LINEAR 4 - Data set: ROAD

INTRODUZIONE

Il data set contiene contiene 26 osservazioni e le seguenti 7 variabili.

1. STATE: nome dello stato
2. DEATHS: numero di morti per incidenti stradali
3. DRIVERS: numero di automobilisti (in 10000s)
4. POPDEN: densità di popolazione per miglio gradrato
5. RURAL: lunghezza delle strade di tipo rurali
6. FUEL: consumo di carburante (in 10 000 000 galloni americani per anno)

Variabile dipendente: DEATHS

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare e polinomiale

```
#-- 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\\3.lin(5)\\4.linear\\road.txt"),sep=" ")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("deaths","drivers","popden","rural","temp","fuel")

#-- print delle prime 6 righe del dataset
pander(head(d))
```

country	deaths	drivers	popden	rural	temp	fuel
Alabama	968	158	64	66	62	119
Alaska	43	11	0.4	5.9	30	6.2
Arizona	588	91	12	33	64	65
Arkanas	640	92	34	73	51	74
Calif	4743	952	100	118	65	105
Colo	566	109	17	73	42	78

STATISTICHE DESCRITTIVE

Si propongono la matrice di correlazione tra le variabili e alcune descrittive di base.

```
#-- R CODE
pander(summary(d[,VAR_NUMERIC])) #-- statistiche descrittive
```

Table 2: Table continues below

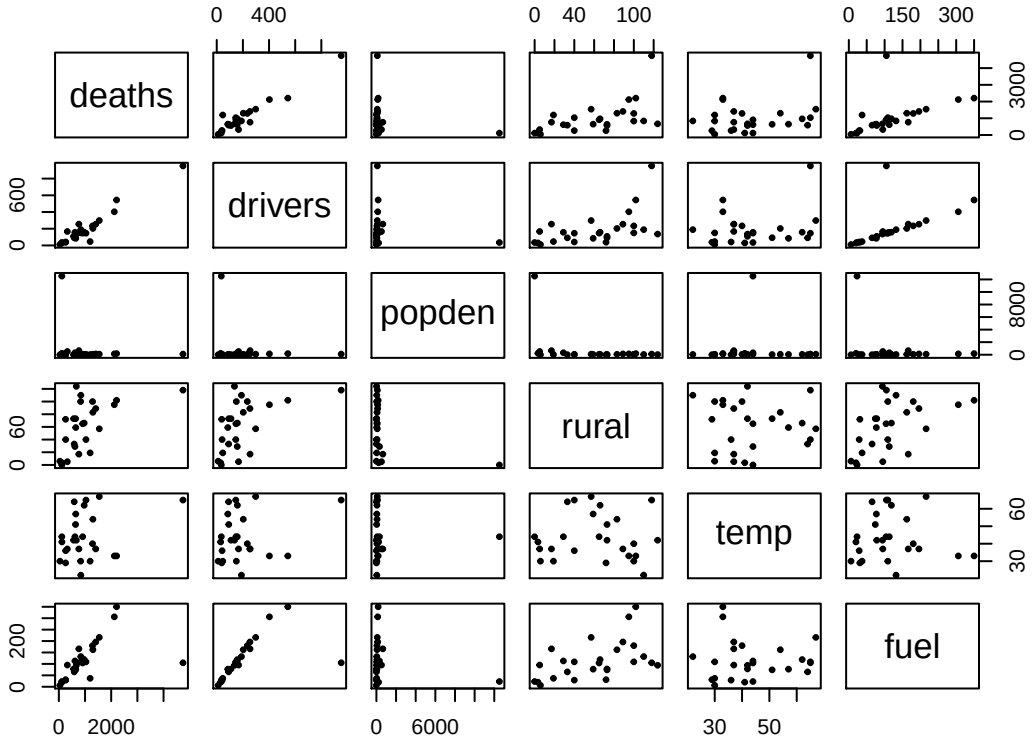
deaths	drivers	popden	rural
Min. : 43.0	Min. : 11.0	Min. : 0.40	Min. : 0.00
1st Qu.: 571.5	1st Qu.: 86.5	1st Qu.: 31.75	1st Qu.: 30.00
Median : 799.5	Median :148.5	Median : 66.00	Median : 65.50
Mean :1000.7	Mean :191.2	Mean : 595.74	Mean : 60.71
3rd Qu.:1265.8	3rd Qu.:226.2	3rd Qu.: 135.00	3rd Qu.: 93.50
Max. :4743.0	Max. :952.0	Max. :12524.00	Max. :124.00

temp	fuel
Min. :22.00	Min. : 6.20
1st Qu.:33.75	1st Qu.: 67.25
Median :41.50	Median :104.50
Mean :43.69	Mean :115.24
3rd Qu.:53.25	3rd Qu.:154.50
Max. :67.00	Max. :350.00

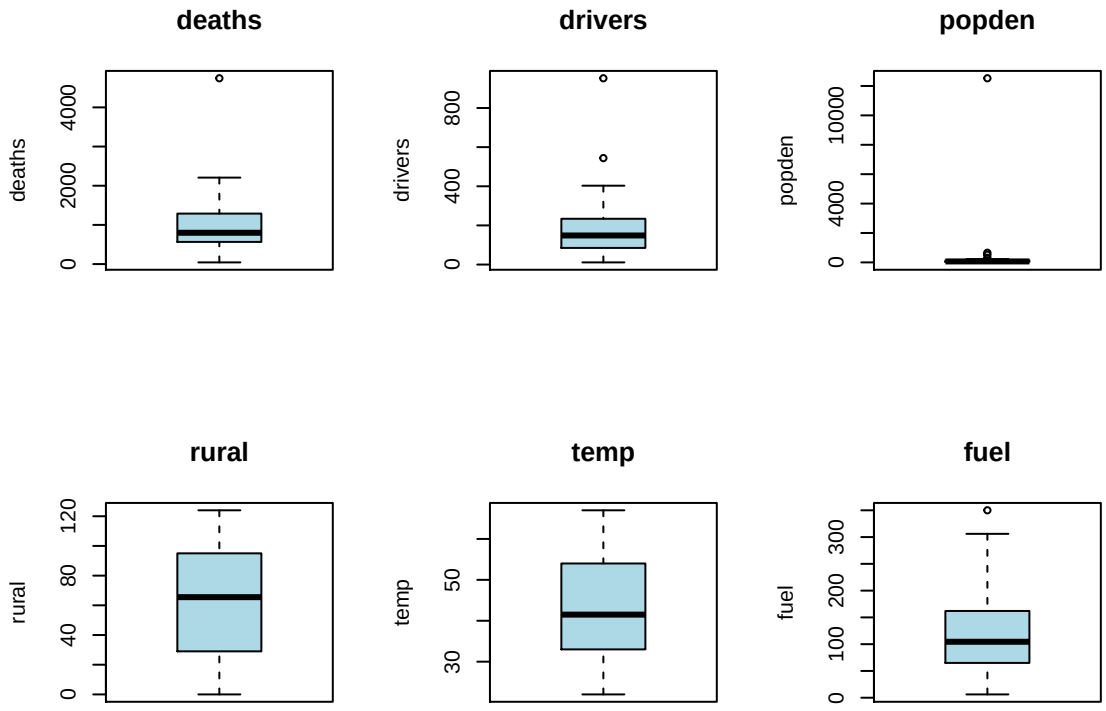
```
pander(cor(d[,VAR_NUMERIC])) #-- matrice di correlazione
```

	deaths	drivers	popden	rural	temp	fuel
deaths	1	0.9555	-0.1924	0.5629	0.3034	0.521
drivers	0.9555	1	-0.1512	0.532	0.2383	0.5895
popden	-0.1924	-0.1512	1	-0.347	-0.003083	-0.2093
rural	0.5629	0.532	-0.347	1	-0.01977	0.5107
temp	0.3034	0.2383	-0.003083	-0.01977	1	-0.01121
fuel	0.521	0.5895	-0.2093	0.5107	-0.01121	1

