NORM_COL 1 - Data set: CUPS

INTRODUZIONE

Il data set contiene performance di misura e caratteristiche di 209 CPUs. Le variabili sono le seguenti:

1. NAME: produttore del modello
2. SYCT: cycle time in nanosecondi
3. MMIN: minimim main memory in KB
4. MMAX: maximum main memory in KB
5. CACH: cache size in KB
6. CHMIN: minimum number of channels
7. CHMAX: maximum number of channels
8. PERF: performance della CPU comparata con il modello IBM 370/158-3
9. ESTPERF: stima della performance

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)\\1.Norm-Col\\cpus.txt")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- names(d)[3:ncol(d)]
```

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

obs	name	syct	mmin	mmax	cach	chmin	chmax	perf	estperf
1	ADVISOR 32/60	125	256	6,000	256	16	128	198	199
2	AMDAHL 470V/7	29	8,000	32,000	32	8	32	269	253
3	AMDAHL 470/7A	29	8,000	32,000	32	8	32	220	253
4	AMDAHL 470V/7B	29	8,000	32,000	32	8	32	172	253
5	AMDAHL 470V/7C	29	8,000	16,000	32	8	16	132	132
6	AMDAHL 470V/8	26	8,000	32,000	64	8	32	318	290

STATISTICHE DESCRITTIVE

Si presentano innanzitutto le statistiche descrittive.

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

Table 2: Table continues below

syct	mmin	mmax	cach
Min. : 17.0	Min. : 64	Min. : 64	Min. : 0.00
1st Qu.: 50.0	1st Qu.: 768	1st Qu.: 4000	1st Qu.: 0.00
Median : 110.0	Median : 2000	Median : 8000	Median : 8.00
Mean : 203.8	Mean : 2868	Mean :11796	Mean : 25.21
3rd Qu.: 225.0	3rd Qu.: 4000	3rd Qu.:16000	3rd Qu.: 32.00
Max. :1500.0	Max. :32000	Max. :64000	Max. :256.00

chmin	chmax	perf	estperf
Min. : 0.000	Min. : 0.00	Min. : 6.0	Min. : 15.00
1st Qu.: 1.000	1st Qu.: 5.00	1st Qu.: 27.0	1st Qu.: 28.00
Median : 2.000	Median : 8.00	Median : 50.0	Median : 45.00
Mean : 4.699	Mean : 18.27	Mean : 105.6	Mean : 99.33
