# LINEAR 9 - Data set: AIRQUALITY

## INTRODUZIONE

Il data set contiene contiene 154 osservazioni con 6 variabili.

1. OZONO: concentrazioni di Ozono (parti per milione misurata a Roosevelt Island)
2. SOLAR.R: radiazione solare (misurata al Central Park)
3. WIND: velocità media del vento (misurata all'aeroporto LaGuardia)
4. TEMP: temperatura in F (misurata all'aeroporto LaGuardia)
5. MONTH: mese
6. DAY: giorno del mese

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare
3. Diagnostiche (QQ-plot, residui)

```r
#-- R CODE

library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
panderOptions('knitr.auto.asis', FALSE)

#-- White test function
white.test <- function(lmod,data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM,"P value"=p.value)
}

#-- funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVARION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\4.tutto(4)\\3.tutto\\airquality.txt"),sep=

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("Ozone","Solar.R","Wind","Temp")

#-- print delle prime 6 righe del dataset
pander(head(d),big.mark=",")
```

| id | Ozone | Solar.R | Wind | Temp | Month | Day |
|---|---|---|---|---|---|---|
| 1 | 41 | 190 | 7 | 67 | 5 | 1 |
| 2 | 36 | 118 | 8 | 72 | 5 | 2 |
| 3 | 12 | 149 | 13 | 74 | 5 | 3 |
| 4 | 18 | 313 | 12 | 62 | 5 | 4 |
| 5 | 20 | 178 | 14 | 56 | 5 | 5 |
| 6 | 28 | 193 | 15 | 66 | 5 | 6 |

## STATISTICHE DESCRITTIVE

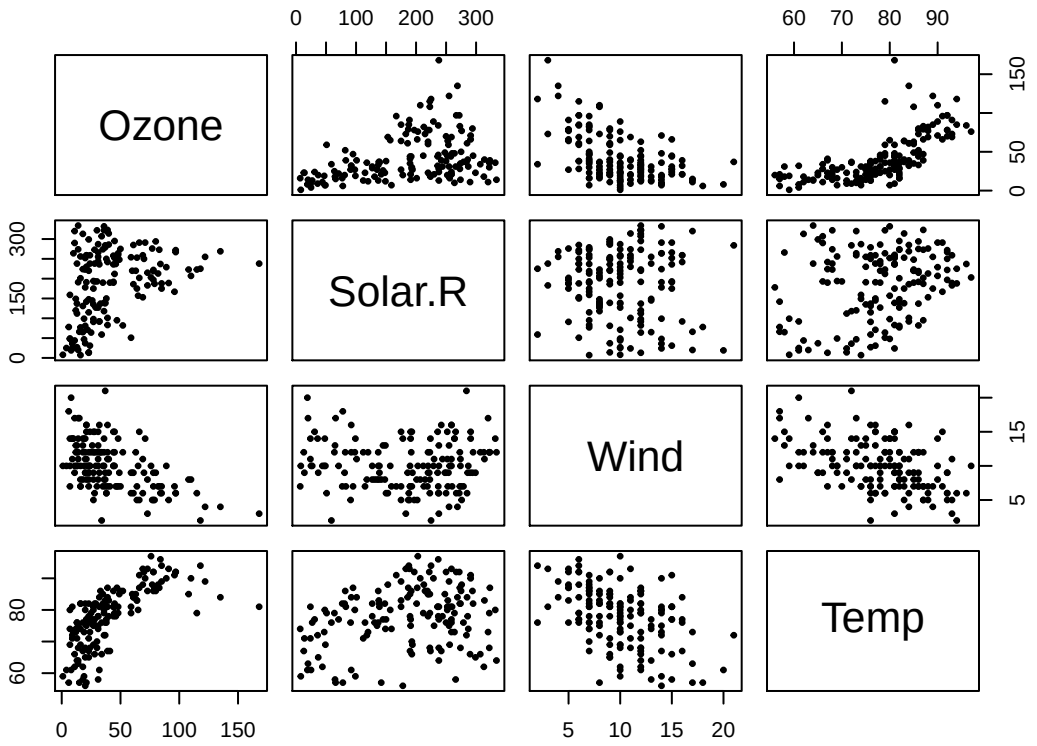
```
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

| Ozone | Solar.R | Wind | Temp |
|---|---|---|---|
| Min. : 1.00 | Min. : 7.0 | Min. : 2.00 | Min. :56.00 |
| 1st Qu.: 20.00 | 1st Qu.:120.0 | 1st Qu.: 7.00 | 1st Qu.:72.00 |
| Median : 33.00 | Median :201.0 | Median :10.00 | Median :79.00 |
| Mean : 41.63 | Mean :185.8 | Mean :10.02 | Mean :77.88 |
| 3rd Qu.: 60.00 | 3rd Qu.:256.0 | 3rd Qu.:12.00 | 3rd Qu.:85.00 |
| Max. :168.00 | Max. :334.0 | Max. :21.00 | Max. :97.00 |

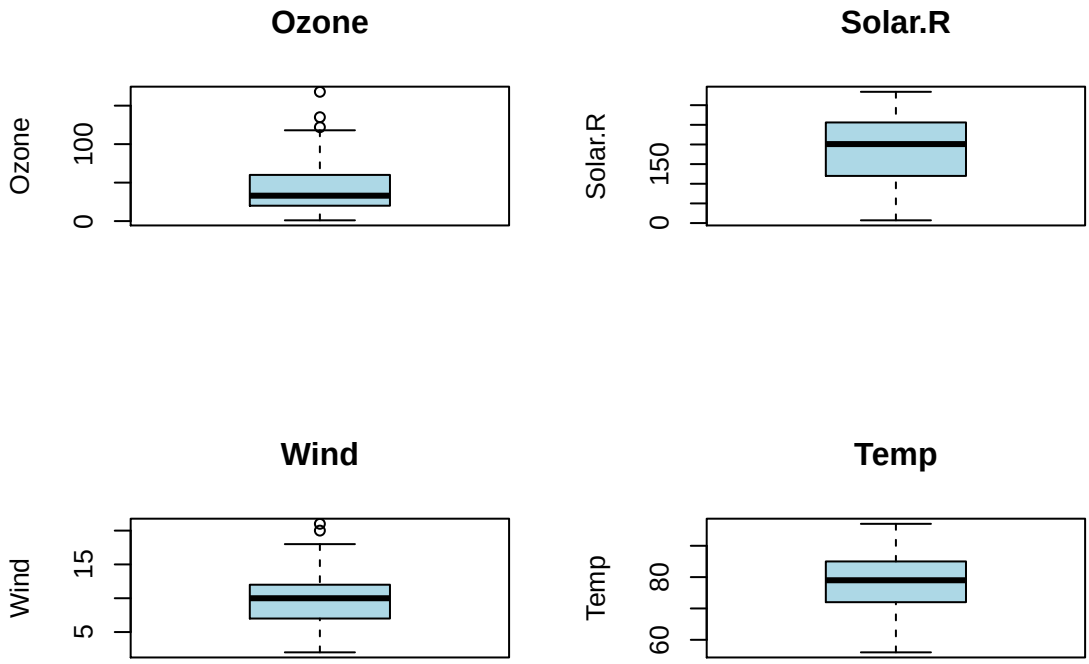
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

| | Ozone | Solar.R | Wind | Temp |
|---|---|---|---|---|
| **Ozone** | 1 | 0.3608 | -0.5403 | 0.6878 |
| **Solar.R** | 0.3608 | 1 | -0.04474 | 0.2744 |
| **Wind** | -0.5403 | -0.04474 | 1 | -0.4555 |
| **Temp** | 0.6878 | 0.2744 | -0.4555 | 1 |

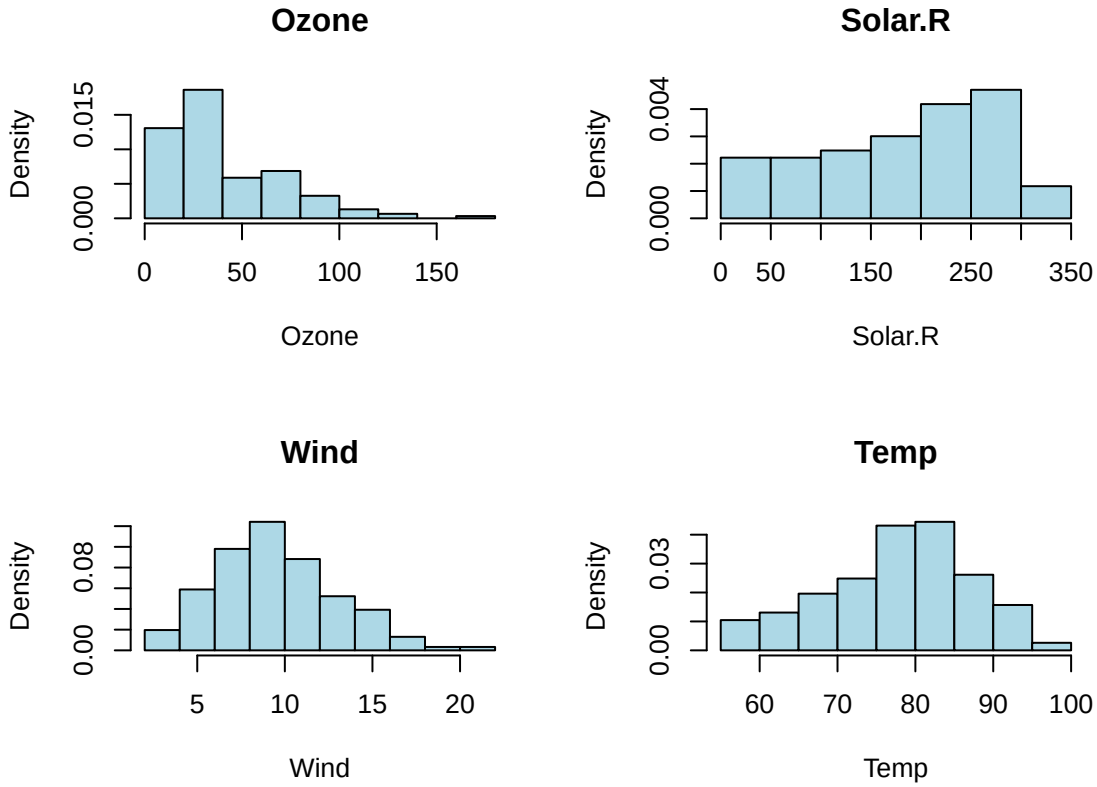
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```




