(1,2))
```

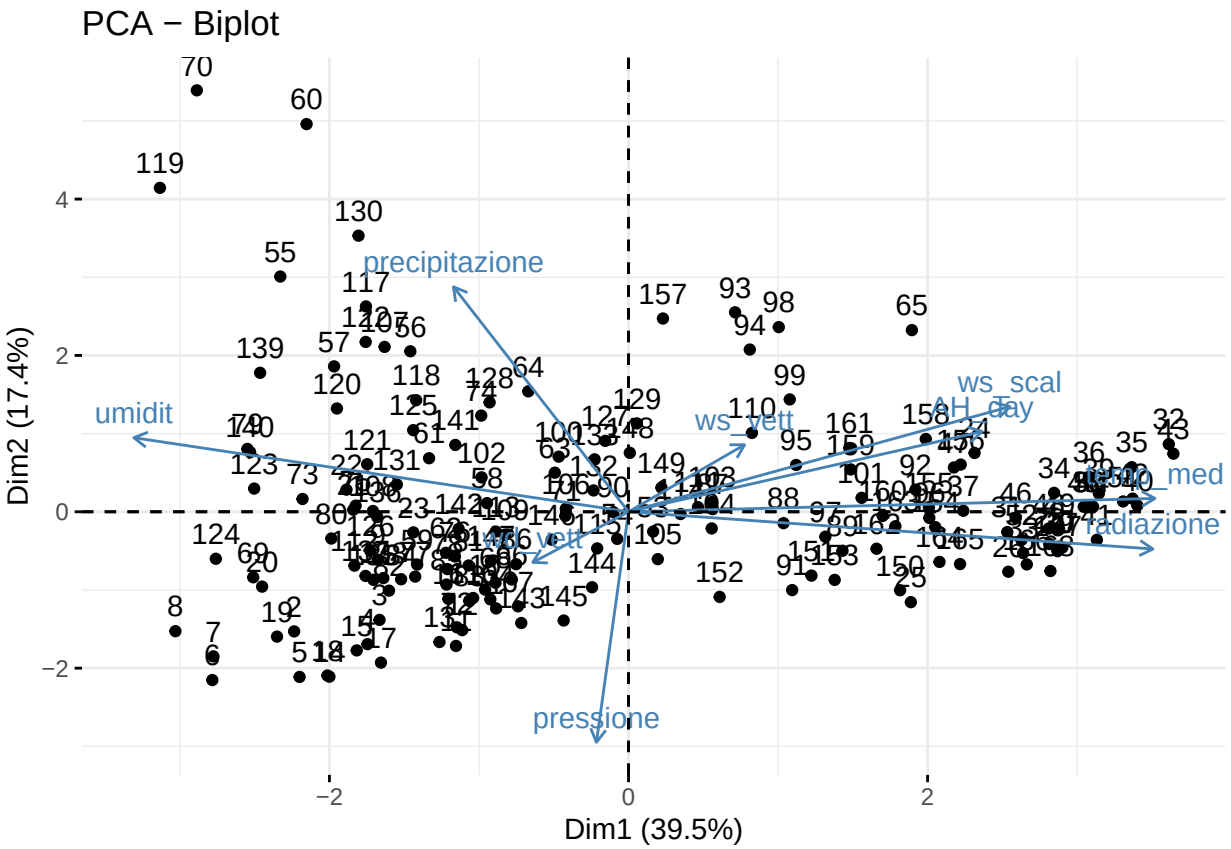
```
barplot(mean_radizioni, ylab = "Radiazione", xlab = "Campagna", main = "Bar plot di Radiaz vs Campagna")
barplot(mean_O3, ylab = "O3", xlab = "Campagna",main = "Bar plot di O3 vs Campagna")
```



Si osservano valori maggiori di radiazione e ozono nelle campagne 2, 5 e 8

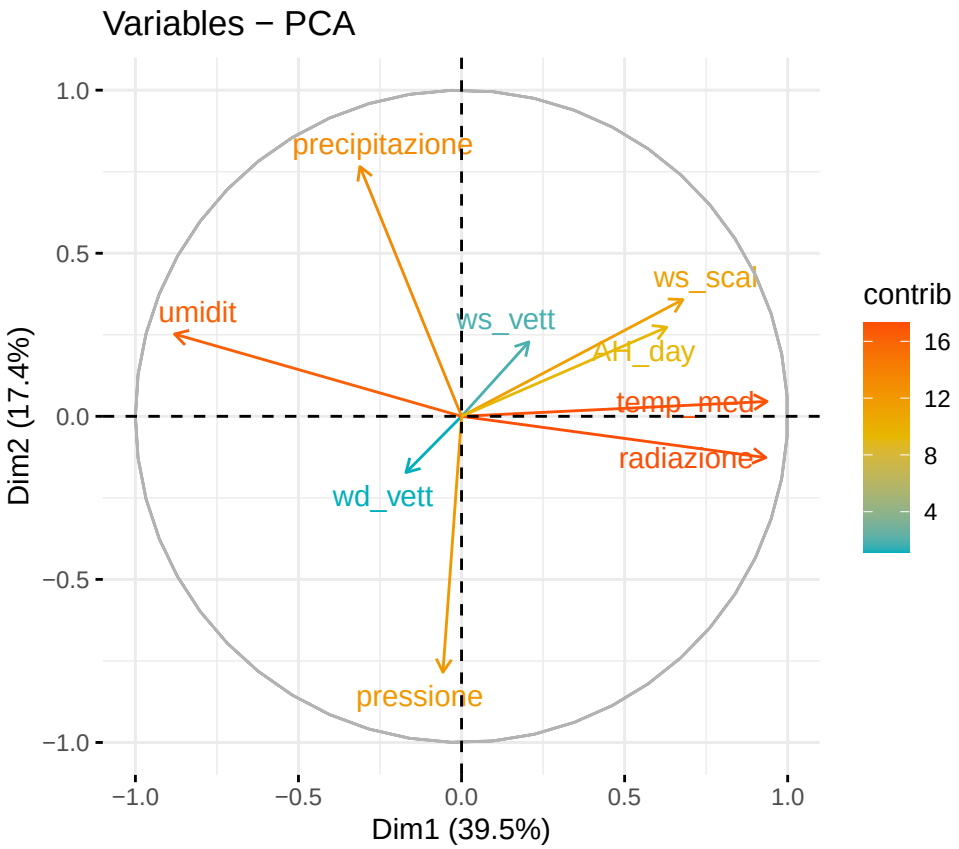
Realizzo le componenti principali per studiare la presenza di ozono Creo quindi un ulteriore dataframe senza dati O3

```
dati_no_O3 <- dati[,c(3:11)]
pca <- PCA(X = dati_no_O3, scale.unit = T, graph = FALSE,ncp = 11)
fviz_pca_biplot(pca)
```



Il primo grafico posiziona ogni giornata in funzione delle prime due componenti principali, le frecce indicano la correlazione tra ogni variabile e le componenti principali.

