GLS 3 - Data set: HARTNAGEL

INTRODUZIONE

Il dataset contiene 38 osservazioni e 7 variabili, con dati raccolti dal 1931 al 1968. Le variabili sono le seguenti:

1. YEAR: 1931-1968
2. TFR: tasso di fertilità totale per 1000 donne
3. PARTIC: forza lavoro femminile per 1000
4. DEGREES: grado di studio di scuola secondaria per 10.000
5. FCONVICT: tasso (femminile) di offese subite per 100.000
6. FTHEFT: tasso (femminile) di furti subiti per 100.000
7. MCONVICT: tasso (maschile) di offese subite per 100.000
8. MTHEFT: tasso (maschile) di furti subiti per 100.000

Analisi proposte:

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

```
#-- 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)\\3.Error-GLS\\Hartnagel

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("year","tfr","partic","degrees","fconvict","ftheft","mconvict","mtheft")
```

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

id	year	tfr	partic	degrees	fconvict	ftheft	mconvict	mtheft
1	1,931	3,200	234	12.4	77.1	NA	778.7	NA
2	1,932	3,084	234	12.9	92.9	NA	745.7	NA
3	1,933	2,864	235	13.9	98.3	NA	768.3	NA
4	1,934	2,803	237	13.6	88.1	NA	733.6	NA
5	1,935	2,755	238	13.2	79.4	20.4	765.7	247.1
6	1,936	2,696	240	13.2	91	22.1	816.5	254.9

STATISTICHE DESCRITTIVE

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

Table 2: Table continues below

year	tfr	partic	degrees	fconvict
Min. :1931	Min. :2441	Min. :232.0	Min. :11.10	Min. : 40.20
1st Qu.:1940	1st Qu.:2817	1st Qu.:240.0	1st Qu.:12.68	1st Qu.: 55.00
Median :1950	Median :3287	Median :245.0	Median :18.40	Median : 78.25
Mean :1950	Mean :3265	Mean :265.8	Mean :25.57	Mean : 84.74
3rd Qu.:1959	3rd Qu.:3708	3rd Qu.:290.0	3rd Qu.:27.52	3rd Qu.:101.08
Max. :1968	Max. :3935	Max. :339.0	Max. :90.40	Max. :157.30
NA	NA	NA	NA	NA

ftheft	mconvict	mtheft
Min. :15.20	Min. : 633.7	Min. :166.8
1st Qu.:20.10	1st Qu.: 743.4	1st Qu.:199.0
Median :22.20	Median : 779.9	Median :250.3
Mean :29.13	Mean : 798.7	Mean :237.2
3rd Qu.:28.88	3rd Qu.: 839.4	3rd Qu.:274.1
Max. :73.00	Max. :1035.7	Max. :296.9
NA's :4	NA	NA's :4

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

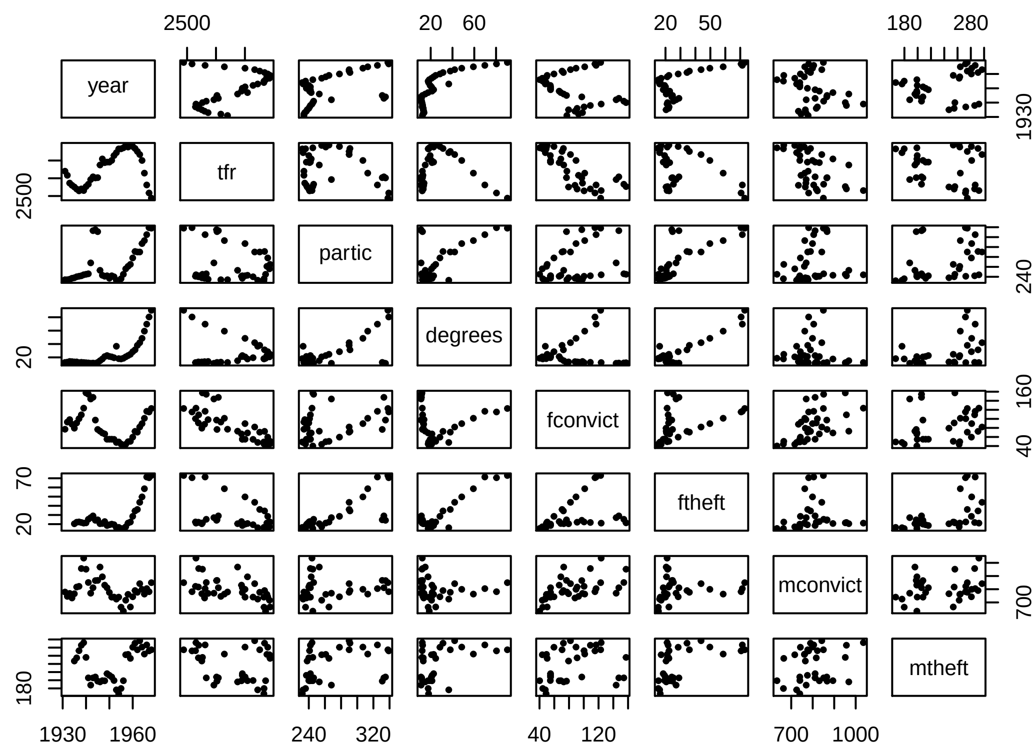
Table 4: Table continues below

	year	tfr	partic	degrees	fconvict	ftheft
year	1	0.4549	0.5503	0.7784	-0.2672	NA
tfr	0.4549	1	-0.1976	-0.1046	-0.7893	NA
partic	0.5503	-0.1976	1	0.631	0.3499	NA

