LINEAR 5 - Data set: NAZIONI

INTRODUZIONE

Nel dataset in oggetto sono riportati i risultati di un'indagine effettuata nel 1995 su 66 nazioni riguardanti alcuni fra gli aspetti socio-demografici prevalenti. Le variabili presenti nel dataset sono le seguenti:

1. DENSITA': densità di popolazione
2. URBANA: percentuale di popolazione residente nelle città
3. VITAFEM: speranza di vita alla nascita delle donne
4. VITAMAS: speranza di vita alla anascita dei maschi
5. ALFABET: percentuale di alfabetizzati sul totale della popolazione
6. PIL: prodotto interno lordo pro-capite
7. RELIG: religione prevalente nella nazione (1=cattolica, 2=ortodossa, 3=protestante)

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare e polinomiale

```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\\3.lin(5)\\5.linear\\nazioni.csv"),sep=";")
d$pil <- as.numeric(gsub(",","",paste(d$pil))) #-- trasformo pil in variabile numerica

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("densita","urbana","vitafem","vitamas","alfabet","pil")
```

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

nazione	densita	urbana	vitafem	vitamas	alfabet	pil	relig
Argentina	12	86	75	68	95	3,408	1
Armenia	126	68	75	68	98	5,000	2
Australia	2	85	80	74	100	16,848	3
Austria	94	58	79	73	99	18,396	1
Barbados	605	45	78	73	99	6,950	3
Belgio	329	96	79	73	99	17,912	1

STATISTICHE DESCRITTIVE

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

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

Table 2: Table continues below

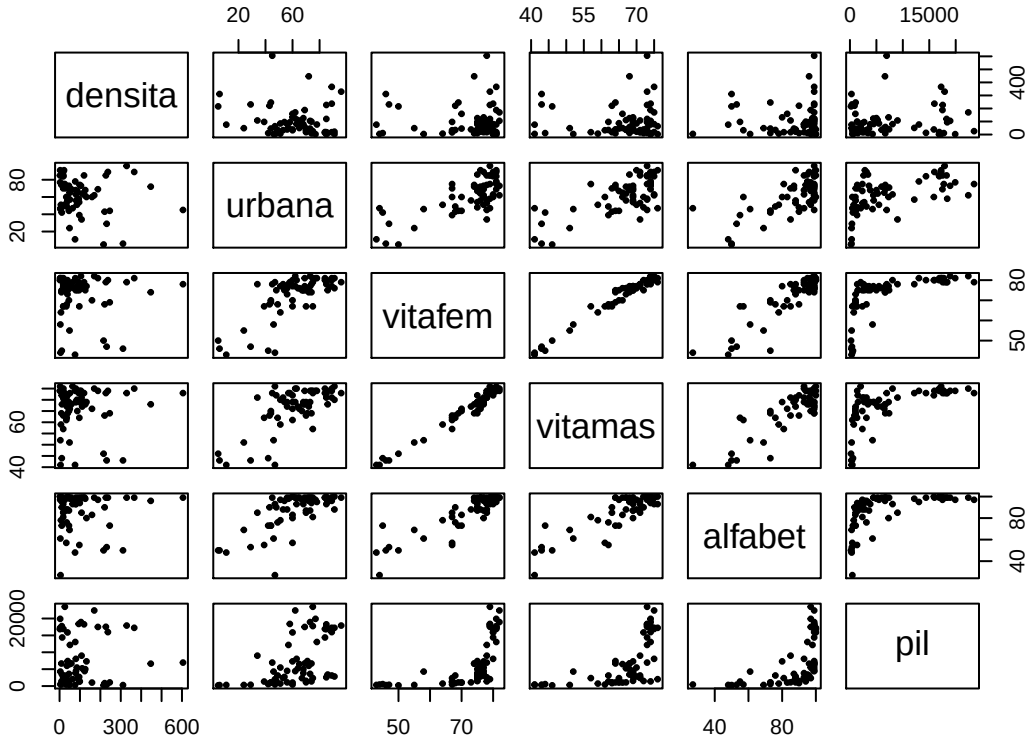
densita	urbana	vitafem	vitamas
Min. : 2.00	Min. : 5.00	Min. :43.00	Min. :41.00
1st Qu.: 19.75	1st Qu.:49.50	1st Qu.:70.00	1st Qu.:64.00
Median : 61.00	Median :64.50	Median :76.00	Median :69.00
Mean :100.15	Mean :62.18	Mean :72.74	Mean :66.58
3rd Qu.:122.25	3rd Qu.:75.00	3rd Qu.:79.00	3rd Qu.:73.00
Max. :605.00	Max. :96.00	Max. :82.00	Max. :76.00

alfabet	pil
Min. : 27.00	Min. : 208
1st Qu.: 83.50	1st Qu.: 1412
Median : 95.50	Median : 4464
Mean : 87.58	Mean : 7303
3rd Qu.: 99.00	3rd Qu.:14048
Max. :100.00	Max. :23474

