GLS 7 - Data set: CRIME

INTRODUZIONE

Il seguente dataset contiene il seguente set di variabili:

1. M: percentuale di maschi in età 14-24 anni
2. SO: dummy che indica se lo stato è del sud
3. ED: media degli anni trascorsi a suola
4. PO1: spese per la polizia nel 1960
5. PO2: spese per la polizia nel 1959
6. LF: tasso di forza lavoro
7. M.F: numero di maschi per 1000 femmine
8. POP: popolazione dello stato
9. NW: numero di individui non bianchi
10. U1: tasso di occupazione dei maschi in età 14-24
11. U2: tasso di occupazione dei maschi in età 35-39
12. GDP: gross domestic product per head
13. INEQ: income inequality
14. PROB: probabilità di essere imprigionato
15. TIME: tempo medio trascorso nelle prigioni dello stato
16. Y: tasso di crimini

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))
```

```r
}

#-- import dei dati
d <- UScrime

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("Ed","Po1","M.F","Pop","U1","U2","GDP","Time","y")

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

Table 1: Table continues below

M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq
151	1	91	58	56	510	950	33	301	108	41	394	261
143	0	113	103	95	583	1,012	13	102	96	36	557	194
142	1	89	45	44	533	969	18	219	94	33	318	250
136	0	121	149	141	577	994	157	80	102	39	673	167
141	0	121	109	101	591	985	18	30	91	20	578	174
121	0	110	118	115	547	964	25	44	84	29	689	126

Prob	Time	y
0.0846	26.2	791
0.0296	25.3	1,635
0.0834	24.3	578
0.0158	29.9	1,969
0.0414	21.3	1,234
0.0342	21	682

STATISTICHE DESCRITTIVE

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

Table 3: Table continues below

Ed	Po1	M.F	Pop
Min. : 87.0	Min. : 45.0	Min. : 934.0	Min. : 3.00
1st Qu.: 97.5	1st Qu.: 62.5	1st Qu.: 964.5	1st Qu.: 10.00
Median :108.0	Median : 78.0	Median : 977.0	Median : 25.00
Mean :105.6	Mean : 85.0	Mean : 983.0	Mean : 36.62
3rd Qu.:114.5	3rd Qu.:104.5	3rd Qu.: 992.0	3rd Qu.: 41.50
Max. :122.0	Max. :166.0	Max. :1071.0	Max. :168.00

U1	U2	GDP	Time	y
Min. : 70.00	Min. :20.00	Min. :288.0	Min. :12.20	Min. : 342.0
1st Qu.: 80.50	1st Qu.:27.50	1st Qu.:459.5	1st Qu.:21.60	1st Qu.: 658.5
Median : 92.00	Median :34.00	Median :537.0	Median :25.80	Median : 831.0
Mean : 95.47	Mean :33.98	Mean :525.4	Mean :26.60	Mean : 905.1
3rd Qu.:104.00	3rd Qu.:38.50	3rd Qu.:591.5	3rd Qu.:30.45	3rd Qu.:1057.5
Max. :142.00	Max. :58.00	Max. :689.0	Max. :44.00	Max. :1993.0

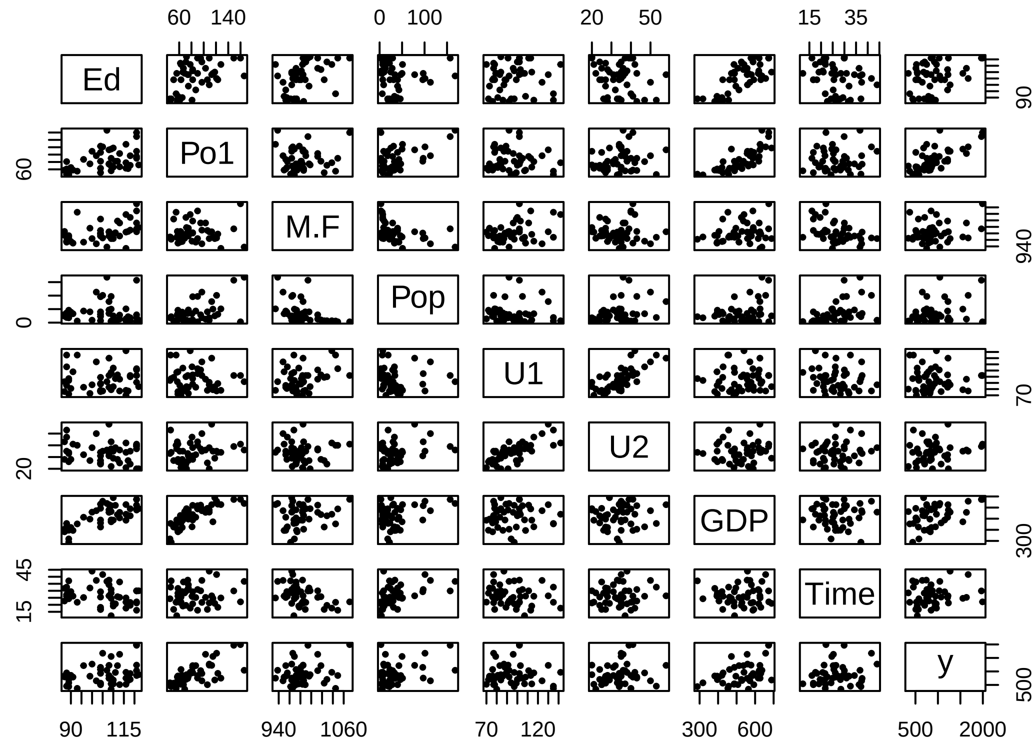
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 5: Table continues below

