NORM_COL 3 - Data set: COLLEGES

INTRODUZIONE

Il data set contiene informazioni riguardanti 521 università americane alla fine dell'anno accademico 1993/1994. Le variabili contenute sono:

1. AVE_MAT: indicatore qualitativo della preparazione nelle discipline matematiche
2. APPL_RIC: numero di domande di iscrizione ricevute all'inizio dell'anno
3. APPL_ACC: numero di domande di iscrizione accettate all'inizio dell'anno
4. P_STUD10: percentuale di studenti procenienti dalle prime 10 scuole superiori americane
5. COSTI_V: costi medi pro-capite per vitto, alloggio sostenuti nell'anno (dollari)
6. COSTI_B: costi medi pro-capite per l'acquisto di libri di testo sostenuti nell'anno (dollari)
7. TASSE: tasse universitarie medie pro-capite versate durante l'anno
8. STUD_DOC: numero di studenti per docente
9. P_LAUR: percentuale di laureati alla fine dell'anno sul totale degli iscritti al primo anno

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\\2.Norm-Col copy(3)\\3.Norm-Col\\colleges.cs

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- names(d)[2:ncol(d)]
```

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

Table 1: Table continues below

id_numb	ave_MAT	appl_ric	appl_acc	p_stud10	costi_v	costi_b
1,061	490	193	146	16	4,120	800