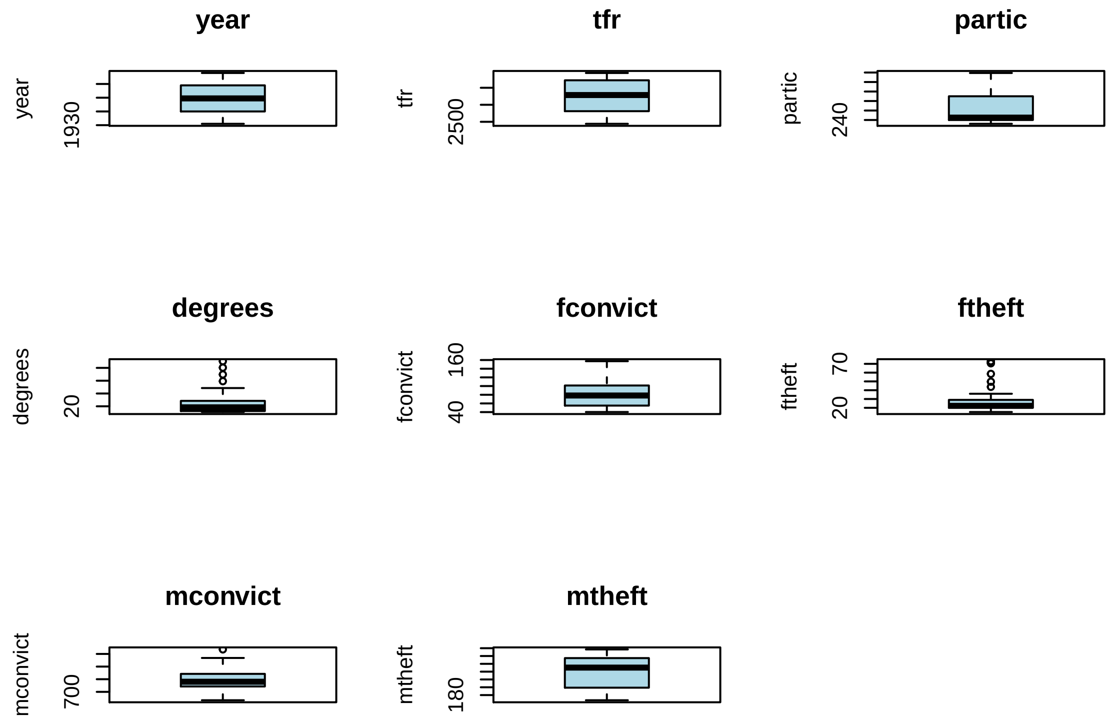
	year	tfr	partic	degrees	fconvict	ftheft
degrees	0.7784	-0.1046	0.631	1	0.1053	NA
fconvict	-0.2672	-0.7893	0.3499	0.1053	1	NA
ftheft	NA	NA	NA	NA	NA	1
mconvict	-0.2835	-0.5256	0.1188	-0.08986	0.5478	NA
mtheft	NA	NA	NA	NA	NA	NA

	mconvict	mtheft
year	-0.2835	NA
tfr	-0.5256	NA
partic	0.1188	NA
degrees	-0.08986	NA
fconvict	0.5478	NA
ftheft	NA	NA
mconvict	1	NA
mtheft	NA	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)
}
```



Non esistono correlazioni elevate tra le variabili.

REGRESSIONE

Si regredisce la variabile "ftheft" su "partic", "degrees", "mtheft"

```
#-- R CODE
mod1 <- lm(ftheft ~ partic + degrees + mtheft, d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-36.06	8.114	-4.445	0.0001112
partic	0.1414	0.02849	4.965	2.574e-05
degrees	0.5386	0.0538	10.01	4.454e-11
mtheft	0.05282	0.02279	2.318	0.02747

Table 7: Fitting linear model: ftheft ~ partic + degrees + mtheft

Observations	Residual Std. Error	R^2	Adjusted R^2
34	4.754	0.9244	0.9168

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

Table 8: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
partic	1	5,256	5,256	232.5	1.132e-15
degrees	1	2,912	2,912	128.8	2.217e-12
mtheft	1	121.4	121.4	5.372	0.02747
Residuals	30	678.1	22.6	NA	NA