```
fviz_pca_var(pca,
             col.var = "contrib", # Colore illustra il contributo delle variabili
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Evita sovrapposizione
             )
```



Il secondo grafico riporta solamente le correlazioni tra ogni variabile e le componenti principali, inoltre il colore illustra il contributo delle variabili nelle variabilita' spiegata dalle componenti in percentuale.

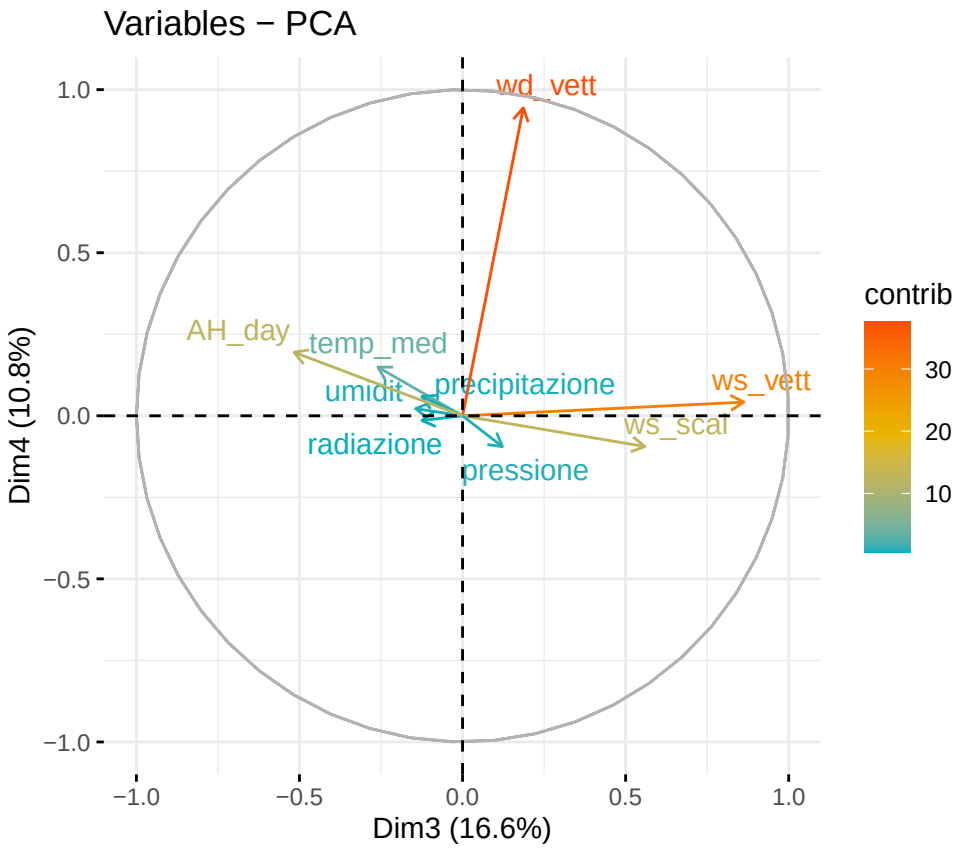
Con pca*var*coord restituiamo le correlazioni tra le variabili e le componenti principali (potenzialmente utilizzabili per interpretarle)

```
round(pca$var$coord,2)
```

```
##                Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7 Dim.8 Dim.9
## wd_vett        -0.17 -0.17  0.19  0.94 -0.09  0.08  0.04  0.01  0.00
## ws_vett         0.21  0.23  0.86  0.04  0.29 -0.20 -0.19  0.03  0.00
## ws_scal         0.68  0.36  0.56 -0.09 -0.02  0.09  0.29 -0.01  0.00
## umidit         -0.88  0.25 -0.14  0.02  0.26 -0.19  0.12  0.15  0.04
## temp_med        0.94  0.05 -0.26  0.15  0.13 -0.02 -0.05 -0.06  0.08
## radiazione      0.93 -0.13 -0.12 -0.01 -0.13  0.19 -0.05  0.21 -0.01
## pressione      -0.06 -0.78  0.12 -0.09  0.52  0.28  0.05 -0.01  0.00
## precipitazione -0.31  0.77 -0.12  0.06  0.19  0.51 -0.08 -0.01  0.00
## AH_day          0.63  0.27 -0.52  0.19  0.40 -0.25  0.05 -0.02 -0.05
```

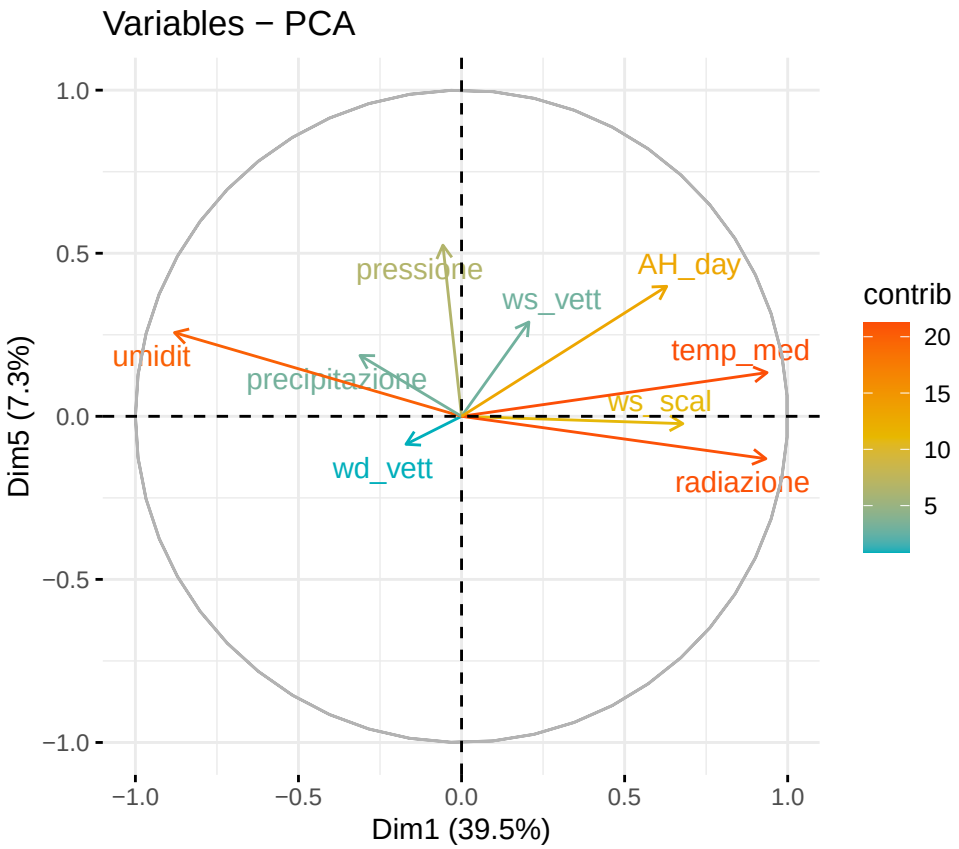
Guardando le frecce (quindi le correlazioni) possiamo interpretare le componenti principali: La prima componente presenta alta correlazione positiva con la temperatura e le radiazioni solari e negativo con l'umidita', Possiamo quindi interpretarla come "Giornata di sole" La seconda componente presenta alta correlazione positiva con le precipitazioni e inversamente con la pressione, possiamo quindi interpretarla come "Giornata di maltempo".

```
fviz_pca_var(pca,
             axes = c(3,4),
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
             )
```



La terza componente principale presenta correlazione positiva alta con la velocita' vettoriale del vento, correlazione positiva ma inferiore con la velocita' scalare e correlazione negativa con la pressione assoluta. Possiamo quindi interpretarla come "Giornata di vento" La quarta componente principale risulta essere quasi solamente correlata con la direzione del vento.

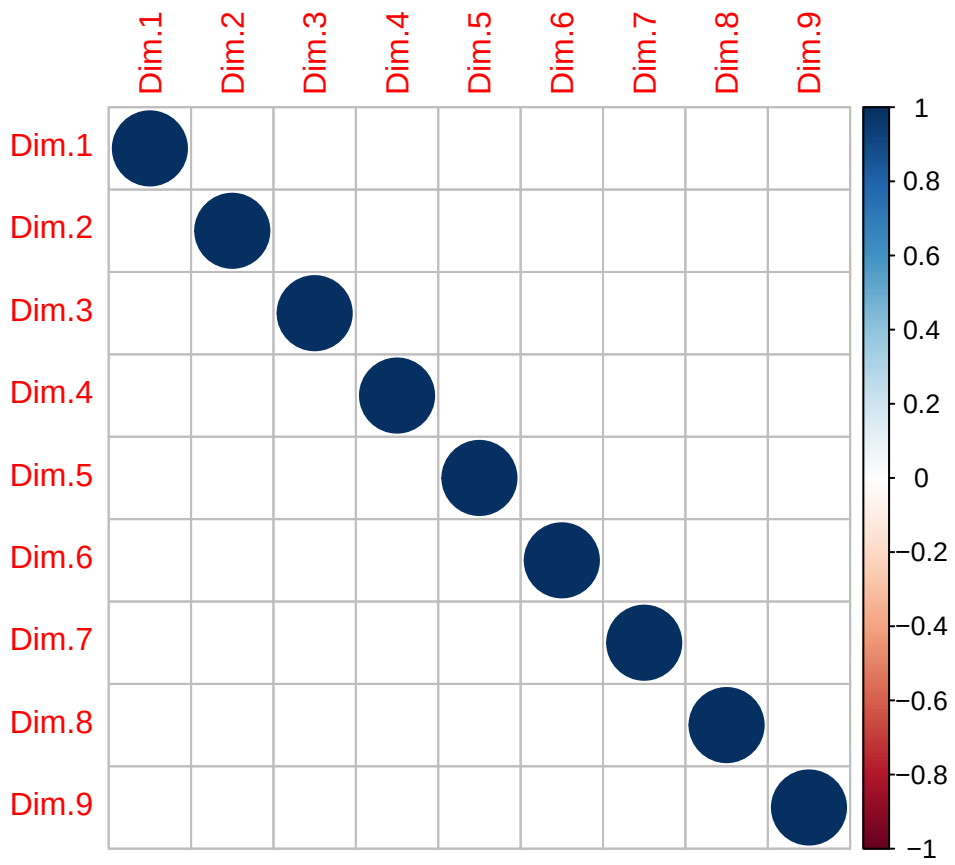
```
fviz_pca_var(pca,
             axes = c(1,5),
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
             )
```



La quinta componente risulta parzialmente correlata con la pressione e l'umidita' assoluta .

Da costruzione le componenti principali devono essere tra loro incorrelate

