LINEAR 10 - Data set: PM10

INTRODUZIONE

Il data set contiene informazioni relative all'inquinamento dell'aria rilevato in corrispondenza delle strade e degli snodi stradali principali, e contiene le seguenti variabili:

1. PM10: concentrazione di particelle di pm10
2. CARS: numero di auto che transitano in un'ora
3. TEMP: temperatura misurata a 2 metri dal suolo
4. WIND: velocità del vento (metri/secondo)
5. D_TEMP: differenza tra temperatura misurata a 25 metri dal suolo e a 2 metri dal suolo
6. HOURS: numero di ore trascorse dalla mezzanotte del giorno di rilevazione

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare

```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)\\4.tutto\\PM_10.csv"),sep=";")
d$cars <- as.numeric(gsub(",","",d$cars))

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("pm10","cars","temp","wind","d_temp")

#-- print delle prime 6 righe del dataset
```

```r
pander(head(d),big.mark=",")
```

id_road	pm10	cars	temp	wind	d_temp	hours
1	39	2,308	-4.4	4.2	0	19
2	21	3,084	-5.7	4.8	-0.3	9
3	41	110	-13.5	4.3	0.2	3
4	19	1,854	1.4	3	0.1	22
5	58	2,351	4.1	5.6	1.1	7
6	40	2,662	5.8	2.3	-0.1	9

STATISTICHE DESCRITTIVE

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

Table 2: Table continues below

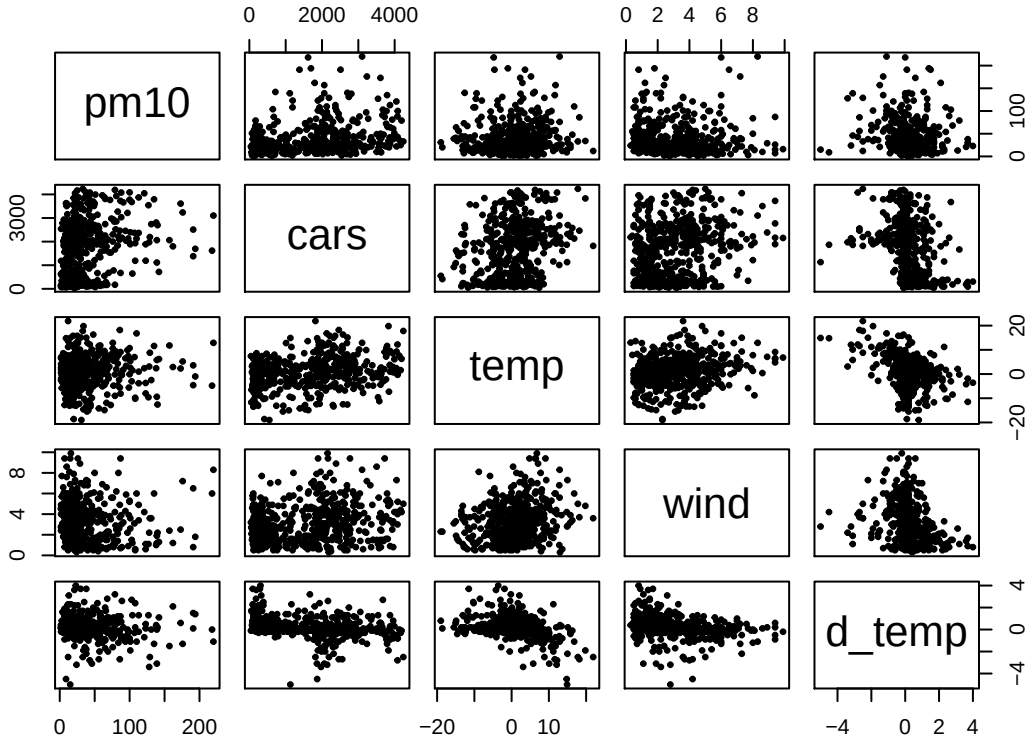
pm10	cars	temp	wind
Min. : 2.00	Min. : 45.0	Min. :-19.000	Min. :0.300
1st Qu.: 15.00	1st Qu.: 523.5	1st Qu.: -3.125	1st Qu.:1.700
Median : 27.00	Median :1851.5	Median : 1.000	Median :3.000
Mean : 37.88	Mean :1683.2	Mean : 0.787	Mean :3.211
3rd Qu.: 48.00	3rd Qu.:2509.5	3rd Qu.: 4.225	3rd Qu.:4.300
Max. :220.00	Max. :4239.0	Max. : 21.900	Max. :9.900

d_temp
Min. :-5.0000
1st Qu.:-0.2000
Median : 0.1000
Mean : 0.1548
3rd Qu.: 0.6000
Max. : 4.0000