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

Test.statistic	P.value
0.1096	0.9467

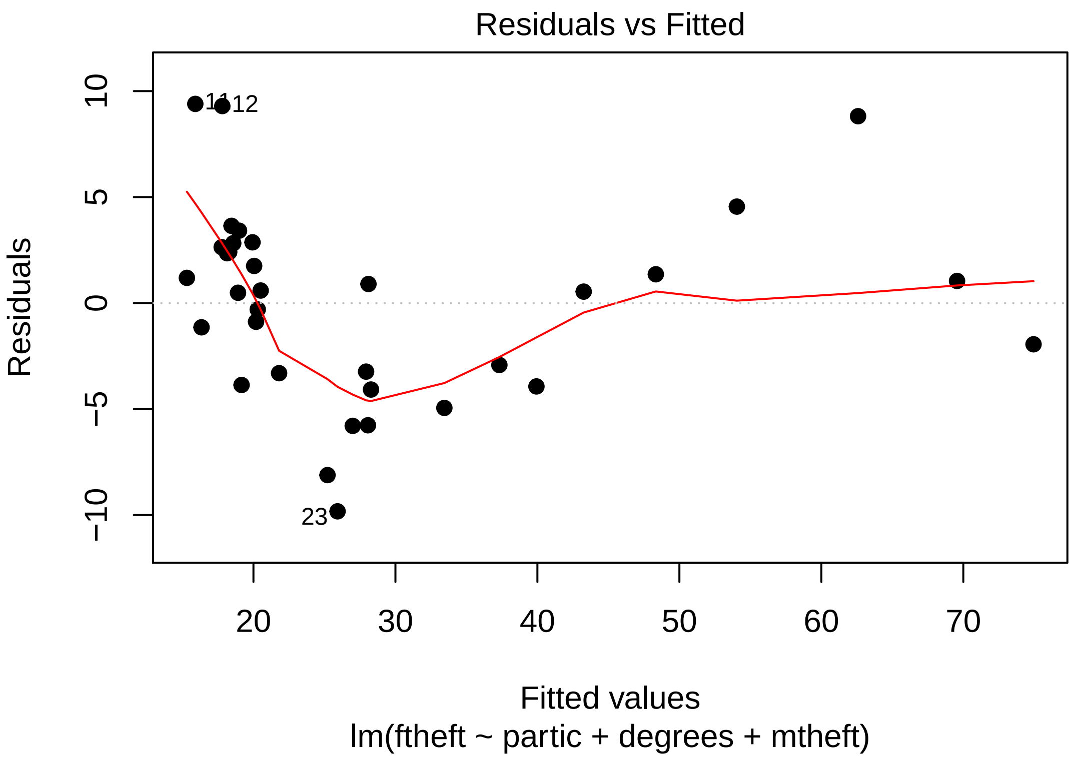
```r
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

Table 10: Durbin-Watson test: `mod1`

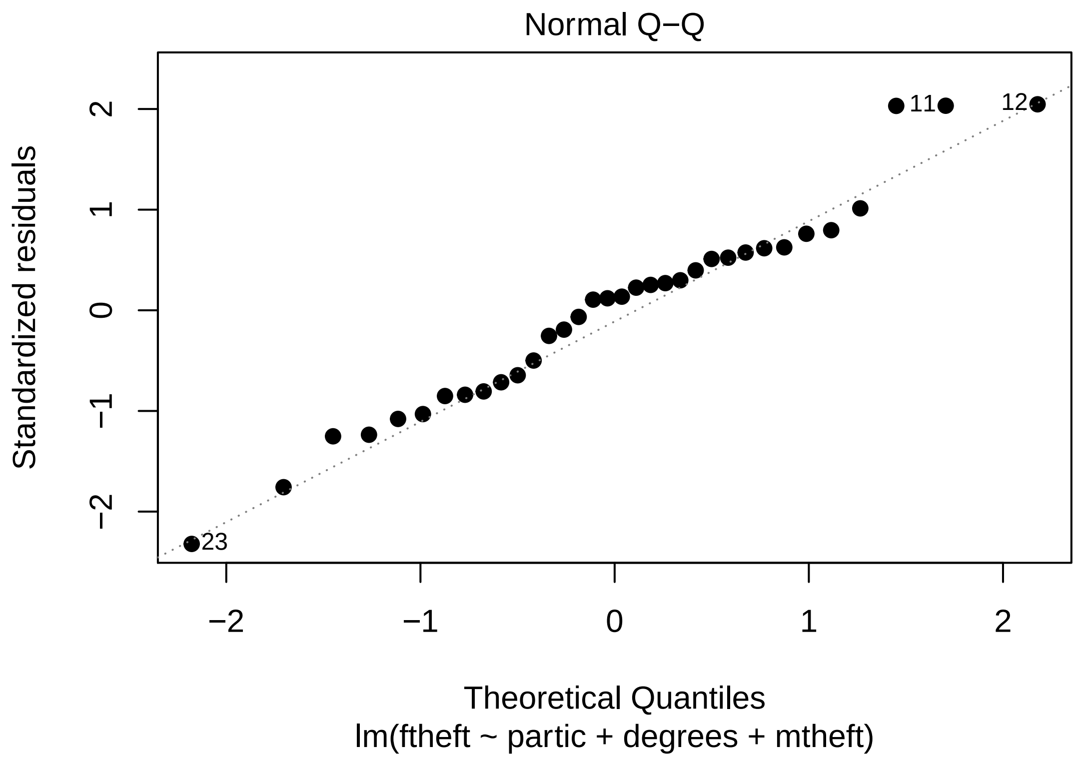
Test statistic	P value	Alternative hypothesis
0.9051	2.545e-05 * * *	true autocorrelation is greater than 0

Il modello interpreta bene la variabile dipendente e il fitting è molto elevato. I parametri significativi sono quelli relativi a "partic" e "degrees". Gli errori sono normali come si evince dalla distribuzione dei residui.

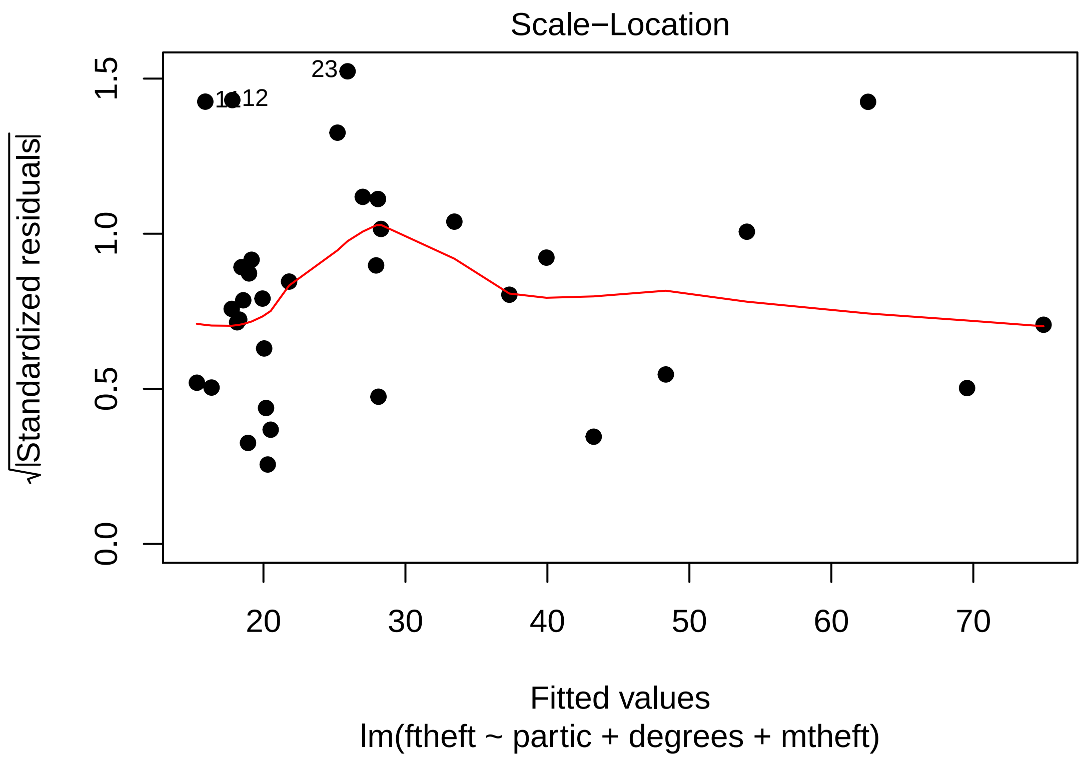
```r
#-- R CODE
plot(mod1,which=1,pch=19)
```



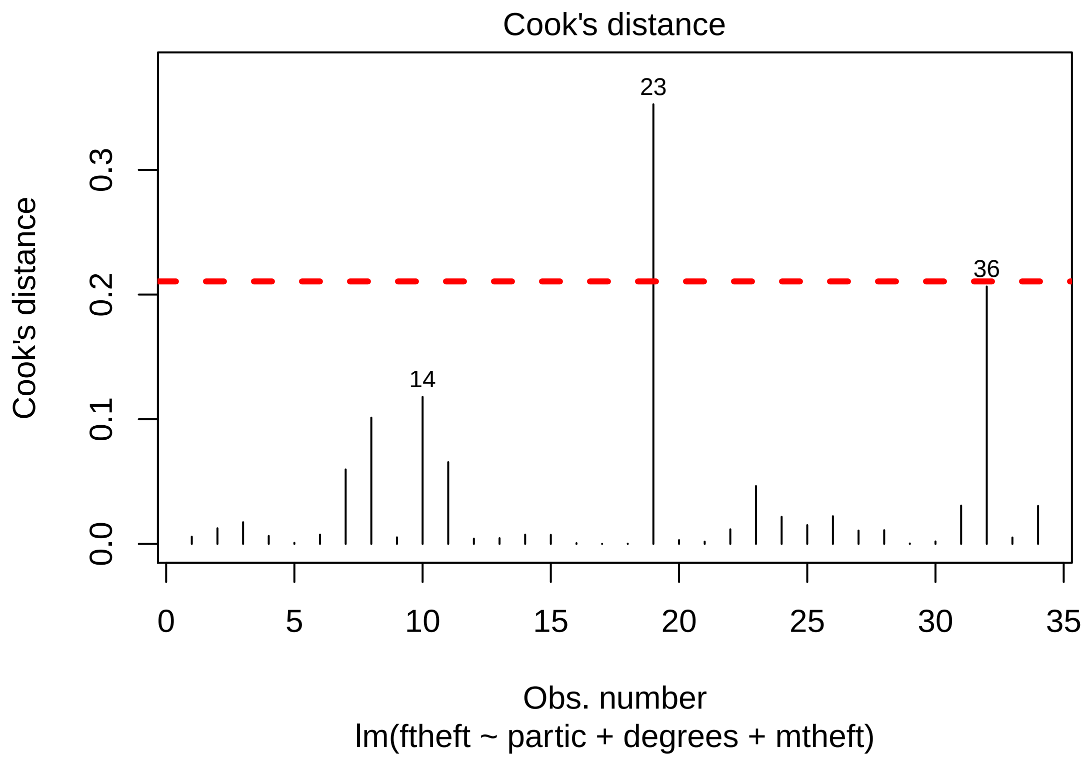