```
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```




La variabile "ozono" che sarà scelta come variabile dipendente è maggiormente correlata con "temp", "solar" e "wind" in senso decrescente.

## REGRESSIONE

Si effettua ora la regressione multipla di "ozono" su "temp", "solar" e "wind".

```
#-- R CODE
mod1 <- lm(Ozone ~ Wind + Temp + Solar.R,d)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -66.66 | 17.67 | -3.772 | 0.0002327 |
| **Wind** | -2.569 | 0.4998 | -5.139 | 8.518e-07 |
| **Temp** | 1.548 | 0.197 | 7.858 | 7.186e-13 |
| **Solar.R** | 0.07234 | 0.01886 | 3.835 | 0.0001845 |

Table 5: Fitting linear model: Ozone ~ Wind + Temp + Solar.R

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 153 | 19.62 | 0.5797 | 0.5712 |

```
pander(anova(mod1),big.mark=",")
```

Table 6: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **Wind** | 1 | 39,823 | 39,823 | 103.5 | 8.711e-19 |
| **Temp** | 1 | 33,588 | 33,588 | 87.28 | 1.282e-16 |
| **Solar.R** | 1 | 5,661 | 5,661 | 14.71 | 0.0001845 |
| **Residuals** | 149 | 57,338 | 384.8 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 6.915 | 0.03151 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 8: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.7 | 0.02467 * | true autocorrelation is greater than 0 |

Sia il modello che le singole variabili risultano significative e l'$R^2$ è sufficientemente elevato (0.58). Si passa ora all'esame della collinearità.

Sia l'indice di tolleranza che il Vif escludono la collinearità tuttavia il condition index vicino alla soglia (28.56) e la proporzione di varianza di "temp" per il 4° autovalore indicano una qualche anomalia.

```
#-- R CODE
pander(ols_eigen_cindex(mod1),big.mark=",")
```

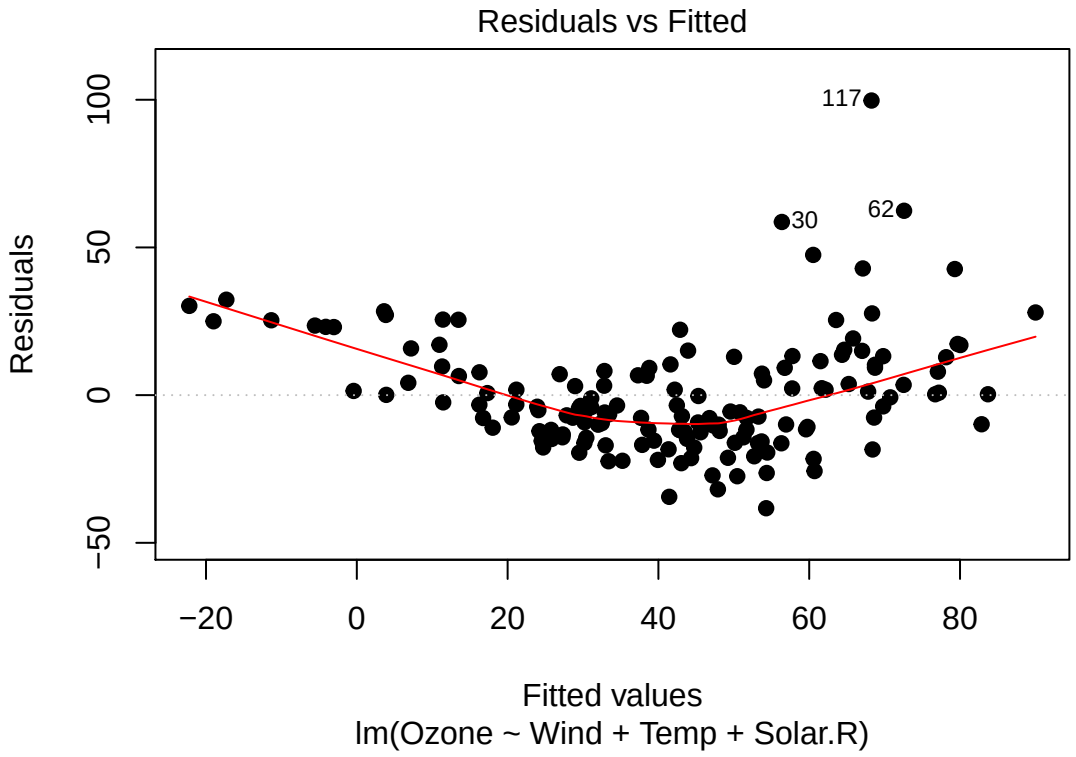
| Eigenvalue | Condition Index | intercept | Wind | Temp | Solar.R |
|---|---|---|---|---|---|
| 3.757 | 1 | 0.0005601 | 0.005749 | 0.0007252 | 0.01056 |
| 0.1592 | 4.857 | 0.00118 | 0.2021 | 1.273e-05 | 0.6459 |
| 0.07882 | 6.905 | 0.0157 | 0.4244 | 0.04309 | 0.3135 |
| 0.004606 | 28.56 | 0.9826 | 0.3677 | 0.9562 | 0.02996 |