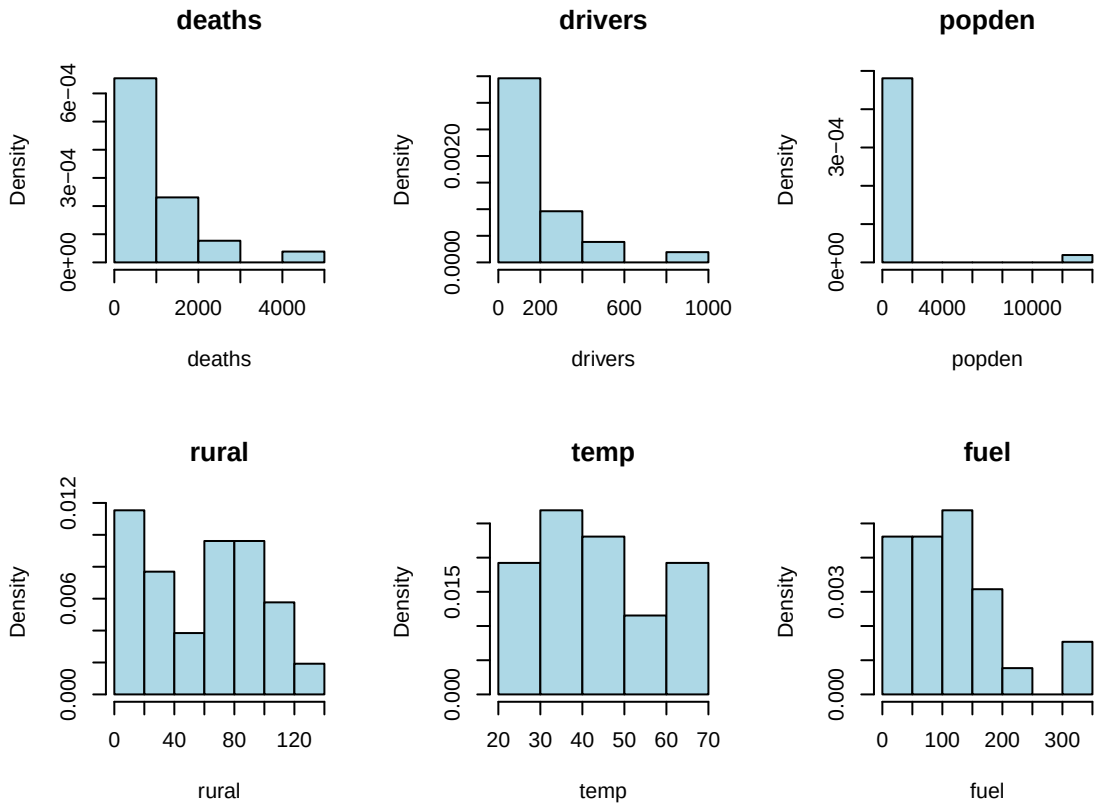
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```r
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```




```
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```



Si nota la fortissima correlazione fra "deaths" e "drivers", come era ragionevole aspettarsi.

REGRESSIONE

Si propone una regressione di "deaths" su "rural".

```
#-- R CODE
mod1 <- lm(deaths~rural,d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	157.6	297.3	0.5302	0.6008
rural	13.89	4.162	3.336	0.002757

Table 6: Fitting linear model: deaths ~ rural

Observations	Residual Std. Error	R^2	Adjusted R^2
26	798.7	0.3168	0.2884

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

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rural	1	7,100,906	7,100,906	11.13	0.002757
Residuals	24	15,311,856	637,994	NA	NA

```
pander(white.test(mod1),big.mark=",") #-- White test (per dettagli ?bptest)
```