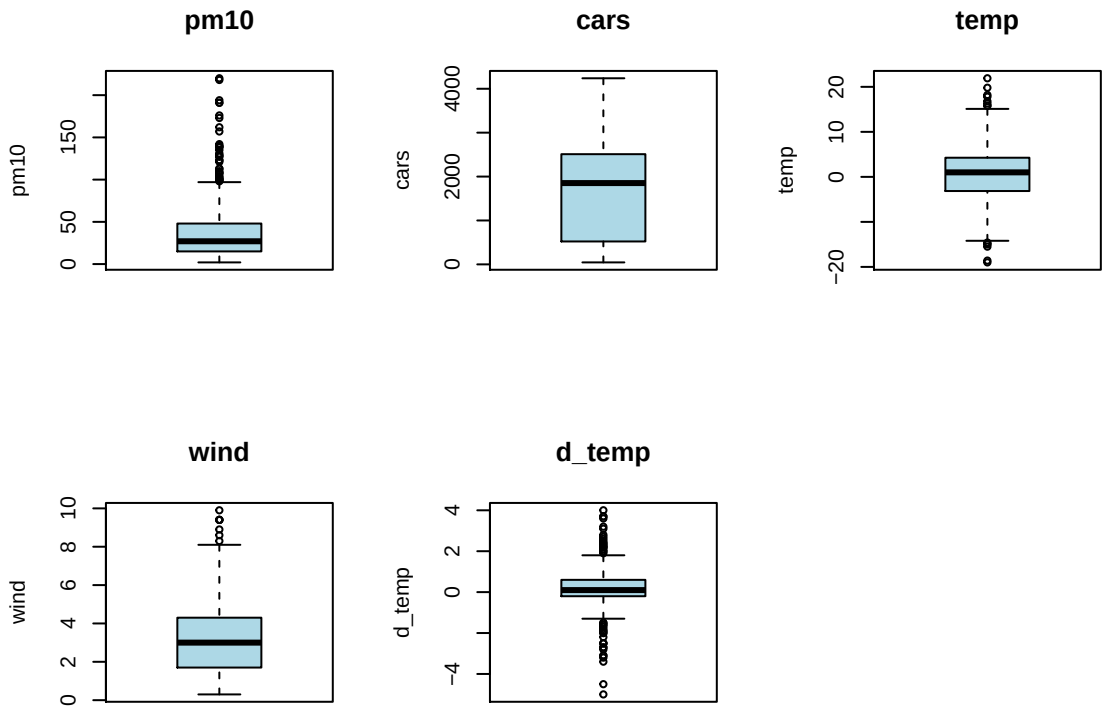
```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

	pm10	cars	temp	wind	d_temp
pm10	1	0.3009	0.06322	-0.06704	-0.07717
cars	0.3009	1	0.2641	0.2026	-0.3154
temp	0.06322	0.2641	1	0.2136	-0.358
wind	-0.06704	0.2026	0.2136	1	-0.2721
d_temp	-0.07717	-0.3154	-0.358	-0.2721	1

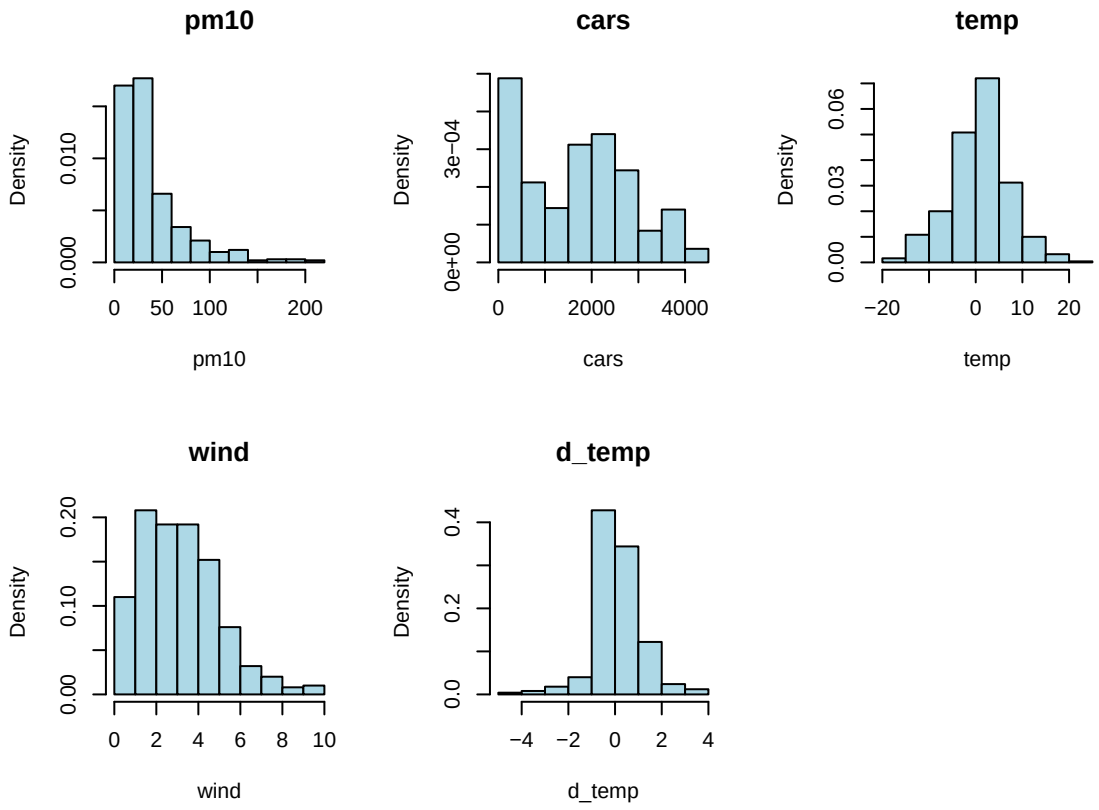
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```r
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
par(mfrow=c(2,3))
```