```
R_coord <- cor(pca$ind$coord)
corrplot(R_coord)
```



Le componenti si dimostrano incorrelate tra loro

Effettuiamo quindi la scelta del numero delle componenti.

Proporzione di variabilita' e regola di Kaiser presente in $eig

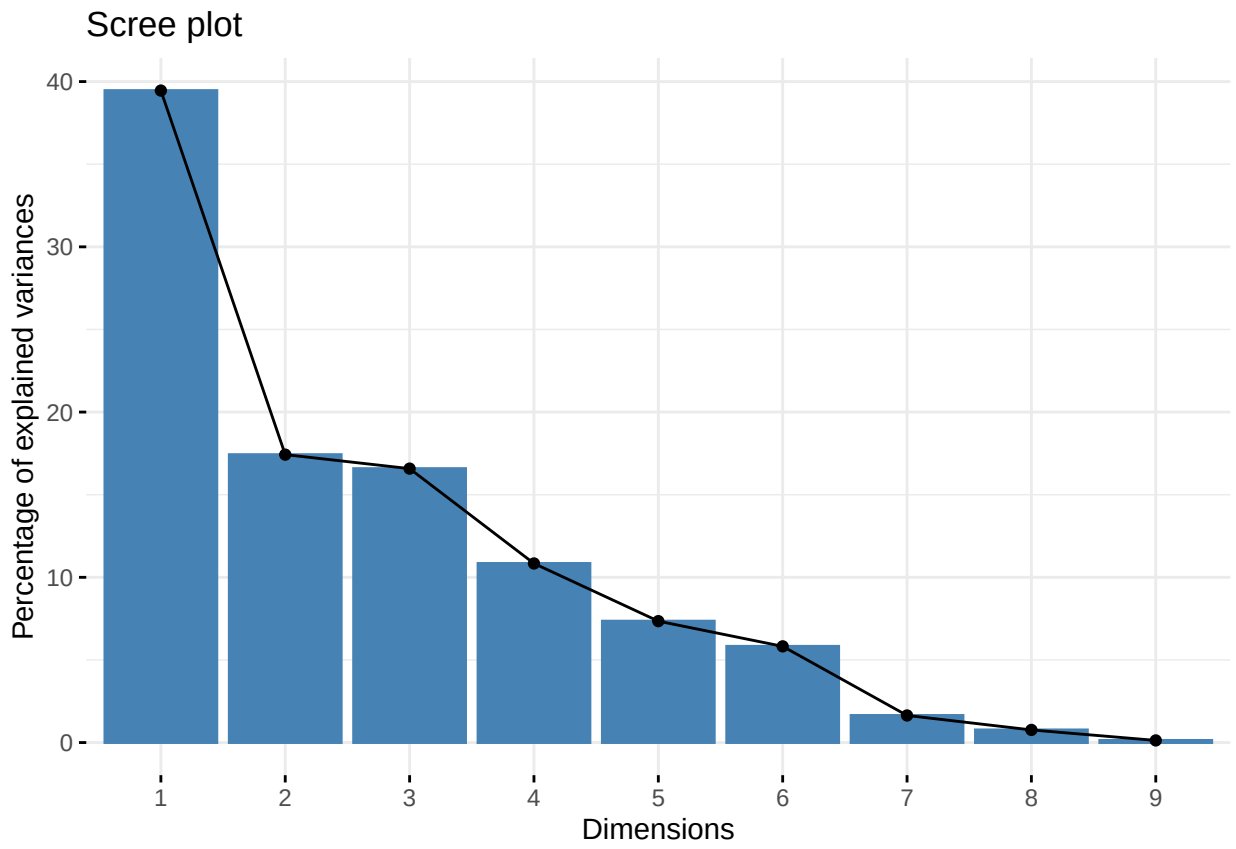
```
pca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 3.55097348             39.4552609                           39.45526
## comp 2 1.56817720             17.4241911                           56.87945
## comp 3 1.49203352             16.5781502                           73.45760
## comp 4 0.97546426             10.8384918                           84.29609
## comp 5 0.66121790              7.3468656                           91.64296
## comp 6 0.52408567              5.8231741                           97.46613
## comp 7 0.14758490              1.6398322                           99.10597
## comp 8 0.06880629              0.7645144                           99.87048
## comp 9 0.01165678              0.1295198                          100.00000
```

Per la regola di Kaiser valutiamo l'inserimento delle variabili con eigenvalue superiore o pari a 1, valutiamo pero' l'inserimento della quarta componente (leggermente minore di 1). Per la regola di proporzione di variabilita' inseriremo solo le prime tre componenti. Decidiamo in ogni caso di inserire la quarta componente visto il valore prossimo a 1 del suo eigenvalue

Produco il grafico a gomito (scree diagram):

```
fviz_eig(pca)
```



Il grafico suggerirebbe l'inserimento di 3 componenti

Con pca*ind*coord produciamo le nuove coordinate delle unita' in base alle componenti principali, e le utilizziamo per costruire un modello di regressione lineare che spieghi la varibile ozono

```
# pca$ind$coord
```

```
pca_coord <- as.data.frame(pca$ind$coord)
pca_coord[,10] <- dati[,12]
colnames(pca_coord) <- c("Dim.1", "Dim.2", "Dim.3", "Dim.4", "Dim.5", "Dim.6", "Dim.7", "Dim.8", "Dim.9

mod_completo <- lm(formula = O3 ~ ., data = pca_coord)
summary(mod_completo)
```

```
##
## Call:
## lm(formula = O3 ~ ., data = pca_coord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.433  -7.132  -0.617   5.751  34.594
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 38.69274    0.78622  49.213  < 2e-16 ***
## Dim.1       17.05622    0.41723  40.880  < 2e-16 ***
## Dim.2       -0.36371    0.62784  -0.579  0.56321
## Dim.3       -3.26652    0.64366  -5.075 1.08e-06 ***
## Dim.4        2.08768    0.79605   2.623  0.00958 **
## Dim.5       -1.29062    0.96688  -1.335  0.18385
## Dim.6        7.07823    1.08604   6.517 9.12e-10 ***
## Dim.7        0.03308    2.04656   0.016  0.98713
## Dim.8        7.35599    2.99731   2.454  0.01521 *
## Dim.9       10.97895    7.28209   1.508  0.13364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.19 on 158 degrees of freedom
## Multiple R-squared:  0.9175, Adjusted R-squared:  0.9128
## F-statistic: 195.2 on 9 and 158 DF,  p-value: < 2.2e-16
```

Per teoria scegliamo le prime 4 componenti e le utilizziamo per regressione Poi selezione tramite stepwise nuovo modello

```
mod_1 <- lm(formula = O3 ~ Dim.1 + Dim.2 + Dim.3 + Dim.4, data = pca_coord)
mod_2 <- stepAIC(mod_1, direction = "both",trace = FALSE)
summary(mod_1)
```

```
##
## Call:
## lm(formula = O3 ~ Dim.1 + Dim.2 + Dim.3 + Dim.4, data = pca_coord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.071  -7.784  -0.392   7.229  35.562
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6927     0.8936  43.301  < 2e-16 ***
## Dim.1        17.0562     0.4742  35.968  < 2e-16 ***
## Dim.2        -0.3637     0.7136  -0.510   0.6110
## Dim.3        -3.2665     0.7316  -4.465 1.49e-05 ***
## Dim.4         2.0877     0.9048   2.307   0.0223 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.58 on 163 degrees of freedom
## Multiple R-squared:   0.89,  Adjusted R-squared:  0.8873
## F-statistic: 329.8 on 4 and 163 DF,  p-value: < 2.2e-16
```