1,009	575	7,548	6,791	25	3,933	600
1,012	575	805	588	67	4,325	400
1,019	513	608	520	26	3,920	500
1,047	510	1,471	1,281	18	2,570	300
1,099	564	823	721	52	3,195	500

tasse	stud_doc	p_laur
10,922	11.9	15
6,642	16.7	69
8,649	14	72
7,703	11.4	44
4,295	23	48
8,588	13.1	63

STATISTICHE DESCRITTIVE

Si presentano innanzitutto le statistiche descrittive.

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

Table 3: Table continues below

ave_MAT	appl_ric	appl_acc	p_stud10	costi_v
Min. :320.0	Min. : 77	Min. : 61	Min. : 1.00	Min. :1780
1st Qu.:470.0	1st Qu.: 738	1st Qu.: 588	1st Qu.:14.00	1st Qu.:3680
Median :509.0	Median : 1456	Median : 1074	Median :21.00	Median :4240
Mean :506.4	Mean : 2821	Mean : 1999	Mean :23.64	Mean :4389
3rd Qu.:541.0	3rd Qu.: 3500	3rd Qu.: 2424	3rd Qu.:30.00	3rd Qu.:4960
Max. :660.0	Max. :48094	Max. :26330	Max. :94.00	Max. :7782

costi_b	tasse	stud_doc	p_laur
Min. : 96.0	Min. : 3190	Min. : 2.50	Min. : 10.0
1st Qu.:450.0	1st Qu.: 6735	1st Qu.:12.20	1st Qu.: 52.0
Median :500.0	Median : 8135	Median :14.10	Median : 63.0
Mean :530.6	Mean : 8532	Mean :14.73	Mean : 62.3
3rd Qu.:600.0	3rd Qu.: 9995	3rd Qu.:16.90	3rd Qu.: 74.0
Max. :900.0	Max. :22704	Max. :28.80	Max. :118.0

costi_b	tasse	stud_doc	p_laur