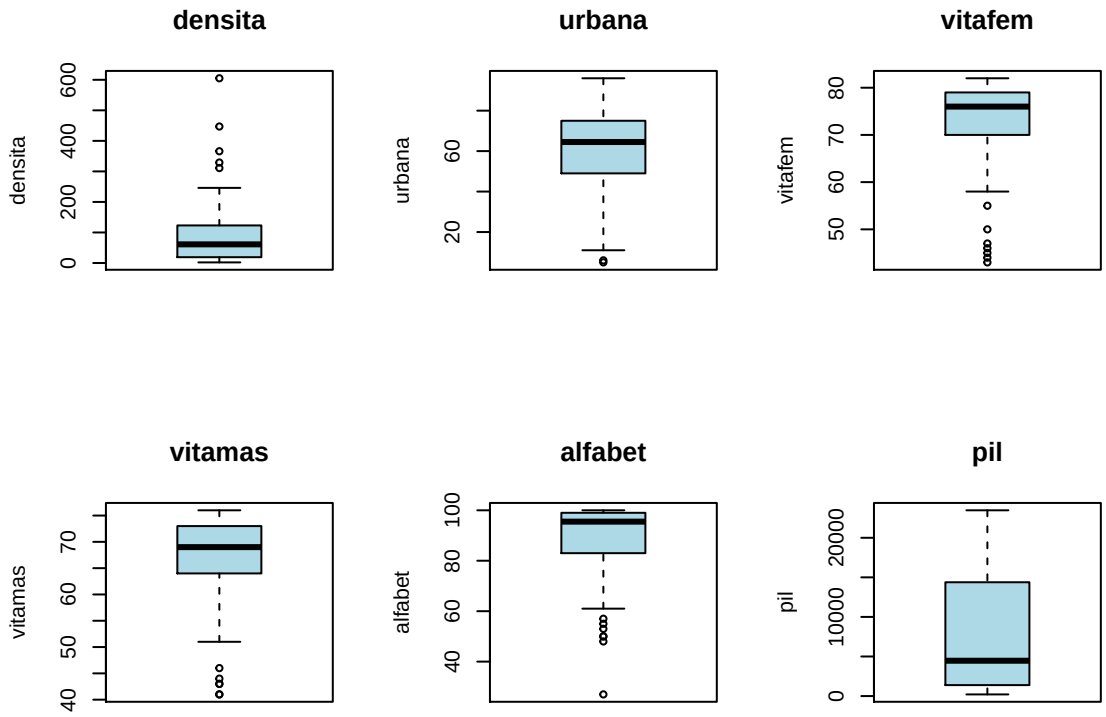
```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

	densita	urbana	vitafem	vitamas	alfabet	pil
densita	1	-0.1501	-0.01275	0.01848	0.02142	0.09363
urbana	-0.1501	1	0.7317	0.7043	0.7054	0.54
vitafem	-0.01275	0.7317	1	0.9836	0.8874	0.601
vitamas	0.01848	0.7043	0.9836	1	0.8628	0.6039
alfabet	0.02142	0.7054	0.8874	0.8628	1	0.5629
pil	0.09363	0.54	0.601	0.6039	0.5629	1

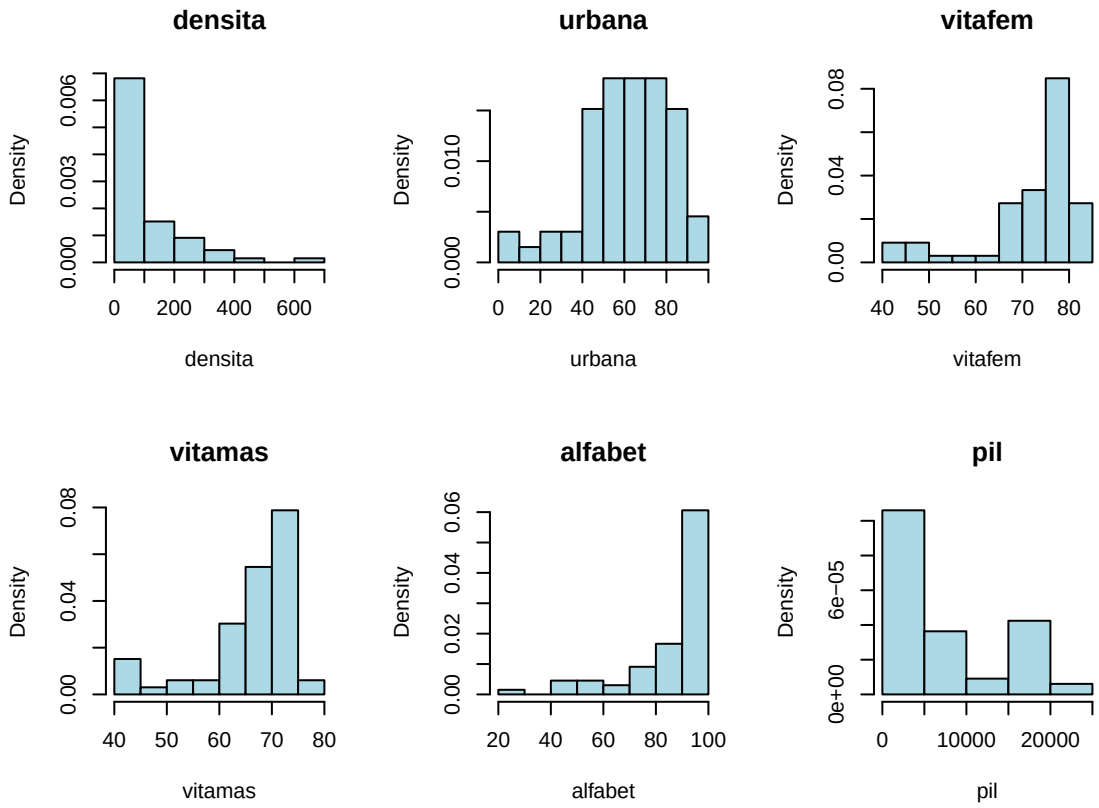
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```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 "urbana" su "pil".

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50.7	3.085	16.43	1.191e-24
pil	0.001572	0.0003062	5.133	2.871e-06

Table 6: Fitting linear model: urbana ~ pil

Observations	Residual Std. Error	R^2	Adjusted R^2
66	17.27	0.2916	0.2806

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

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pil	1	7,856	7,856	26.35	2.871e-06
Residuals	64	19,080	298.1	NA	NA

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

Test.statistic	P.value
7.729	0.02098

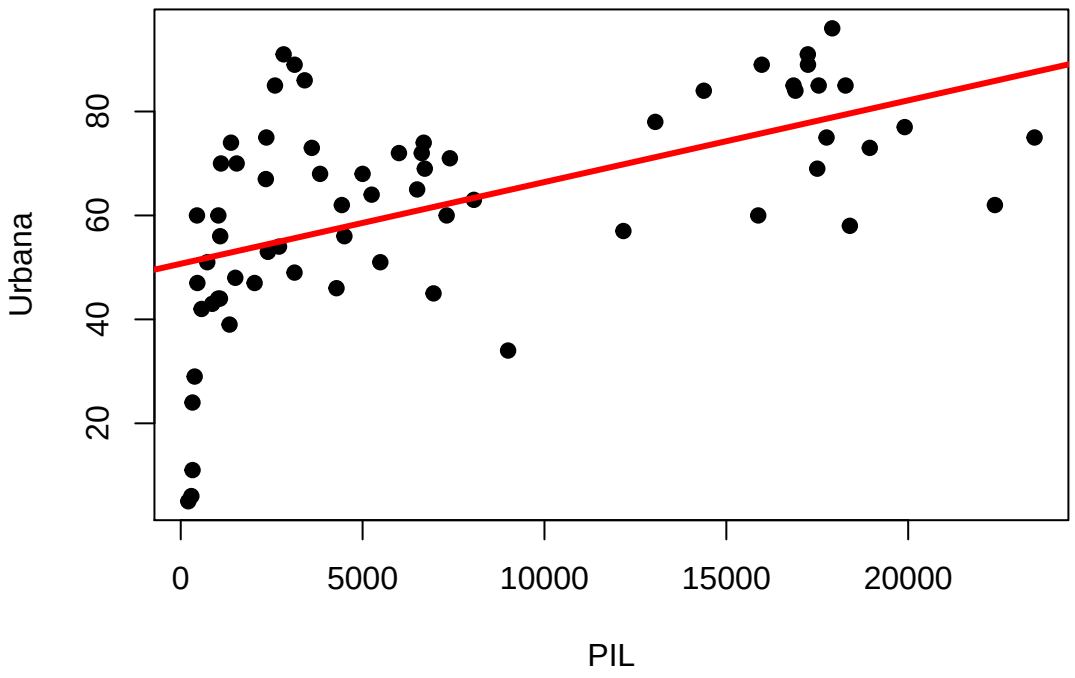
```
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

Table 9: Durbin-Watson test: `mod1`

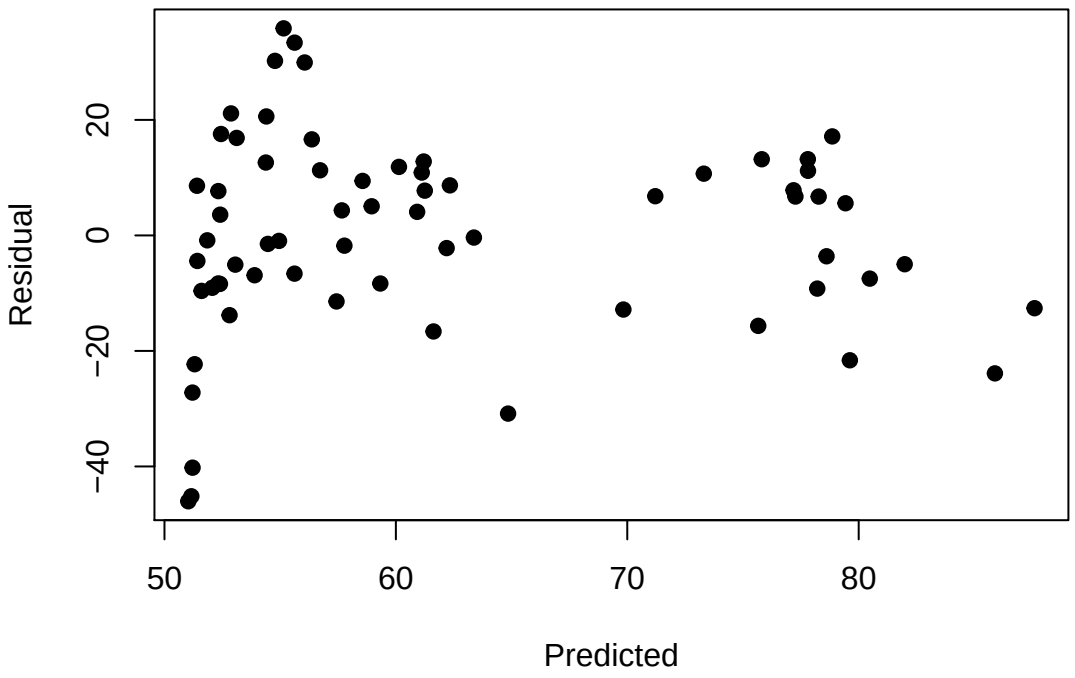
Test statistic	P value	Alternative hypothesis
1.893	0.3235	true autocorrelation is greater than 0