3rd Qu.: 6.000	3rd Qu.: 24.00	3rd Qu.: 113.0	3rd Qu.: 101.00
Max. :52.000	Max. :176.00	Max. :1150.0	Max. :1238.00

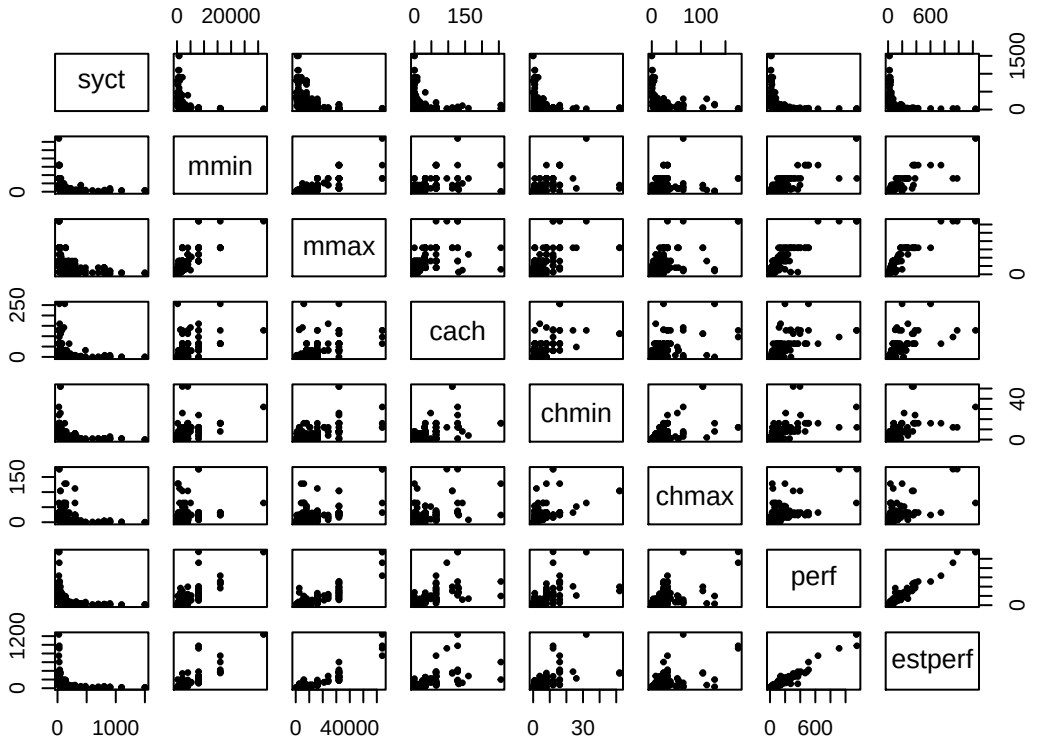
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 4: Table continues below

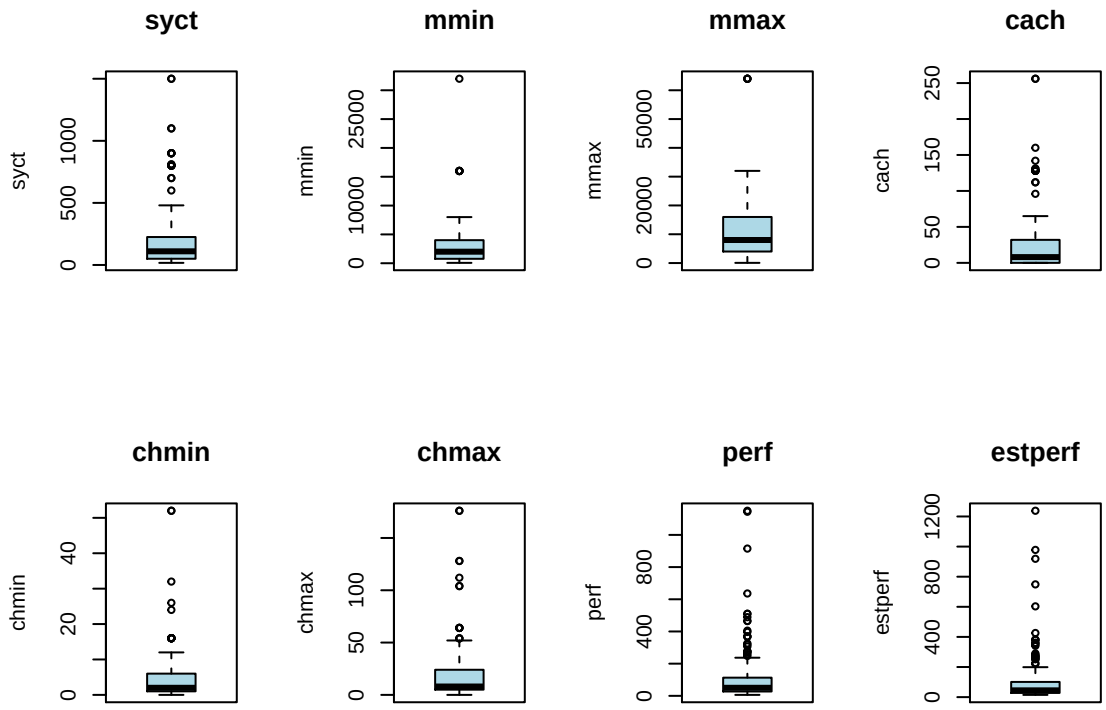
	syct	mmin	mmax	cach	chmin	chmax
syct	1	-0.3356	-0.3786	-0.321	-0.3011	-0.2505
mmin	-0.3356	1	0.7582	0.5347	0.5172	0.2669
mmax	-0.3786	0.7582	1	0.538	0.5605	0.5272
cach	-0.321	0.5347	0.538	1	0.5822	0.4878
chmin	-0.3011	0.5172	0.5605	0.5822	1	0.5483
chmax	-0.2505	0.2669	0.5272	0.4878	0.5483	1
perf	-0.3071	0.7949	0.863	0.6626	0.6089	0.6052
estperf	-0.2884	0.8193	0.9012	0.6486	0.6106	0.5922

	perf	estperf
syct	-0.3071	-0.2884
mmin	0.7949	0.8193
mmax	0.863	0.9012
cach	0.6626	0.6486
chmin	0.6089	0.6106
chmax	0.6052	0.5922
perf	1	0.9665
estperf	0.9665	1

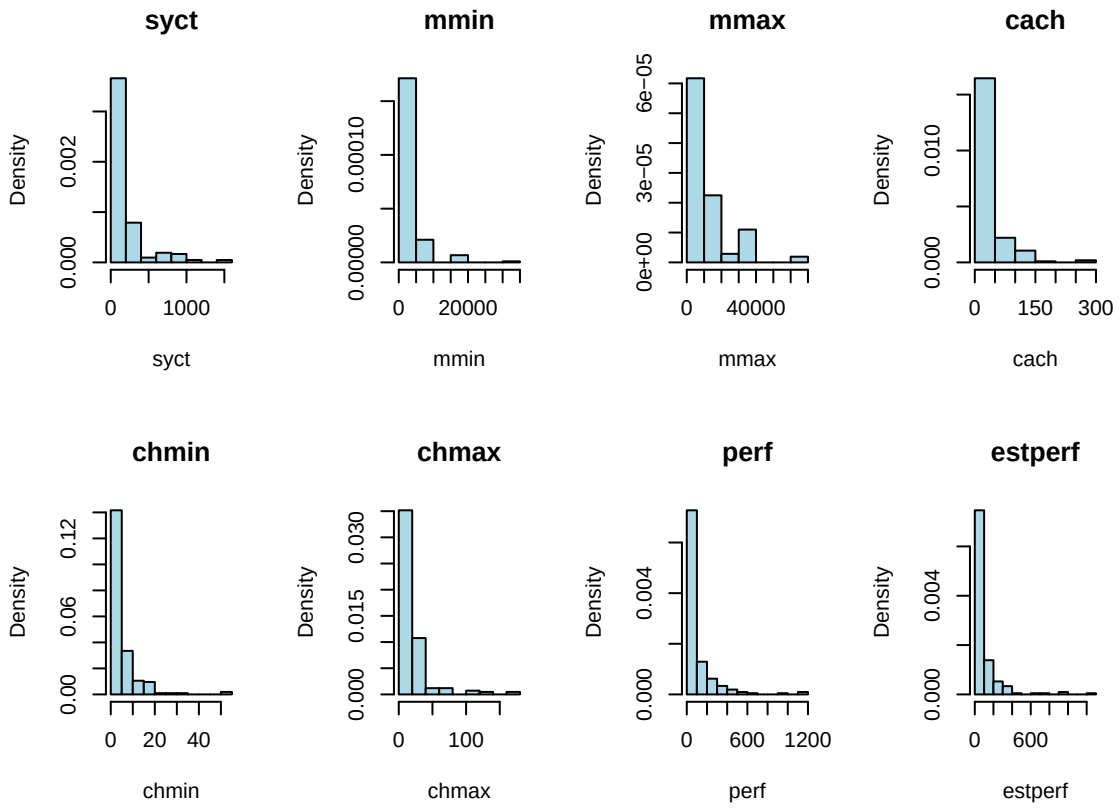
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```