```r
plot(mod1,which=2,pch=19)
```



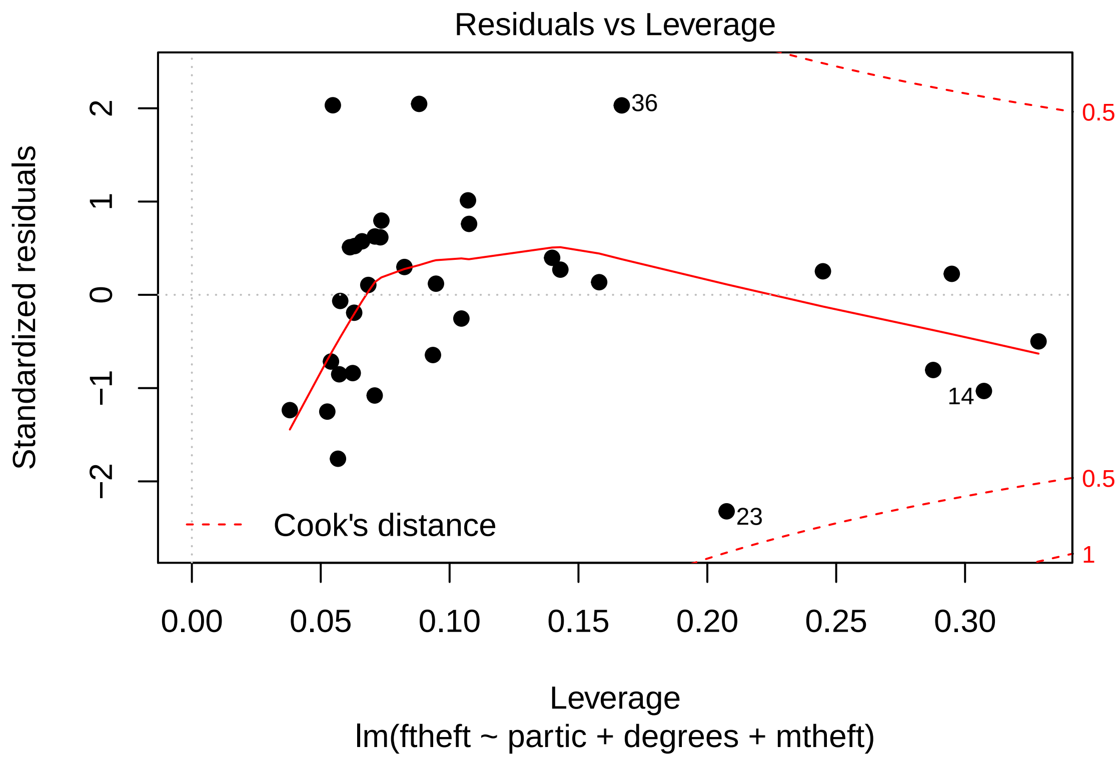

```
plot(mod1,which=3,pch=19)
```




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



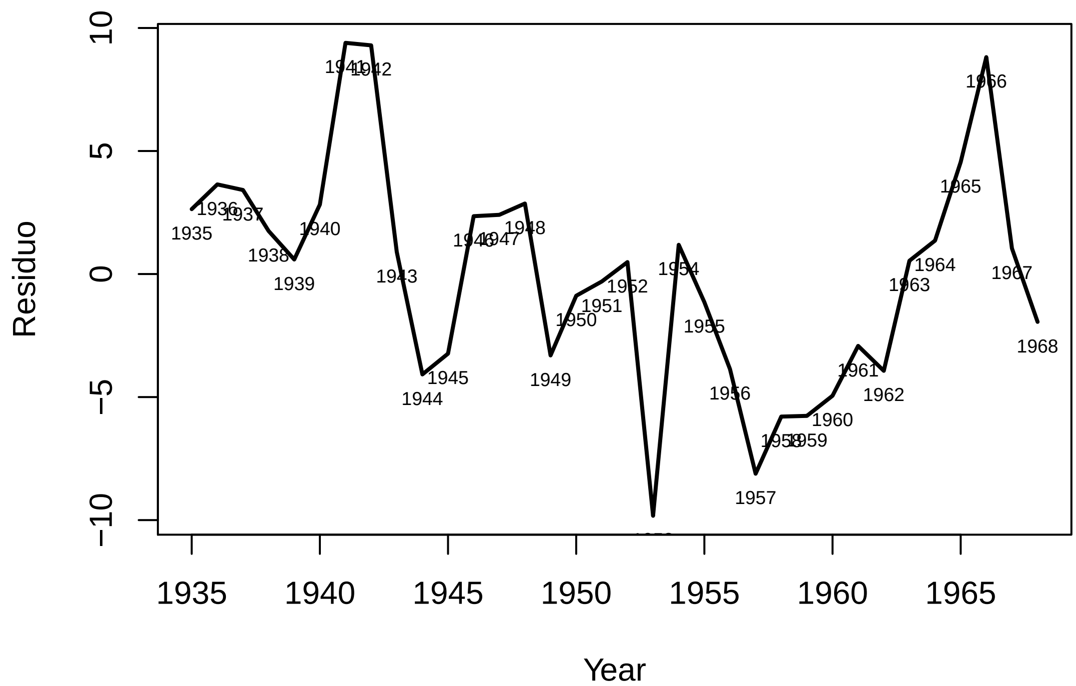