Pil è significativo ma il fitting è modesto ($R^2 = 0.291$). Dai test si osserva che gli errori sono non correlati ma eteroschedastci. Lo si vede anche dai seguenti grafici:

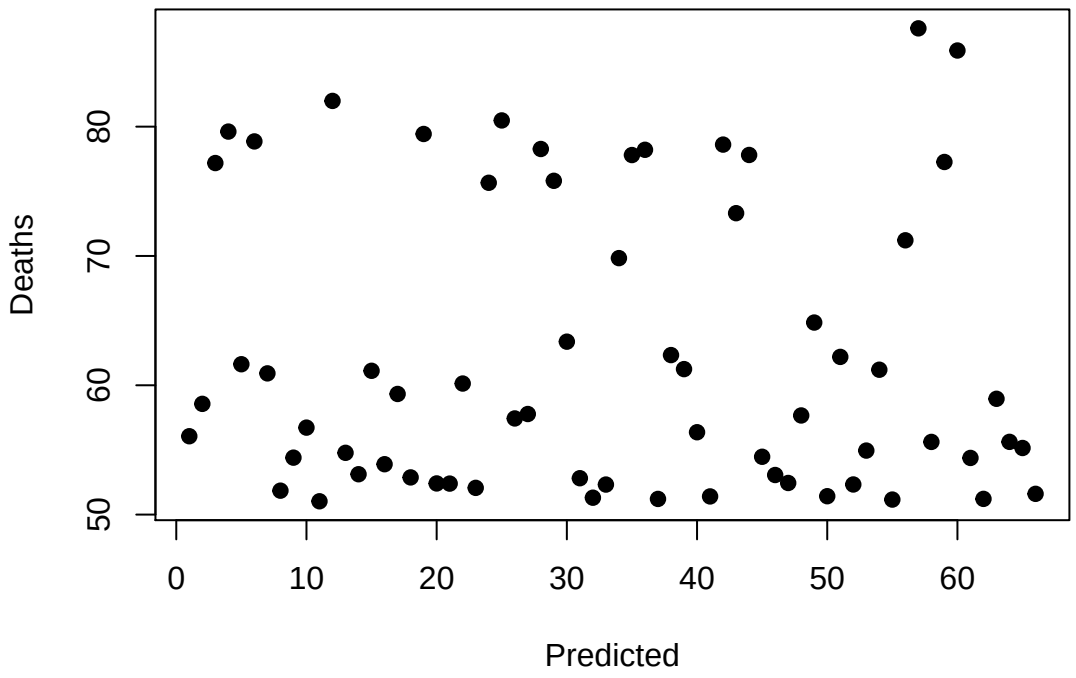
```
#-- R CODE
plot(d$pil,d$urbana,pch=19,xlab="PIL",ylab="Urbana")
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
```




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




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



Si verifica ora se sono più appropriati modelli non lineari. Iniziamo con modelli di grado 2, 3 e 4.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43.58	3.97	10.98	2.904e-16
pil	0.004717	0.00121	3.898	0.0002381
I(pil^2)	-1.561e-07	5.828e-08	-2.678	0.009429

Table 11: Fitting linear model: urbana ~ pil + I(pil^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
66	16.49	0.364	0.3439

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

Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pil	1	7,856	7,856	28.89	1.187e-06

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(pil^2)	1	1,950	1,950	7.173	0.009429
Residuals	63	17,130	271.9	NA	NA

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

Test.statistic	P.value
5.53	0.06296

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

Table 14: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.842	0.2623	true autocorrelation is greater than 0

```
#-- R CODE
mod3 <- lm(urbana~pil+I(pil^2)+I(pil^3),d)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.5	4.769	8.492	5.574e-12
pil	0.007012	0.002319	3.023	0.003634
I(pil^2)	-4.46e-07	2.569e-07	-1.736	0.08747
I(pil^3)	9.199e-12	7.939e-12	1.159	0.251

Table 16: Fitting linear model: urbana ~ pil + I(pil^2) + I(pil^3)

Observations	Residual Std. Error	R^2	Adjusted R^2
66	16.44	0.3775	0.3474

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

Table 17: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pil	1	7,856	7,856	29.05	1.162e-06
I(pil^2)	1	1,950	1,950	7.212	0.009282
I(pil^3)	1	363.1	363.1	1.343	0.251
Residuals	62	16,767	270.4	NA	NA

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

Test.statistic	P.value
4.294	0.1168

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

Table 19: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
1.901	0.3483	true autocorrelation is greater than 0

```r
#-- R CODE
mod4 <- lm(urbana~pil+I(pil^2)+I(pil^3)+I(pil^4),d)
pander(summary(mod4),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.05	5.318	5.087	3.729e-06
pil	0.02335	0.004416	5.288	1.763e-06
I(pil^2)	-4.213e-06	9.288e-07	-4.536	2.743e-05
I(pil^3)	2.773e-10	6.446e-11	4.302	6.228e-05
I(pil^4)	-5.872e-15	1.403e-15	-4.184	9.33e-05

Table 21: Fitting linear model: urbana ~ pil + I(pil^2) + I(pil^3) + I(pil^4)

Observations	Residual Std. Error	R^2	Adjusted R^2
66	14.61	0.5163	0.4846

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

Table 22: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pil	1	7,856	7,856	36.78	9.092e-08
I(pil^2)	1	1,950	1,950	9.132	0.003669
I(pil^3)	1	363.1	363.1	1.7	0.1972
I(pil^4)	1	3,739	3,739	17.51	9.33e-05
Residuals	61	13,028	213.6	NA	NA