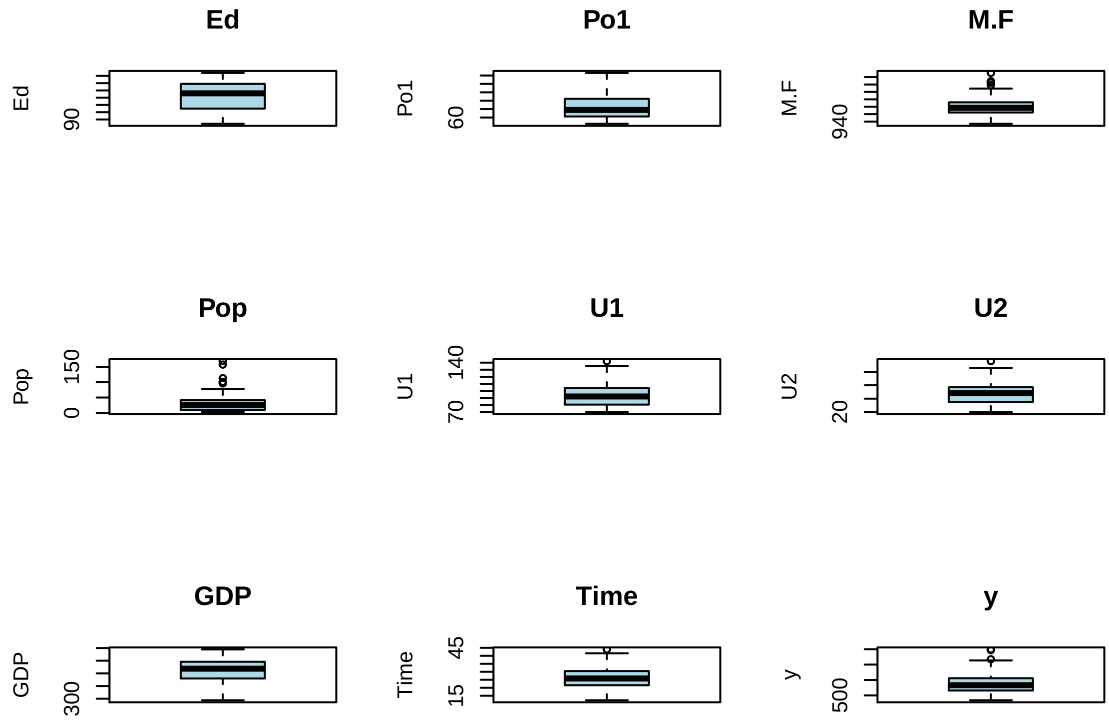
	Ed	Po1	M.F	Pop	U1	U2
Ed	1	0.483	0.4369	-0.01723	0.0181	-0.2157
Po1	0.483	1	0.03376	0.5263	-0.0437	0.1851
M.F	0.4369	0.03376	1	-0.4106	0.3519	-0.01869
Pop	-0.01723	0.5263	-0.4106	1	-0.03812	0.2704
U1	0.0181	-0.0437	0.3519	-0.03812	1	0.7459
U2	-0.2157	0.1851	-0.01869	0.2704	0.7459	1
GDP	0.736	0.7872	0.1796	0.3083	0.04486	0.09207
Time	-0.254	0.1034	-0.4277	0.4642	-0.1699	0.1014
y	0.3228	0.6876	0.2139	0.3375	-0.05048	0.1773

