GLS 8 - Data set: QUAKES

INTRODUZIONE

Il dataset riguarda i terremoti rilevati vicino a Fiji. Le osservazioni rappresentano i movimenti sismici rilevati nel 1964 con magnitudo maggiore di 4. Le variabili sono:

1. LAT: latitudine
2. LONG: longitudine
3. DEPTH: profondità
4. MAG: magnitudo
5. STATIONS: stazione

Analisi proposte:

1. Statistiche descrittive
2. Regressione
3. Gestione dell'autocorrelazione

```r
#-- R CODE

library(Hmisc)
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\\1.Error-GLS copy(8)\\8.Error-GLS\\QUAKES.T

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("lat","long","depth","mag")

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

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

id	lat	long	depth	mag	stations
1	-20.42	181.6	562	4.8	41
2	-20.62	181	650	4.2	15
3	-26	184.1	42	5.4	43
4	-17.97	181.7	626	4.1	19
5	-20.42	182	649	4	11
6	-19.68	184.3	195	4	12

STATISTICHE DESCRITTIVE

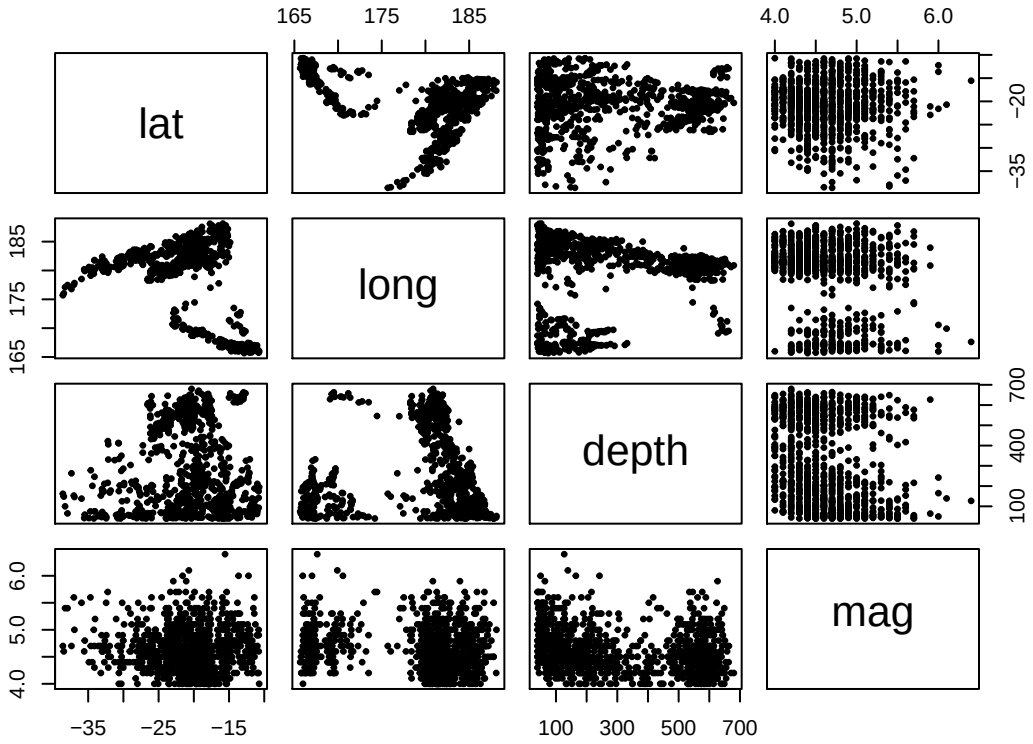
```
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

lat	long	depth	mag
Min. :-38.59	Min. :165.7	Min. : 40.0	Min. :4.00
1st Qu.:-23.47	1st Qu.:179.6	1st Qu.: 99.0	1st Qu.:4.30
Median :-20.30	Median :181.4	Median :247.0	Median :4.60
Mean :-20.64	Mean :179.5	Mean :311.4	Mean :4.62
3rd Qu.:-17.64	3rd Qu.:183.2	3rd Qu.:543.0	3rd Qu.:4.90
Max. :-10.72	Max. :188.1	Max. :680.0	Max. :6.40