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

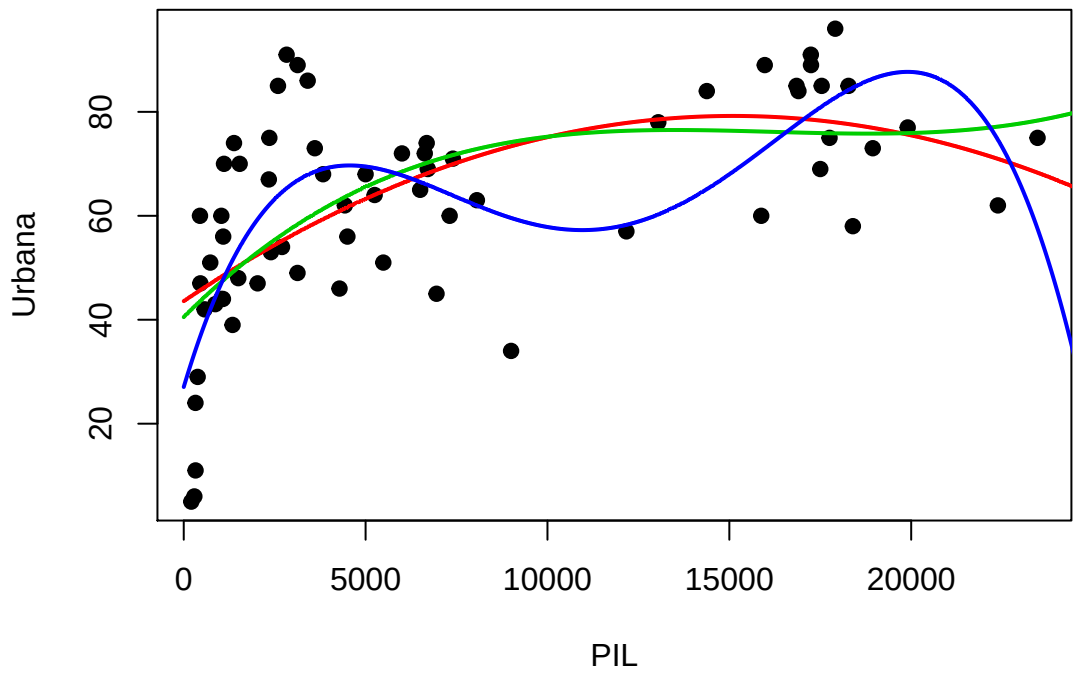
Test.statistic	P.value
4.185	0.1234

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

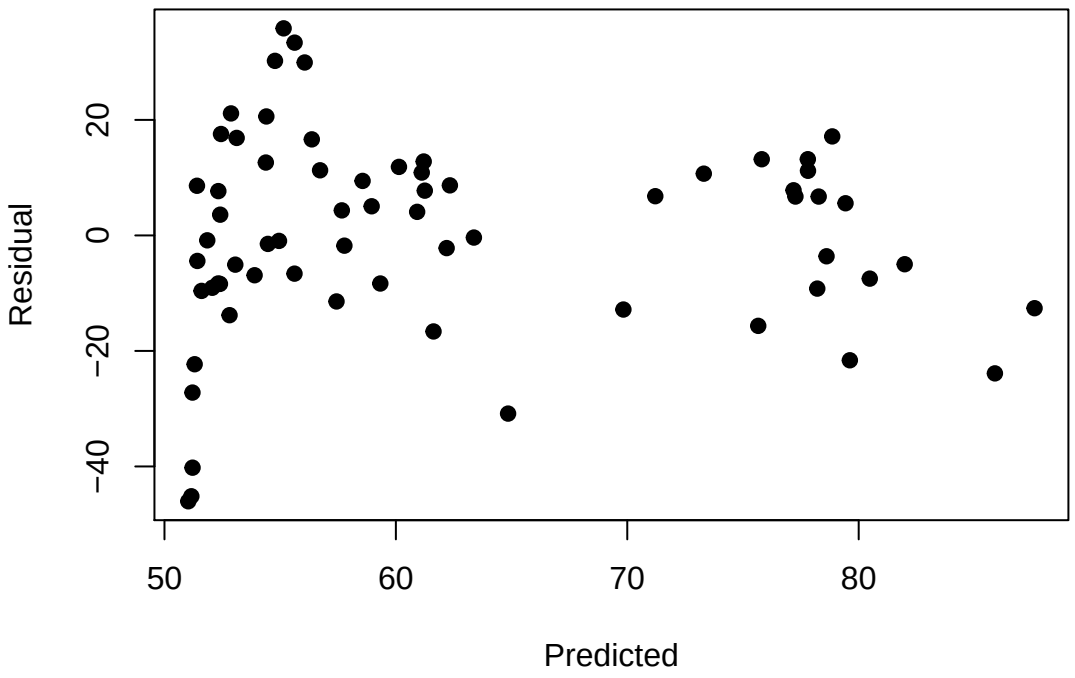
Table 24: Durbin-Watson test: `mod4`

Test statistic	P value	Alternative hypothesis
1.962	0.4415	true autocorrelation is greater than 0

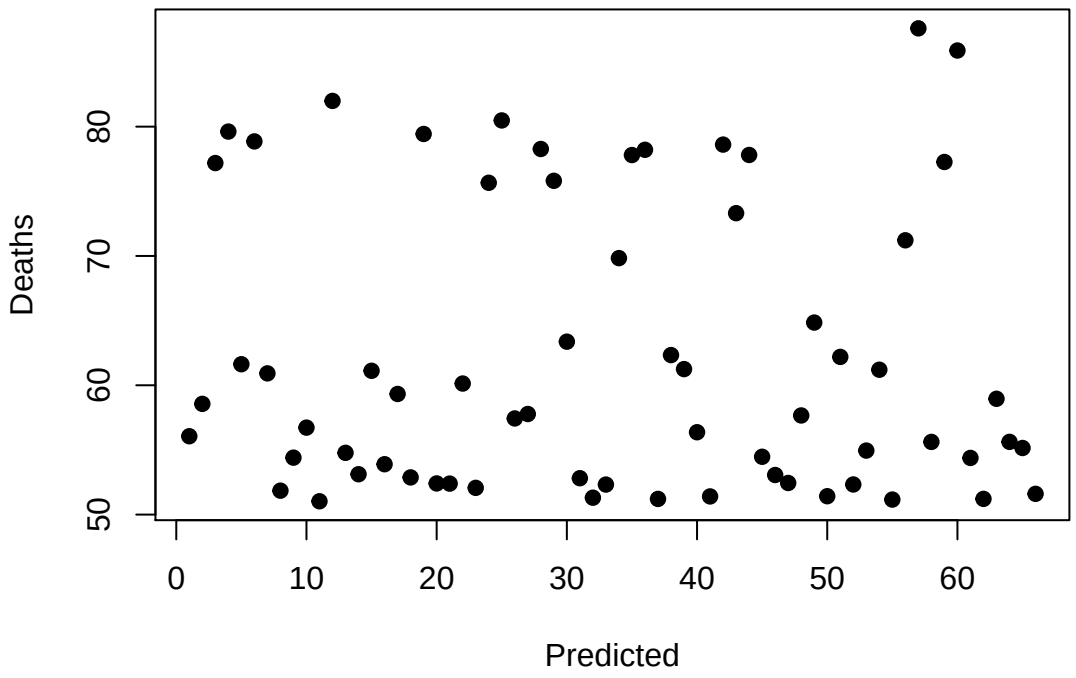
```
#-- R CODE
plot(d$pil,d$urbana,pch=19,xlab="PIL",ylab="Urbana")
lines(seq(0,200000,1),predict(mod2,data.frame(pil=seq(0,200000,1))),col=2,lwd=2)
lines(seq(0,200000,1),predict(mod3,data.frame(pil=seq(0,200000,1))),col=3,lwd=2)
lines(seq(0,200000,1),predict(mod4,data.frame(pil=seq(0,200000,1))),col="blue",lwd=2)
```




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




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



Si considerino ora i modelli logaritmici: lin-log (variabile esplicativa $log(PIL)$), log-lineare (variabile dipendente $log(Urbana)$), log-log (variabile dipendente $log(Urbana)$; variabile esplicativa $log(PIL)$).

```
#-- R CODE
mod5 <- lm(urbana~I(log(pil)),d)
pander(summary(mod5),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-26.99	11.65	-2.318	0.02367
I(log(pil))	10.81	1.394	7.752	8.838e-11

Table 26: Fitting linear model: urbana ~ I(log(pil))

Observations	Residual Std. Error	R^2	Adjusted R^2
66	14.73	0.4842	0.4762

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

Table 27: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(pil))	1	13,043	13,043	60.09	8.838e-11

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	64	13,893	217.1	NA	NA

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

Test.statistic	P.value
2.446	0.2943

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

Table 29: Durbin-Watson test: `mod5`

Test statistic	P value	Alternative hypothesis
1.873	0.3016	true autocorrelation is greater than 0

```
#-- R CODE
mod6 <- lm(I(log(urbana))~pil,d)
pander(summary(mod6),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.782	0.08838	42.79	8.307e-49
pil	3.446e-05	8.772e-06	3.928	0.0002124

Table 31: Fitting linear model: I(log(urbana)) ~ pil

Observations	Residual Std. Error	R^2	Adjusted R^2
66	0.4946	0.1943	0.1817

```
pander(anova(mod6),big.mark=",")
```

Table 32: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pil	1	3.775	3.775	15.43	0.0002124
Residuals	64	15.66	0.2446	NA	NA

```
pander(white.test(mod6),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
7.965	0.01864

```r
pander(dwtest(mod6),big.mark=",") #-- Durbin-Whatson test
```

Table 34: Durbin-Watson test: `mod6`

Test statistic	P value	Alternative hypothesis
2.07	0.6046	true autocorrelation is greater than 0