```
summary(mod_2)
```

```
##
## Call:
## lm(formula = O3 ~ Dim.1 + Dim.3 + Dim.4, data = pca_coord)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.826  -7.673  -0.085   7.238  35.724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6927     0.8916  43.399  < 2e-16 ***
## Dim.1        17.0562     0.4731  36.050  < 2e-16 ***
## Dim.3        -3.2665     0.7299  -4.475 1.42e-05 ***
## Dim.4         2.0877     0.9027   2.313    0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.56 on 164 degrees of freedom
## Multiple R-squared:  0.8899, Adjusted R-squared:  0.8878
## F-statistic: 441.7 on 3 and 164 DF,  p-value: < 2.2e-16
```

In base a Adjusted R-squared scegliamo il secondo modello, in quanto l'omissione di "Dim.2" non compromette la variabilita' spiegata dal modello La rimozione della seconda componente principale risulta logica siccome è correlata con variabili non correlate con la presenza di ozono

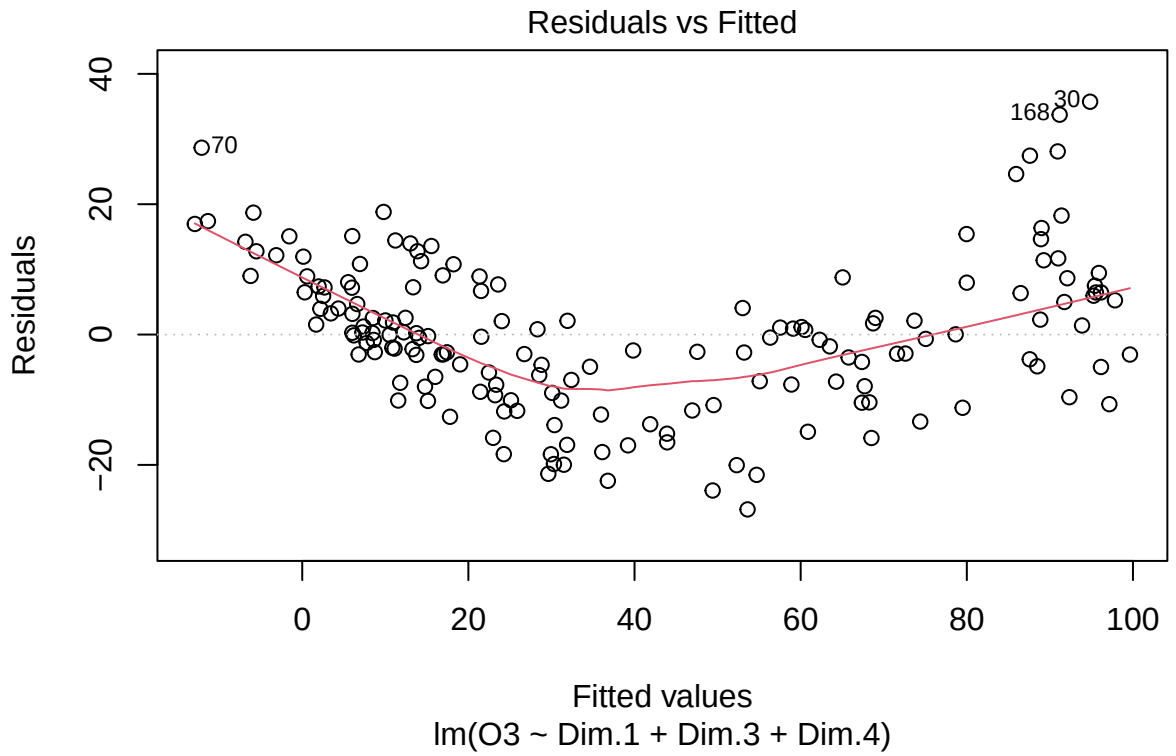
Effettuiamo il metodo di selezione stepwise anche per il modello completo

```
mod_3 <- stepAIC(mod_completo, direction = "both", trace = FALSE)
summary(mod_3)
```

```
##
## Call:
## lm(formula = O3 ~ Dim.1 + Dim.3 + Dim.4 + Dim.6 + Dim.8 + Dim.9,
##     data = pca_coord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.703  -7.426  -0.680   5.504  34.932
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6927     0.7841  49.349  < 2e-16 ***
## Dim.1        17.0562     0.4161  40.992  < 2e-16 ***
## Dim.3        -3.2665     0.6419  -5.089 9.94e-07 ***
## Dim.4         2.0877     0.7939   2.630  0.00937 **
## Dim.6         7.0782     1.0831   6.535 7.97e-10 ***
## Dim.8         7.3560     2.9891   2.461  0.01491 *
## Dim.9        10.9789     7.2621   1.512  0.13254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.16 on 161 degrees of freedom
## Multiple R-squared:  0.9164, Adjusted R-squared:  0.9133
## F-statistic:   294 on 6 and 161 DF,  p-value: < 2.2e-16
```

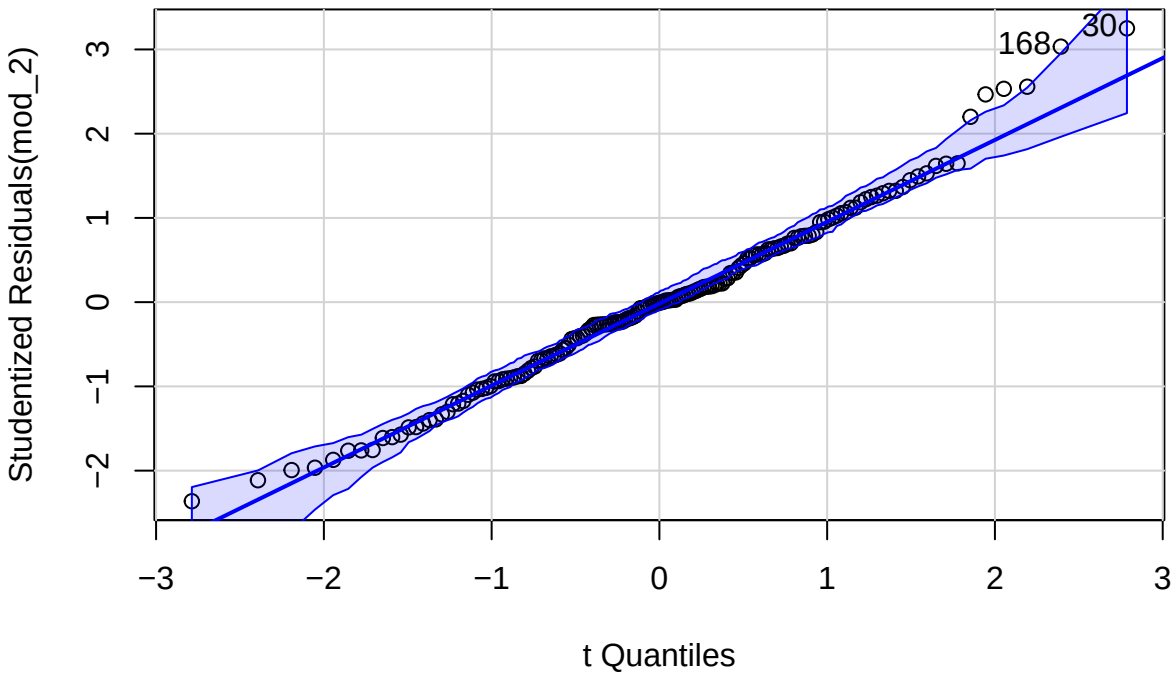
Adjusted R-squared presenta valore maggiore di quelli gia' visualizzati, tuttavia la presenza delle componenti 6,8,9 non segue la teoria legata alla componenti principali