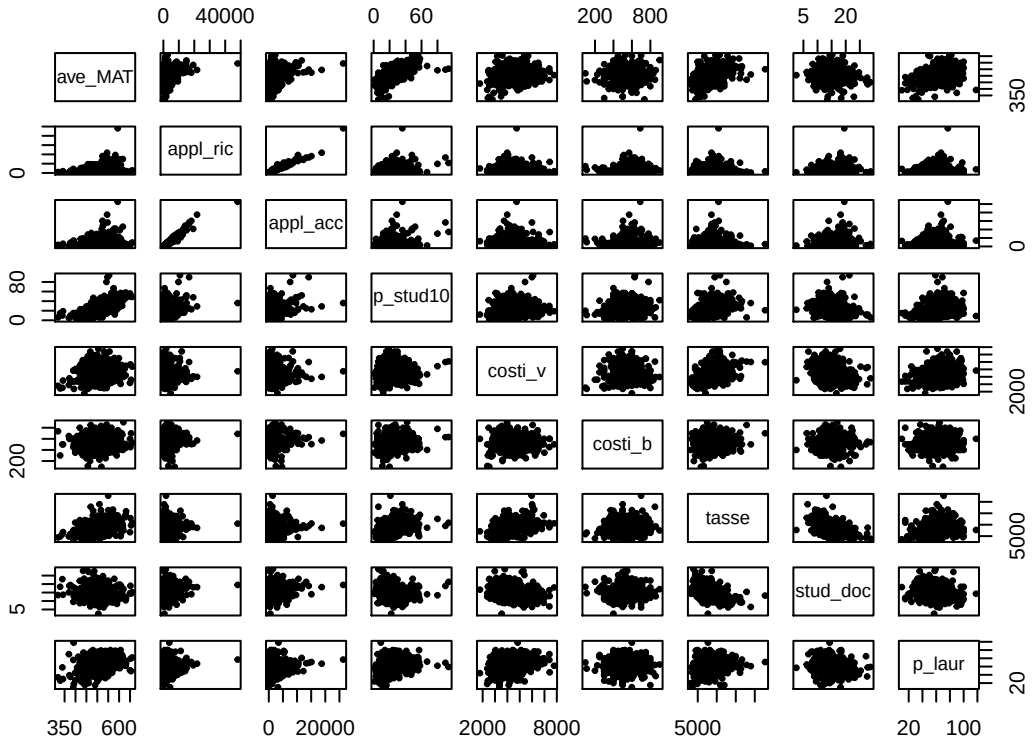
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 5: Table continues below

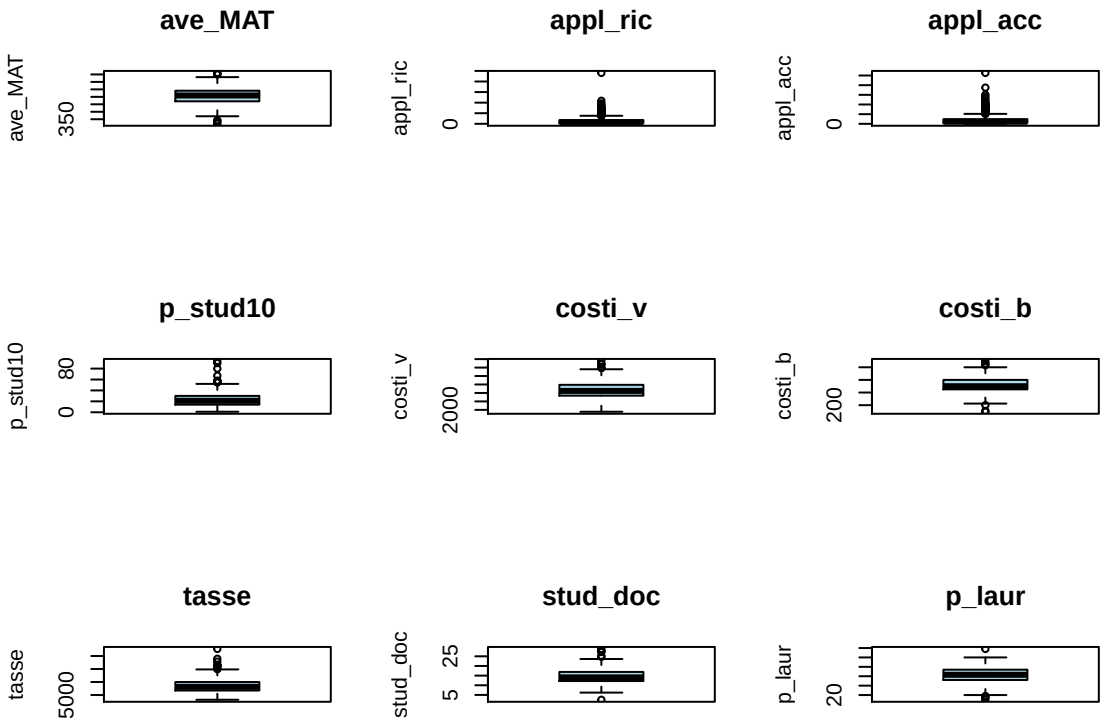
	ave_MAT	appl_ric	appl_acc	p_stud10	costi_v	costi_b
ave_MAT	1	0.2747	0.27	0.6926	0.1911	0.1146
appl_ric	0.2747	1	0.9703	0.1733	0.02063	0.1885
appl_acc	0.27	0.9703	1	0.1885	0.02449	0.1896
p_stud10	0.6926	0.1733	0.1885	1	0.1285	0.1425
costi_v	0.1911	0.02063	0.02449	0.1285	1	0.0624
costi_b	0.1146	0.1885	0.1896	0.1425	0.0624	1
tasse	0.434	0.008766	0.02133	0.3932	0.4434	0.09308
stud_doc	-0.08992	0.3196	0.3027	-0.1596	-0.1918	0.05982
p_laur	0.3847	0.006031	0.01002	0.3162	0.3457	-0.05675

	tasse	stud_doc	p_laur
ave_MAT	0.434	-0.08992	0.3847
appl_ric	0.008766	0.3196	0.006031
appl_acc	0.02133	0.3027	0.01002
p_stud10	0.3932	-0.1596	0.3162
costi_v	0.4434	-0.1918	0.3457
costi_b	0.09308	0.05982	-0.05675
tasse	1	-0.5183	0.2604
stud_doc	-0.5183	1	-0.2121
p_laur	0.2604	-0.2121	1

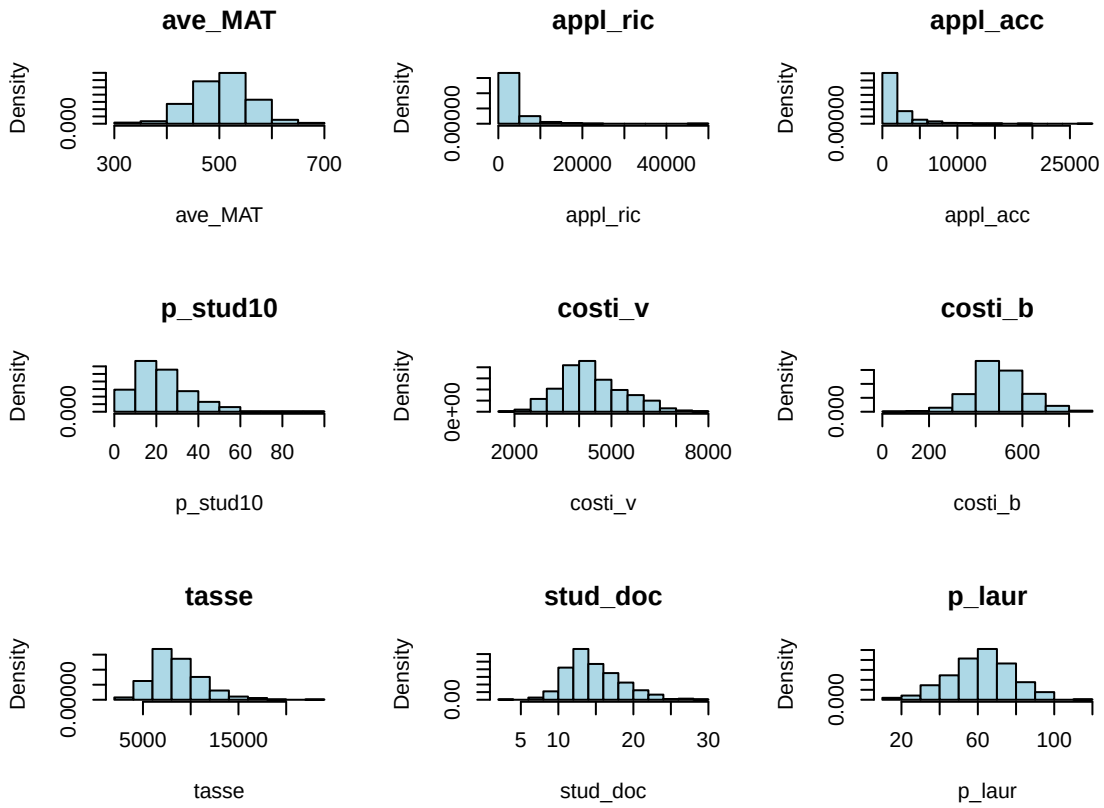
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




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




```
par(mfrow=c(3,3))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```



Si effettua quindi la regressione della variabile "appl_acc" rispetto ai regressori prescelti.