```
par(mfrow=c(2,4))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```




```
par(mfrow=c(2,4))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```



Si costruisce quindi un modello lineare in cui la variabile dipendente "perf" viene regredita rispetto alle variabili esplicative.

REGRESSIONE

```r
#-- R CODE
mod1 <- lm(perf ~ syct + mmin + mmax + cach + chmin + chmax + estperf, d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.907	6.78	1.019	0.3096
syct	-0.01345	0.01251	-1.075	0.2835
mmin	0.001777	0.001511	1.176	0.241
mmax	-0.0006548	0.000591	-1.108	0.2692
cach	0.1741	0.09905	1.757	0.08039
chmin	-0.1073	0.5787	-0.1853	0.8531
chmax	0.3479	0.1658	2.099	0.0371
estperf	0.9447	0.06087	15.52	3.168e-36

Table 7: Fitting linear model: perf ~ syct + mmin + mmax + cach + chmin + chmax + estperf

Observations	Residual Std. Error	R^2	Adjusted R^2
209	40.56	0.9385	0.9364

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

Table 8: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
syct	1	507,352	507,352	308.4	1.875e-42
mmin	1	2,902,264	2,902,264	1,764	1.833e-101
mmax	1	855,686	855,686	520.1	1.161e-57
cach	1	210,391	210,391	127.9	2.897e-23
chmin	1	14,136	14,136	8.592	0.003768
chmax	1	163,396	163,396	99.31	2.895e-19
estperf	1	396,286	396,286	240.9	3.168e-36
Residuals	201	330,716	1,645	NA	NA

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

Test.statistic	P.value
38.37	4.664e-09

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

Table 10: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.465	1.992e-05 * * *	true autocorrelation is greater than 0

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

Variables	Tolerance	VIF
syct	0.7464	1.34
mmin	0.2302	4.345
mmax	0.1647	6.072
cach	0.4884	2.047
chmin	0.5084	1.967
chmax	0.4259	2.348
estperf	0.08913	11.22

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

Table 12: Table continues below

Eigenvalue	Condition Index	intercept	syct	mmin	mmax
5.229	1	0.003721	0.001694	0.004093	0.002634
1.222	2.069	0.03064	0.2558	0.002072	0.000156
0.5315	3.137	0.0001023	0.000718	0.09295	0.007672
0.3515	3.857	4.73e-06	0.01579	6.794e-06	0.02394
0.2864	4.273	0.008689	0.01904	0.003912	0.0002204
0.2435	4.634	0.3761	0.5007	0.000975	0.004573
0.1017	7.171	0.002433	0.0001392	0.6389	0.3101
0.03456	12.3	0.5783	0.2061	0.2571	0.6507

cach	chmin	chmax	estperf
0.008791	0.009021	0.006759	0.001961
0.009347	0.004108	0.0004476	0.001877
0.01794	0.05529	0.2457	0.00535
0.5374	0.07934	0.1772	0.008721
0.3112	0.7502	0.0474	0.008881
0.01065	0.06863	0.001546	0.03333
0.01564	0.02899	0.3562	0.02143
0.08906	0.004394	0.1648	0.9185

I modello risulta significativo ma solo la variabile “estperf” ha associato un parametro che cade nella regione di rifiuto per cui è respinta l’ipotesi nulla di non significatività.

Si esamina quindi la collinearità; come si può notare l’indice di tolleranza è molto piccolo e l’inflation indice è molto grande proprio per “estperf” l’unica variabile significativa, per cui la quota di varianza risulta altresì molto elevata per l’8^ autovalore. “estperf” risulta quindi multicollineare con le altre variabili e viene quindi eliminata. Si effettua una nuova regressione escludendo “estperf”.