```r
#-- R CODE
mod7 <- lm(I(log(urbana))~I(log(pil)),d)
pander(summary(mod7),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.754	0.3263	5.374	1.153e-06
I(log(pil))	0.2763	0.03907	7.071	1.394e-09

Table 36: Fitting linear model: I(log(urbana)) ~ I(log(pil))

Observations	Residual Std. Error	R^2	Adjusted R^2
66	0.4129	0.4386	0.4298

```r
pander(anova(mod7),big.mark=",")
```

Table 37: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(pil))	1	8.523	8.523	50	1.394e-09
Residuals	64	10.91	0.1704	NA	NA

```r
pander(white.test(mod7),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
28.5	6.472e-07

```r
pander(dwtest(mod7),big.mark=",") #-- Durbin-Whatson test
```

Table 39: Durbin-Watson test: `mod7`

Test statistic	P value	Alternative hypothesis
1.95	0.4193	true autocorrelation is greater than 0

In tutti e 3 i modelli le variabili esplicative sono significative ma il log-log risulta avere un miglior fitting

anche se inferiore nettamente al modello di 4 grado. Si sceglie quindi il modello di 4 grado anche perché nel modello log-log si ha ancora eteroschedasticità degli errori.

Si propone ora di studiare il "Pil" in funzione della percentuale di alfabetizzati cominciando con il modello lineare:

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13,690	3,920	-3.493	0.0008728
alfabet	239.7	44	5.448	8.686e-07

Table 41: Fitting linear model: pil ~ alfabet

Observations	Residual Std. Error	R^2	Adjusted R^2
66	5826	0.3168	0.3061

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

Table 42: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
alfabet	1	1.007e+09	1.007e+09	29.68	8.686e-07
Residuals	64	2.172e+09	33,939,618	NA	NA

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

Test.statistic	P.value
13.83	0.0009944

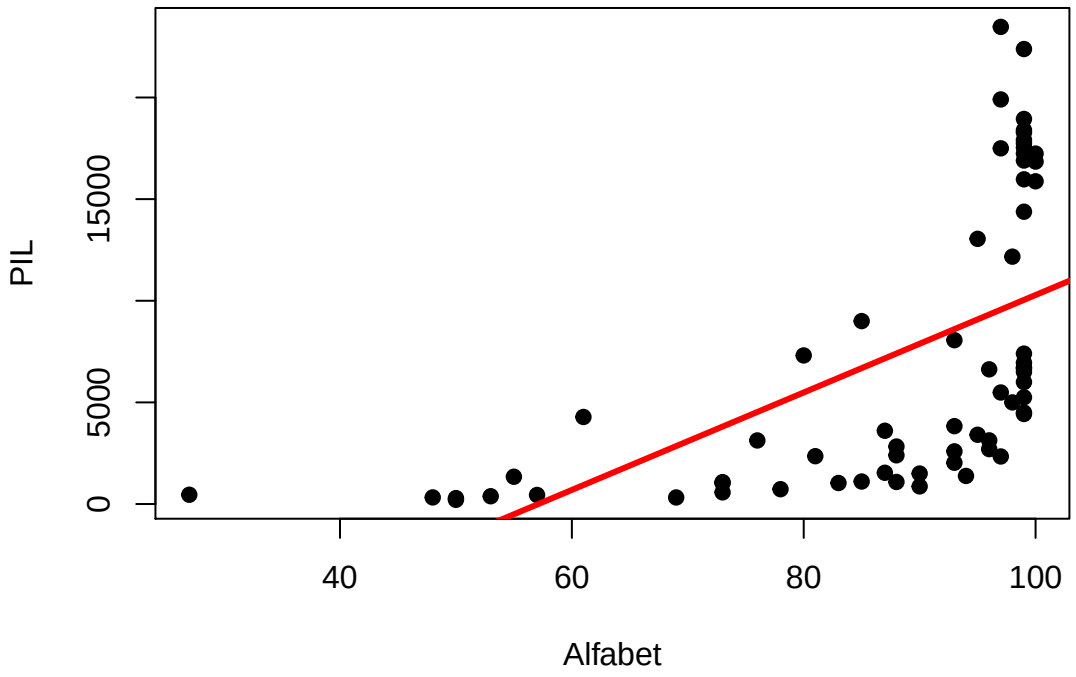
```
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

Table 44: Durbin-Watson test: `mod1`

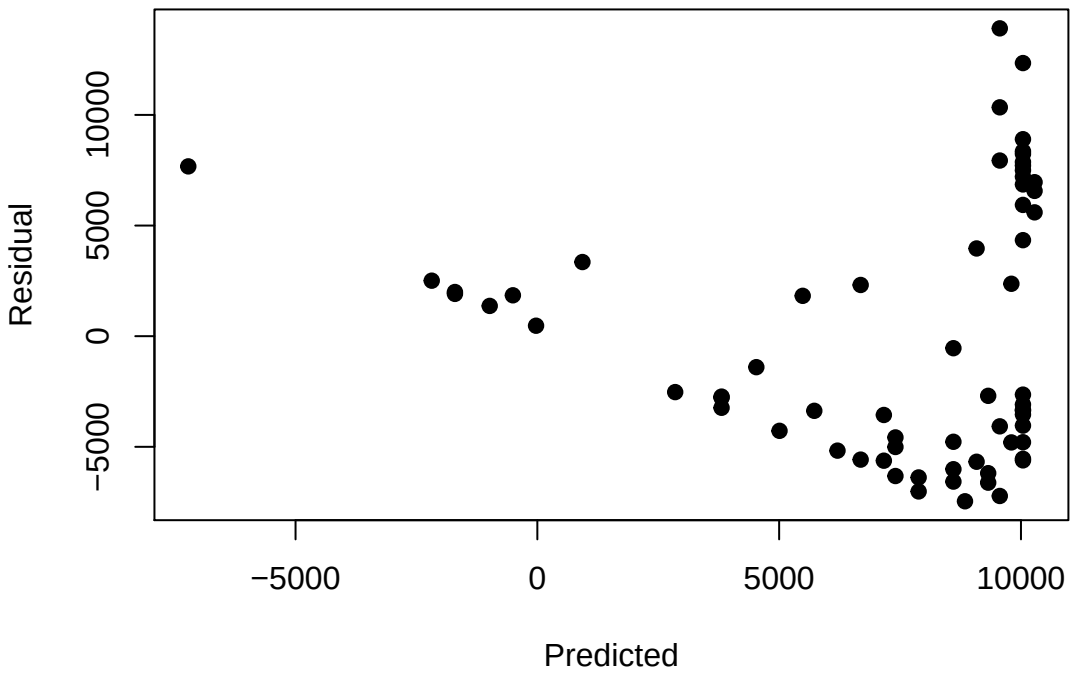
Test statistic	P value	Alternative hypothesis
1.625	0.061	true autocorrelation is greater than 0

"Alfabet" è significativa ma il fitting è basso ($R^2 = 0.3168$). Oltretutto gli errori sono nettamente eteroschedatici come si vede dai grafici e del test di White.