REGRESSIONE

```r
#-- R CODE
mod1 <- lm(appl_acc ~ ave_MAT + appl_ric + p_stud10, d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	485	320	1.515	0.1303
ave_MAT	-0.9691	0.7223	-1.342	0.1803
appl_ric	0.6542	0.007427	88.09	1.699e-313
p_stud10	6.75	2.884	2.341	0.01964

Table 8: Fitting linear model: appl_acc ~ ave_MAT + appl_ric + p_stud10

Observations	Residual Std. Error	R^2	Adjusted R^2
521	618.7	0.9421	0.9418

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

Table 9: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ave_MAT	1	249,252,364	249,252,364	651.1	1.473e-93
appl_ric	1	2.969e+09	2.969e+09	7,754	2.014e-313
p_stud10	1	2,097,218	2,097,218	5.478	0.01964
Residuals	517	197,927,054	382,838	NA	NA

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

Test.statistic	P.value
389.5	0

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

Table 11: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.691	0.0001907 * * *	true autocorrelation is greater than 0

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

Variables	Tolerance	VIF
ave_MAT	0.4957	2.017
appl_ric	0.924	1.082
p_stud10	0.52	1.923

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

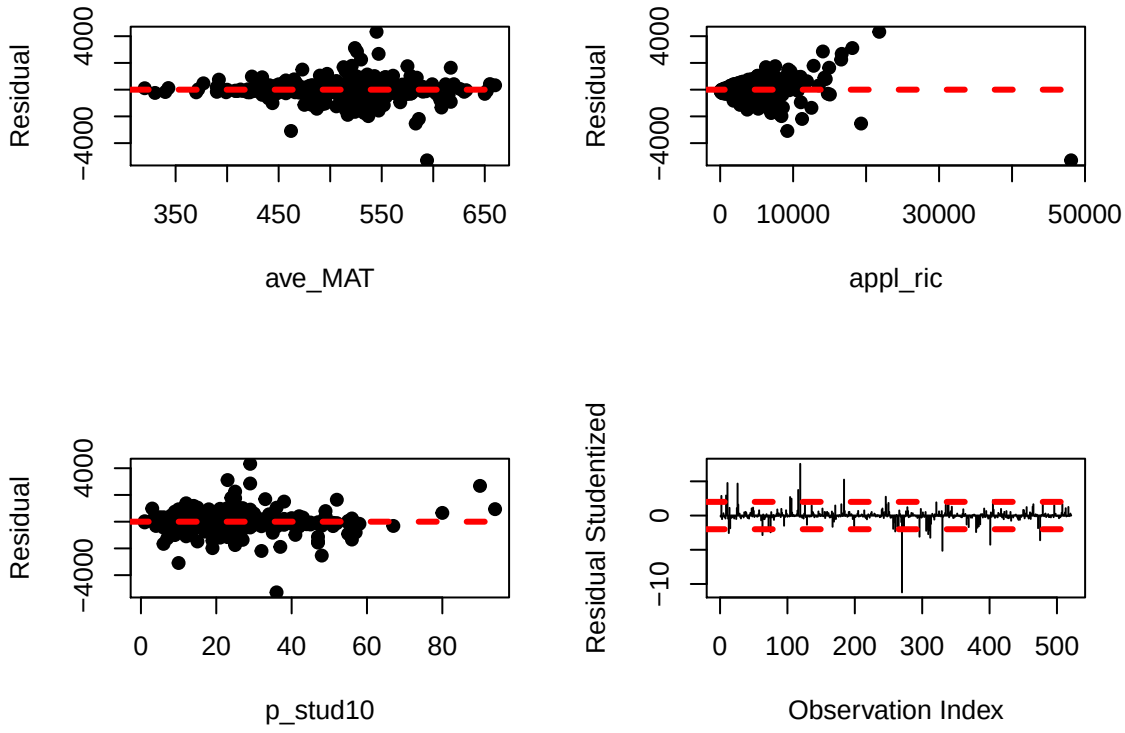
Eigenvalue	Condition Index	intercept	ave_MAT	appl_ric	p_stud10
3.319	1	0.0006063	0.0004713	0.03007	0.009646
0.5333	2.494	0.0008757	0.0005552	0.9289	0.01101
0.1451	4.783	0.01139	0.003456	0.0007515	0.566
0.003083	32.81	0.9871	0.9955	0.04032	0.4133

Si verifica ora l'omoschedasticità e incorrelazione degli errori cominciando con le rappresentazioni grafiche. Sia nel grafico dei valori osservati-previsti della variabile dipendente che in quello dei valori residui-previsti si nota una configurazione non omogenea della nuvola di punti a segnalare la probabile presenza di eteroschedasticità dei residui.

Tale eteroschedasticità sembra confermata dai grafici dei residui-valori osservati delle regressioni semplici con una sola variabile esplicativa per volta in cui esistono molti punti che si discostano dalla nuvola di punti. Si passa ora a esaminare i test sulla sfericità dei residui.

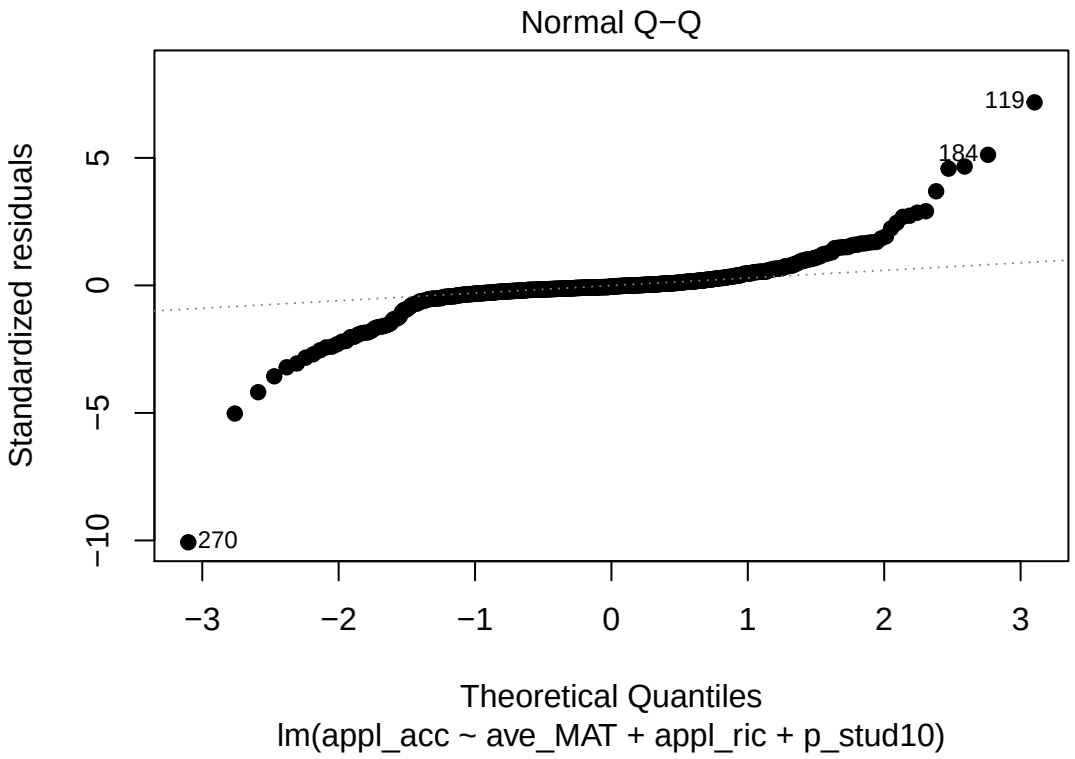
Il test di White porta a rigettare l'ipotesi di omoschedasticità. Si può quindi concludere che gli errori sono eteroschedastici e correlati.