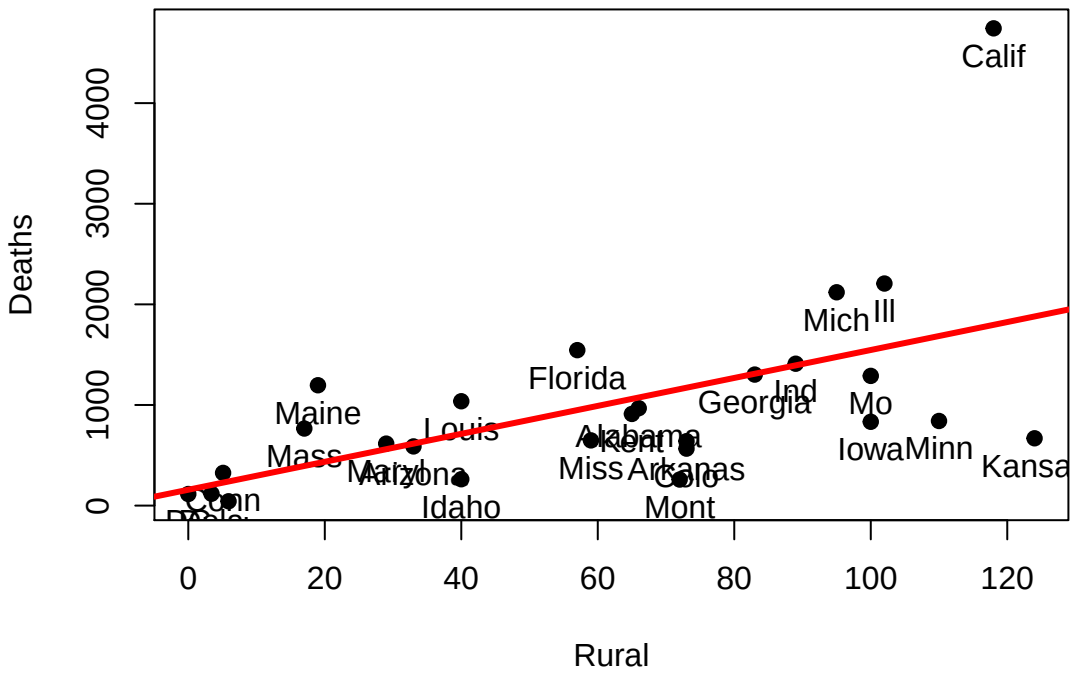
Test.statistic	P.value
8.063	0.01775

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

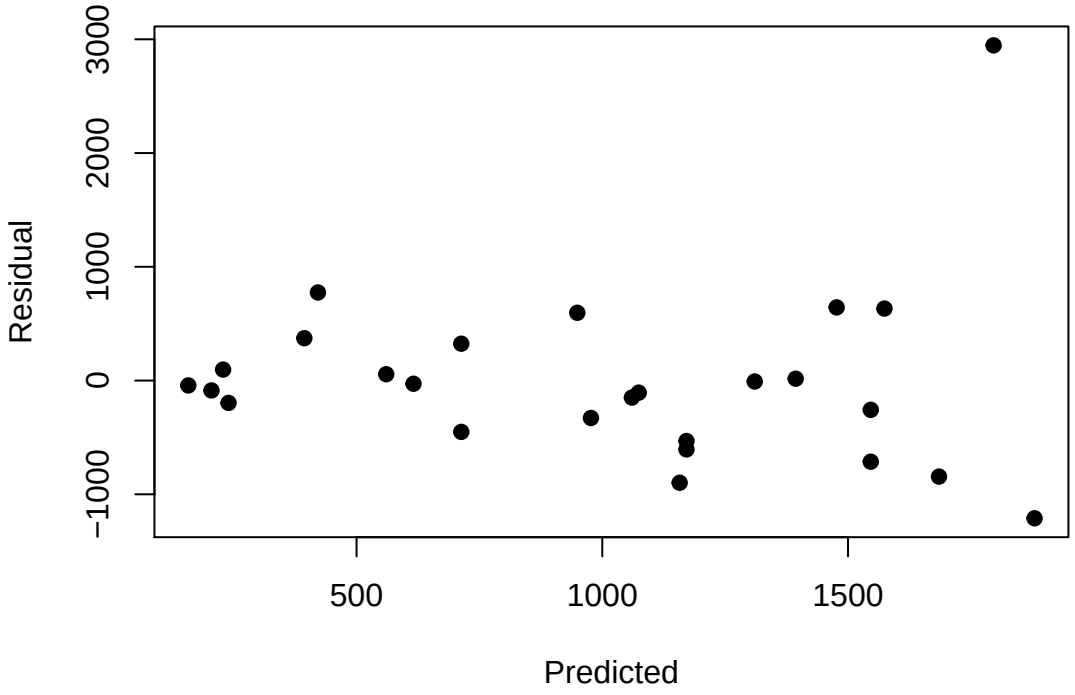
Table 9: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
2.219	0.6793	true autocorrelation is greater than 0

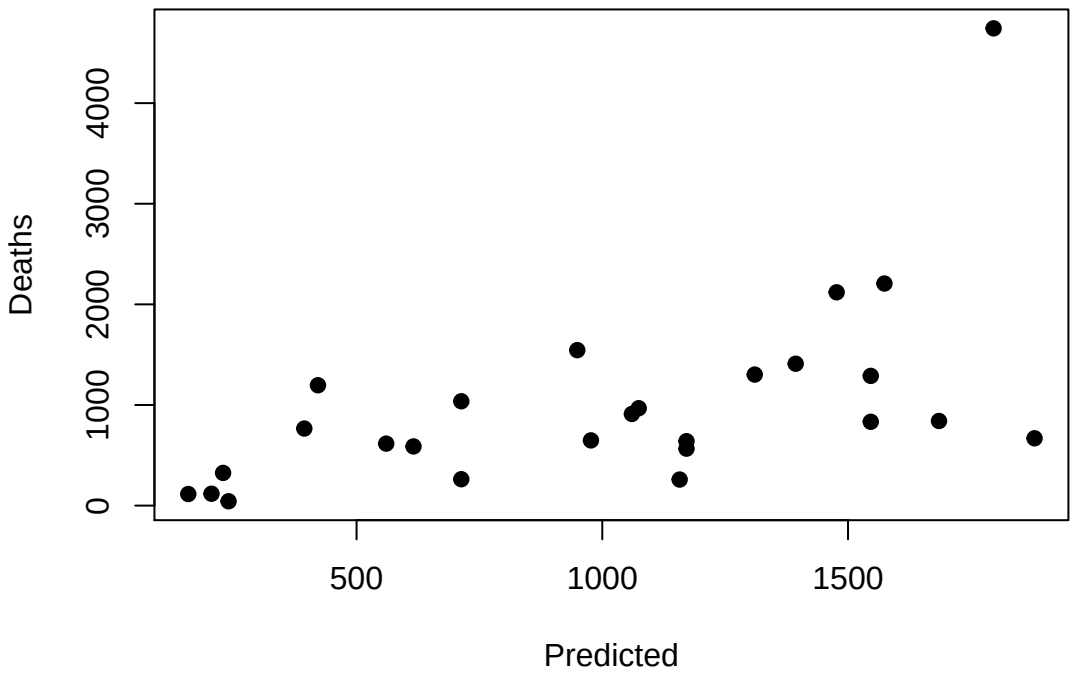
```
#-- R CODE
plot(d$rural,d$deaths,pch=19,xlab="Rural",ylab="Deaths")
text(d$rural,d$deaths,d$country,pos=1)
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
```




```
#-- R CODE
plot(fitted(mod1),resid(mod1),pch=19,xlab="Predicted",ylab="Residual")
```




```
plot(fitted(mod1),d$deaths,pch=19,xlab="Predicted",ylab="Deaths")
```

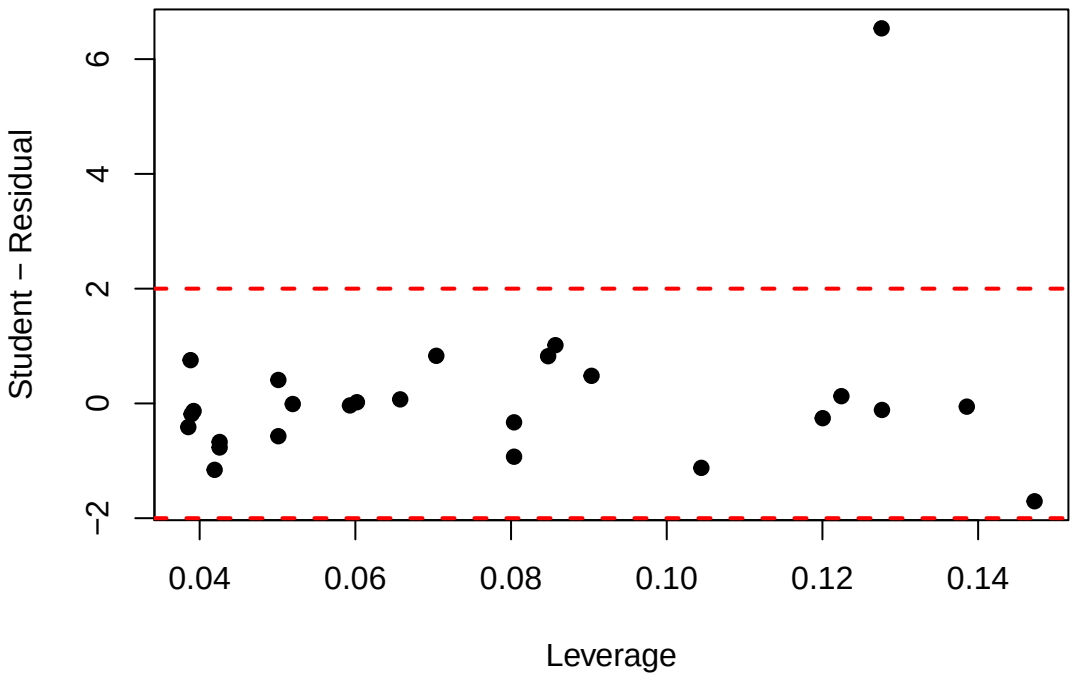


Il fitting è modesto ($R^2 = 0.3168$) e il parametro associato alla variabile "rural" è significativo. Le rappresentazioni grafiche mostrano che i residui sono omoschedastici ma esiste un outlier che andrebbe eliminato la California.