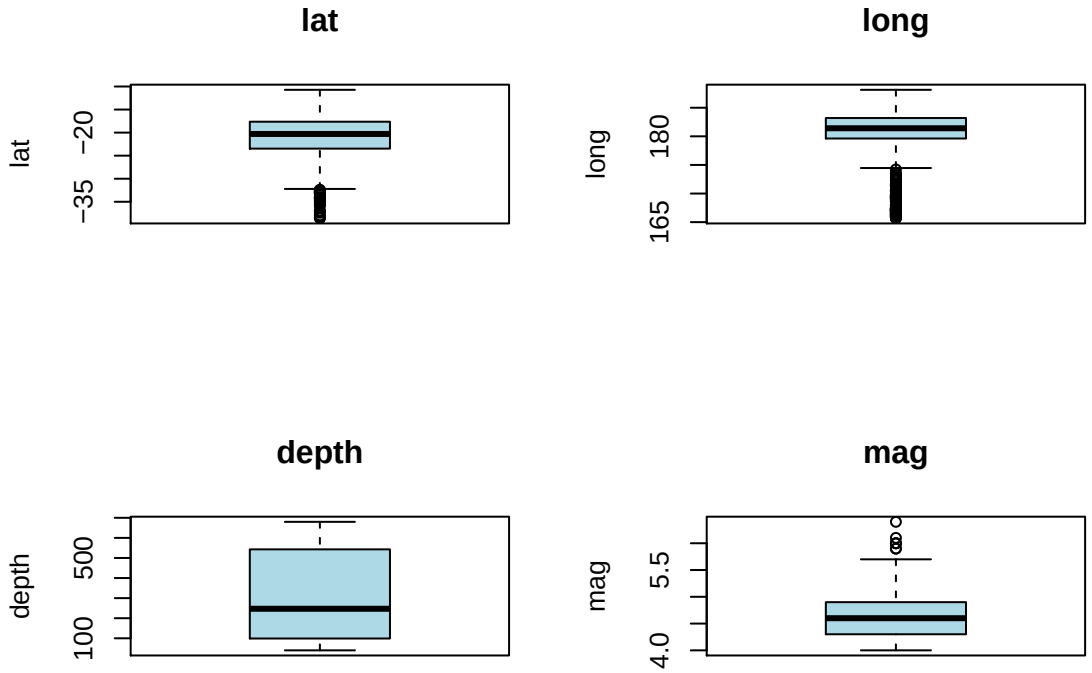
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

	lat	long	depth	mag
lat	1	-0.3645	0.03103	-0.05046
long	-0.3645	1	0.1444	-0.1731
depth	0.03103	0.1444	1	-0.2306
mag	-0.05046	-0.1731	-0.2306	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)
}
```



REGRESSIONE

```
#-- R CODE
mod1 <- lm(mag ~ stations + depth + long + lat , d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.731	0.1878	30.51	7.085e-145
stations	0.01531	0.0002795	54.78	1.251e-302
depth	-0.0002726	2.878e-05	-9.473	1.913e-20
long	-0.009452	0.001096	-8.627	2.468e-17
lat	-0.00769	0.001308	-5.879	5.626e-09

Table 5: Fitting linear model: mag ~ stations + depth + long + lat

Observations	Residual Std. Error	R^2	Adjusted R^2
1000	0.1928	0.7719	0.7709

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

Table 6: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
stations	1	117.4	117.4	3,160	4.481e-311
depth	1	4.602	4.602	123.9	3.386e-27
long	1	1.786	1.786	48.06	7.426e-12
lat	1	1.284	1.284	34.56	5.626e-09
Residuals	995	36.97	0.03716	NA	NA