```
plot(mod1,which=5,pch=19)
```



Il grafico Q-Q plot mostra la presenza di outlier ai valori estremi della distribuzione; i grafici residui studentizzati, leverage, Distanza di Cook confermano tale ipotesi. Tuttavia non operiamo alcuna correzione per quel che concerne gli outlier. Il test di White non respinge l'ipotesi di omoschedasticità dei residui.

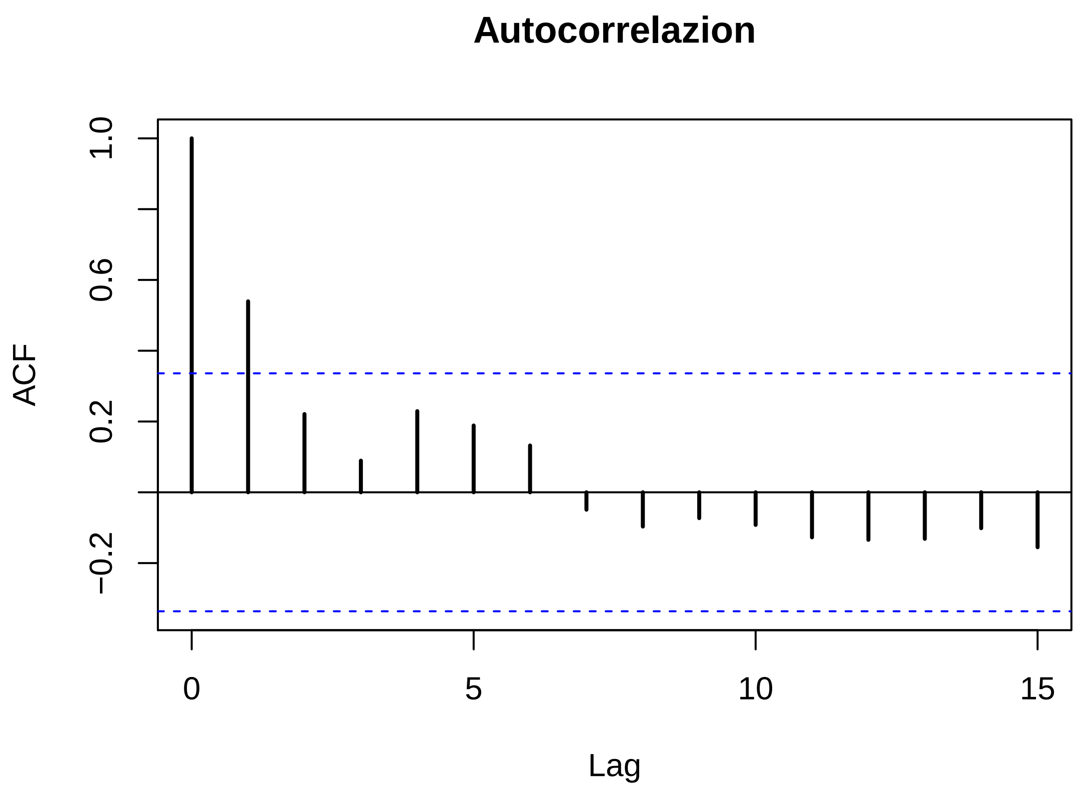
```
#-- R CODE
index <- as.numeric(row.names(mod1$model))

plot(d$year[index],resid(mod1),pch=19,xlab="Year",ylab="Residuo",type="l",col=1,lwd=2)
text(d$year[index],resid(mod1),d$year[index],pos=1,cex=.6)
```



Il test respinge senza alcun dubbio l'ipotesi nulla di non correlazione dei residui. Ora si regrediscono i residui con i residui ritardati per ottenere il coefficiente di autocorrelazione seriale di primo grado. Si parte dalle statistiche descrittive.

```
#-- R CODE
autocorr <- acf(resid(mod1),main="Autocorrelazion",lwd=2)
```




```
pander(data.frame(LAG=autocorr$lag,VALUE=autocorr$acf)[1:5,])
```

LAG	VALUE
0	1
1	0.5395
2	0.2211
3	0.08943
4	0.2293

```
#-- metodo alternativo per ottenere il corff. di autocorrelazione
cor(resid(mod1),c(NA,resid(mod1)[1:(length(resid(mod1))-1)]),use="pairwise.complete.obs")
```

[1] 0.5441966

Ora si regrediscono OLS "res" su OLS "res_1":

```
#-- R CODE
d1 <- data.frame(
  mod1$model,
  resid=resid(mod1),
  resid_l1=c(NA,resid(mod1)[1:(length(resid(mod1))-1)]) #-- residui ritardati
)
mod2 <- lm(resid ~ resid_l1,d1)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1119	0.6795	-0.1647	0.8702
resid_l1	0.5429	0.1503	3.612	0.001061

Table 13: Fitting linear model: resid ~ resid_l1

Observations	Residual Std. Error	R^2	Adjusted R^2
33	3.903	0.2961	0.2734

```
pander(anova(mod2),big.mark=",")
```

Table 14: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
resid_l1	1	198.7	198.7	13.04	0.001061
Residuals	31	472.2	15.23	NA	NA

```
pander(white.test(mod2),big.mark=",") #-- white test
```

Test.statistic	P.value
0.4612	0.7941

```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 16: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.889	0.3346	true autocorrelation is greater than 0

Si propone il modello corretto per la correlazione seriale sostituendo manualmente ad ogni valore Y_t il valore $\hat{Y}_t = Y_t - 0.542Y_{t-1}$ in modo tale che i nuovi residui siano tra loro incorrelati.

```
#-- R CODE
d1 <- data.frame(
  mod1$model,
  resid=resid(mod1)
)

d1$ftheft_l1 <- Lag(d1$ftheft,1)
d1$partic_l1 <- Lag(d1$partic,1)
d1$degrees_l1 <- Lag(d1$degrees,1)
d1$mtheft_l1 <- Lag(d1$mtheft,1)
d1$resid_l1 <- Lag(d1$resid,1)
```

```r
d1$int_tild <- 1-0.542

d1$ftheft_t <- d1$ftheft-0.542*d1$ftheft_l1
d1$partic_t <- d1$partic-0.542*d1$partic_l1
d1$degrees_t <- d1$degrees-0.542*d1$degrees_l1
d1$mtheft_t <- d1$mtheft-0.542*d1$mtheft_l1
d1$resid_t <- d1$resid-0.542*d1$resid_l1


mod3 <- lm(ftheft_t ~ 0 + int_tild + partic_t + degrees_t +  mtheft_t,d1)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
int_tild	-26.85	10.27	-2.616	0.01399
partic_t	0.1182	0.03536	3.343	0.002294
degrees_t	0.5169	0.0695	7.437	3.393e-08
mtheft_t	0.0426	0.03072	1.386	0.1762