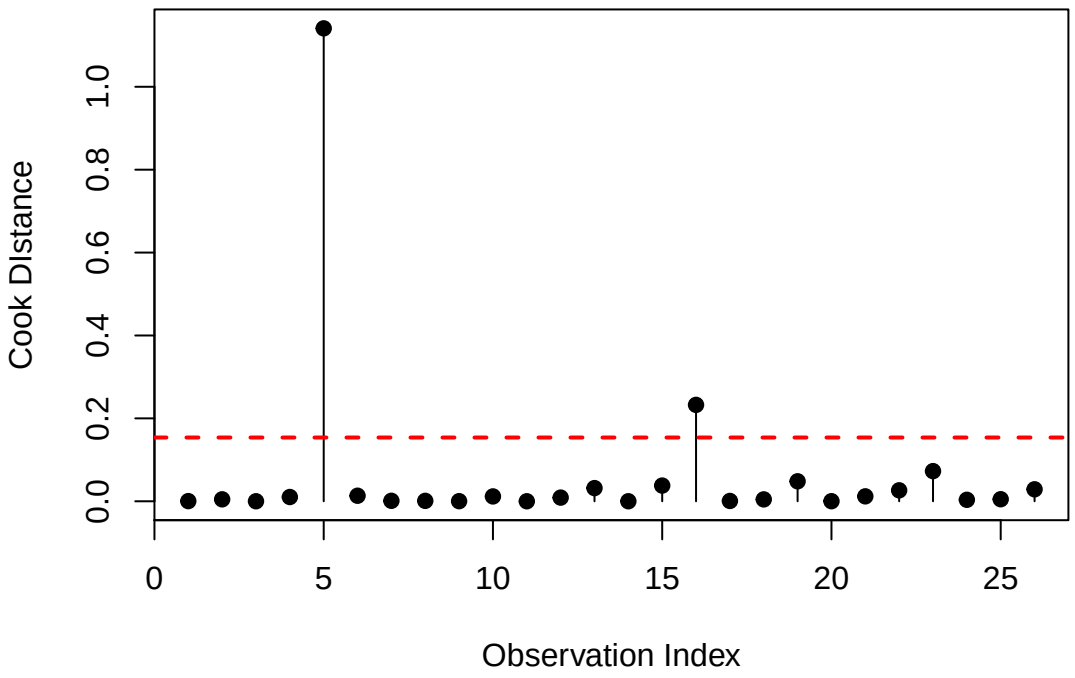
La presenza di tale outlier è confermata da tutti i grafici.

```r
#-- R CODE

plot(hatvalues(mod1),rstudent(mod1),pch=19,xlab="Leverage",ylab="Student - Residual")
abline(h=2,col=2,lty=2,lwd=2)
abline(h=-2,col=2,lty=2,lwd=2)
```



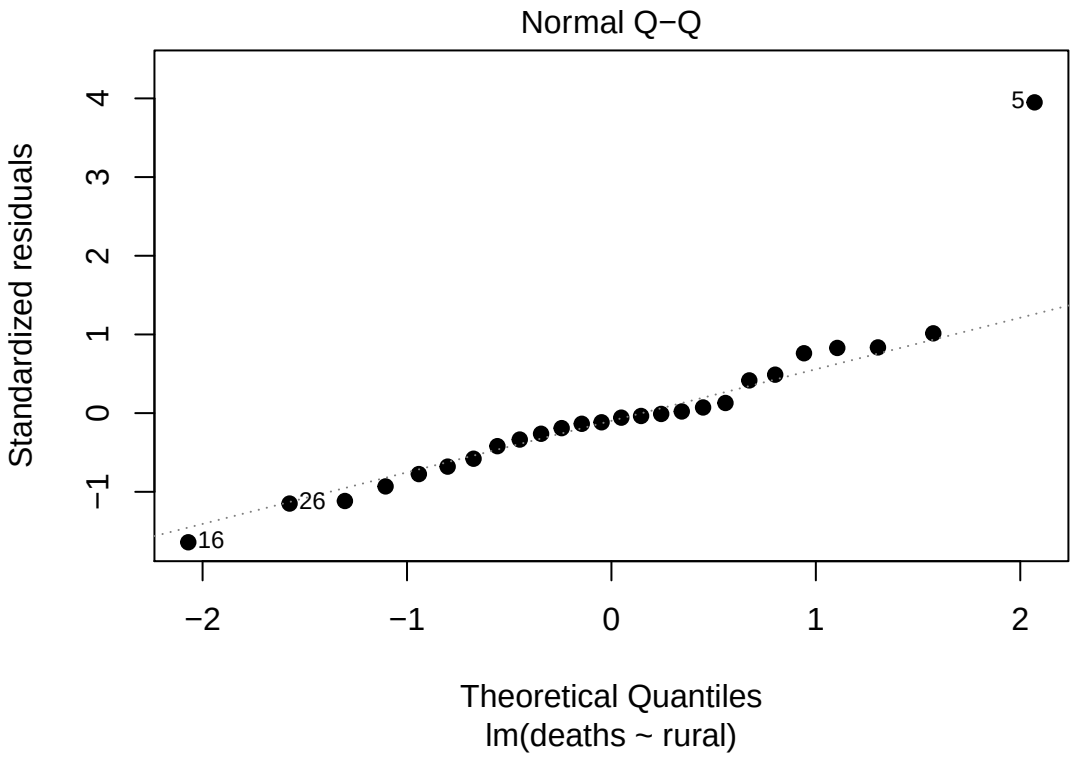

```
plot(cooks.distance(mod1),pch=19,xlab="Observation Index",ylab="Cook DIstance",type="h")
points(cooks.distance(mod1),pch=19)
abline(h=4/nrow(d),col=2,lty=2,lwd=2)
```