	GDP	Time	y
Ed	0.736	-0.254	0.3228
Po1	0.7872	0.1034	0.6876
M.F	0.1796	-0.4277	0.2139
Pop	0.3083	0.4642	0.3375
U1	0.04486	-0.1699	-0.05048
U2	0.09207	0.1014	0.1773
GDP	1	0.0006486	0.4413
Time	0.0006486	1	0.1499
y	0.4413	0.1499	1

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




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



REGRESSIONE

```r
#-- R CODE
mod1 <- lm(y ~ Ed + GDP + U1 + U2 + M.F + Po1, d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4,117	1,466	-2.808	0.007668
Ed	7.37	6.606	1.116	0.2712
GDP	-1.746	0.8762	-1.993	0.05309
U1	-10.28	4.301	-2.389	0.02168
U2	21.89	9.463	2.313	0.02593
M.F	4.584	1.668	2.748	0.008954
Po1	10.49	2.31	4.543	5.016e-05

Table 8: Fitting linear model: y ~ Ed + GDP + U1 + U2 + M.F + Po1

Observations	Residual Std. Error	R^2	Adjusted R^2
47	257.4	0.6147	0.5569

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

Table 9: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ed	1	717,146	717,146	10.82	0.0021
GDP	1	623,064	623,064	9.401	0.003878
U1	1	34,129	34,129	0.5149	0.4772
U2	1	866,003	866,003	13.07	0.0008317
M.F	1	621,536	621,536	9.378	0.003917
Po1	1	1,367,878	1,367,878	20.64	5.016e-05
Residuals	40	2,651,172	66,279	NA	NA

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

Test.statistic	P.value
9.519	0.00857

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

Table 11: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.868	0.3384	true autocorrelation is greater than 0

Gli errori risultano omoschedastici e incorrelati. Si ripropone ora il modello solo con le variabili significative.

```r
#-- R CODE
mod2 <- lm(y ~ U1 + U2 + M.F + Po1, d) #-- stima modello lineare semplice
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4,241	1,500	-2.828	0.00714
U1	-10.42	4.147	-2.513	0.0159
U2	20.18	8.42	2.397	0.02105
M.F	4.906	1.625	3.019	0.004296
Po1	7.446	1.426	5.223	5.145e-06

Table 13: Fitting linear model: y ~ U1 + U2 + M.F + Po1

Observations	Residual Std. Error	R^2	Adjusted R^2
47	263.5	0.5761	0.5357

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

Table 14: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
U1	1	17,533	17,533	0.2524	0.618
U2	1	716,850	716,850	10.32	0.002526
M.F	1	1,334,682	1,334,682	19.22	7.653e-05
Po1	1	1,894,779	1,894,779	27.28	5.145e-06
Residuals	42	2,917,084	69,454	NA	NA