```
pander(ols_vif_tol(mod1),big.mark=",")
```

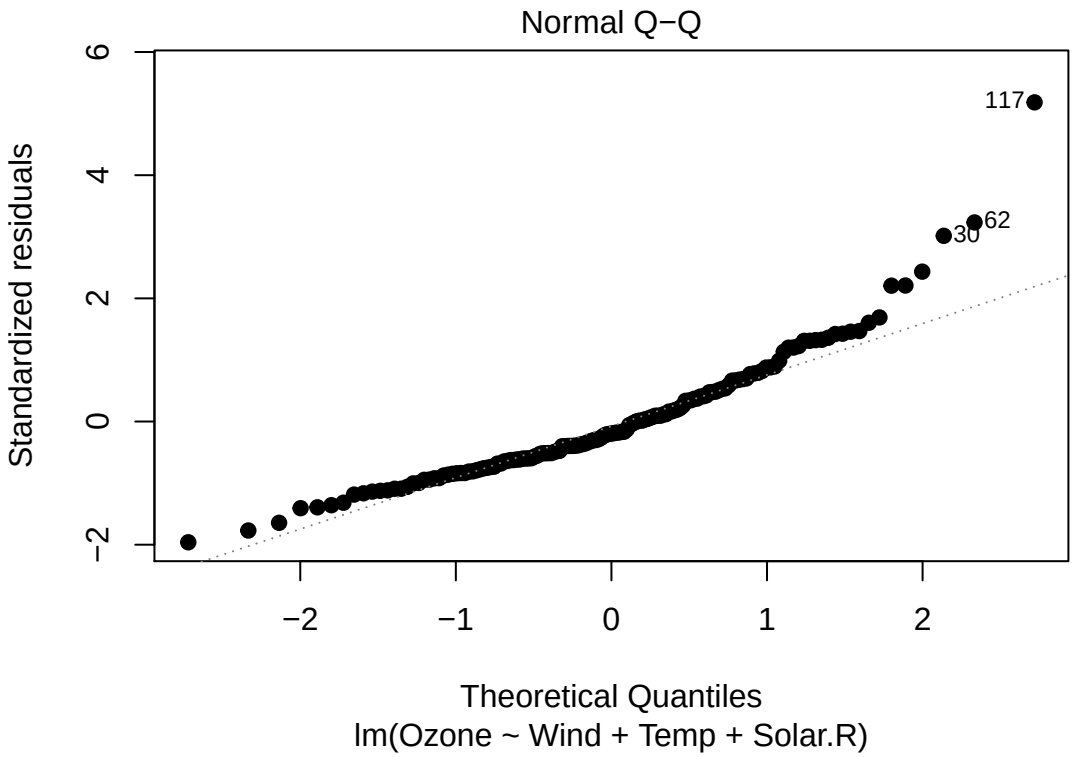
| Variables | Tolerance | VIF |
|---|---|---|
| Wind | 0.7856 | 1.273 |
| Temp | 0.7279 | 1.374 |
| Solar.R | 0.9166 | 1.091 |

Passiamo ora all'esame degli outlier inziando dall'analisi dei grafici inerenti i residui.

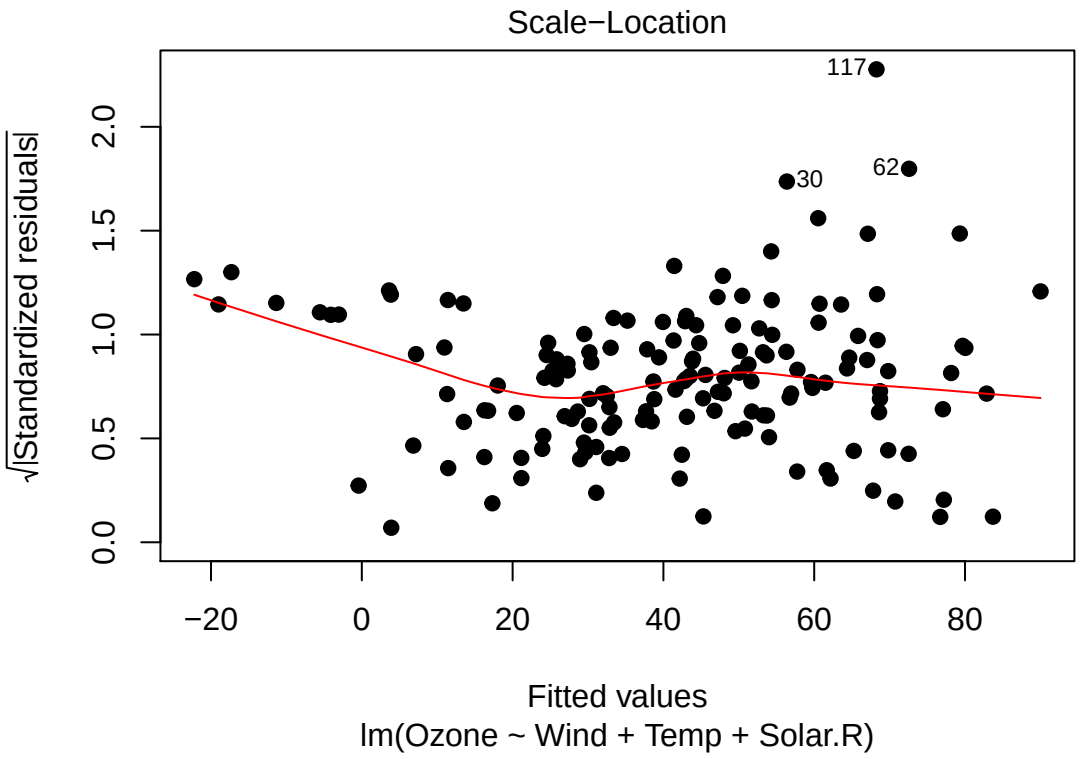
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



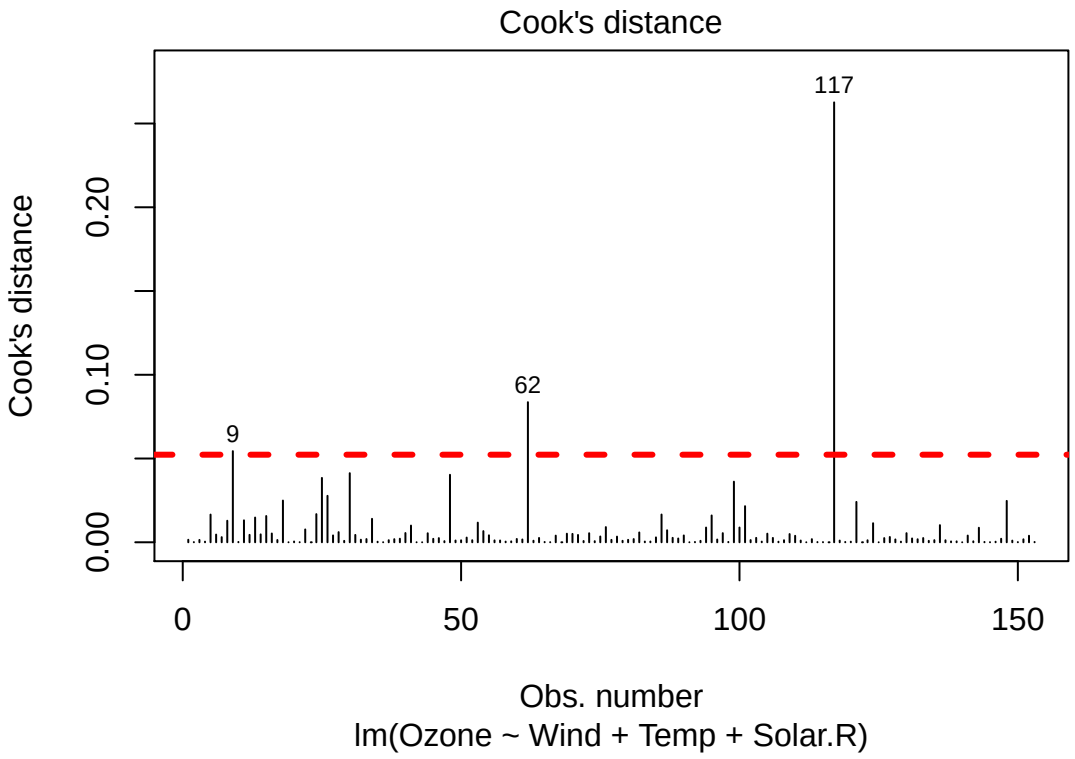