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

Test.statistic	P.value
1.775	0.4118

```
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

Table 8: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.941	0.1751	true autocorrelation is greater than 0

Le 4 variabili esplicative "stations", "depth", "long", "lat" risultano tutte significative. Il valore dell'R^2 è molto buono e il modello interpreta bene la variabile dipendente.

Si verifica ora la multicollinearità delle variabili esplicative.

Per tutti e 4 i valori l'indice di tolleranza è quasi prossimo a uno e quindi mostra che non esiste collinearità.

Il condition index perfeziona tale conclusione perché se risulta debolmente dipendente per il quarto auto valore mentre il quinto assume valore molto elevato andando a spiegare quota di varianza elevata per l'intercetta e la variabile "long".

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

Table 9: Table continues below

Eigenvalue	Condition Index	intercept	stations	depth	long
4.428	1	5.236e-05	0.01179	0.0121	4.778e-05
0.3358	3.631	1.551e-06	0.3853	0.5253	5.971e-07
0.2016	4.687	0.0004574	0.5729	0.4053	0.0004269
0.03429	11.36	0.006639	0.02474	0.03822	0.004663
0.0005048	93.66	0.9928	0.005207	0.019	0.9949

lat
0.002315
0.0002065

lat
0.04157
0.8563
0.09963

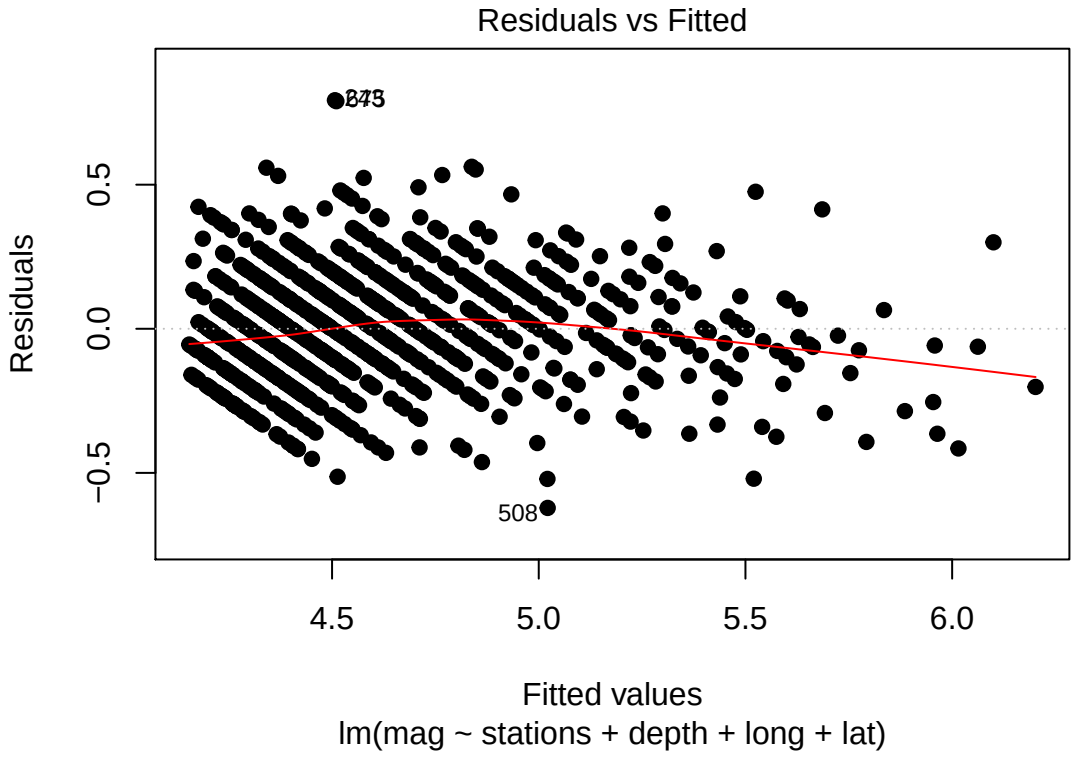
```
pander(ols_vif_tol(mod1),big.mark=",")
```

Variables	Tolerance	VIF
stations	0.9924	1.008
depth	0.967	1.034
long	0.841	1.189
lat	0.8597	1.163

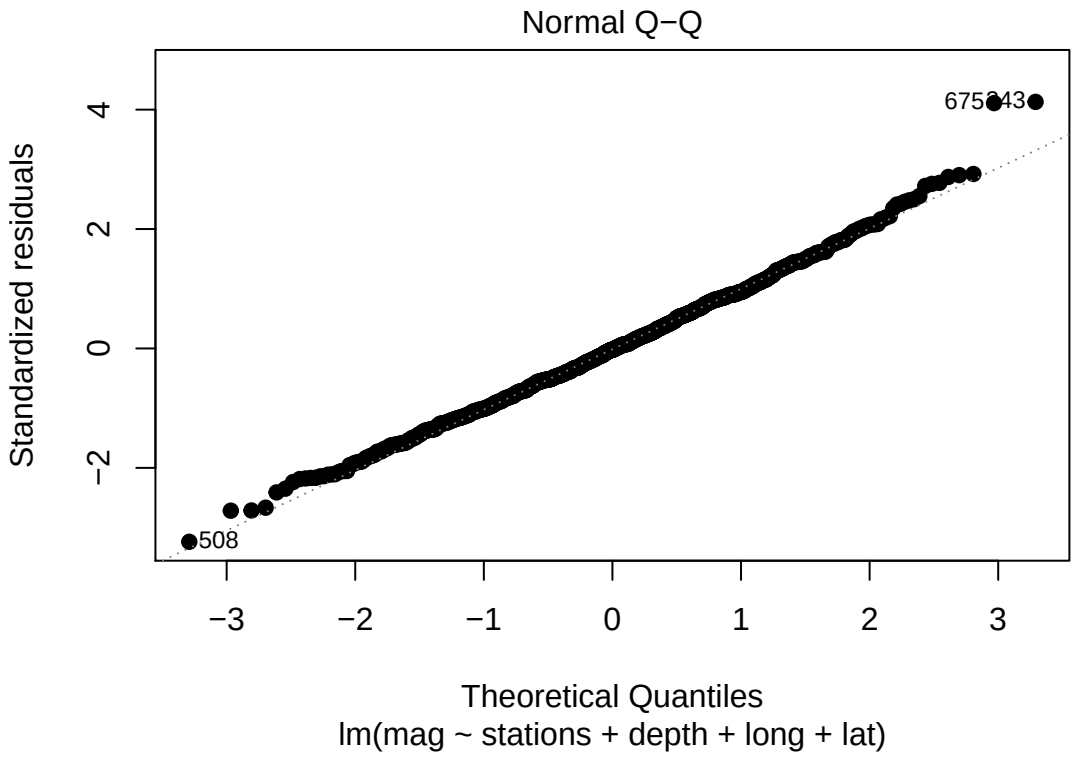