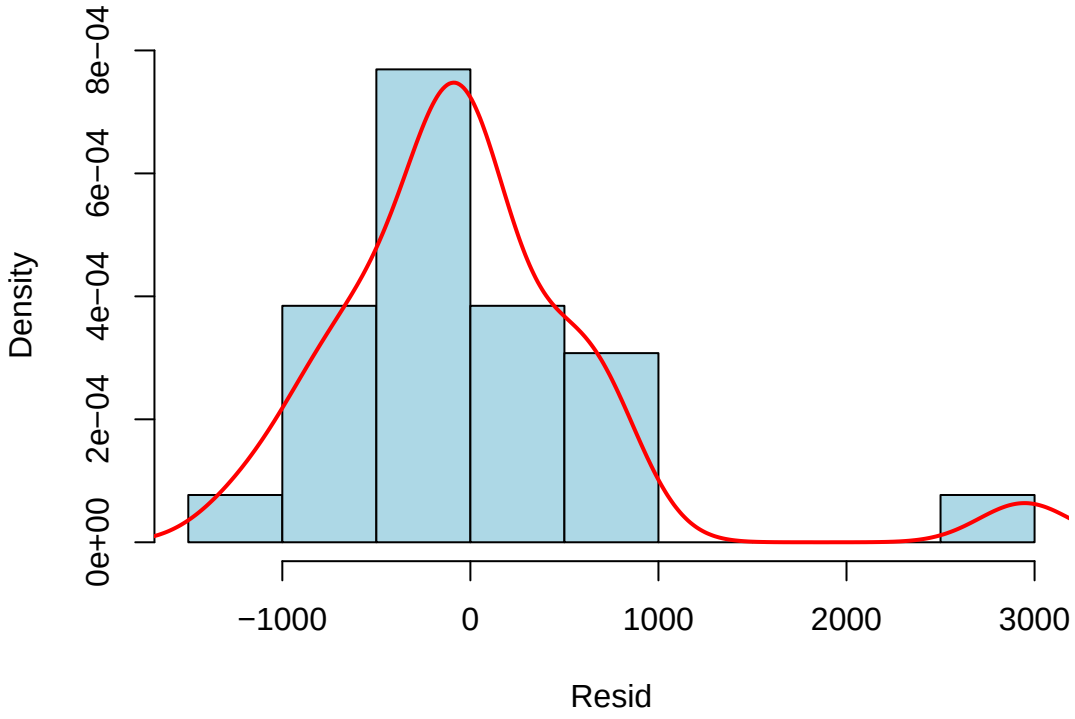
Il Q-Q plot e la distribuzione dei residui mostrano che i residui sono normali eccetto che per la presenza del citato outlier e nelle code.

```
#-- 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)
```



Invece di eliminare l'osservazione California si prova a vedere se una funzione non lineare può interpretarla come può intepretare meglio la varaibile "deaths". Si provi innanzitutto con il modello quadratico.

```
#-- R CODE
mod2 <- lm(deaths~rural+I(rural^2),d)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	359.8	392.6	0.9164	0.369
rural	2.203	15.25	0.1445	0.8864
I(rural^2)	0.09941	0.1248	0.7968	0.4337

Table 11: Fitting linear model: deaths ~ rural + I(rural^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
26	804.9	0.3352	0.2774

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

Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rural	1	7,100,906	7,100,906	10.96	0.003051
I(rural^2)	1	411,303	411,303	0.6349	0.4337
Residuals	23	14,900,553	647,850	NA	NA

```
pander(white.test(mod2),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
12.61	0.001823

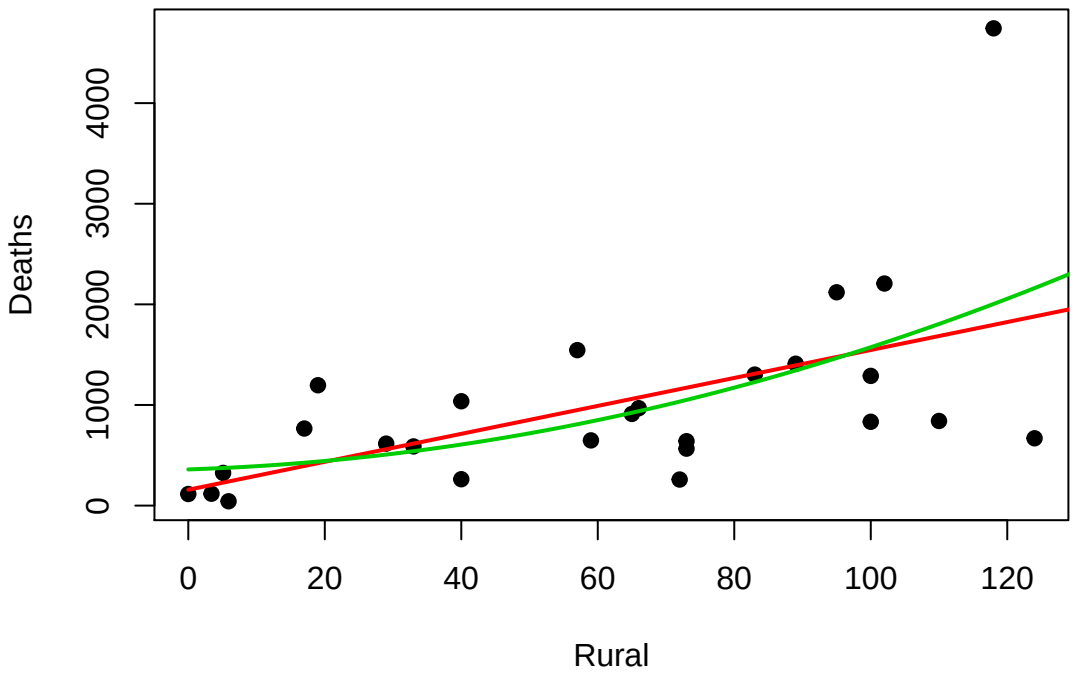
```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 14: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
2.114	0.5928	true autocorrelation is greater than 0

Il fitting migliora leggermente ma "rural" e $rural^2$ non risultano significativi quindi non è adeguato il modello quadratico come si vede anche dalla rappresentazione grafica seguente:

```
#-- R CODE
plot(d$rural,d$deaths,pch=19,xlab="Rural",ylab="Deaths")
lines(seq(0,25000,1),predict(mod1,data.frame(rural=seq(0,25000,1))),col=2,lwd=2)
lines(seq(0,25000,1),predict(mod2,data.frame(rural=seq(0,25000,1))),col=3,lwd=2)
```



Si prova ora con il modello lin-log in cui la variabile esplicativa è $log(Rural)$. Si deve per forza eliminare l'osservazione distretto di Washington perché la lunghezza delle strade rurali è zero e quindi il logaritmo di zero sarebbe infinito che non ha senso.