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



REGRESSIONE

Non appaiono particolari forti correlazioni fra le variabili. Verificando con il condition index si vede del resto che non esiste collinearità.

```
#-- R CODE
mod1 <- lm(pm10~cars + temp + wind + d_temp + hours,d)
pander(ols_eigen_cindex(mod1),big.mark=",")
```

Table 5: Table continues below

Eigenvalue	Condition Index	intercept	cars	temp	wind
3.522	1	0.00877	0.01635	0.004786	0.01514
1.346	1.617	0.001717	0.0006934	0.2666	1.304e-05
0.6293	2.366	0.0001854	0.003958	0.7048	0.002952
0.2538	3.725	0.009812	0.2082	0.0008305	0.5084
0.1646	4.625	0.01945	0.7568	0.006556	0.01598
0.0843	6.464	0.9601	0.01407	0.01645	0.4575

d_temp	hours
0.0001144	0.01381
0.3048	0.001084

d_temp	hours
0.5388	0.00014
0.0001368	0.1458
0.08834	0.5089
0.06778	0.3303

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.31	4.278	5.917	6.136e-09
cars	0.008741	0.001483	5.893	7.037e-09
temp	-0.008703	0.2542	-0.03424	0.9727
wind	-2.497	0.8411	-2.969	0.003137
d_temp	-0.5832	1.689	-0.3453	0.73
hours	0.4792	0.2339	2.049	0.041

Table 8: Fitting linear model: pm10 ~ cars + temp + wind + d_temp + hours

Observations	Residual Std. Error	R^2	Adjusted R^2
500	33.13	0.1153	0.1063

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

Table 9: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cars	1	55,490	55,490	50.55	4.086e-12
temp	1	173.9	173.9	0.1585	0.6907
wind	1	10,317	10,317	9.399	0.00229
d_temp	1	58.43	58.43	0.05323	0.8176
hours	1	4,608	4,608	4.198	0.041
Residuals	494	542,229	1,098	NA	NA