Si verifica ora la normalità. Si analizza innanzitutto la distribuzione dei residui e il box plot. L'istogramma si sovrappone bene alla curva normale teorica.

Per ciò che concerne il box plot dei residui si verifica che c'è simmetria intorno alla media. Anche la distribuzione cumulata dei residui empirici si sovrappone a quella dei residui della distribuzione teorica normale

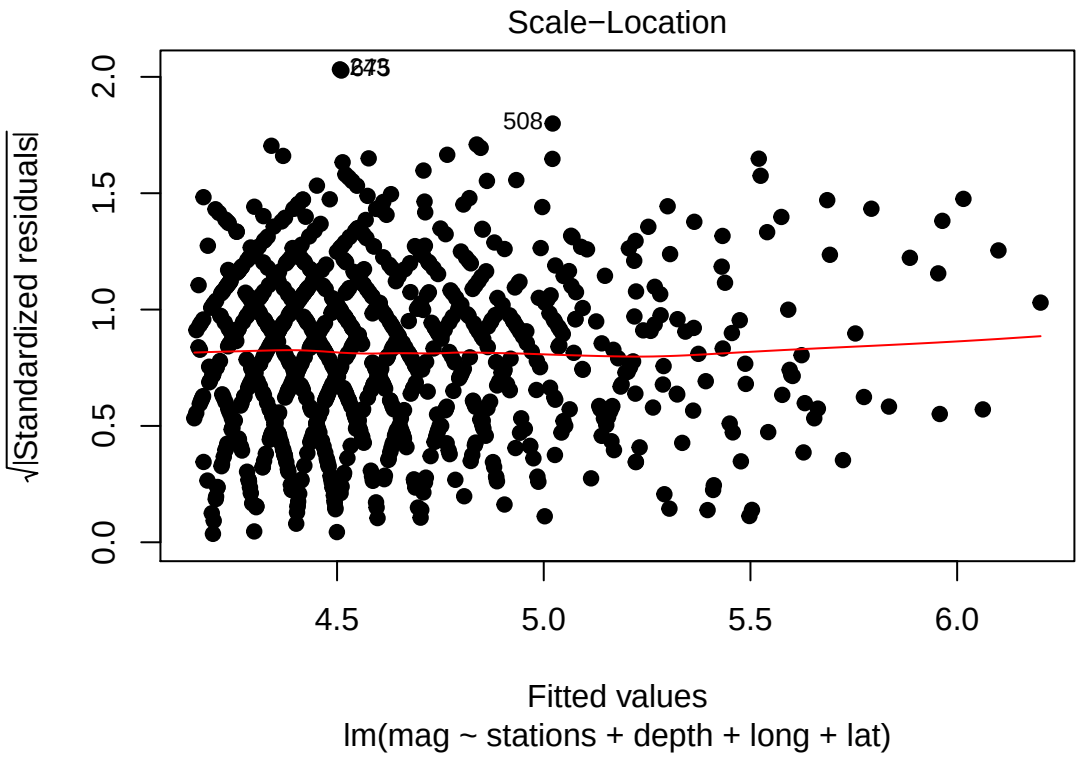
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



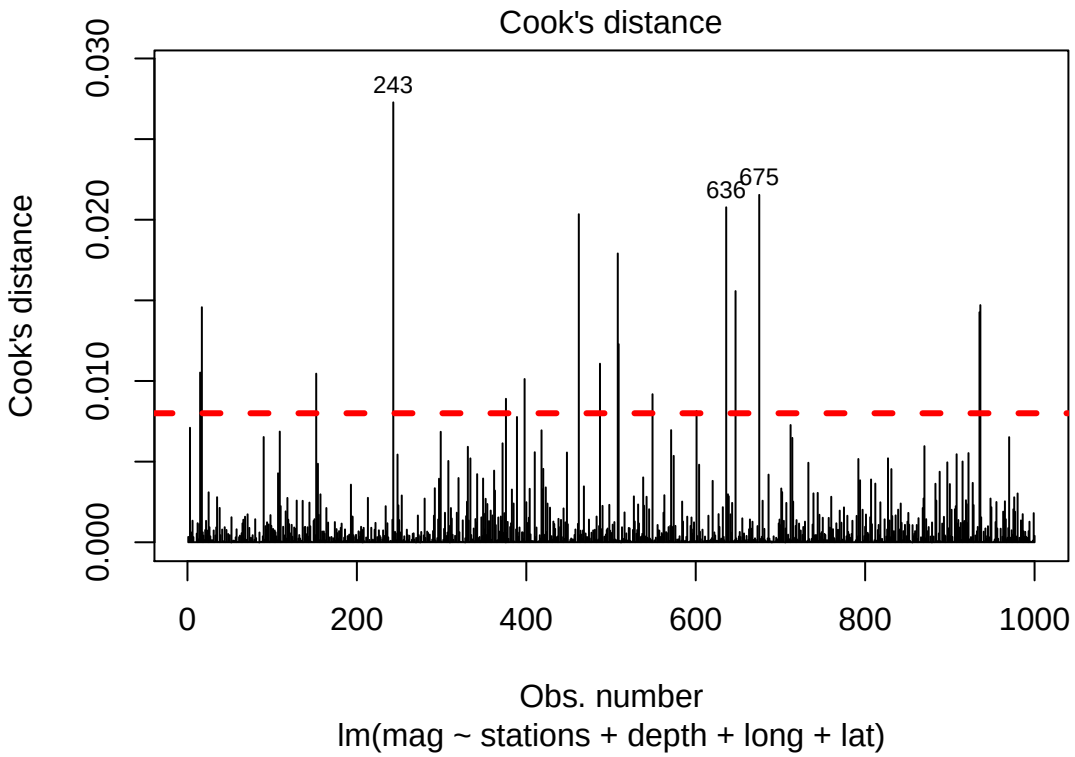
```
plot(mod1,which=2,pch=19)
```



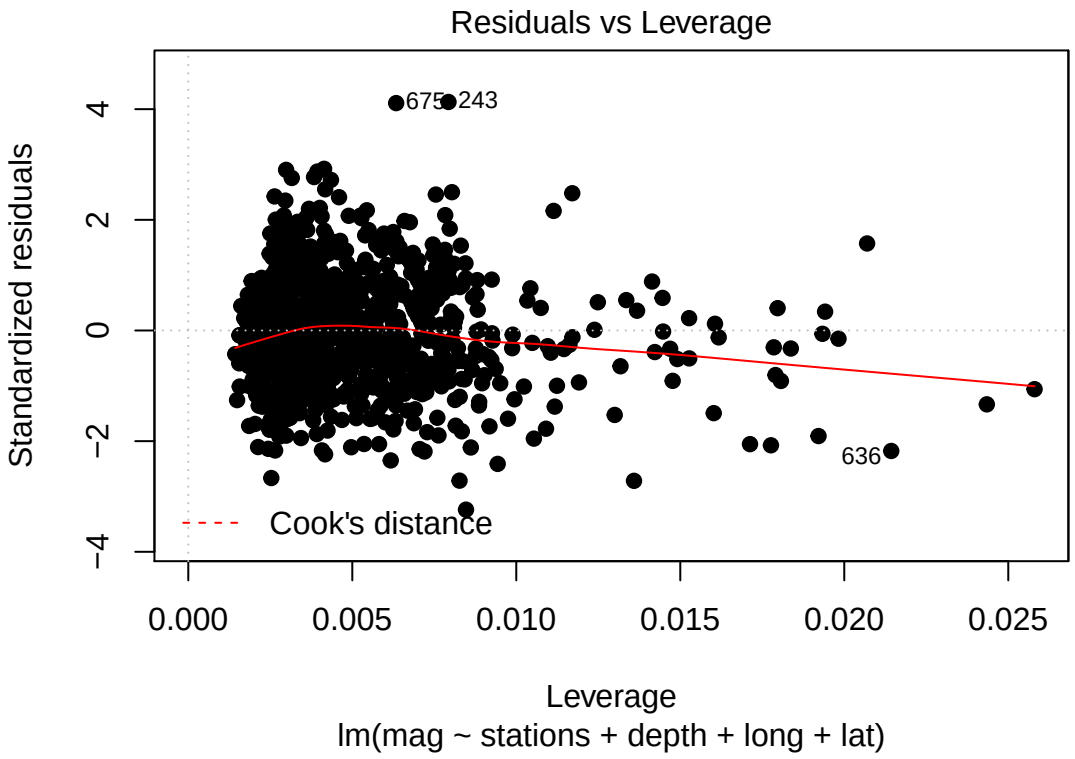