```
plot(mod1,which=2,pch=19)
```



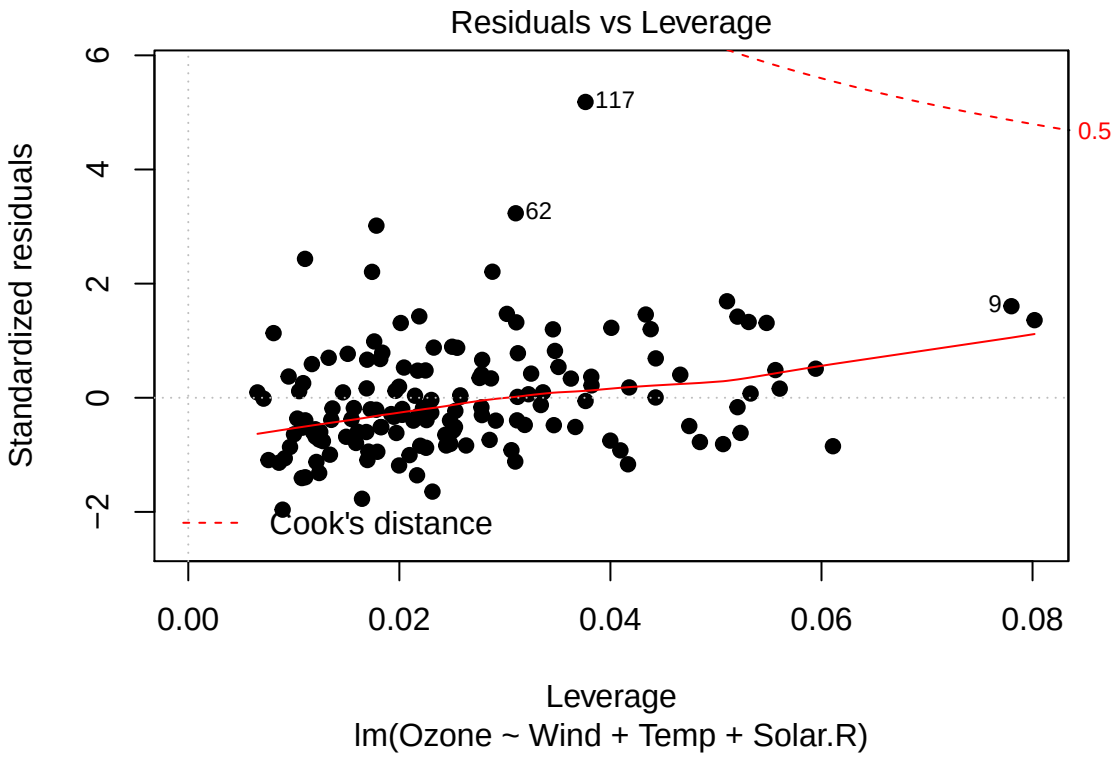

```
plot(mod1,which=3,pch=19)
```



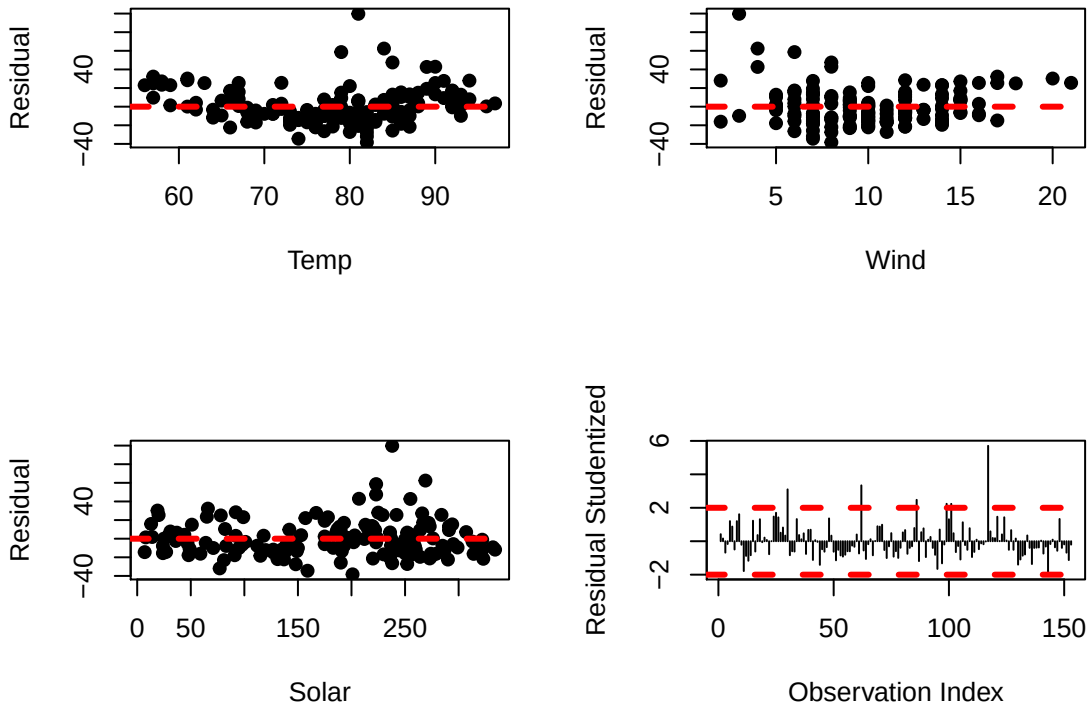

```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d),col=2,lwd=3,lty=2)
```