Si vede come come cambia radicalmente la situazione inerente la significatività delle variabili: “mmin”, “mmax”, “cach”, “chmax” risultano significative. Inoltre nessuna delle variabili è ora collineare.

```
#-- R CODE
mod1 <- lm(perf ~ syct + mmin + mmax + cach + chmin + chmax, d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-55.9	8.045	-6.948	4.987e-11
syct	0.04886	0.01752	2.789	0.005793
mmin	0.01529	0.001827	8.371	9.416e-15
mmax	0.005571	0.0006418	8.68	1.326e-15
cach	0.6412	0.1396	4.594	7.64e-06
chmin	-0.2701	0.8557	-0.3156	0.7526
chmax	1.483	0.2201	6.738	1.64e-10

Table 15: Fitting linear model: perf ~ syct + mmin + mmax + cach + chmin + chmax

Observations	Residual Std. Error	R^2	Adjusted R^2
209	59.99	0.8649	0.8609

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

Table 16: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
syct	1	507,352	507,352	141	5.23e-25
mmin	1	2,902,264	2,902,264	806.4	1.863e-72
mmax	1	855,686	855,686	237.8	5.71e-36
cach	1	210,391	210,391	58.46	8.263e-13
chmin	1	14,136	14,136	3.928	0.04885
chmax	1	163,396	163,396	45.4	1.64e-10
Residuals	202	727,002	3,599	NA	NA

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

Test.statistic	P.value
56.08	6.629e-13

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

Table 18: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.202	1.095e-09 * * *	true autocorrelation is greater than 0

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

Variables	Tolerance	VIF
syct	0.8322	1.202
mmin	0.3446	2.902
mmax	0.3054	3.274
cach	0.5381	1.858
chmin	0.5086	1.966
chmax	0.5287	1.891