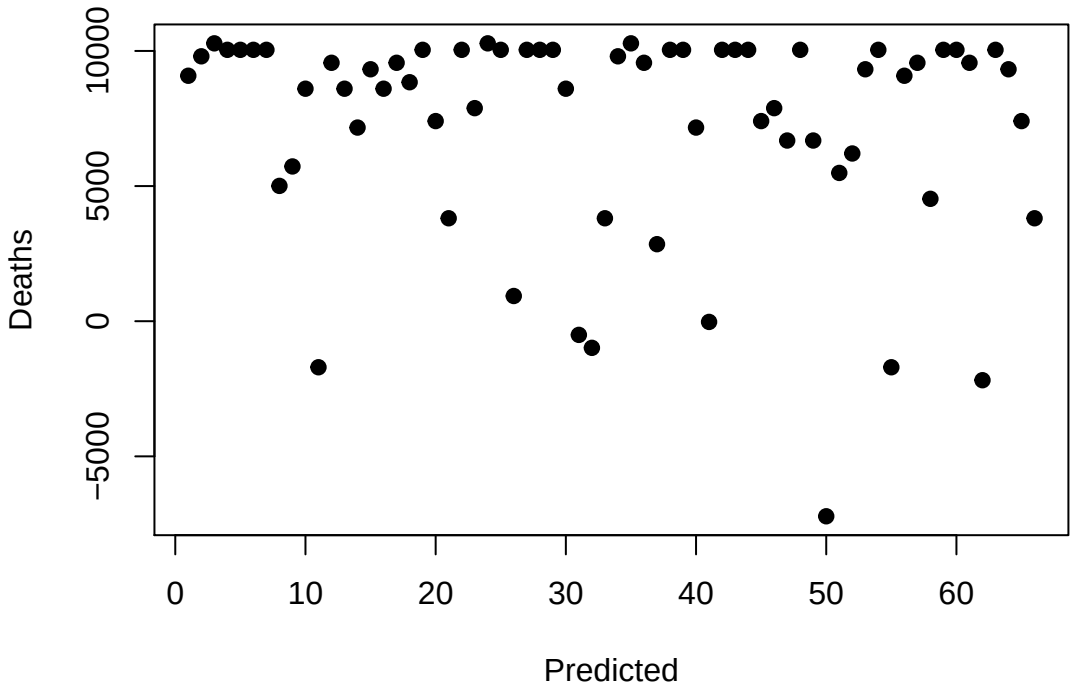
```r
#-- R CODE
plot(d$alfabet,d$pil,pch=19,xlab="Alfabet",ylab="PIL")
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
```




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




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



Si prova a verificare se le cose migliorano con la regressione quadratica.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24,474	11,054	2.214	0.03045
alfabet	-891.7	312.6	-2.853	0.005852
I(alfabet^2)	7.678	2.103	3.65	0.0005338

Table 46: Fitting linear model: pil ~ alfabet + I(alfabet^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
66	5335	0.4361	0.4182

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

Table 47: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
alfabet	1	1.007e+09	1.007e+09	35.39	1.294e-07

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(alfabet^2)	1	379,224,361	379,224,361	13.33	0.0005338
Residuals	63	1.793e+09	28,458,908	NA	NA

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

Test.statistic	P.value
15.64	0.0004022

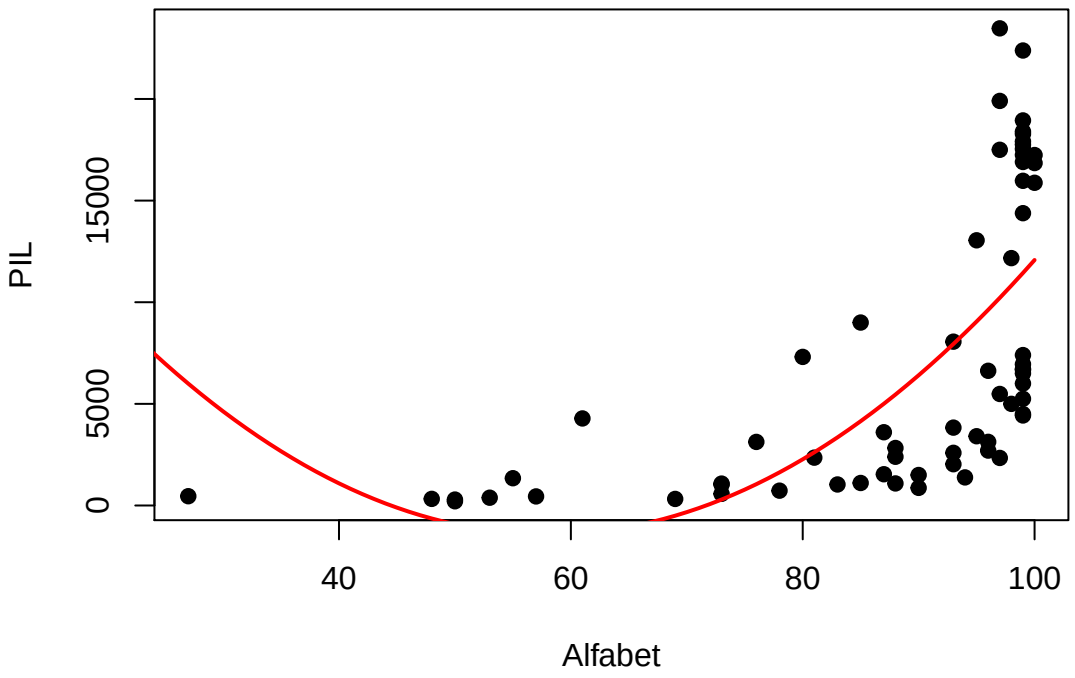
```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 49: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.794	0.2034	true autocorrelation is greater than 0

Il fitting della funzione quadratica che ha la concavità verso il basso migliora e alfabet è significativo al 1° e 2° grado. Tuttavia come si vede dalla rappresentazione grafica gli errori sono chiaramente ancora eteroschedastici.

```
#-- R CODE
plot(d$alfabet,d$pil,pch=19,xlab="Alfabet",ylab="PIL")
lines(seq(0,100,1),predict(mod2,data.frame(alfabet=seq(0,100,1))),col=2,lwd=2)
```



Un rapido esame della rappresentazione grafica delle funzioni di 3° 4° grado (provare per esercizio) mostra che esse sono ancora meno appropriate per interpretare la variabile dipendente. Si verifica ora la bontà delle funzioni lin-log (variabile esplicativa $log(Alfabet)$), log-lineare (variabile dipendente $log(PIL)$), log-log (variabile dipendente $log(PIL)$, variabile esplicativa $log(Alfabet)$). Nel modello lin-log la variabile esplicativa è significativa ma il fitting è molto basso e gli errori ancora eteroschedastici.

```
#-- R CODE
mod3 <- lm(pil~I(log(alfabet)),d)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57,490	13,869	-4.145	0.0001017
I(log(alfabet))	14,565	3,113	4.679	1.543e-05

Table 51: Fitting linear model: pil ~ I(log(alfabet))

Observations	Residual Std. Error	R^2	Adjusted R^2
66	6084	0.2549	0.2432

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

Table 52: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(alfabet))	1	810,273,375	810,273,375	21.89	1.543e-05
Residuals	64	2.369e+09	37,017,763	NA	NA

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

Test.statistic	P.value
14.62	0.0006701

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

Table 54: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
1.621	0.05943	true autocorrelation is greater than 0

Nel modello log-lineare il fitting è nettamente migliore, il migliore di tutti i modelli presentati ($R^2 = 0.6077$); la variabile esplicativa significativa, gli errori normali, omoschedasctici e incorrelati.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.802	0.5566	5.034	4.16e-06
alfabet	0.06222	0.006249	9.958	1.246e-14

Table 56: Fitting linear model: I(log(pil)) ~ alfabet

Observations	Residual Std. Error	R^2	Adjusted R^2
66	0.8273	0.6077	0.6016

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

Table 57: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
alfabet	1	67.87	67.87	99.16	1.246e-14
Residuals	64	43.81	0.6845	NA	NA