```
plot(mod1,which=3,pch=19)
```



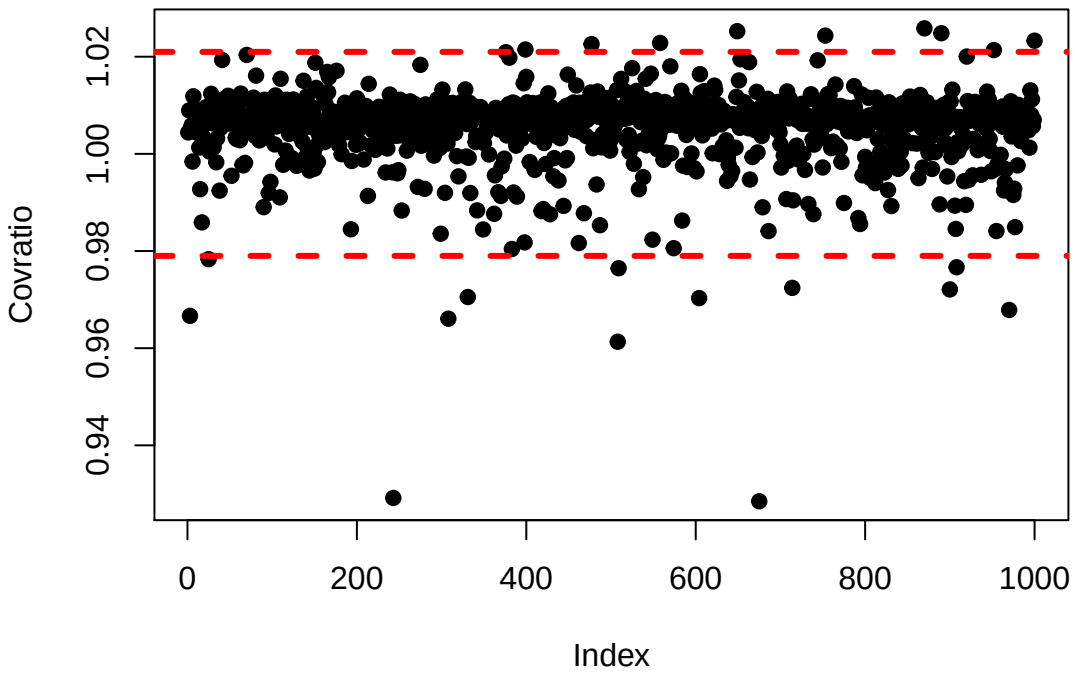

```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d),col=2,lwd=3,lty=2)
```



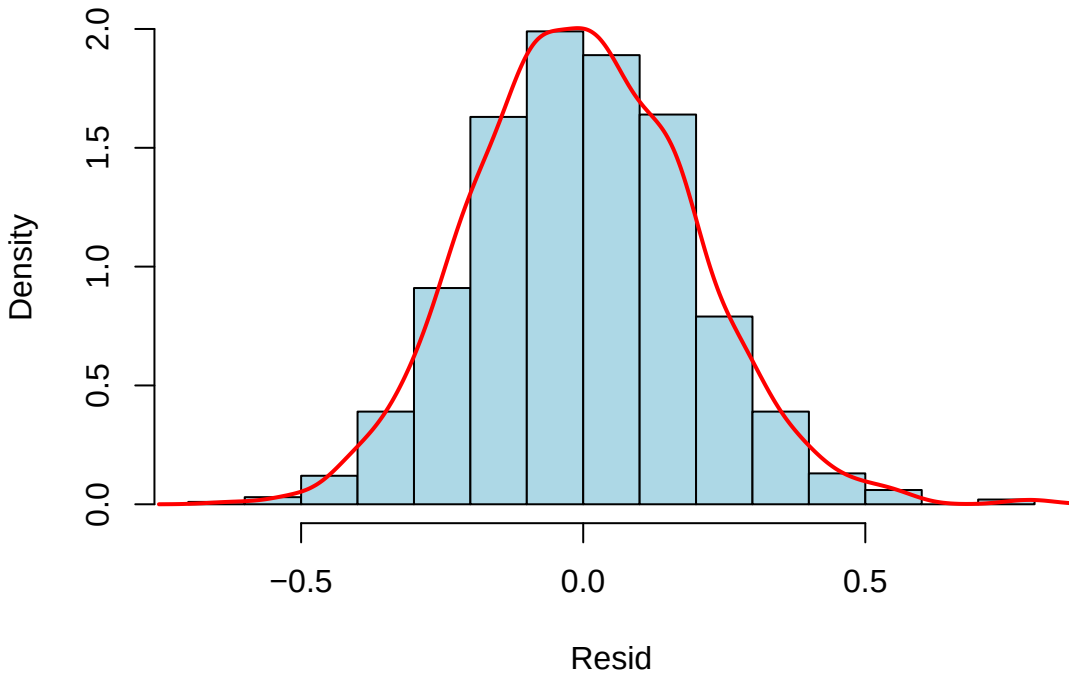

```
plot(mod1,which=5,pch=19)
```




```r
#-- R CODE
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)
```




```
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
0.9962	0.01426 *

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

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

Test statistic	P value	Alternative hypothesis
0.3367	0 * * *	two-sided

I test, in particolare Shapiro-Wilk (vicino a 1 come valore) e Kolmogorov-Smirnov, cadono tutti nella regione di accettazione: non respingo l'ipotesi nulla di normalità dei residui.