```
#-- R CODE

d_log <- d[!is.infinite(log(d$rural)),] #-- elimino le osservazioni che hanno log(rural)=Infinito

mod3 <- lm(deaths~I(log(rural)),d_log)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-646.5	678.9	-0.9523	0.3508
I(log(rural))	440.1	171.8	2.561	0.01747

Table 16: Fitting linear model: deaths ~ I(log(rural))

Observations	Residual Std. Error	R^2	Adjusted R^2
25	854.8	0.2219	0.1881

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

Table 17: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(rural))	1	4,792,153	4,792,153	6.559	0.01747
Residuals	23	16,804,851	730,646	NA	NA

```
pander(white.test(mod3),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
3.359	0.1865

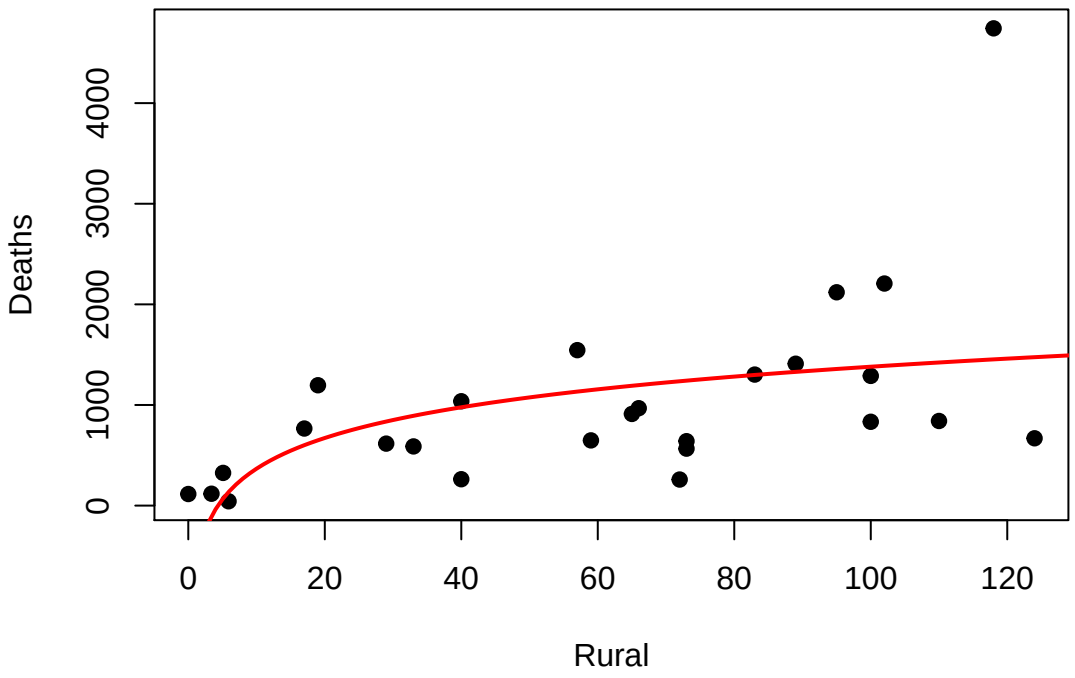
```
pander(dwtest(mod3),big.mark=",") #-- Durbin-Whatson test
```

Table 19: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
2.407	0.8293	true autocorrelation is greater than 0

Ora $log(Rural)$ è significativo, ma il fitting peggiora e quindi il modello non va bene. Lo si ve anche dalla rappresentazione grafica seguente:

```
#-- R CODE
plot(d$rural,d$deaths,pch=19,xlab="Rural",ylab="Deaths")
lines(seq(0,25000,1),predict(mod3,data.frame(rural=seq(0,25000,1))),col=2,lwd=2)
```



Si propone ora il modello log lineare ove la variabile $log(Death)$ dipende da "rural".

```
#-- R CODE
mod4 <- lm(I(log(deaths))~rural,d)
pander(summary(mod4),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.437	0.2845	19.11	5.034e-16
rural	0.01773	0.003984	4.451	0.0001679

Table 21: Fitting linear model: I(log(deaths)) ~ rural

Observations	Residual Std. Error	R^2	Adjusted R^2
26	0.7644	0.4522	0.4293

```
pander(anova(mod4),big.mark=",")
```

Table 22: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rural	1	11.57	11.57	19.81	0.0001679
Residuals	24	14.02	0.5843	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)

```
pander(white.test(mod4),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
5.261	0.07205

```
pander(dwtest(mod4),big.mark=",") #-- Durbin-Whatson test
```

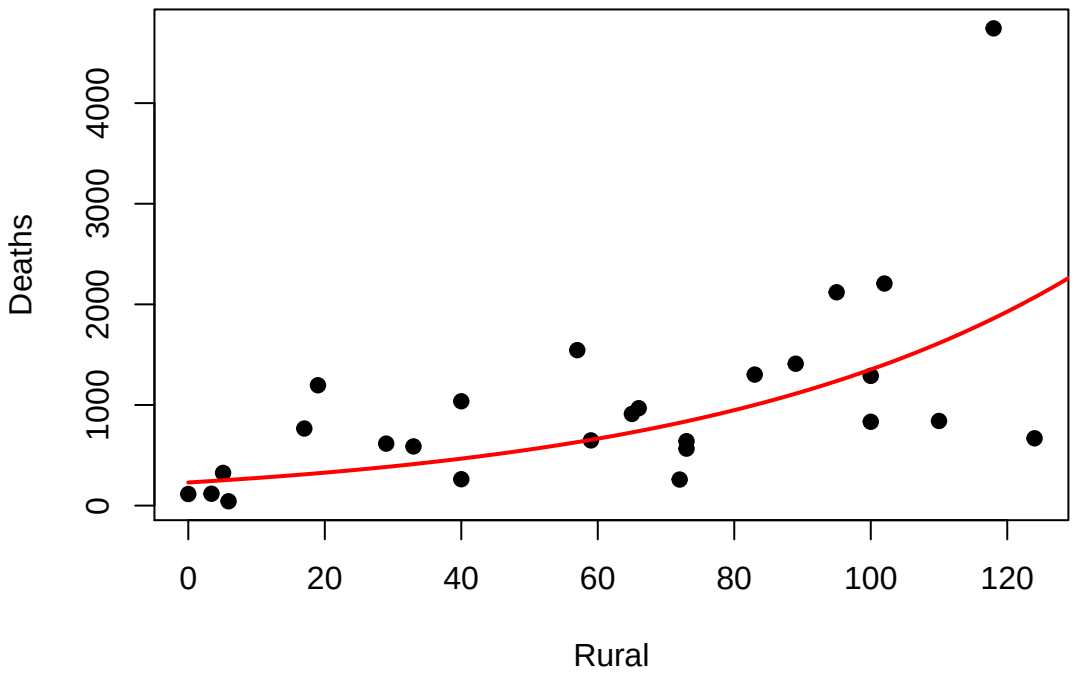
Table 24: Durbin-Watson test: `mod4`

Test statistic	P value	Alternative hypothesis
1.85	0.3139	true autocorrelation is greater than 0

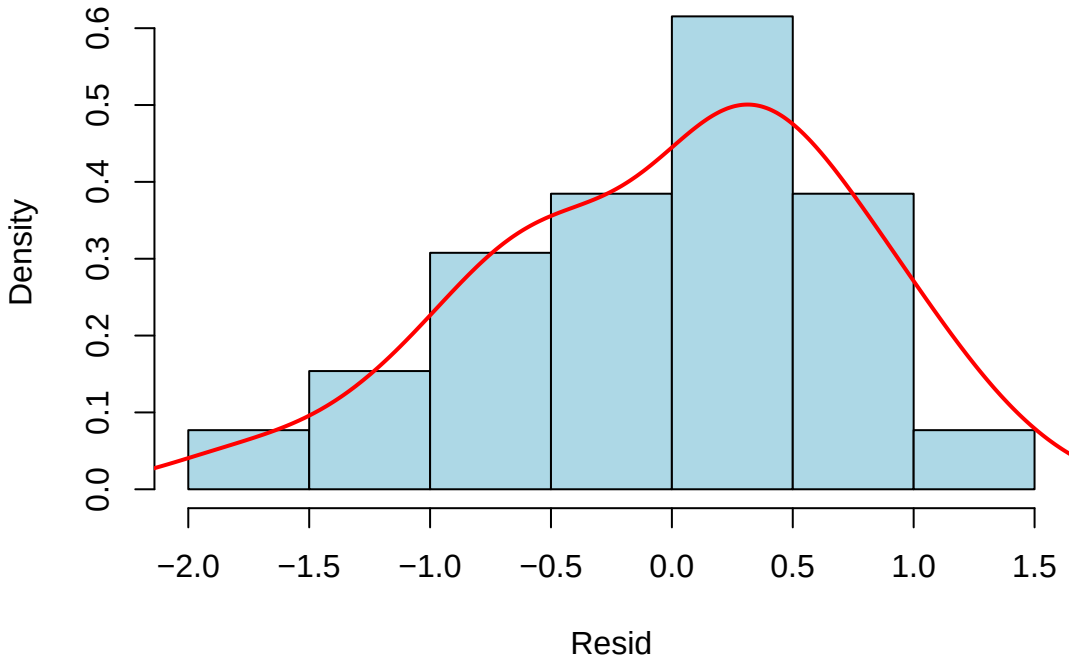
La rappresentazione grafica conferma il buon fitting del modello eccetto che per California.

Gli errori sono chiaramente omoschedatici.