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

Test.statistic	P.value
12.48	0.001945

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

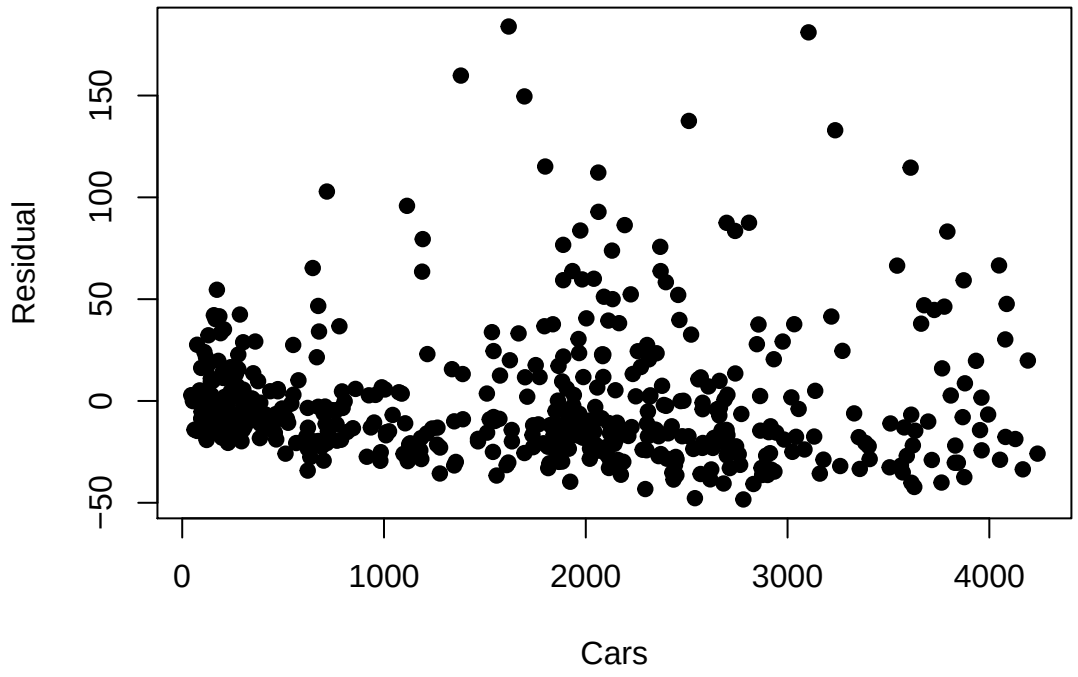
Table 11: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.907	0.1489	true autocorrelation is greater than 0

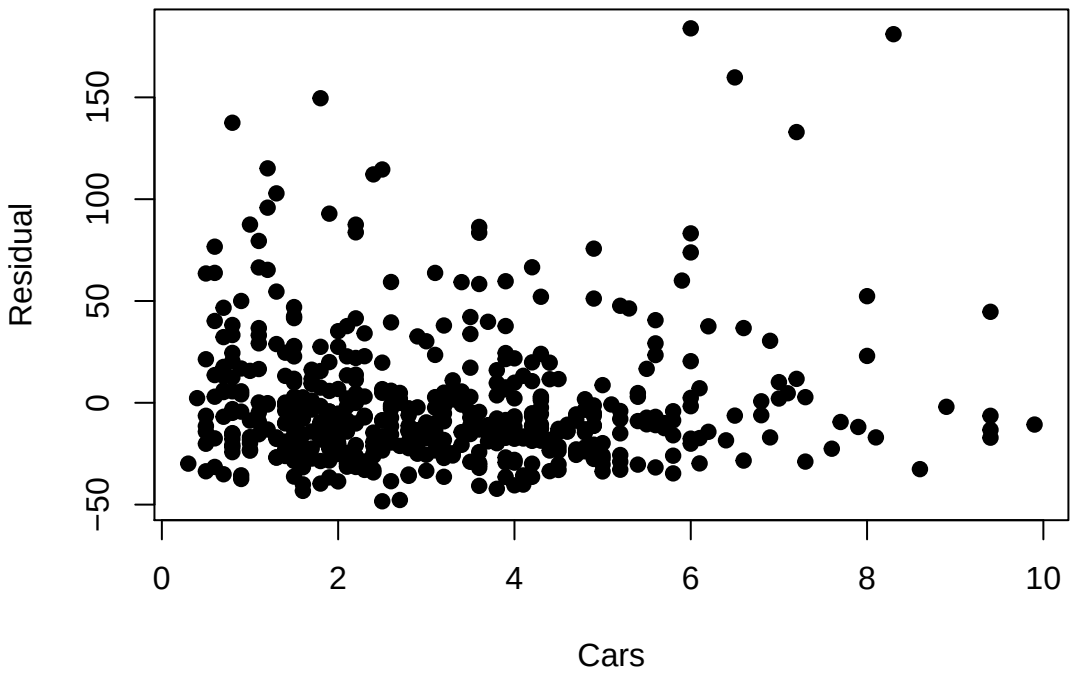
Test statistic	P value	Alternative hypothesis