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

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

plot(d$p_stud10,resid(mod1),pch=19,xlab="p_stud10",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)
```

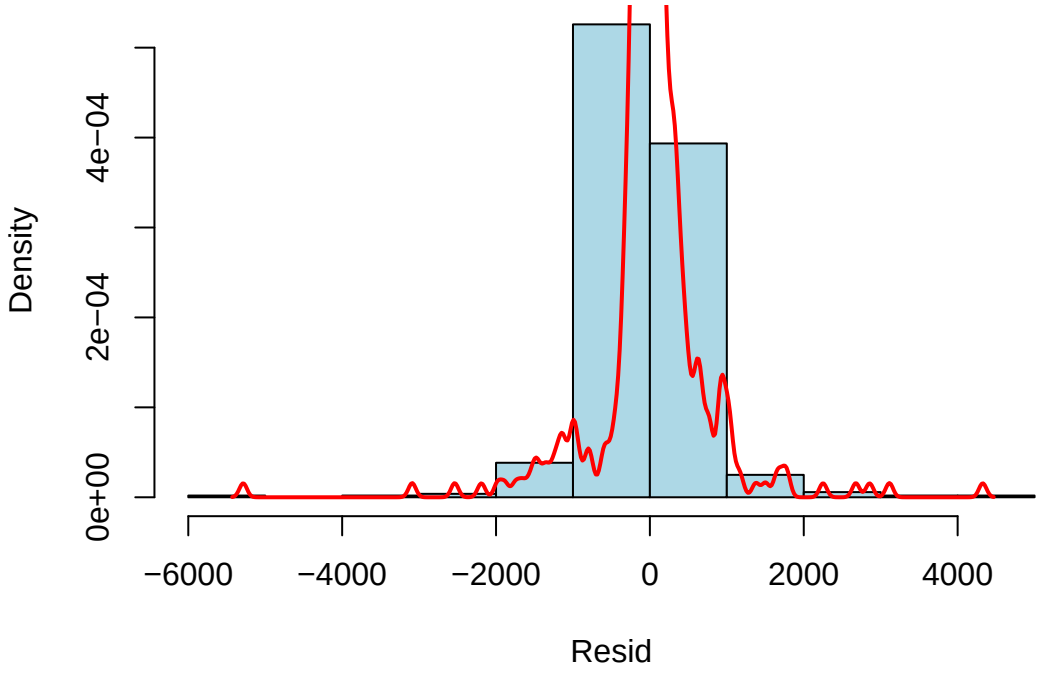


Si esamina ora la normalità dei residui cominciando con le rappresentazioni grafiche.

```
#-- 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 14: Shapiro-Wilk normality test: `resid(mod1)`

Test statistic	P value
0.7163	9.006e-29 * * *

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

Table 15: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

Test statistic	P value	Alternative hypothesis
0.564	0 * * *	two-sided

La distribuzione dei residui e il Q-Q plot mostrano chiaramente una situazione di non normalità confermata dal confronto tra quantili delle distribuzione empirica e teorica normale.

Tal non normalità è confermata dal grafico in cui si confrontano valori residui-predetti. Si nota come la nuvola di punti differisce molto dalla configurazione sferica o elittica tipica di una distribuzione normale degli errori.

Nel complesso quindi si hanno errori eteroschedastici, non normalità dei residui, presenza di outlier: si conclude che non è opportuno usare il modello lineare classico basato su OLS.