Table 18: Fitting linear model: ftheft_t ~ 0 + int_tild + partic_t + degrees_t + mtheft_t

Observations	Residual Std. Error	R^2	Adjusted R^2
33	3.944	0.9518	0.9452

```r
pander(anova(mod3),big.mark=",")
```

Table 19: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
int_tild	1	6,773	6,773	435.4	5.184e-19
partic_t	1	1,115	1,115	71.65	2.508e-09
degrees_t	1	992.4	992.4	63.8	8.276e-09
mtheft_t	1	29.9	29.9	1.922	0.1762
Residuals	29	451.1	15.56	NA	NA

```r
pander(white.test(mod3),big.mark=",") #-- white test
```

Test.statistic	P.value
0.5589	0.7562

```r
pander(dwtest(mod3),big.mark=",") #-- Durbin-Whatson test
```

Table 21: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
1.795	0.1592	true autocorrelation is greater than 0

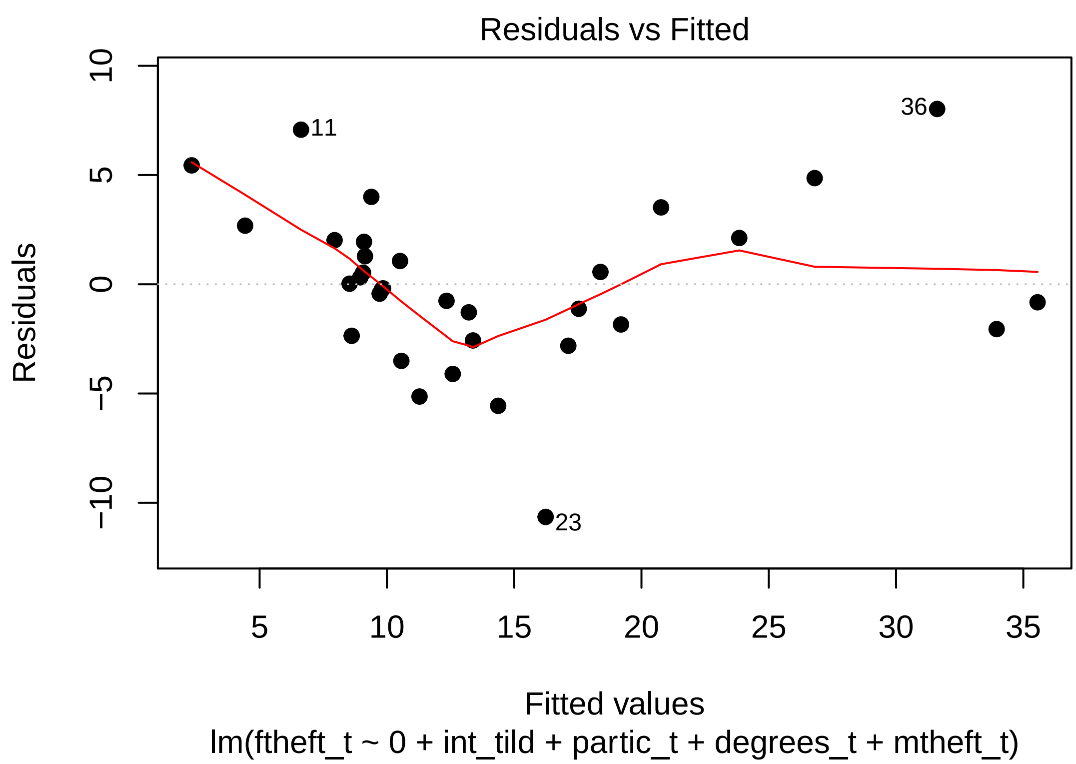
Test statistic	P value	Alternative hypothesis

Il modello interpreta ancora meglio i dati. Ancora "partic" e "degrees" risultano significative a significare che il numero di condanne delle donne per furto aumenta all'aumentare della quota di laureate sulla popolazione e partecipazione alla forza lavoro quasi a significare che un loro aumento di partecipazione alla vita in positivo significa anche un aumento della loro criminalità.

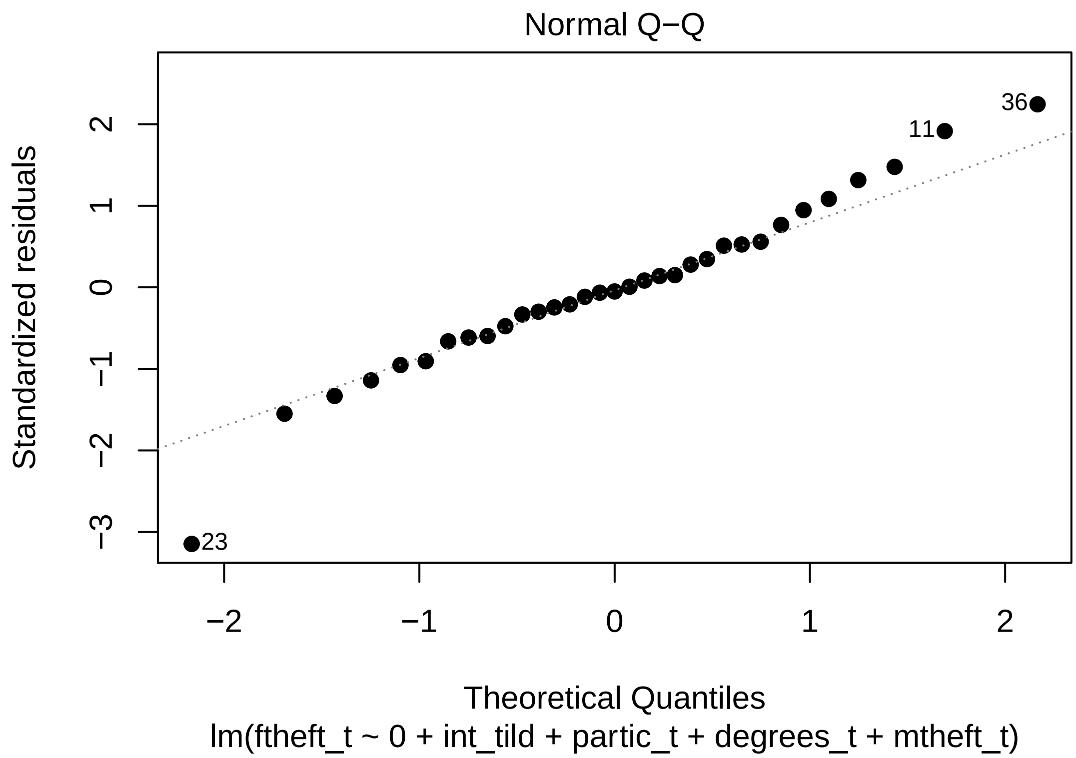
La non correlazione fra gli errori confermata dal test di Durbin-Watson.