Si analizza invece attraverso i residui l'eteroschedasticità degli errori

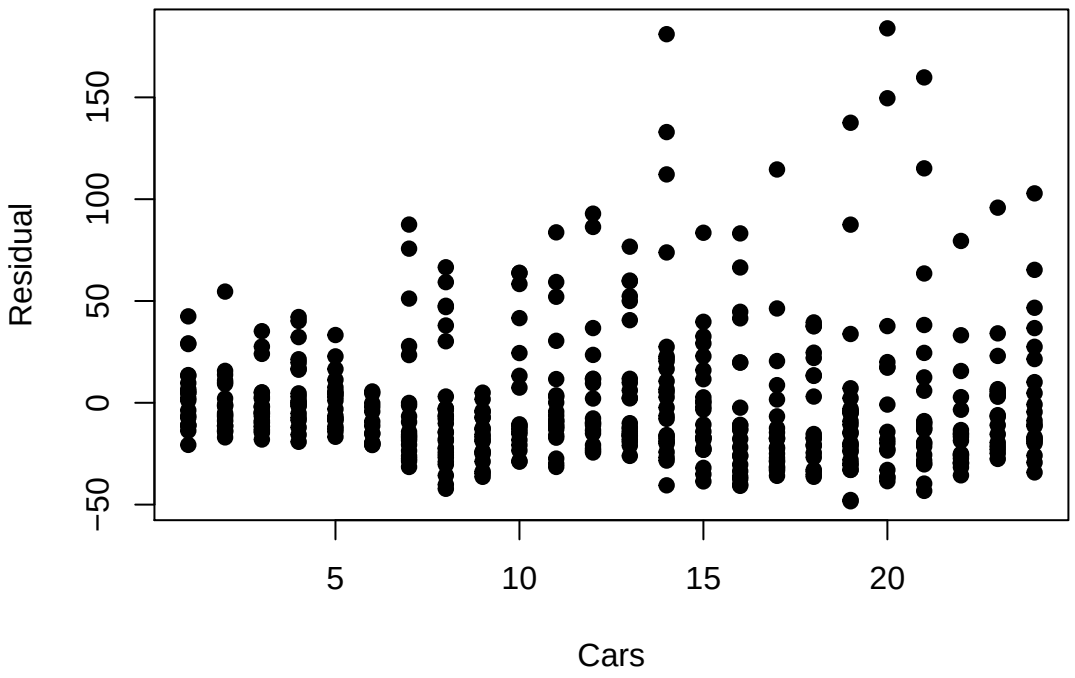
```
#-- R CODE
plot(d$cars,resid(mod1),pch=19,xlab="Cars",ylab="Residual")
```




```
plot(d$wind,resid(mod1),pch=19,xlab="Cars",ylab="Residual")
```




```
plot(d$hours,resid(mod1),pch=19,xlab="Cars",ylab="Residual")
```



La configurazione dei residui è bene lontana da una forma rettangolare porta a ipotizzare eteroschedasticità degli errori stessi.

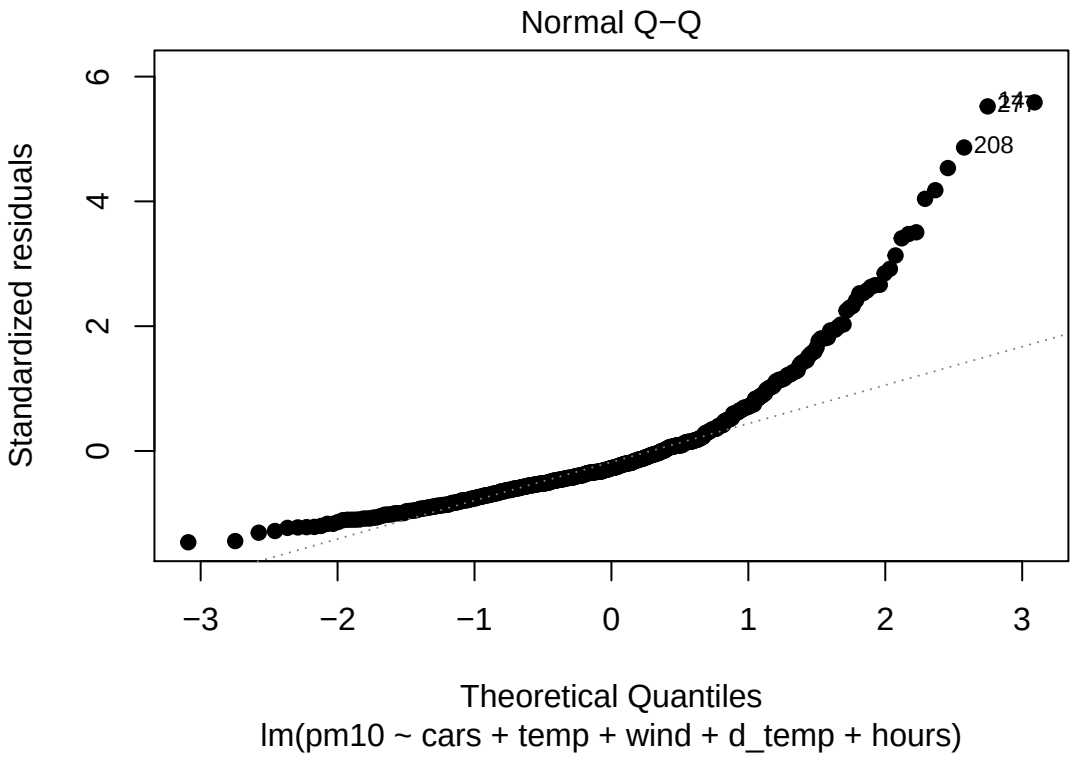
Tale eteroschedasticità è confermata anche dal test di White che respinge nettamente l'ipotesi di omoschdasticità.