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

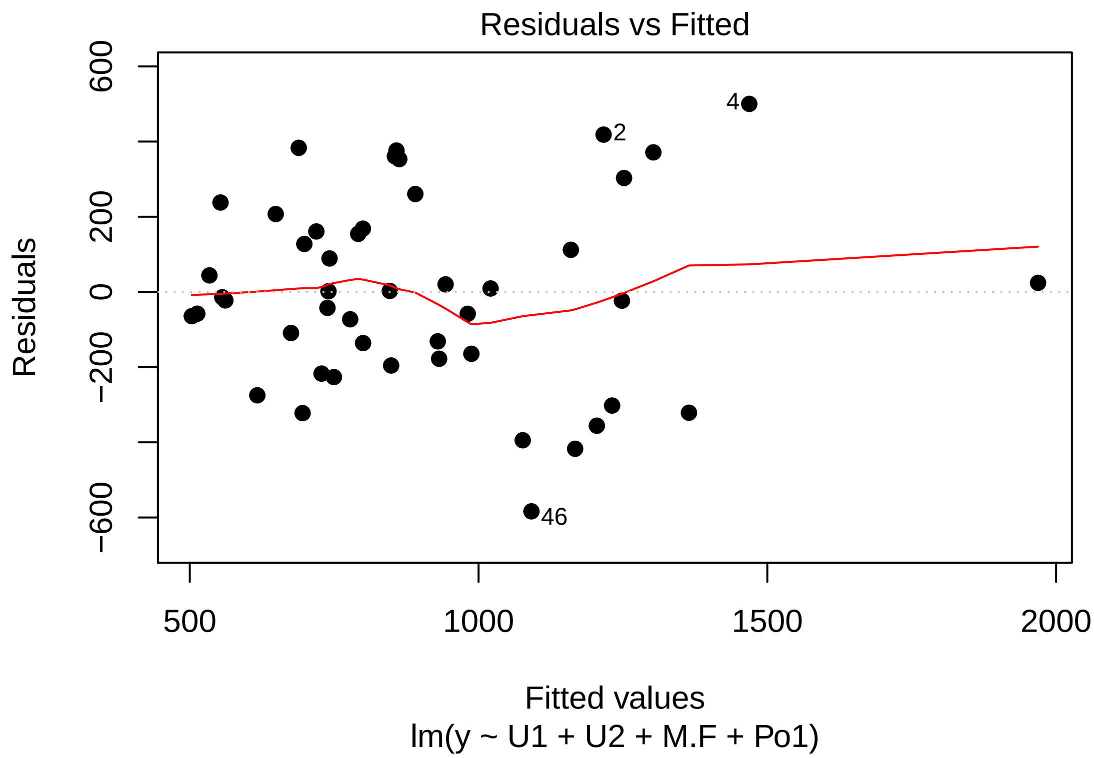
Test.statistic	P.value
10.16	0.00621

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

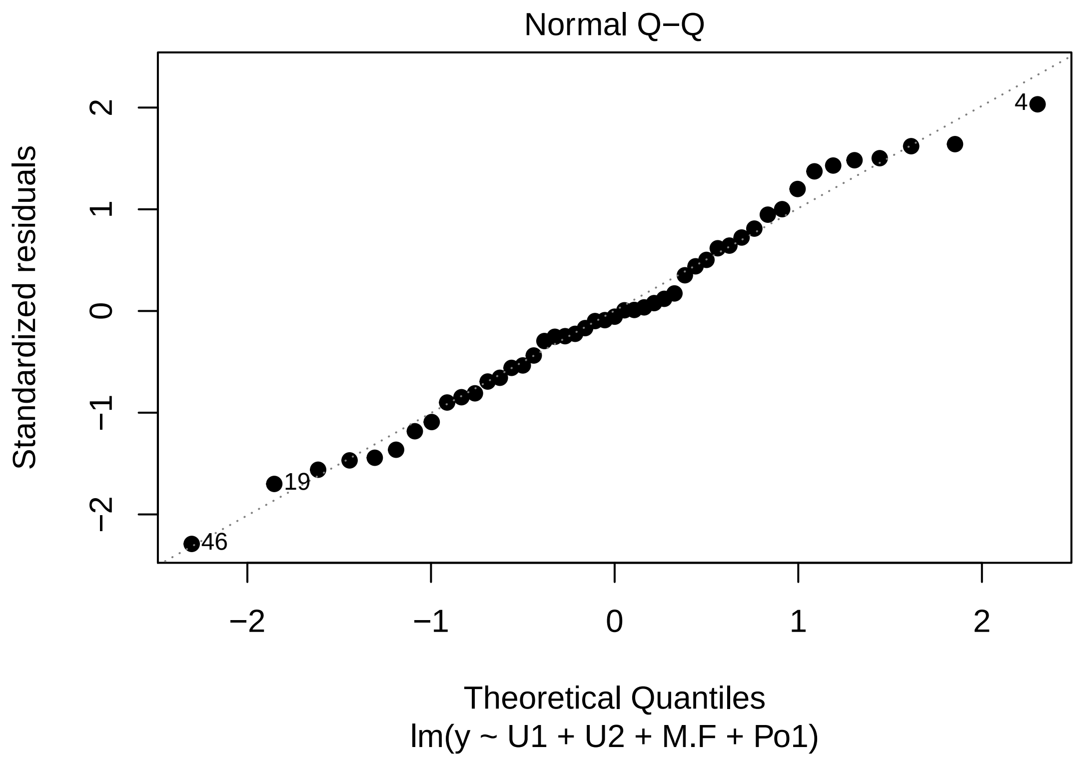
Table 16: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.649	0.1213	true autocorrelation is greater than 0