Si conferma l'omoschedasticità dei residui e diminuiscono nettamente gli outlier.

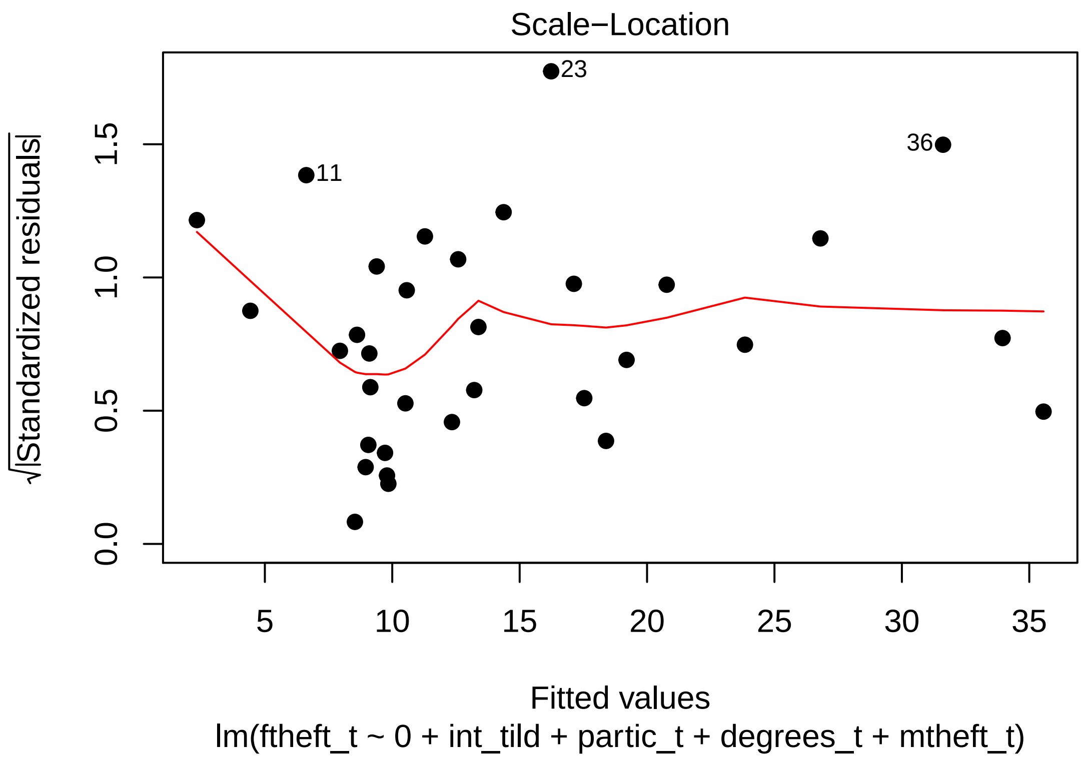
```r
#-- R CODE
plot(mod3,which=1,pch=19)
```