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

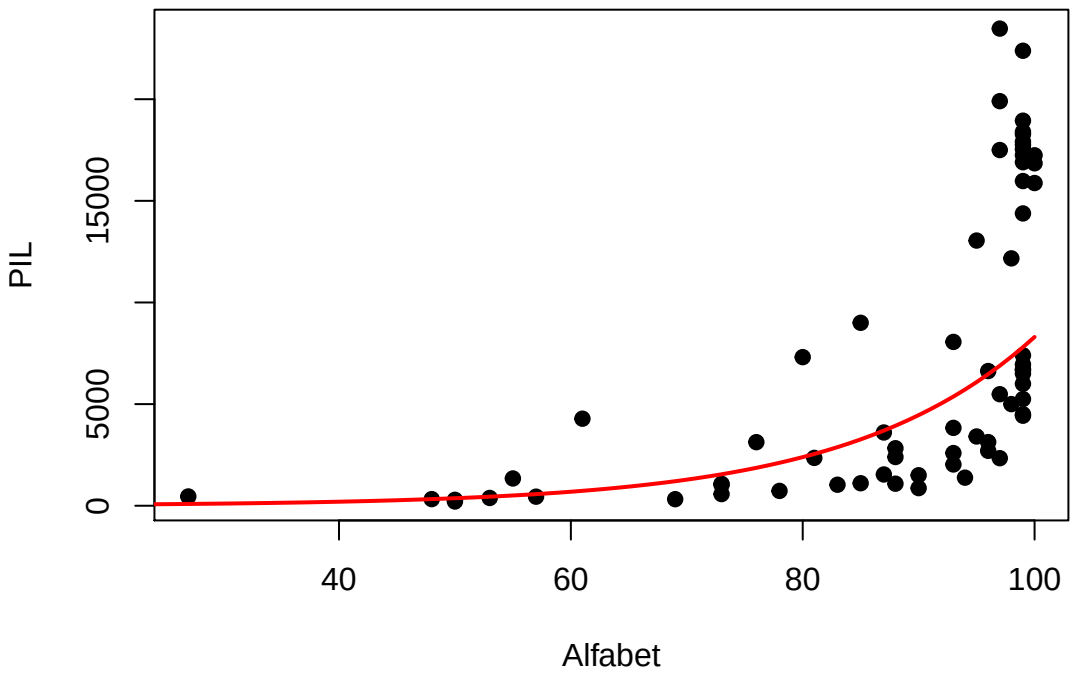
Test.statistic	P.value
3.622	0.1635

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

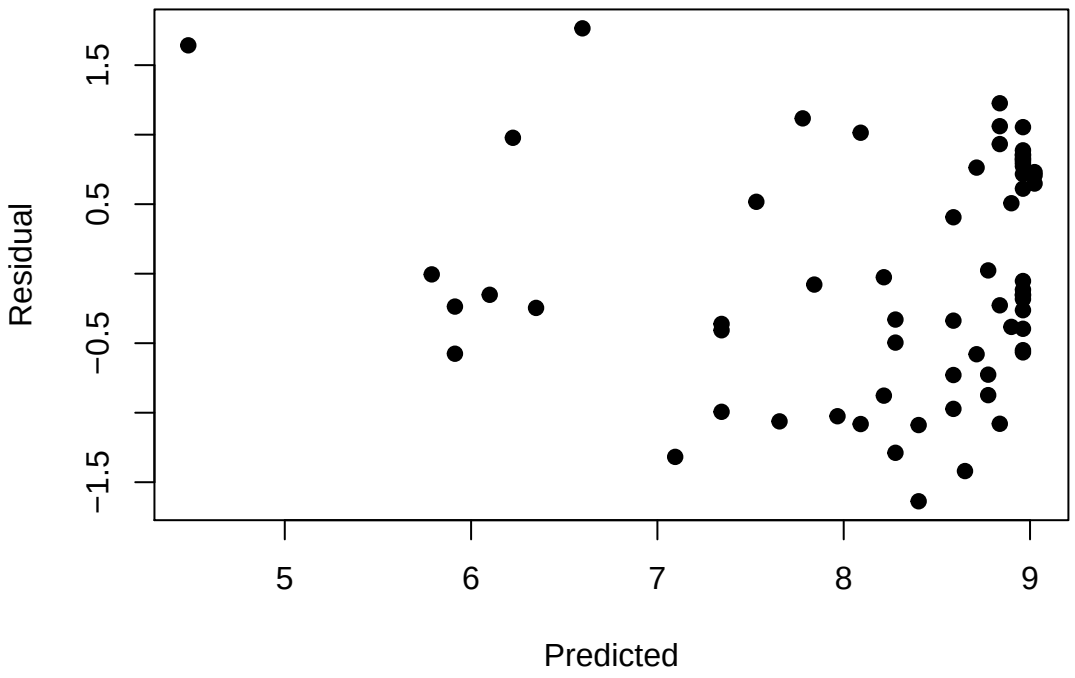
Table 59: Durbin-Watson test: `mod4`

Test statistic	P value	Alternative hypothesis
1.623	0.06008	true autocorrelation is greater than 0

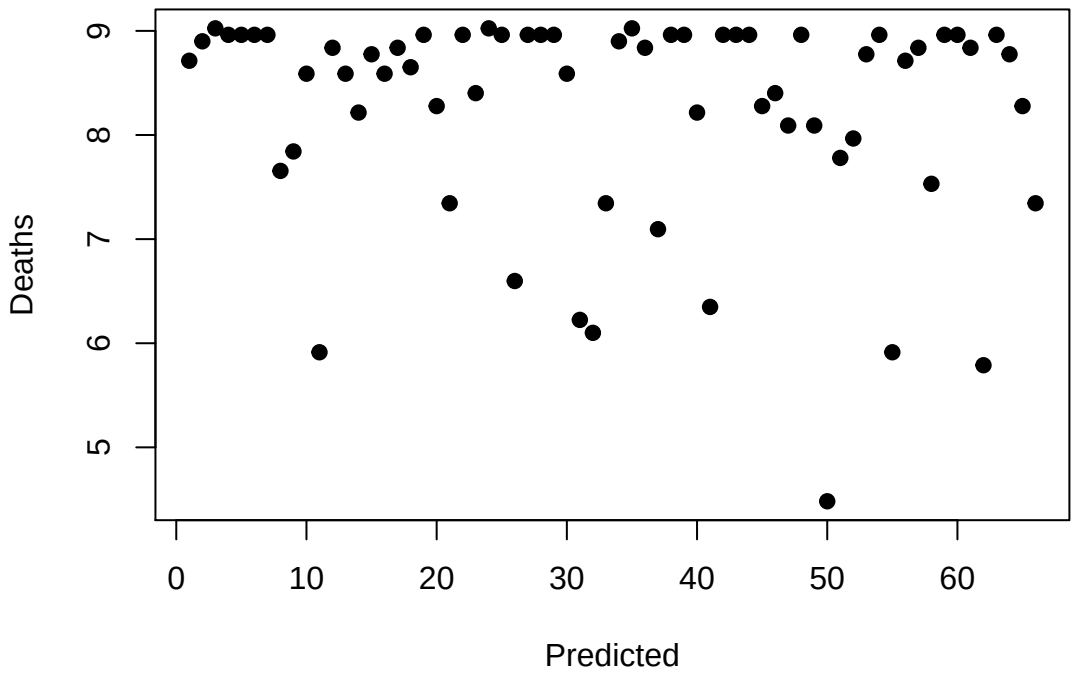
```r
#-- R CODE
plot(d$alfabet,d$pil,pch=19,xlab="Alfabet",ylab="PIL")
lines(seq(0,100,1),exp(predict(mod4,data.frame(alfabet=seq(0,100,1)))),col=2,lwd=2)
```