```
plot(mod_2,which = 1)
```



Residuals vs fitted: L'andamento parabolico del grafico mostra relazione non lineare (ma varianza unitaria costante)

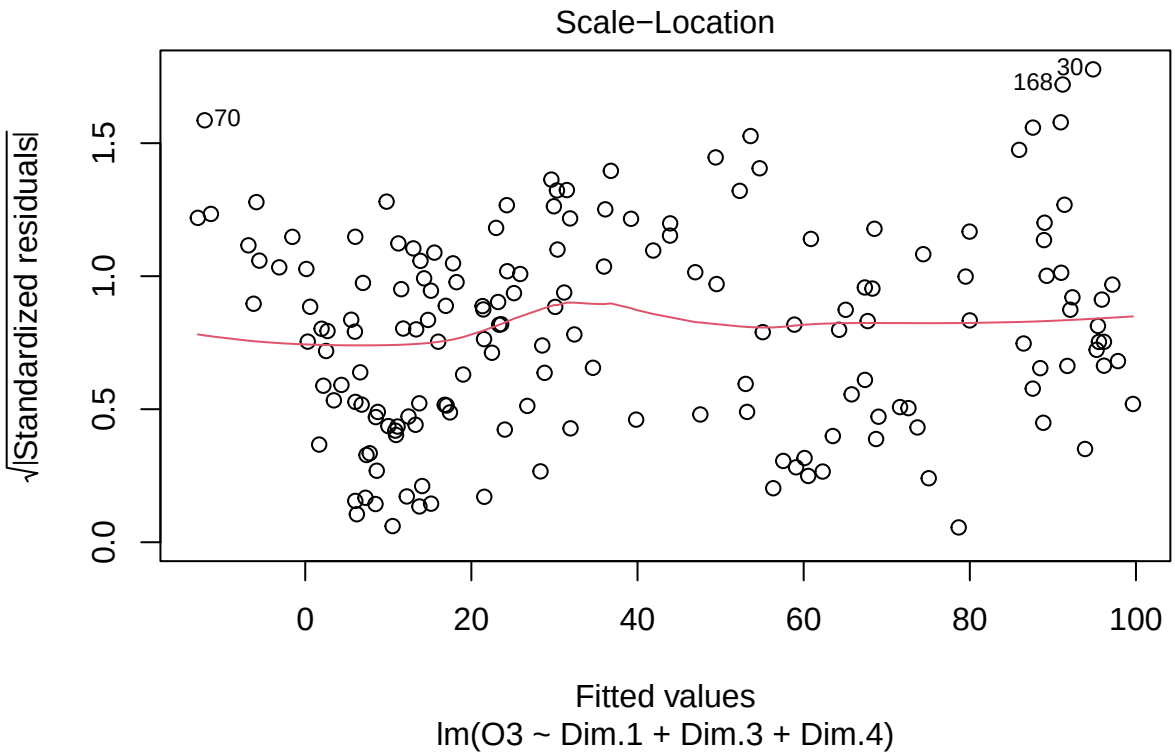
```
qqPlot(mod_2)
```




```
## [1]  30 168
```

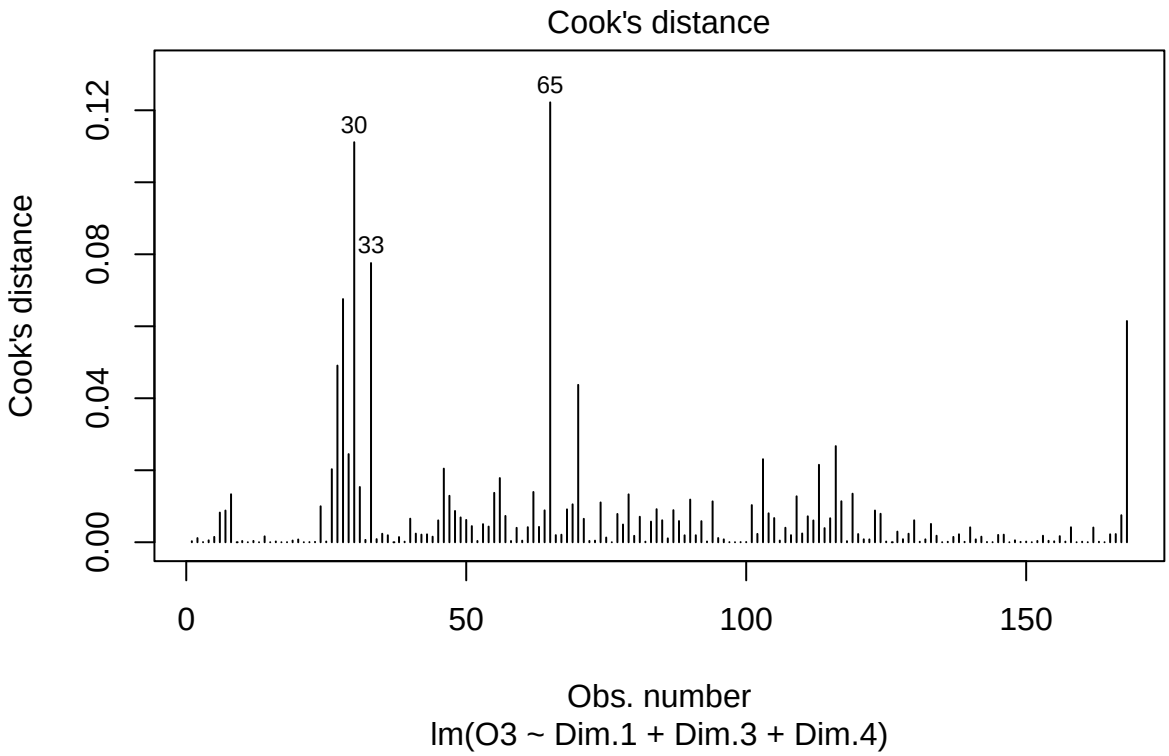
QQ plot: non smentisce la normalita' distributiva dei residui (bande di significativita' confermano)

```
plot(mod_2, which=3)
```



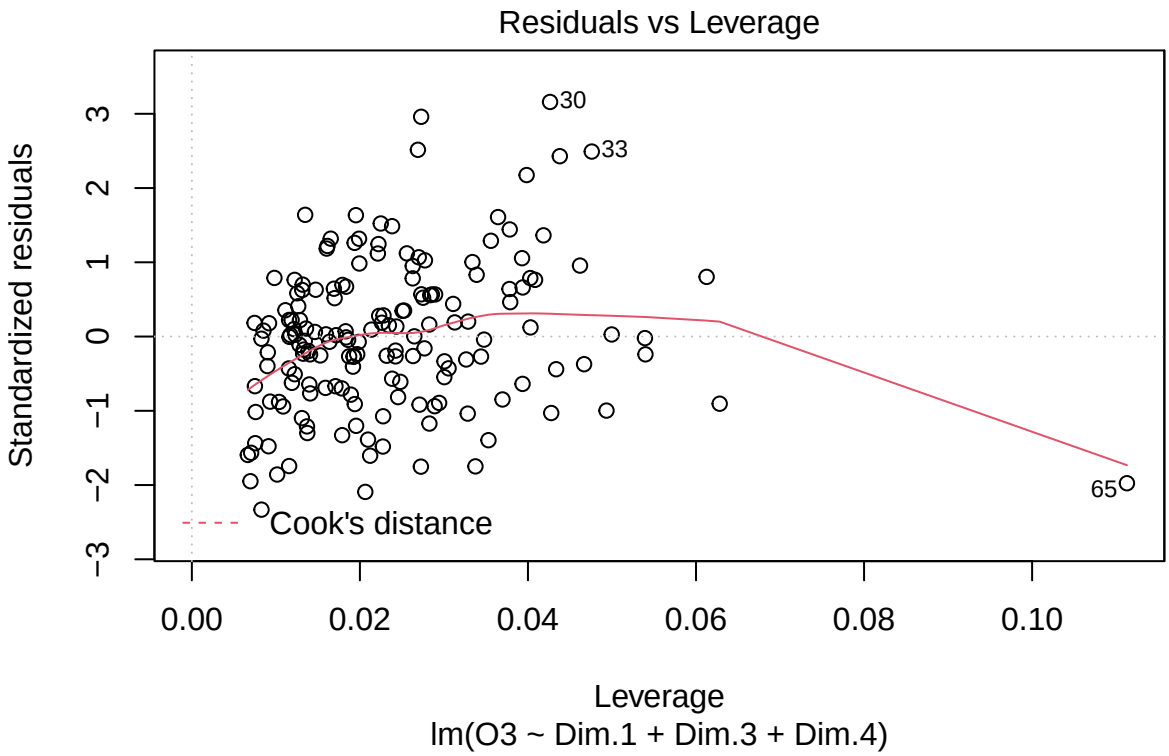
Scale-location: linea praticamente parallela all'asse delle x, omoschedasticita'