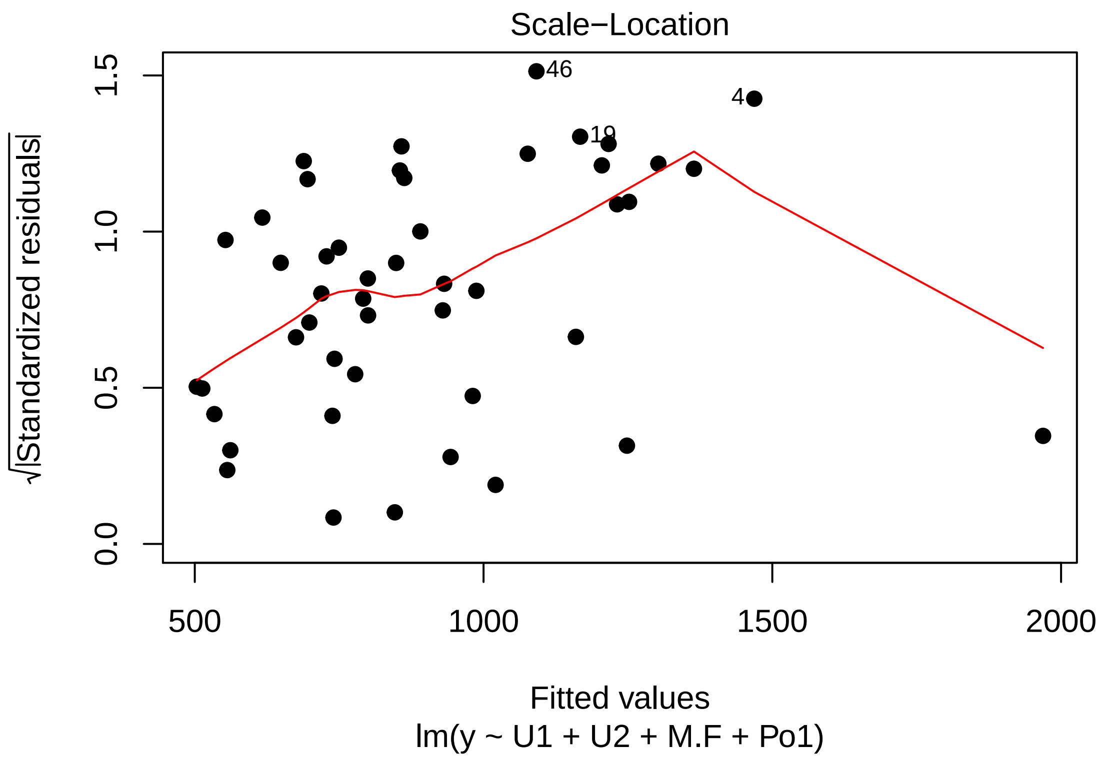
```r
#-- R CODE
plot(mod2,which=1,pch=19)
```