```
#-- R CODE
plot(d$rural,d$deaths,pch=19,xlab="Rural",ylab="Deaths")
lines(seq(0,25000,1),exp(predict(mod4,data.frame(rural=seq(0,25000,1)))),col=2,lwd=2)
```




```
#-- R CODE
hist(resid(mod4),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod4)),col=2,lwd=2)
```



Si propone ora il modello log-log in cui la variabile $log(Deaths)$ dipende da $log(Rural)$. Migliora il fitting ed è il massimo tra i modelli proposti. $Log(Rural)$ è significativo. Le rappresentazioni grafiche confermano che il modello è il più adeguato tra quelli proposti e i residui sono omoschedastici. Le statistiche inerenti gli outlier mostrano che solo California ha valori fuori norma. Il Q-Q plot e la distribuzione dei residui mostrano con chiarezza che ora tali residui sono chiaramente normali.

```
#-- R CODE
mod5 <- lm(I(log(deaths))~I(log(rural)),d_log)
pander(summary(mod5),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.027	0.5564	7.239	2.281e-07
I(log(rural))	0.6686	0.1408	4.747	8.737e-05

Table 26: Fitting linear model: I(log(deaths)) ~ I(log(rural))

Observations	Residual Std. Error	R^2	Adjusted R^2
25	0.7005	0.4949	0.473

```
pander(anova(mod5),big.mark=",")
```

Table 27: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(rural))	1	11.06	11.06	22.54	8.737e-05
Residuals	23	11.29	0.4908	NA	NA

```
pander(white.test(mod5),big.mark=",") #-- White test (per dettagli ?bptest)
```

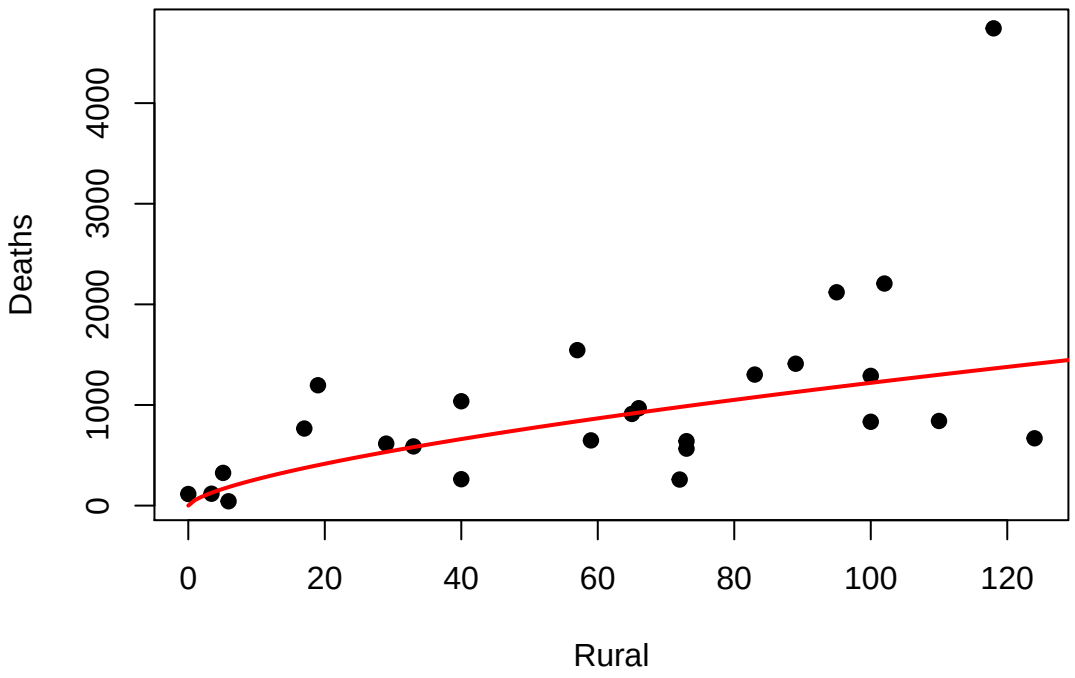
Test.statistic	P.value
1.141	0.5651

```
pander(dwtest(mod5),big.mark=",") #-- Durbin-Whatson test
```