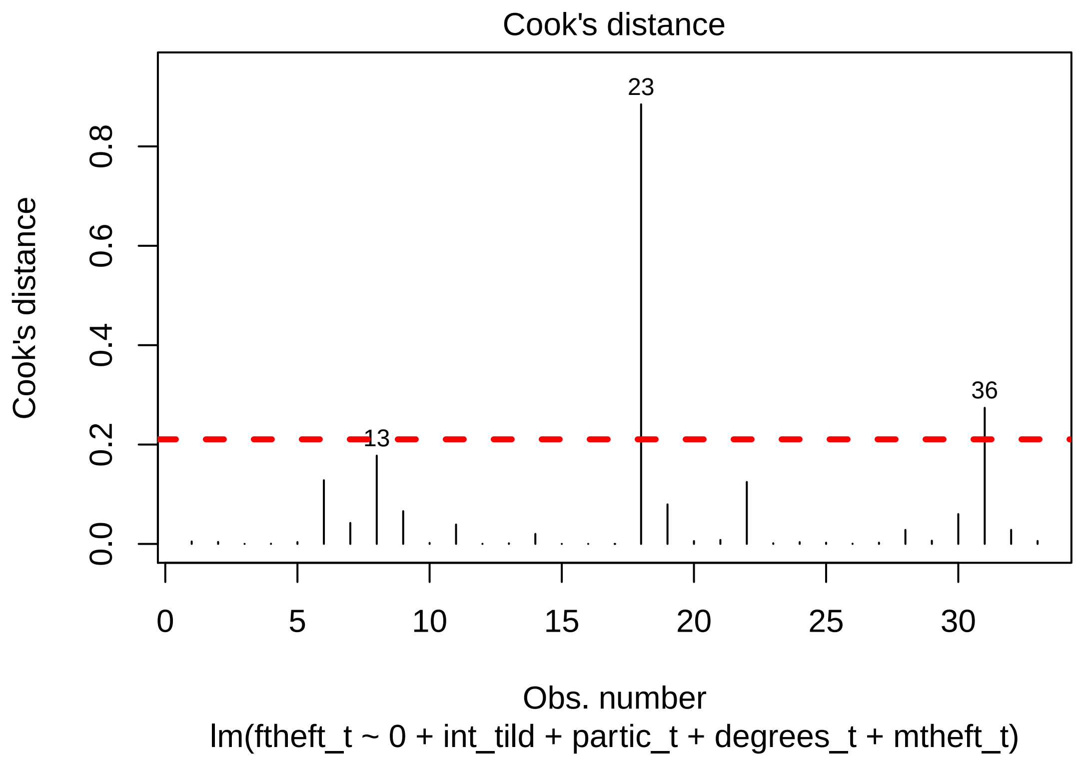

```r
plot(mod3,which=2,pch=19)
```



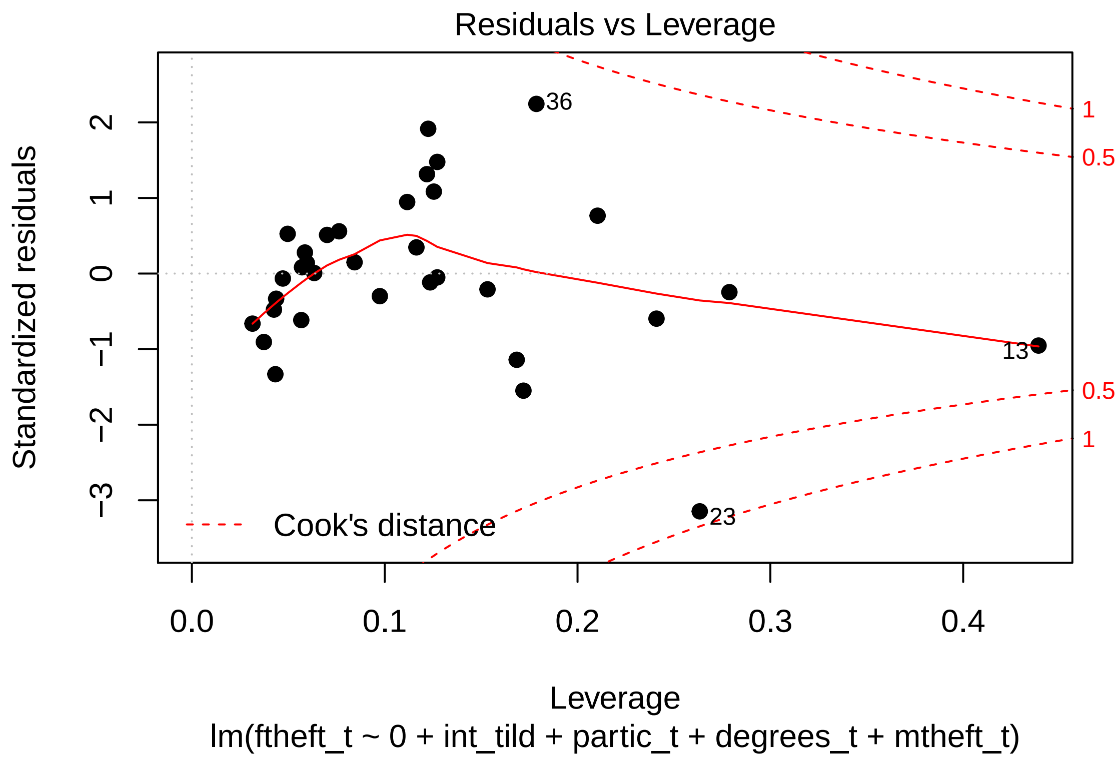

```
plot(mod3,which=3,pch=19)
```




```
plot(mod3,which=4,pch=19)
abline(h=2*4/nrow(d),col=2,lwd=3,lty=2)
```




```r
plot(mod3,which=5,pch=19)
```



Proviamo ora ad utilizzare funzioni che permettono di considerare automaticamente le correlazioni del residui.

```
#-- R CODE
mod4 <- arima(d1$ftheft, order=c(1,0,0), xreg = d1[,c("partic","degrees","mtheft")],method="ML")
mod4
```

```
##
## Call:
## arima(x = d1$ftheft, order = c(1, 0, 0), xreg = d1[, c("partic", "degrees",
##     "mtheft")], method = "ML")
##
## Coefficients:
##          ar1  intercept  partic  degrees  mtheft
##       0.9399    -0.1869  0.0797   0.2962  0.0162
## s.e.  0.0804    16.7414  0.0335   0.1290  0.0312
##
## sigma^2 estimated as 11.93:  log likelihood = -91.46,  aic = 194.93
```

```
coeftest(mod4)
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value Pr(>|z|)
## ar1         0.939946   0.080392 11.6920  < 2e-16 ***
## intercept -0.186927  16.741448 -0.0112  0.99109
```

```
## partic     0.079676   0.033487  2.3793  0.01735 *
## degrees    0.296180   0.129031  2.2954  0.02171 *
## mtheft     0.016240   0.031186  0.5207  0.60255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
durbinWatsonTest(as.numeric(mod4$residuals))
```

```
## [1] 1.849001
```

```r
#-- R CODE
mod5 <- arima(d1$ftheft, order=c(2,0,0), xreg = d1[,c("partic","degrees","mtheft")],method="ML")
```

```
## Warning in log(s2): NaNs produced
```

```r
mod5
```

```
##
## Call:
## arima(x = d1$ftheft, order = c(2, 0, 0), xreg = d1[, c("partic", "degrees",
##     "mtheft")], method = "ML")
##
## Coefficients:
##          ar1     ar2  intercept  partic  degrees  mtheft
##       1.4587  -0.487    12.9791  0.0590   0.0165  0.0325
## s.e.  0.2085   0.209    19.4903  0.0326   0.1249  0.0295
##
## sigma^2 estimated as 10.7:  log likelihood = -90.45,  aic = 194.89
```

```r
coeftest(mod5)
```

```
##
## z test of coefficients:
##
##            Estimate Std. Error z value  Pr(>|z|)
## ar1        1.458714   0.208529  6.9953 2.647e-12 ***
## ar2       -0.487049   0.209031 -2.3300   0.01980 *
## intercept 12.979050  19.490288  0.6659   0.50546
## partic     0.059006   0.032561  1.8122   0.06996 .
## degrees    0.016487   0.124897  0.1320   0.89498
## mtheft     0.032465   0.029529  1.0994   0.27157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
durbinWatsonTest(as.numeric(mod5$residuals), max.lag=2)
```

```
## [1] 2.128935 1.545536
```

```r
#-- R CODE
mod6 <- arima(d1$ftheft, order=c(3,0,0), xreg = d1[,c("partic","degrees","mtheft")],method="ML")
mod6
```

```
##
## Call:
## arima(x = d1$ftheft, order = c(3, 0, 0), xreg = d1[, c("partic", "degrees",
##     "mtheft")], method = "ML")
```

```
##
## Coefficients:
##          ar1      ar2      ar3  intercept  partic  degrees  mtheft
##       1.3113  -0.0252  -0.3348    17.7348  0.0418   0.0035  0.0168
## s.e.  0.1827   0.3130   0.1944    14.8364  0.0308   0.1127  0.0278
##
## sigma^2 estimated as 9.812:  log likelihood = -89.09,  aic = 194.18
```

```r
coeftest(mod6)
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value  Pr(>|z|)
## ar1        1.3113206  0.1826663  7.1788 7.034e-13 ***
## ar2       -0.0252079  0.3130354 -0.0805   0.93582
## ar3       -0.3348318  0.1943683 -1.7227   0.08495 .
## intercept 17.7348161 14.8364159  1.1954   0.23195
## partic     0.0418409  0.0307828  1.3592   0.17407
## degrees    0.0034557  0.1127231  0.0307   0.97554
## mtheft     0.0168085  0.0277591  0.6055   0.54484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
durbinWatsonTest(as.numeric(mod5$residuals), max.lag=3)
```

```
## [1] 2.128935 1.545536 1.427091
```