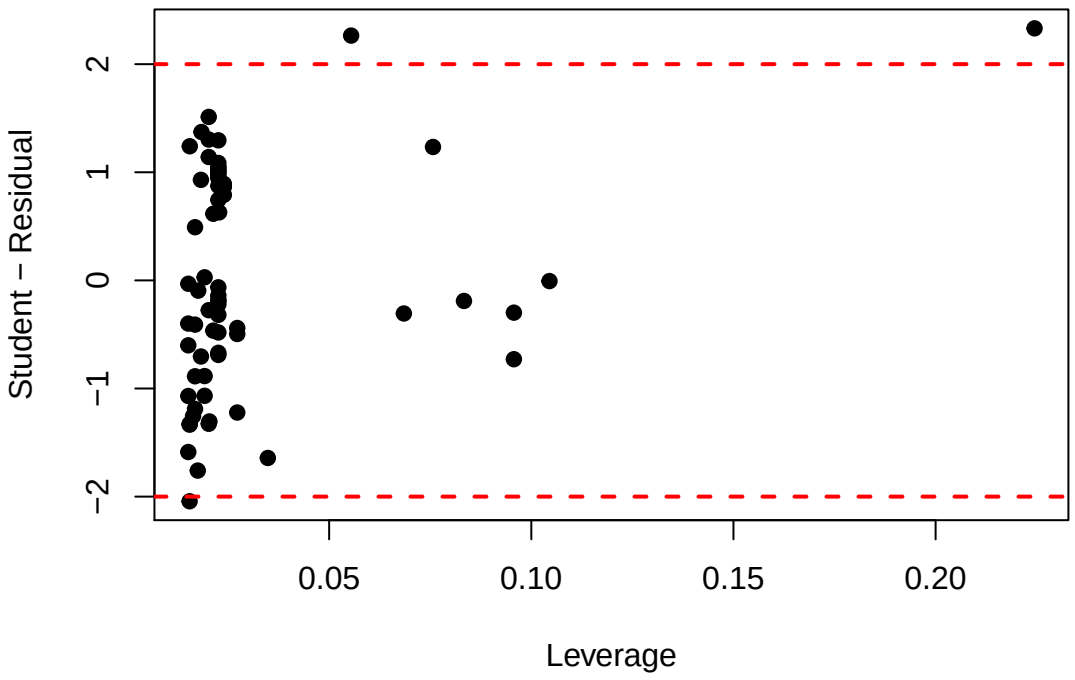

```r
#-- R CODE
plot(fitted(mod4),resid(mod4),pch=19,xlab="Predicted",ylab="Residual")
```



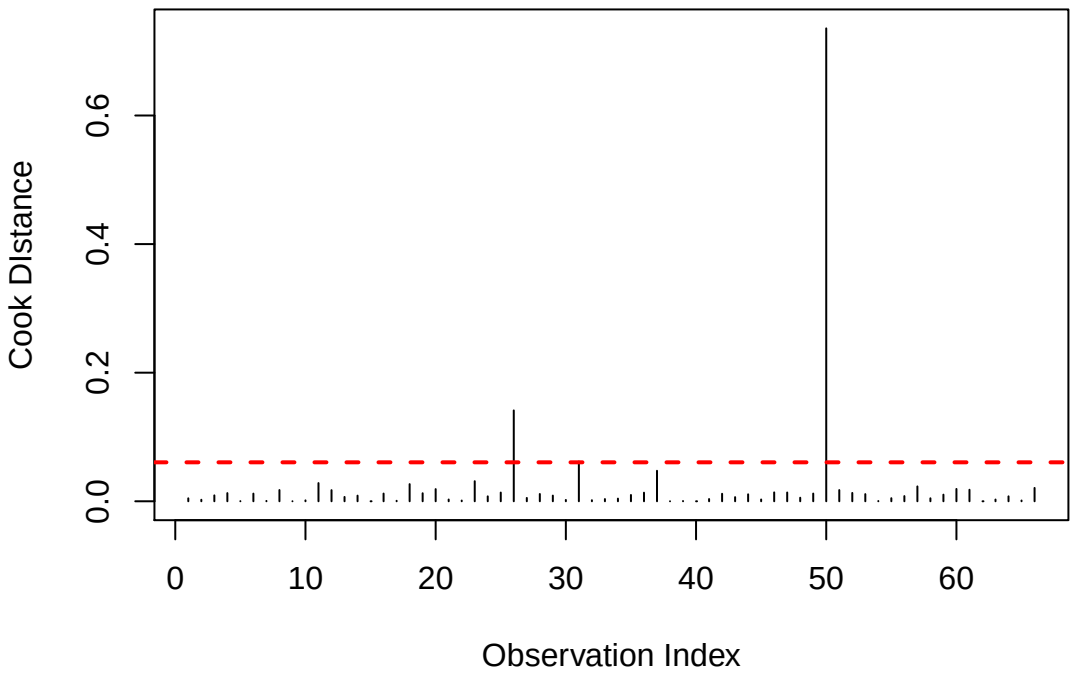

```
plot(fitted(mod4),d$deaths,pch=19,xlab="Predicted",ylab="Deaths")
```



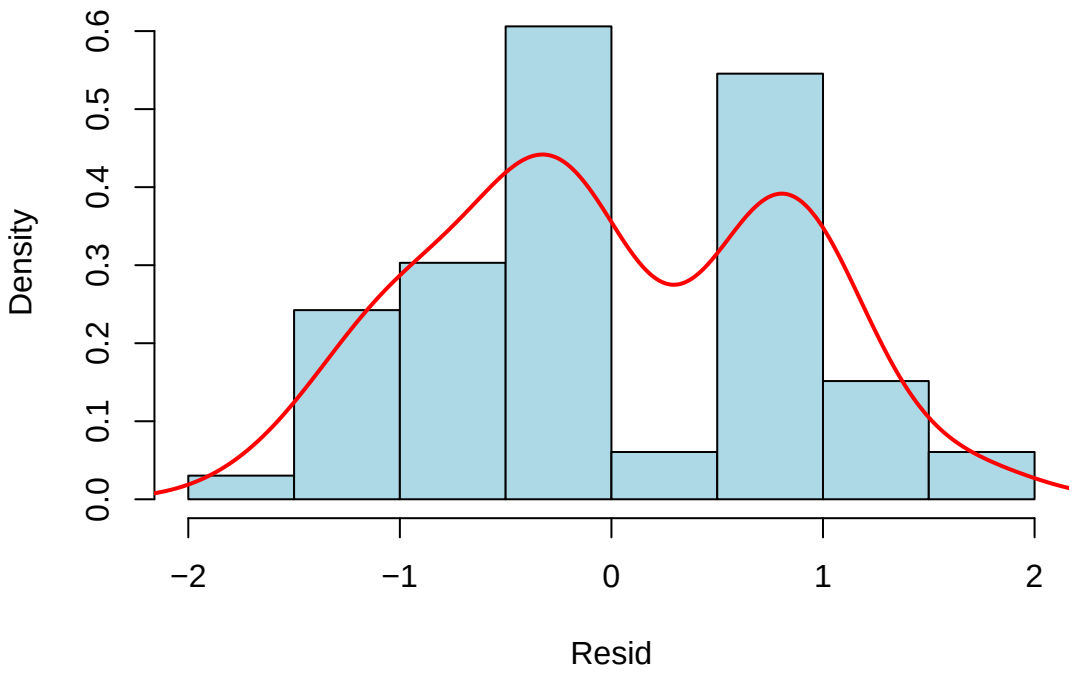

```r
#-- R CODE

plot(hatvalues(mod4),rstudent(mod4),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(mod4),pch=19,xlab="Observation Index",ylab="Cook DIstance",type="h")
abline(h=4/nrow(d),col=2,lty=2,lwd=2)
```




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



Anche nel modello log-log la variabile esplicativa è significativa, gli errori eteroschedastici e incorrelati ma il fitting è peggiore che nel modello log-lineare ($R^2 = 0.5273$).

```
#-- R CODE
mod5 <- lm(I(log(pil))~I(log(alfabet)),d)
pander(summary(mod5),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.217	2.07	-4.452	3.48e-05
I(log(alfabet))	3.927	0.4647	8.45	5.196e-12

Table 61: Fitting linear model: I(log(pil)) ~ I(log(alfabet))

Observations	Residual Std. Error	R^2	Adjusted R^2
66	0.9081	0.5273	0.52

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

Table 62: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(alfabet))	1	58.89	58.89	71.41	5.196e-12

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	64	52.78	0.8247	NA	NA

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

Test.statistic	P.value
22.54	1.275e-05

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

Table 64: Durbin-Watson test: `mod5`

Test statistic	P value	Alternative hypothesis
1.627	0.06225	true autocorrelation is greater than 0

Quindi si opta per il modello log-lineare. Per migliorare i risultati occorrerebbe eliminare i due outiler e applicare il modello loglineare alle osservazioni rimanenti.