```
plot(mod_2, which=4)
```



Cook's distance: notiamo almeno 2 osservazioni potenzialmente influneti (30, 65)

```
plot(mod_2, which=5)
```



Residuals vs Leverage: Misura il peso che ha ogni unita' nel determinare curvatura del coefficiente di regressione: unita' 65 unico punto che potrebbe essere punto di leva.

Effettuiamo test per trovare outliers

```
outlierTest(mod_2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 30 3.250317          0.0014005      0.23529
```

No Studentized residuals with Bonferroni p < 0.05 Non sembrano presenti outliers

Creiamo i residui di student e ne testiamo la distribuzione normale con il test di shapiro

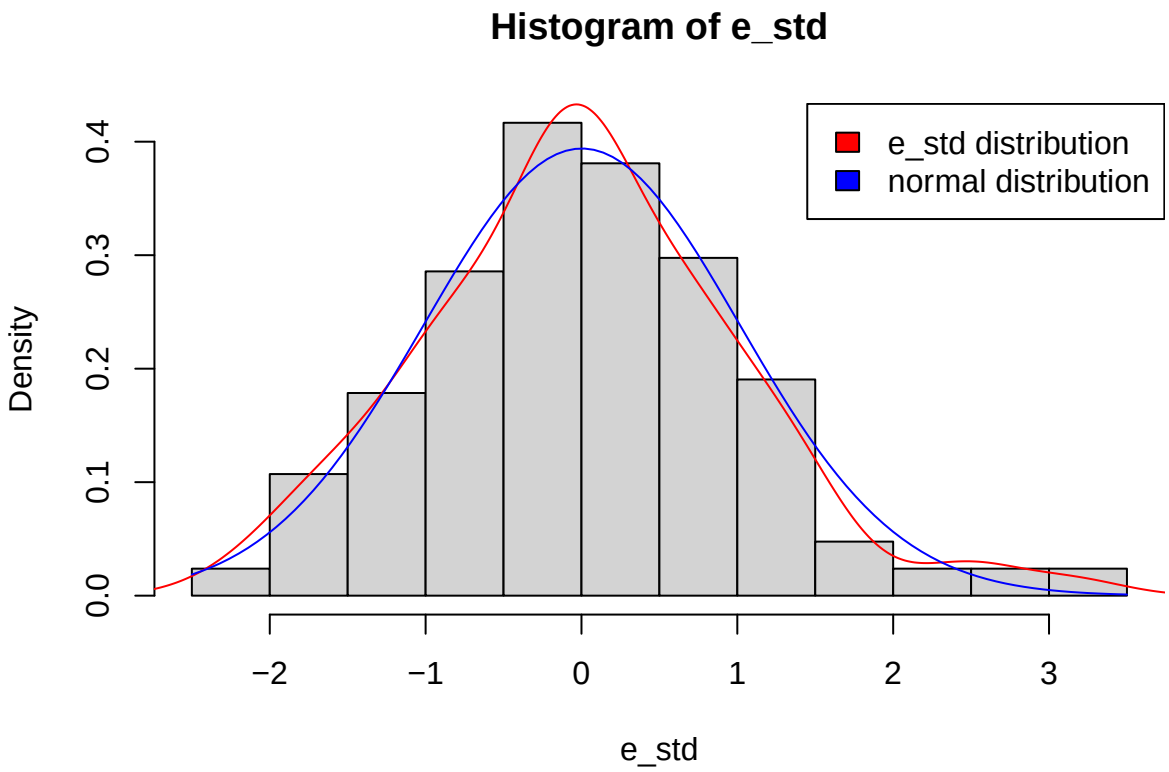
```
e_std <- studres(mod_2)
shapiro.test(e_std)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e_std
## W = 0.98659, p-value = 0.108
```

La normalità distributiva è rispettata

Visualizziamo la distribuzione dei residui tramite istogramma

```
hist(e_std, prob = T,breaks = 20)
lines(x = density(x = e_std), col = "red")
curve(dnorm(x,mean=mean(e_std),sd=sd(e_std)), add=TRUE,col="blue")
legend("topright", c("e_std distribution", "normal distribution"), fill=c("red", "blue"))
```

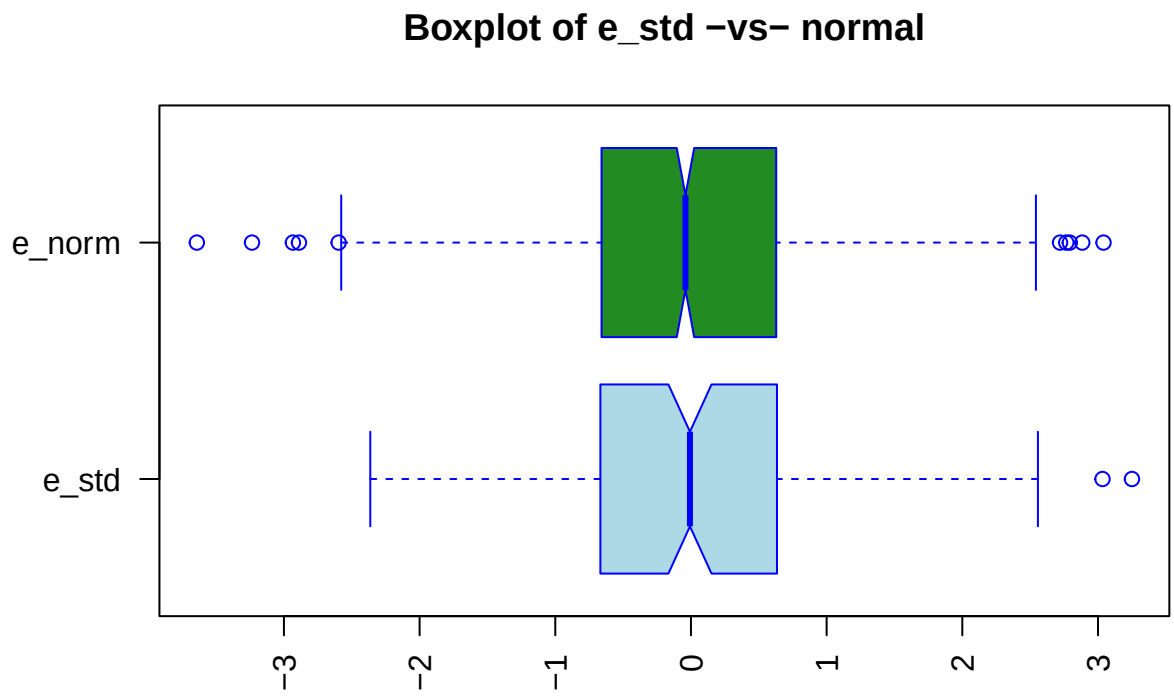


L'Istogramma ricorda la distribuzione normale

Boxplot, genero una distribuzione normale con la stessa media (= 0) e sd dei residui (= 1) e li plotto insieme per confronto

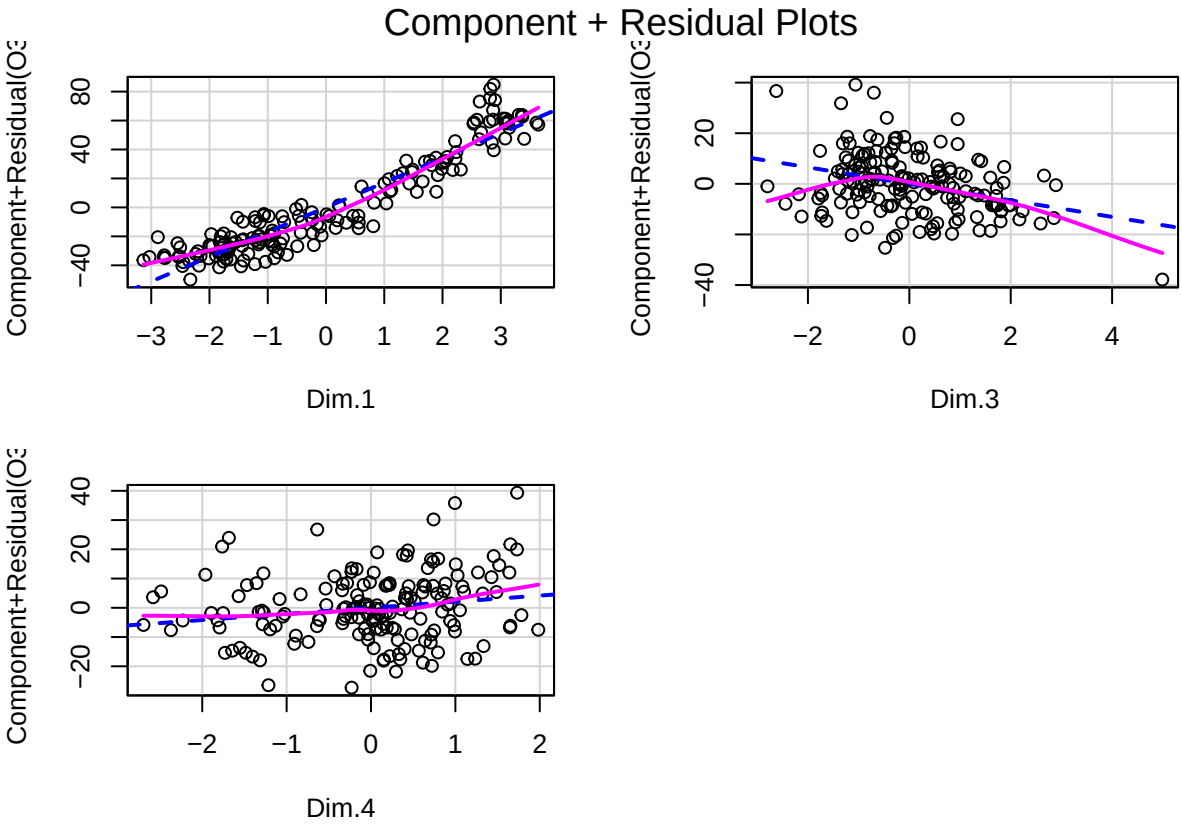
```
e_norm <- rnorm(1000,mean=mean(e_std, na.rm=TRUE), sd=sd(e_std, na.rm=TRUE))
```