```
plot(mod1,which=5,pch=19)
```

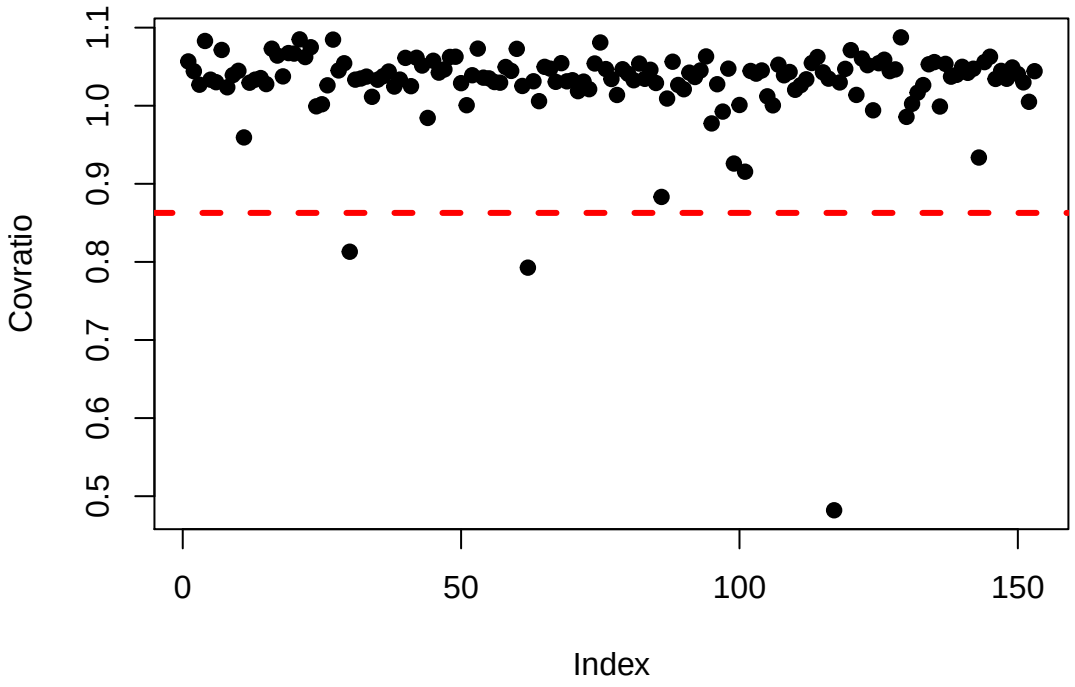



```r
#-- R CODE
par(mfrow=c(2,2))
plot(d$Temp,resid(mod1),pch=19,xlab="Temp",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$Wind,resid(mod1),pch=19,xlab="Wind",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$Solar.R,resid(mod1),pch=19,xlab="Solar",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(1:nrow(d),rstudent(mod1),pch=19,xlab="Observation Index",ylab="Residual Studentized",type="h")
abline(h=2,lwd=3,lty=2,col=2)
abline(h=-2,lwd=3,lty=2,col=2)
```

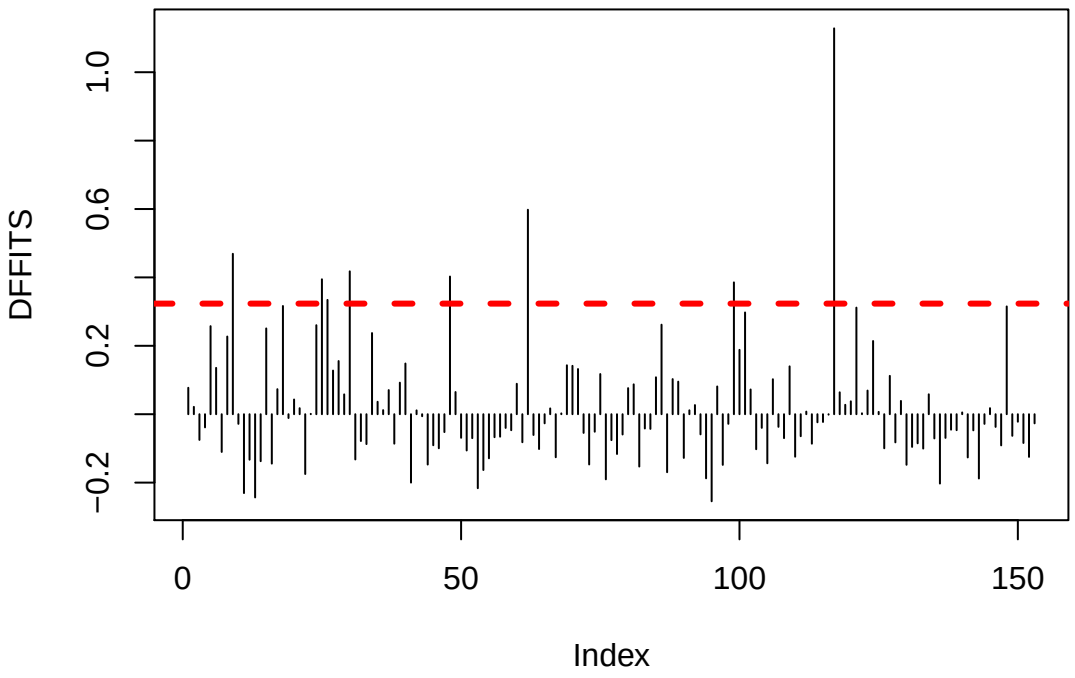



> >

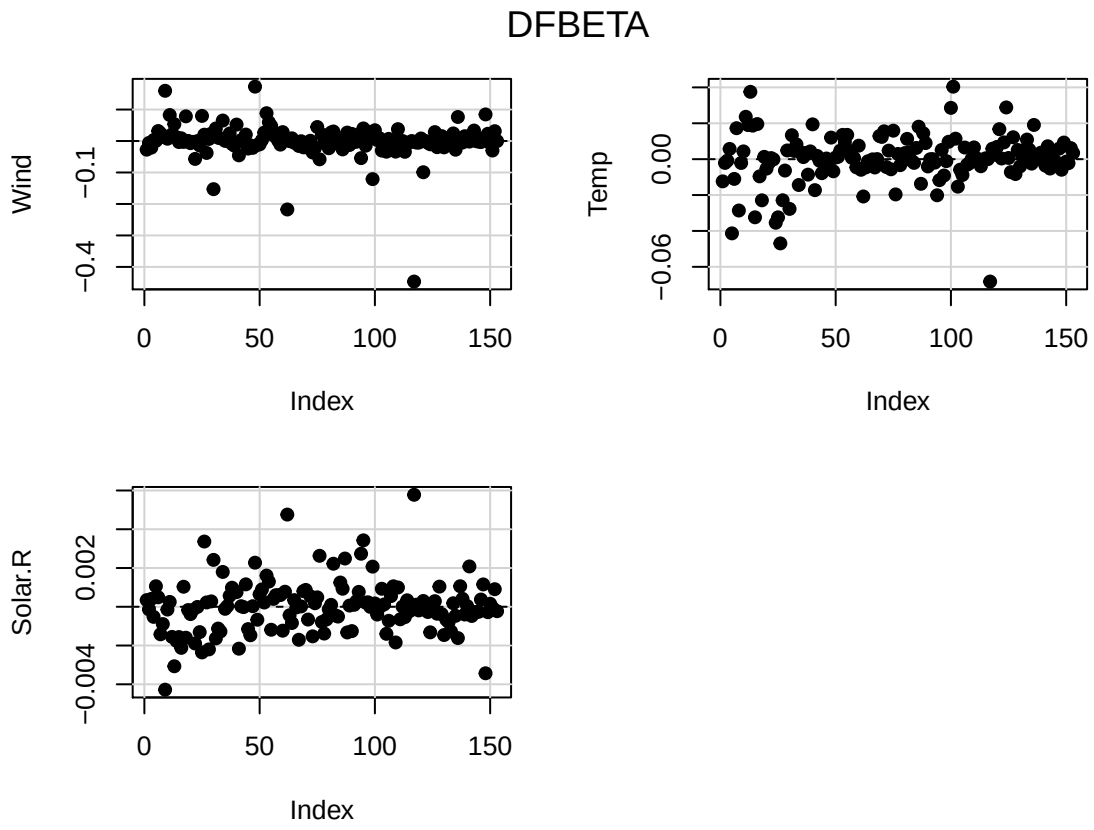
```
#-- R CODE

plot(covratio(mod1),pch=19,ylab="Covratio")
abline(h=1-3*7/nrow(d),lwd=3,col=2,lty=2)
abline(h=1+3*7/nrow(d),lwd=3,col=2,lty=2)
```




```
plot(dffits(mod1),pch=19,ylab="DFFITS",type="h")
abline(h=2*sqrt(4/nrow(d)),lwd=3,col=2,lty=2)
abline(h=-2*sqrt(4/nrow(d)),lwd=3,col=2,lty=2)
```




```
dfbetaPlots(mod1,pch=19,main="DFBETA")
```




Considerazioni generali:

1. Dal QQ-Plot e dagli altri grafici si notano anomalie alle estremità della distribuzione.
2. Dall'analisi del leverage plot si notano alcuni valori al di fuori dalla banda che identifica i valori critici data da 2 volte il numero dei regressori diviso n.
3. Dall'analisi dei residui studentizzati si osservano alcuni valori al di fuori dalla banda che identifica i valori critici.
4. Dall'analisi dei DFITTS (pe misurare l'influenza delle singole osservazioni sul coefficiente di regressione e sulla loro varianza quando è rimosso dal processo di stima) anche in questo caso vi sono valori oltre la soglia di tolleranza.
5. La presenza di valori anomali è confermata anche dai DFBETA

Eliminiamo quindi le osservazioni: 30, 62, 86, 99, 101, 117, 9 e 48.