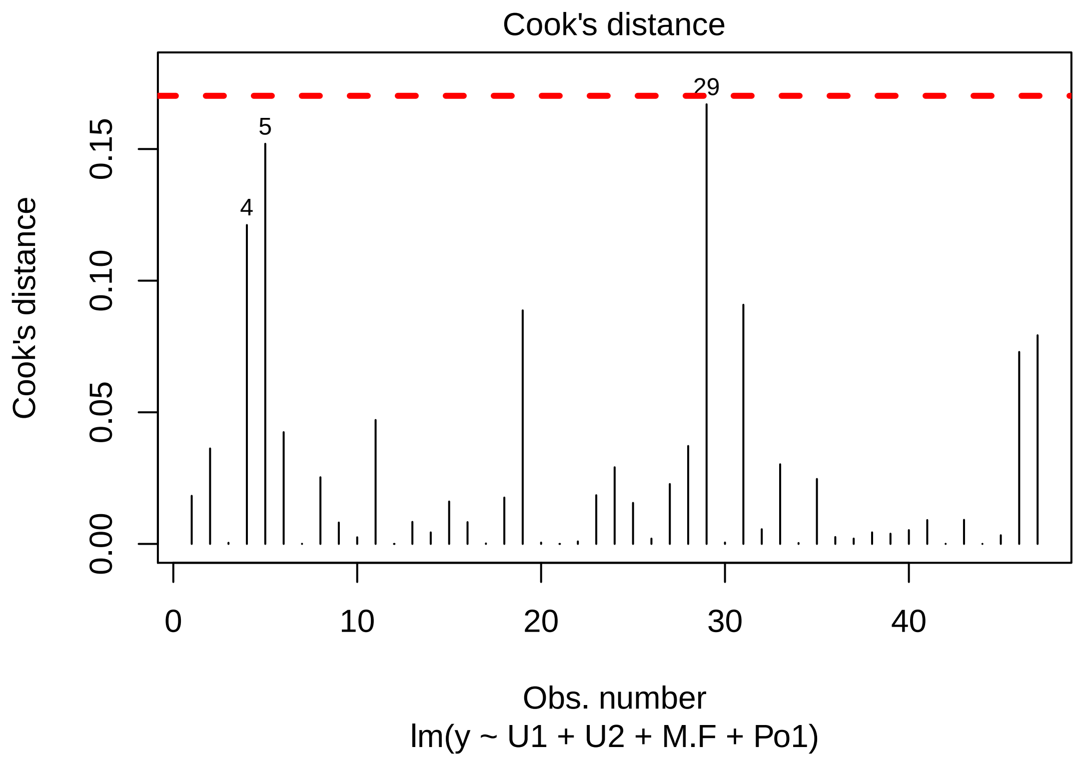

```
plot(mod2,which=2,pch=19)
```



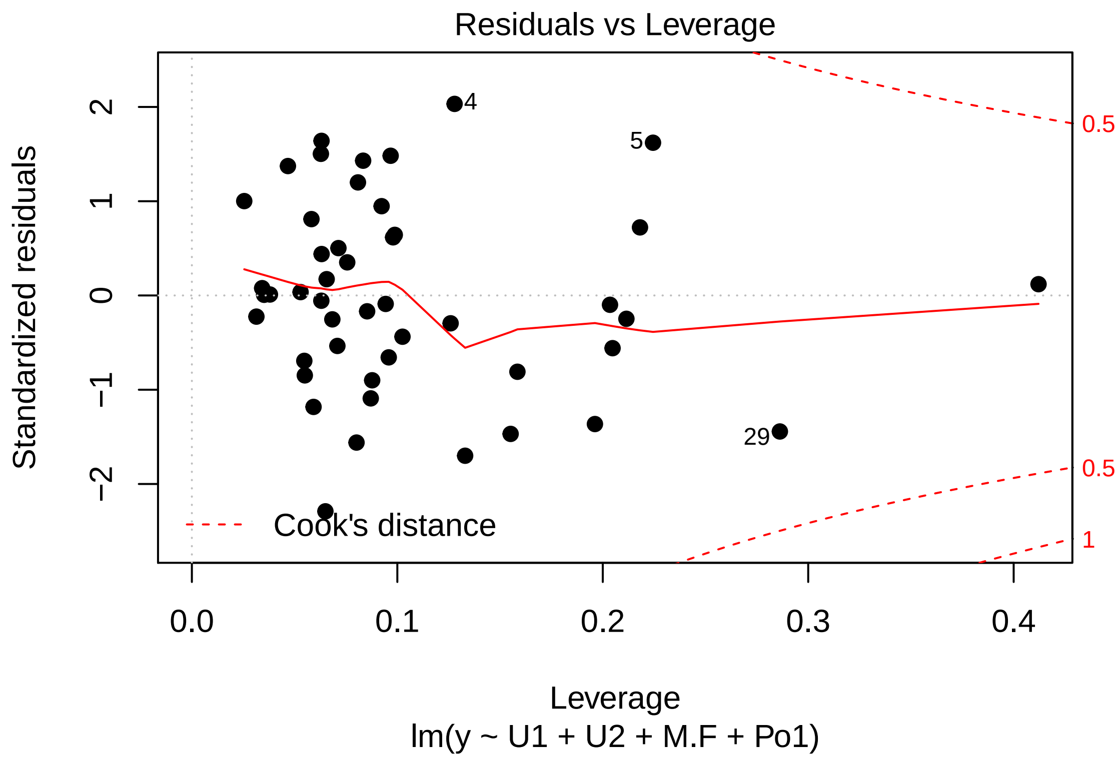

```
plot(mod2,which=3,pch=19)
```




```
plot(mod2,which=4,pch=19)
abline(h=2*4/nrow(d),col=2,lwd=3,lty=2)
```




```
plot(mod2,which=5,pch=19)
```



Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
4
5
29
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
Leverage
lm(y ~ U1 + U2 + M.F + Po1)