```
boxplot(e_std,e_norm,
main = "Boxplot of e_std -vs- normal", at=c(1,2),names = c('e_std','e_norm'),
col = c("light blue","forest green"), las=2,
border = "blue",
horizontal = TRUE,
notch = TRUE
)
```



Il boxplot si avvicina alla distribuzione normale

```
crPlots(mod_2)
```



Permette di visualizzare quanto si allontana il comportamento dei nostri dati dalla linearita'. Nella Dim.1 e Dim.4 i dati si comportano lineramente, tuttavua la DIm.3 sembra assumere comportamento parabolico

Creo modello con log per studiare l'apparente relazione non lineare dei residuals vs fitted

```
mod_11 <- lm(formula = log(O3) ~ Dim.1 + Dim.2 + Dim.3 + Dim.4, data = pca_coord)
mod_12 <- stepAIC(mod_11, direction = "both",trace = FALSE)
```

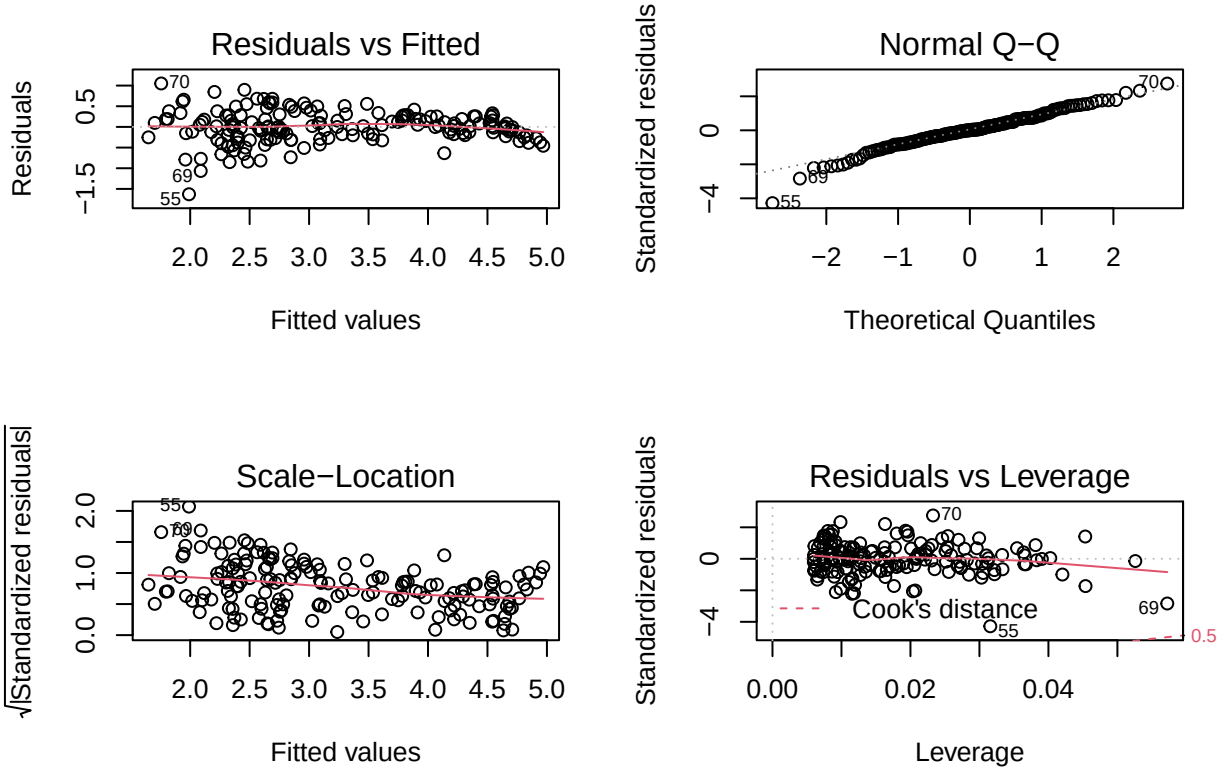
```
summary(mod_12)
```

```
##
## Call:
## lm(formula = log(O3) ~ Dim.1 + Dim.4, data = pca_coord)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6292 -0.2037  0.0023  0.2505  1.0510
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.20634    0.02983 107.473   <2e-16 ***
## Dim.1        0.49097    0.01583  31.011   <2e-16 ***
## Dim.4       -0.04429    0.03021  -1.466    0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3867 on 165 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8521
## F-statistic: 481.9 on 2 and 165 DF,  p-value: < 2.2e-16
```

L'Adjusted R-squared risulta minore rispetto al modello non logaritmico (mod_2 Adjusted R-squared: 0.8878) Viene inoltre rimossa la Dim.3

```
par(mfrow=c(2,2))
plot(mod_12)
```



I residui risultano lineari, tuttavia sembra presente eteroschedasticità, le osservazioni 55 e 69 risultano potenzialmente influenti.

```
outlierTest(mod_12)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 55 -4.527259         1.1439e-05    0.0019217
```

il test conferma che l'osservazione 55 potrebbe essere un outlier

```
suppressMessages(library(lmtest))
resettest(mod_2)
```

```
##
##  RESET test
```

```
##
## data:  mod_2
## RESET = 67.917, df1 = 2, df2 = 162, p-value < 2.2e-16
```

rifuto ipotesi nulla di relazione lineare

```
resettest(mod_12)
```

```
##
##  RESET test
##
## data:  mod_12
## RESET = 4.2609, df1 = 2, df2 = 163, p-value = 0.01571
```

il modello logaritmico migliora la relazione lineare

```
cooksD2 <- cooks.distance(mod_12)
influential <- cooksD2[(cooksD2 > (3 * mean(cooksD2, na.rm = TRUE)))]
influential2 <- cooksD2[(cooksD2 > (4/nrow(pca_coord)))]


influential
```

```
##         55         57         69         70         73         74         77
## 0.19921660 0.02692378 0.16288945 0.06007129 0.02857570 0.03166109 0.02079924
##        121        139
## 0.04755369 0.02962784
```

```
influential2
```

```
##         55         57         69         70         73         74        121
## 0.19921660 0.02692378 0.16288945 0.06007129 0.02857570 0.03166109 0.04755369
##        139
## 0.02962784
```

i punti 55 e 69 evidenziano una distanza di Cook elevata (calcolati come: maggiori di 3 volte la media oppure maggiori di 4/n)