```r
#-- R CODE
d1 <- d[-c(30, 62, 86, 99, 101, 117, 9, 48),]
mod1 <- lm(Ozone ~ Wind + Temp + Solar.R,d1)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -77.91 | 13.48 | -5.781 | 4.571e-08 |
| **Wind** | -1.773 | 0.4096 | -4.328 | 2.831e-05 |
| **Temp** | 1.591 | 0.1487 | 10.7 | 6.333e-20 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **Solar.R** | 0.05667 | 0.01439 | 3.937 | 0.0001293 |

Table 12: Fitting linear model: Ozone ~ Wind + Temp + Solar.R

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 145 | 14.65 | 0.6564 | 0.6491 |

```
pander(anova(mod1),big.mark=",")
```

Table 13: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **Wind** | 1 | 22,058 | 22,058 | 102.7 | 1.797e-18 |
| **Temp** | 1 | 32,457 | 32,457 | 151.1 | 4.6e-24 |
| **Solar.R** | 1 | 3,328 | 3,328 | 15.5 | 0.0001293 |
| **Residuals** | 141 | 30,279 | 214.7 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 21.01 | 2.741e-05 |

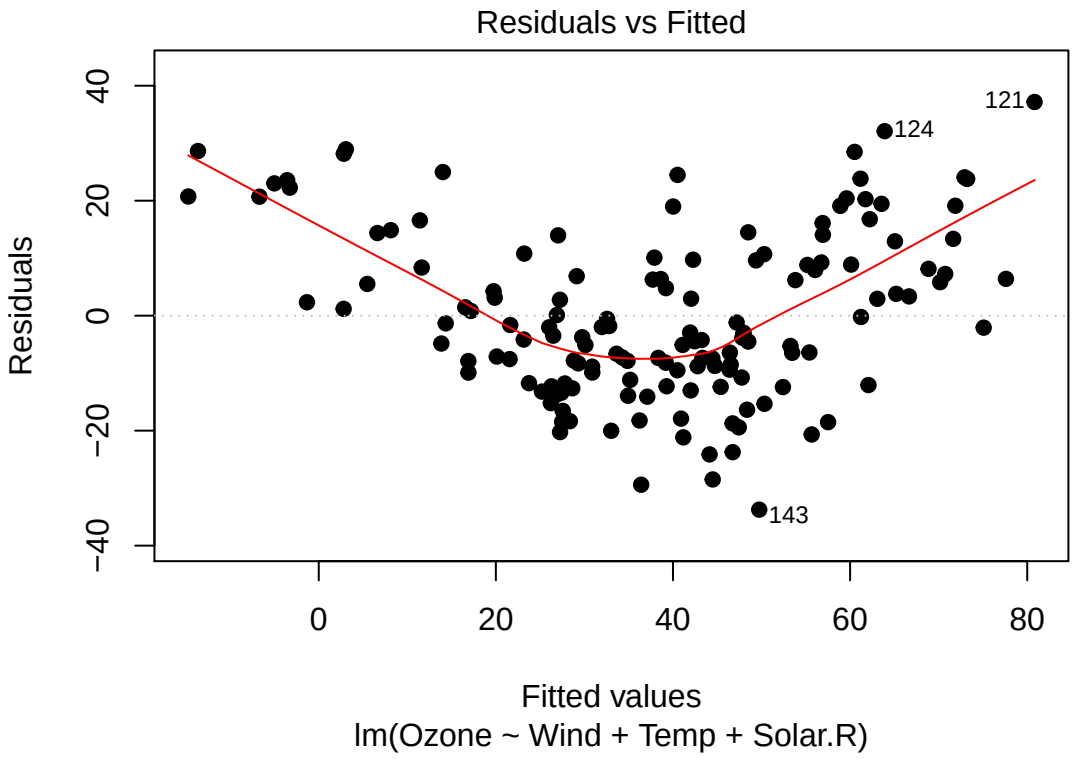
```
pander(dwtest(mod1),big.mark=",")
```

Table 15: Durbin-Watson test: `mod1`

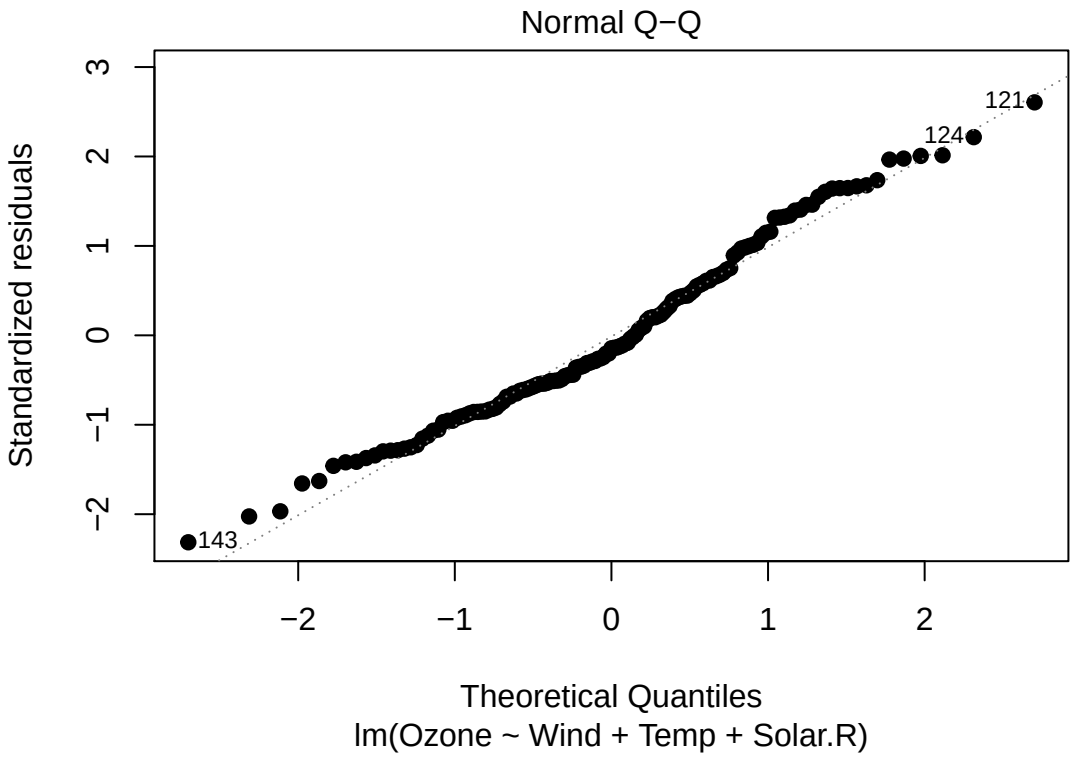
| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.585 | 0.004512 * * | true autocorrelation is greater than 0 |

Migliora l'$R^2$ che sale a 0.65. Si verifica ora su questo modello con 145 osservazioni la normalità dei residui analizzando la distribuzione dei residui e il loro box-plot.

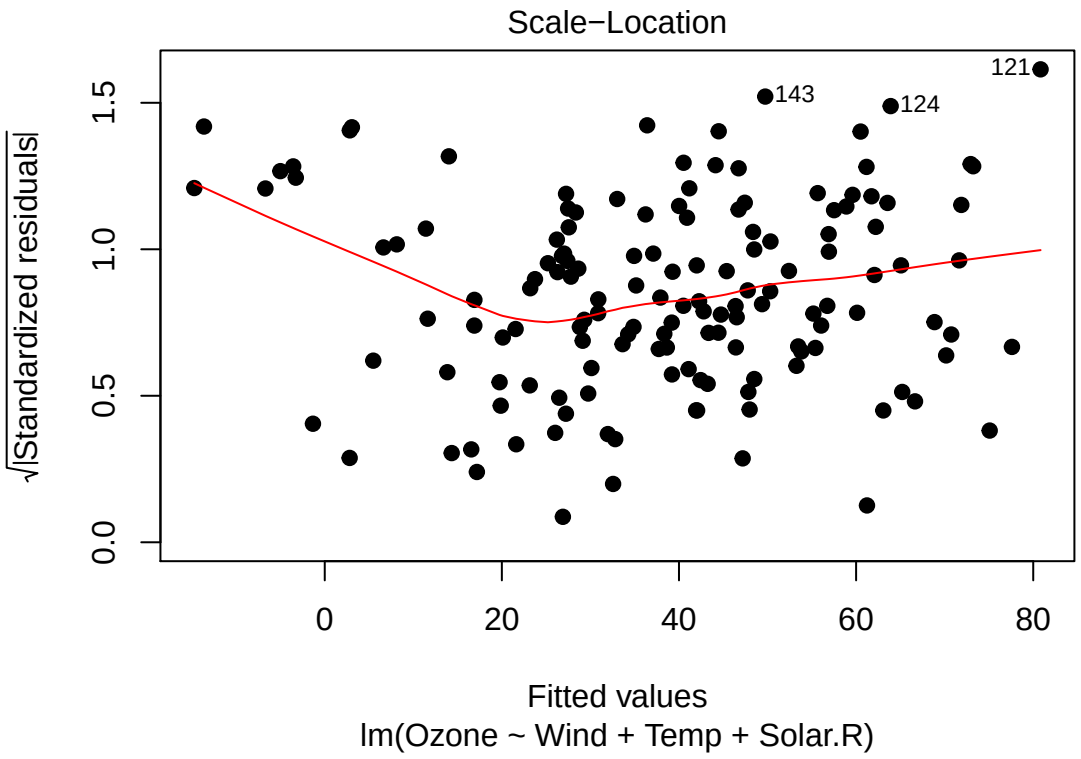
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