Il test Durbin Watson non respinge invece l'ipotesi di incorrelazione fra gli errori come si vede dal valore della statistica test e dal p-value.

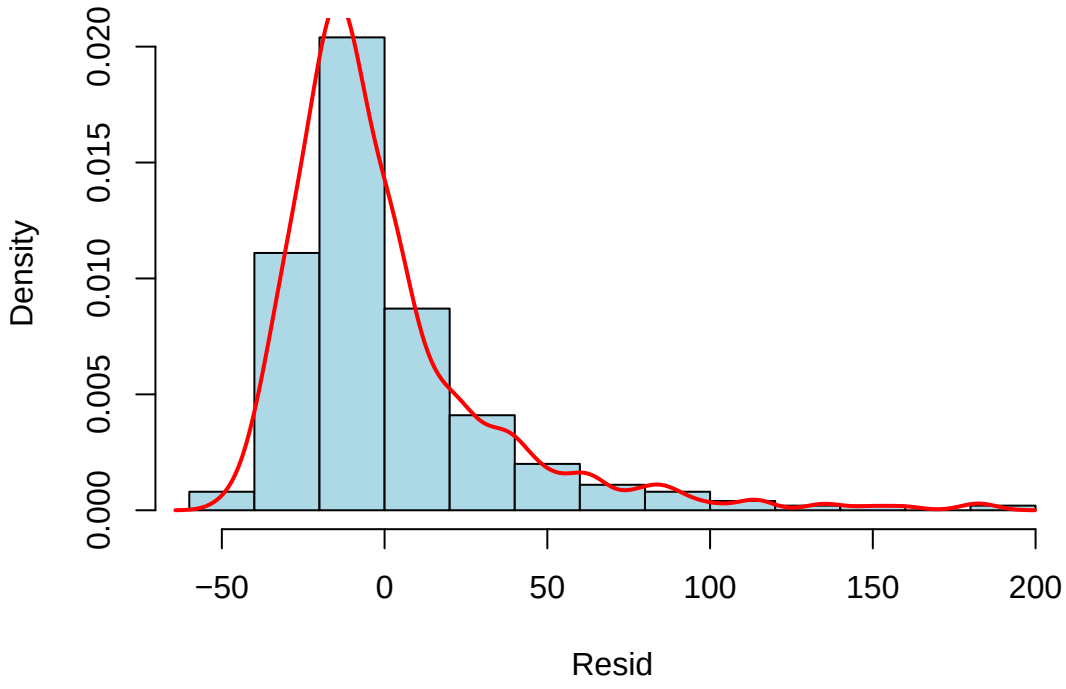
Per ciò che concerne la normalità sia la distribuzione dei residui, che il Q-Q plot che i test sulla normalità mostrano che i residui stessi non sono normali.

```
#-- R CODE

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




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




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

Table 12: Shapiro-Wilk normality test: `resid(mod1)`

Test statistic	P value
0.8084	6.627e-24 * * *

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

Table 13: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

Test statistic	P value	Alternative hypothesis
0.6069	0 * * *	two-sided