```
pca_coord2 <- pca_coord[-c(55,69),]
```

```
mod_1_no_out <- lm(formula = log(O3) ~ Dim.1 + Dim.2 + Dim.3 + Dim.4, data = pca_coord2)
mod_2_no_out <- stepAIC(mod_1_no_out, direction = "both",trace = FALSE)
```

```
summary(mod_2_no_out)
```

```
##
## Call:
## lm(formula = log(O3) ~ Dim.1 + Dim.4, data = pca_coord2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.88825 -0.19451 -0.00329  0.24850  1.00265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.22290    0.02771 116.320   <2e-16 ***
## Dim.1        0.47977    0.01476  32.501   <2e-16 ***
## Dim.4       -0.04503    0.02872  -1.568    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3569 on 163 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8651
## F-statistic: 529.9 on 2 and 163 DF,  p-value: < 2.2e-16
```

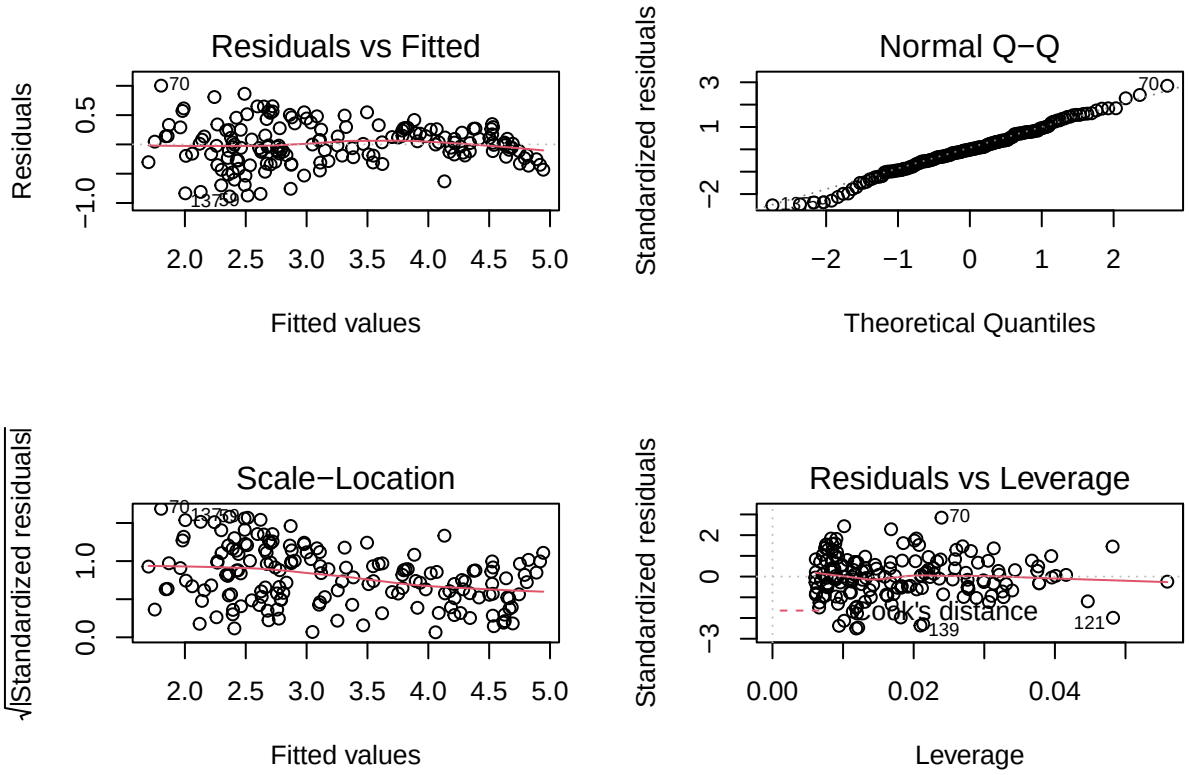
Adjusted R-squared risulta maggiore di del modello log con gli outliers (Adjusted R-squared: 0.8521)

```
resettest(mod_2_no_out)
```

```
##
##  RESET test
##
## data:  mod_2_no_out
## RESET = 4.7098, df1 = 2, df2 = 161, p-value = 0.01028
```

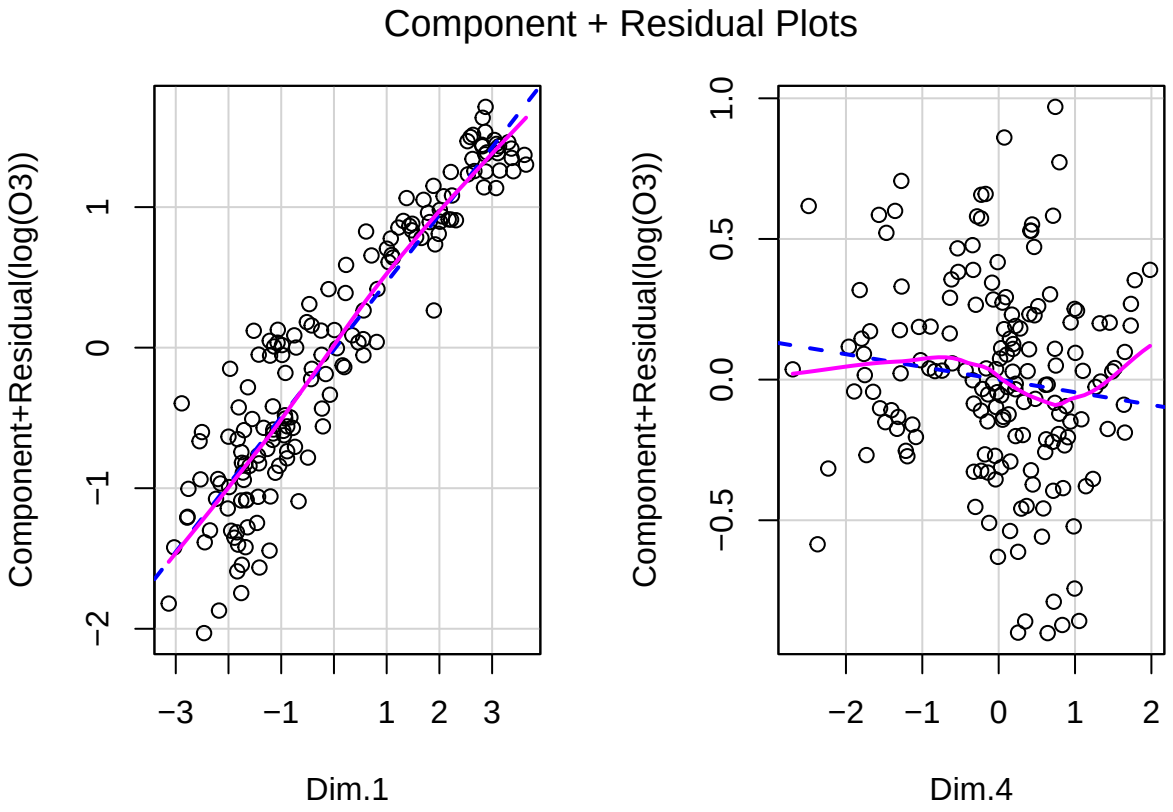
Il test reset restituisce un p.value minore rispetto al modello logaritmico con tutte le osservazioni

```
par(mfrow=c(2,2))
plot(mod_2_no_out)
```



La rimozione degli outliers migliora il modello come evidenziato dal plot

```
crPlots(mod_2_no_out)
```



I grafici delle Dim.1 e Dim.4

Dimostrazione teoria:

```
pca_test <- prcomp(dati_no_03)
```

varianza

```
pca_test.var <- pca_test$sdev ^ 2
pca_test.var
```

```
## [1] 1.134157e+04 7.105120e+03 8.637693e+01 4.811166e+01 1.779030e+01
## [6] 5.756607e-01 1.419122e-01 8.841367e-02 2.625916e-02
```

E' rispettata la proprietà per la quale le componenti principali sono ordinate per la variabilità che riescono a sintetizzare

```
sum(pca_test.var)
```

```
## [1] 18599.8
```

```
total.var<-sum(diag(cov(dati_no_03)))
total.var
```

```
## [1] 18599.8
```

E’ rispettata la proprietà per la quale la variabilità complessiva dei due sistemi di variabili coincide