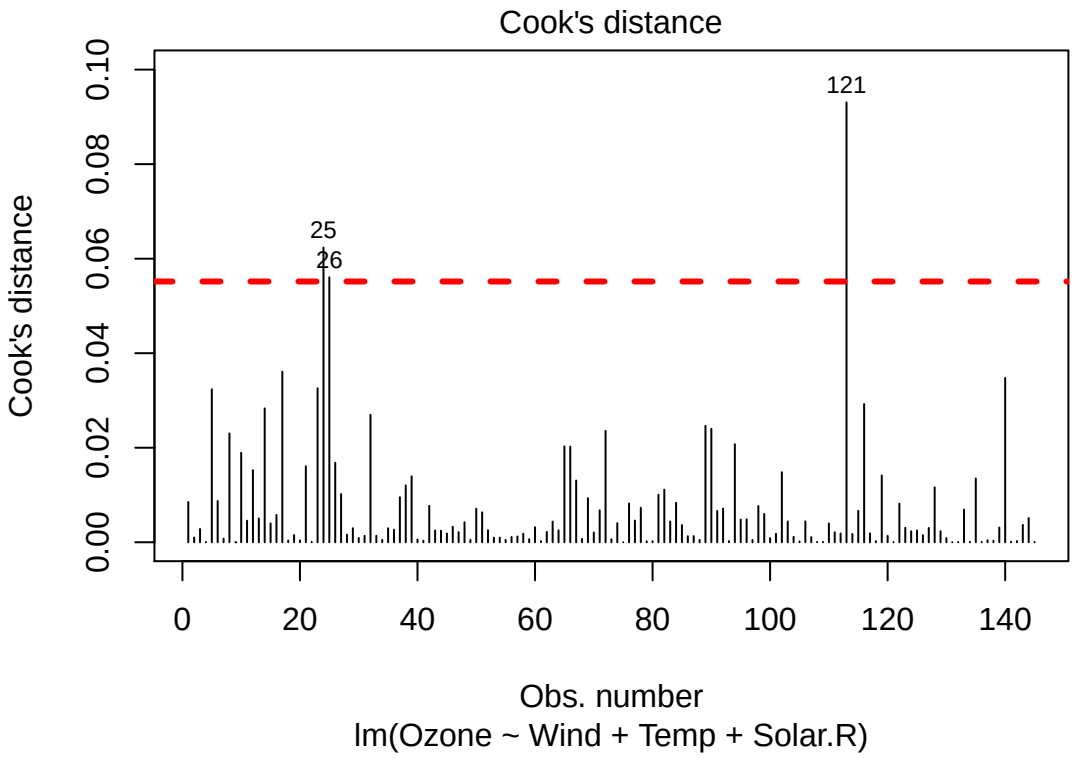

```
plot(mod1,which=2,pch=19)
```



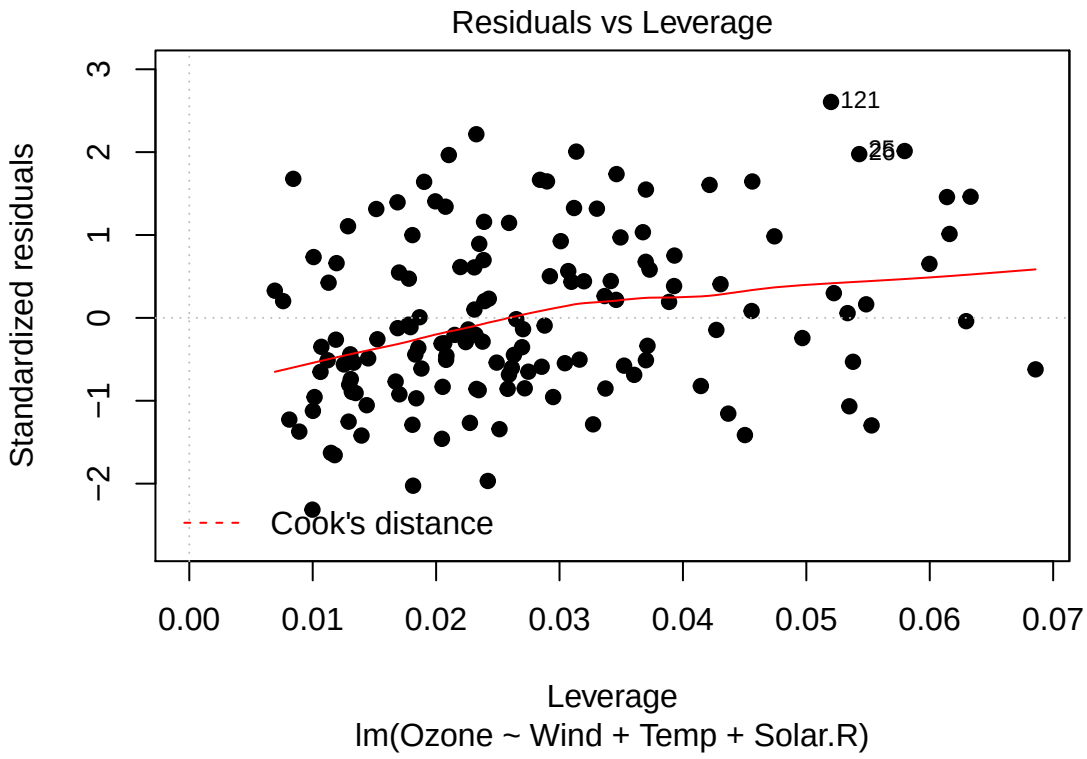

```
plot(mod1,which=3,pch=19)
```



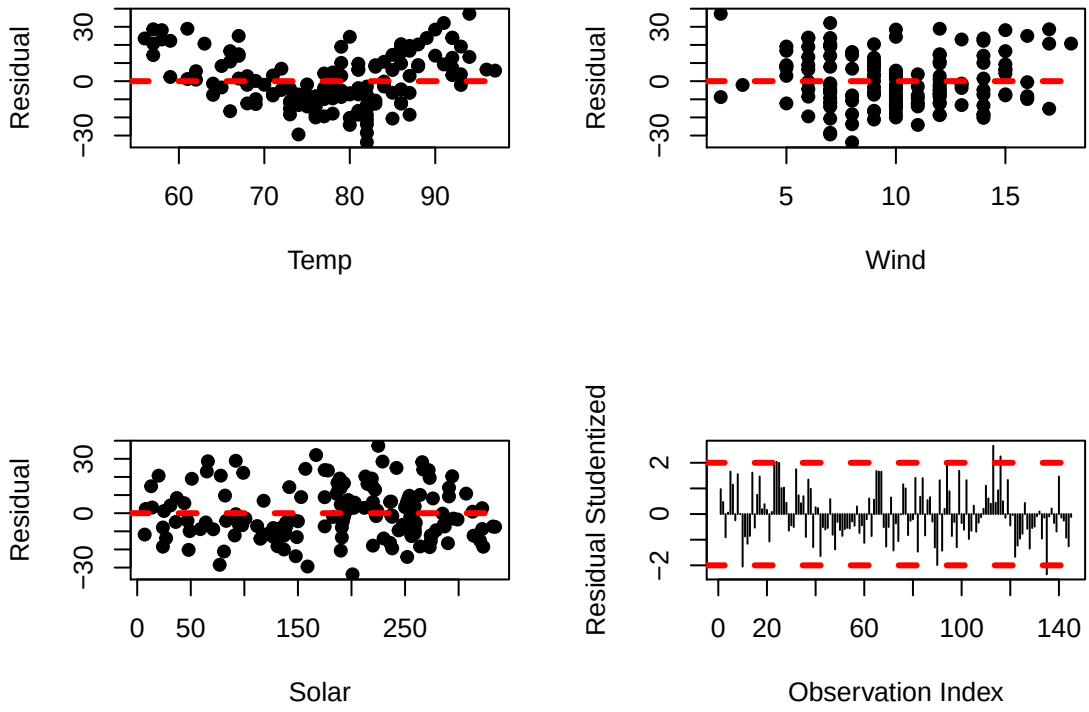

```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d1),col=2,lwd=3,lty=2)
```




```
plot(mod1,which=5,pch=19)
```