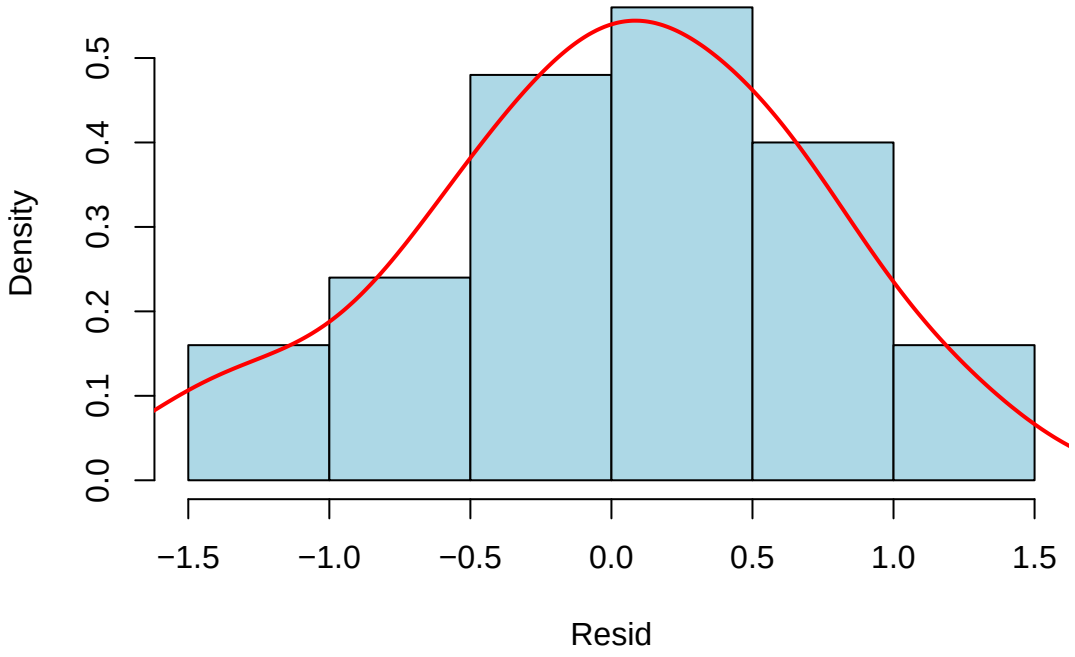
Table 29: Durbin-Watson test: `mod5`

Test statistic	P value	Alternative hypothesis
2.04	0.508	true autocorrelation is greater than 0

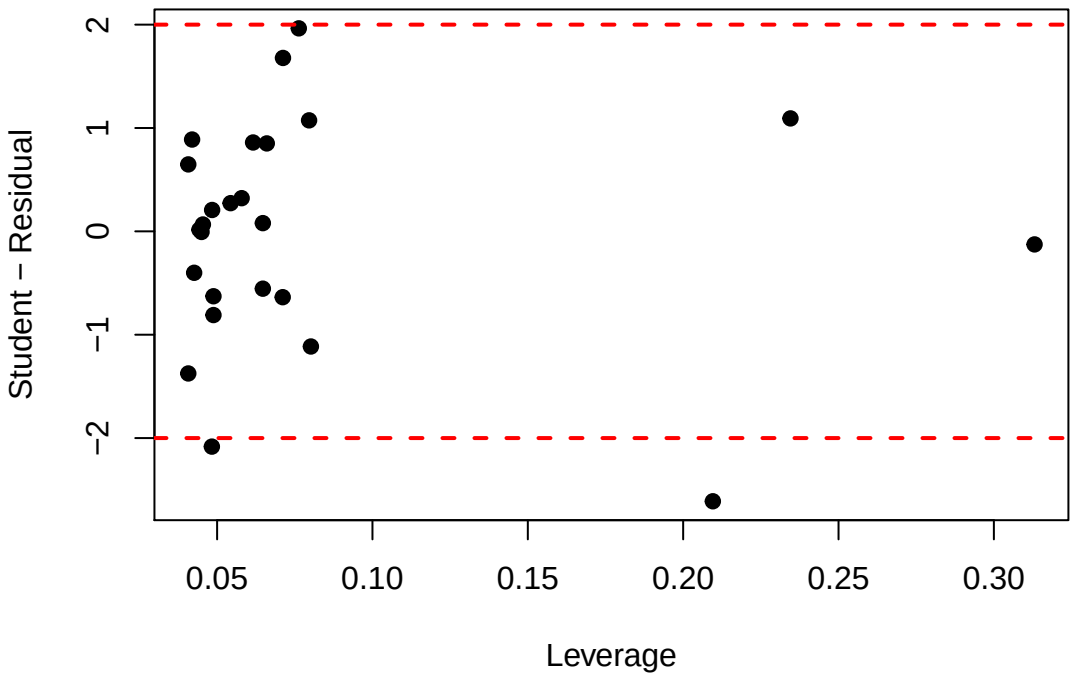
```
#-- R CODE
plot(d$rural,d$deaths,pch=19,xlab="Rural",ylab="Deaths")
lines(seq(0,25000,1),exp(predict(mod5,data.frame(rural=seq(0,25000,1)))),col=2,lwd=2)
```



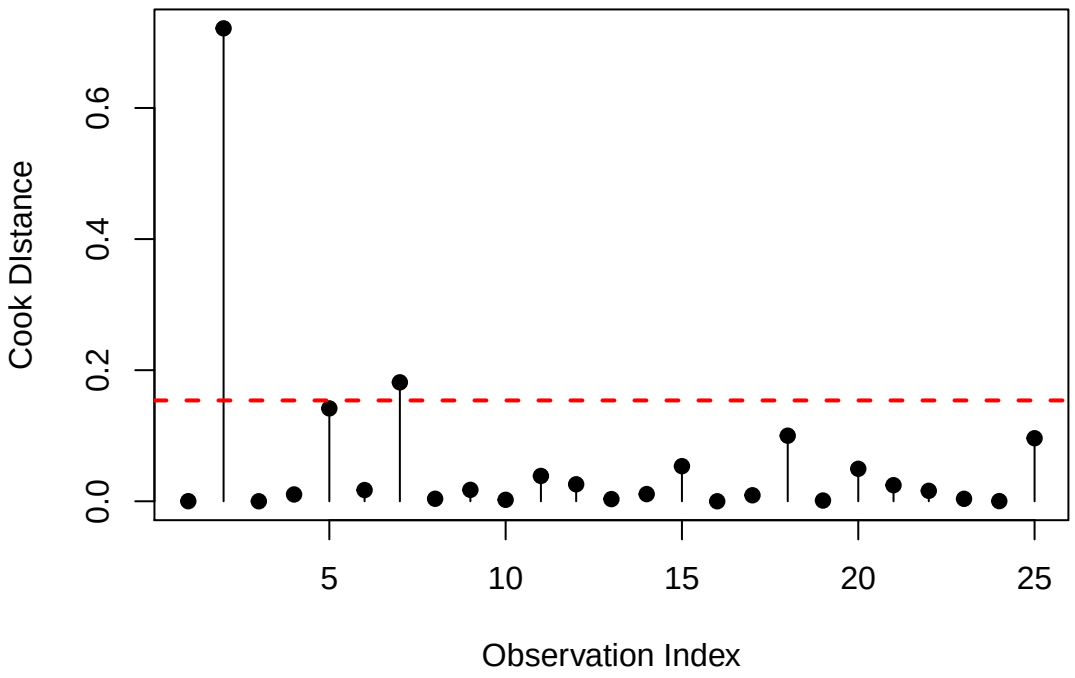

```
#-- R CODE
#plot(mod5,which=2,pch=19)
hist(resid(mod5),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod5)),col=2,lwd=2)
```




```r
#-- R CODE
plot(hatvalues(mod5),rstudent(mod5),pch=19,xlab="Leverage",ylab="Student - Residual")
abline(h=2,col=2,lty=2,lwd=2)
abline(h=-2,col=2,lty=2,lwd=2)
```




```
plot(cooks.distance(mod5),pch=19,xlab="Observation Index",ylab="Cook DIstance",type="h")
points(cooks.distance(mod5),pch=19)
abline(h=4/nrow(d),col=2,lty=2,lwd=2)
```



In definitiva il modello migliore è quello log-log. Per migliorare i risultati occorre ora eliminare l'osservazione California.