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

Table 20: Table continues below

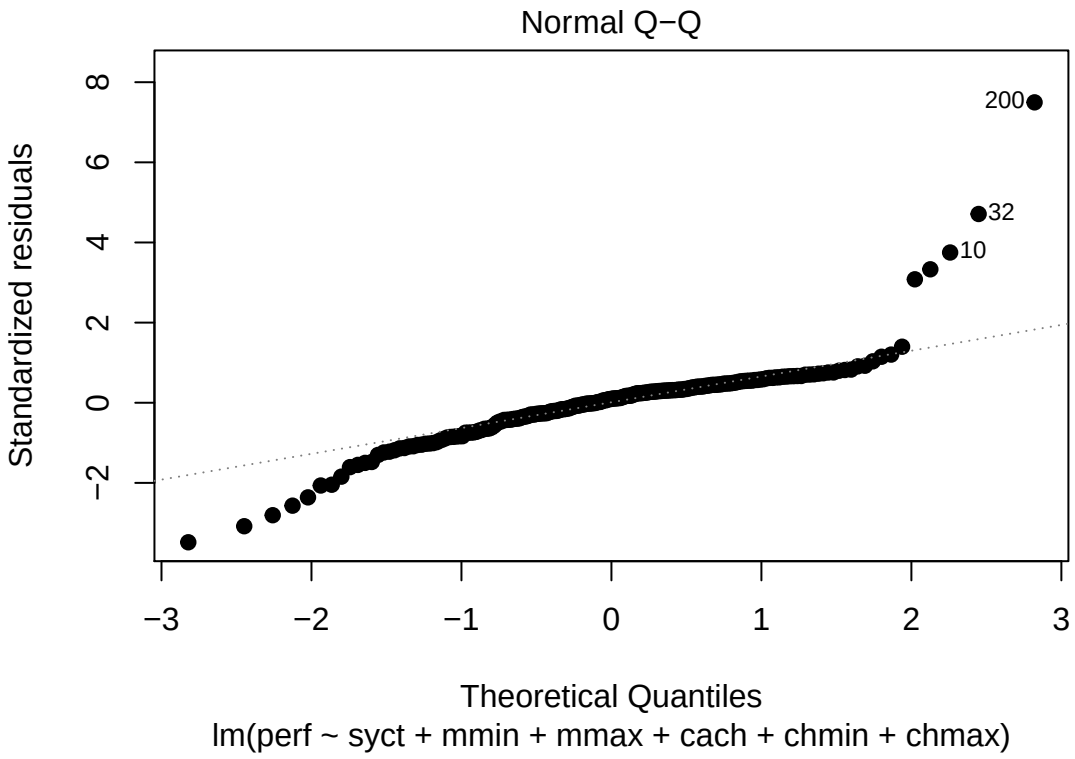
Eigenvalue	Condition Index	intercept	syct	mmin	mmax
4.411	1	0.00899	0.003418	0.008168	0.006632
1.173	1.939	0.04065	0.3022	0.004667	0.0008586
0.5093	2.943	0.002391	0.001543	0.1747	0.01694
0.3382	3.612	0.01553	0.07993	0.007724	0.06429
0.2784	3.981	0.01555	0.01743	0.0004515	0.01137
0.1938	4.77	0.8161	0.5666	0.1189	0.008159
0.09652	6.76	0.1008	0.02886	0.6854	0.8918

cach	chmin	chmax
0.01352	0.01287	0.01186
0.01879	0.01015	0.002861
0.003788	0.02724	0.3411
0.6926	0.02682	0.149
0.2411	0.8894	0.0771
0.00235	0.0186	0.1212
0.02794	0.01494	0.2969

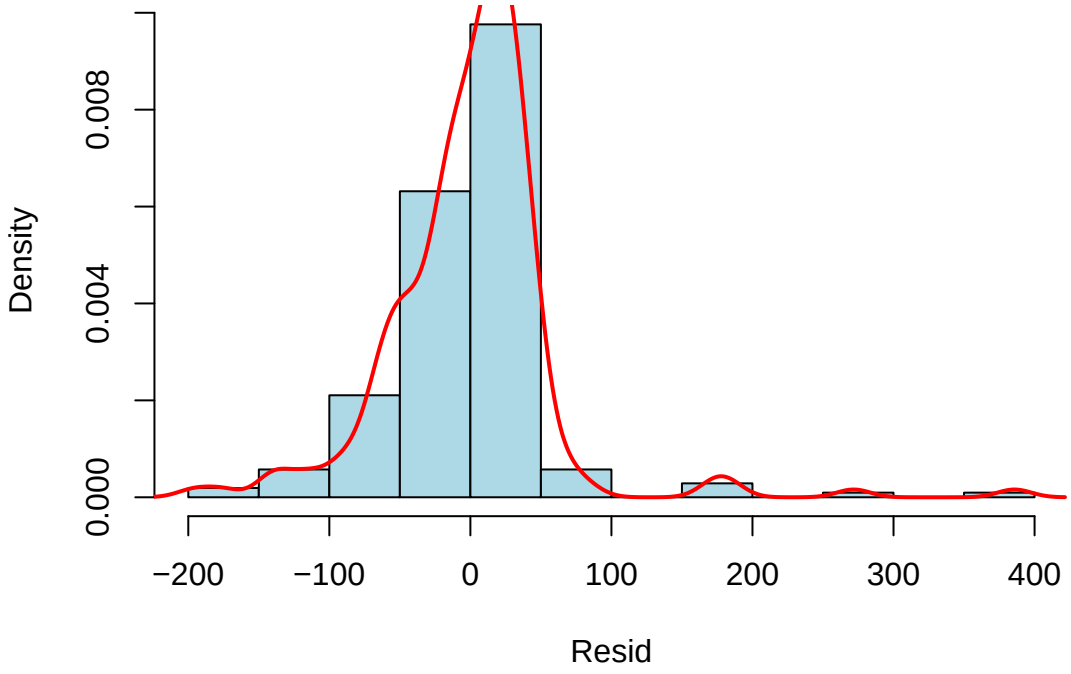
Si analizza ora la normalità dei residui considerando il modello con le sole variabili significative “mmin”, “mmax”, “cach”, “chmax”. Si inizia studiando il valore degli indici di asimmetria e curtosi, la distribuzione dei residui e il box plot.

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

Test statistic	P value
0.8313	2.607e-14 * * *

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

Table 23: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

Test statistic	P value	Alternative hypothesis
0.5248	0 * * *	two-sided

La distribuzione dei residui sembra respingere l'ipotesi di normalità e tutti i test respingono l'ipotesi nulla di normalità. Sia dal grafico del Q-Q plot che dal confronto dei quantili della distribuzione normale teorica e osservata si vede la forte discrepanza tra tali distribuzioni. E' una ulteriore prova della non normalità dei residui.