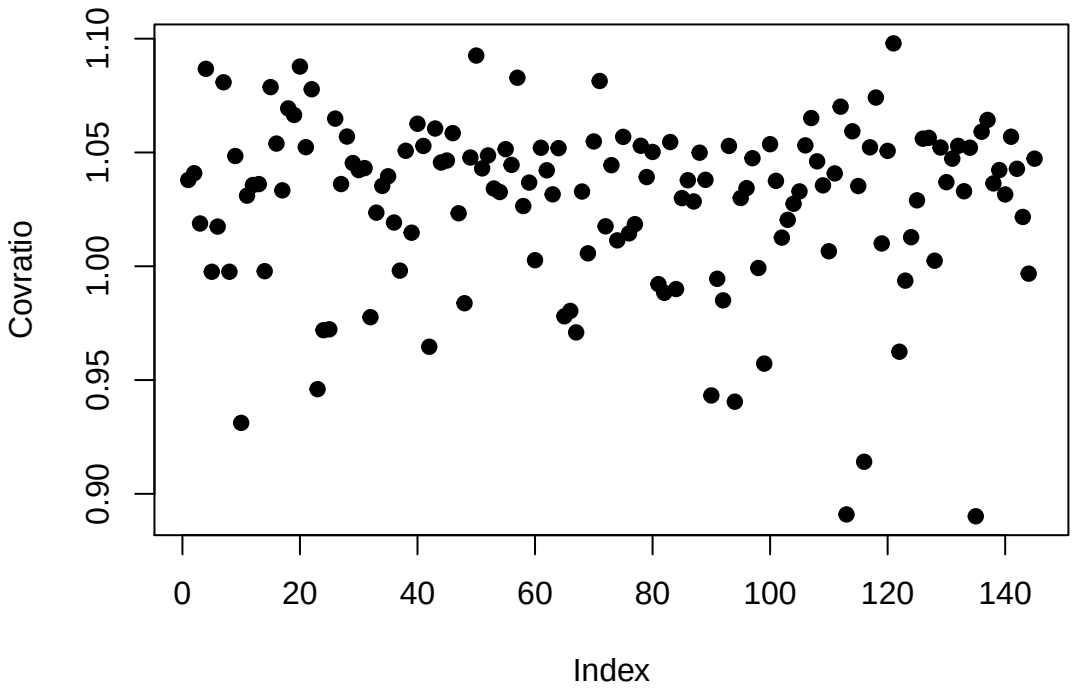



```r
#-- R CODE
par(mfrow=c(2,2))
plot(d1$Temp,resid(mod1),pch=19,xlab="Temp",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d1$Wind,resid(mod1),pch=19,xlab="Wind",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d1$Solar.R,resid(mod1),pch=19,xlab="Solar",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(1:nrow(d1),rstudent(mod1),pch=19,xlab="Observation Index",ylab="Residual Studentized",type="h")
abline(h=2,lwd=3,lty=2,col=2)
abline(h=-2,lwd=3,lty=2,col=2)
```

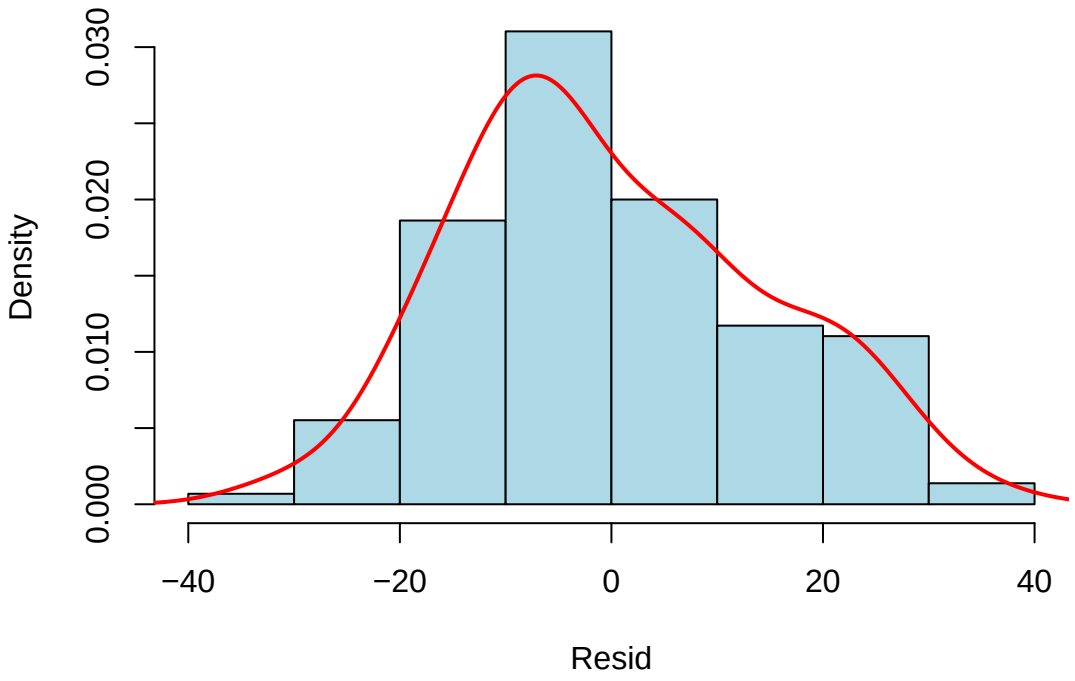



> >

```
#-- R CODE

plot(covratio(mod1),pch=19,ylab="Covratio")
abline(h=1-3*7/nrow(d1),lwd=3,col=2,lty=2)
abline(h=1+3*7/nrow(d1),lwd=3,col=2,lty=2)
```




```
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```




```
pander(shapiro.test(resid(mod1)))
```

Table 16: Shapiro-Wilk normality test: `resid(mod1)`

| Test statistic | P value |
|---|---|
| 0.9805 | 0.03651 * |

```
pander(ks.test(resid(mod1),"pnorm"))
```

Table 17: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 0.4949 | 0 * * * | two-sided |

Si può quindi concludere che la distribuzione si discosta dalla normalità ma in modo non rilevante e quindi si può accettare l'ipotesi di normalità se non si ritiene di non voler essere troppo stringenti nelle condizioni per accettare la normalità.

L'analisi residui-valori predetti e quella dei residui inerenti regressioni uni variate rispetto ai singoli regressori mostrano residui che si collocano in modo non regolare intorno allo 0, non certo secondo una forma rettangolare. Per le osservazioni estreme i residui sembrano molto discosti dal valore zero a differenza che in centro della distribuzione a segnalare la probabile non sfericità degli errori.

Il p-value del p-value del test di Dubin Whatson ci porta a rifiutare l'ipotesi nulla di incorrelazione fra i

residui.

```
#-- R CODE
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 21.01 | 2.741e-05 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 19: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.585 | 0.004512 * * | true autocorrelation is greater than 0 |