LINEAR 3 - Data set: PRESTIGE

INTRODUZIONE

Il data set contiene contiene 102 osservazioni e le seguenti 6 variabili.

1. EDUCATION: istruzione media (in anni) dei lavoratori nel 1971
2. INCOME: reddito medio (in dollari) dei lavoratori nel 1971
3. WOMEN: percentuale di lavoratori donne nel 1971
4. PRESTIGE: punteggio di Pineo-Porter relativo al prestigio delle occupazioni, ottenuto tramite sondaggio sociale condotto a metà del 1960.
5. CENSUS: codice dell'occupazione nel censimento canadese
6. TYPE: tipologia di occupazione (variabile categoriale).

Variabile dipendente: PRESTIGE.

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)\\3.linear\\prestige.txt"),sep=" "

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("education","income","women","prestige")
```

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

name	education	income	women	prestige	census	type
GOV.ADMINISTRATORS	13.11	12351	11.16	68.8	1113	prof
GENERAL.MANAGERS	12.26	25879	4.02	69.1	1130	prof
ACCOUNTANTS	12.77	9271	15.7	63.4	1171	prof
PURCHASING.OFFICERS	11.42	8865	9.11	56.8	1175	prof
CHEMISTS	14.62	8403	11.68	73.5	2111	prof
PHYSICISTS	15.64	11030	5.13	77.6	2113	prof

STATISTICHE DESCRITTIVE

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

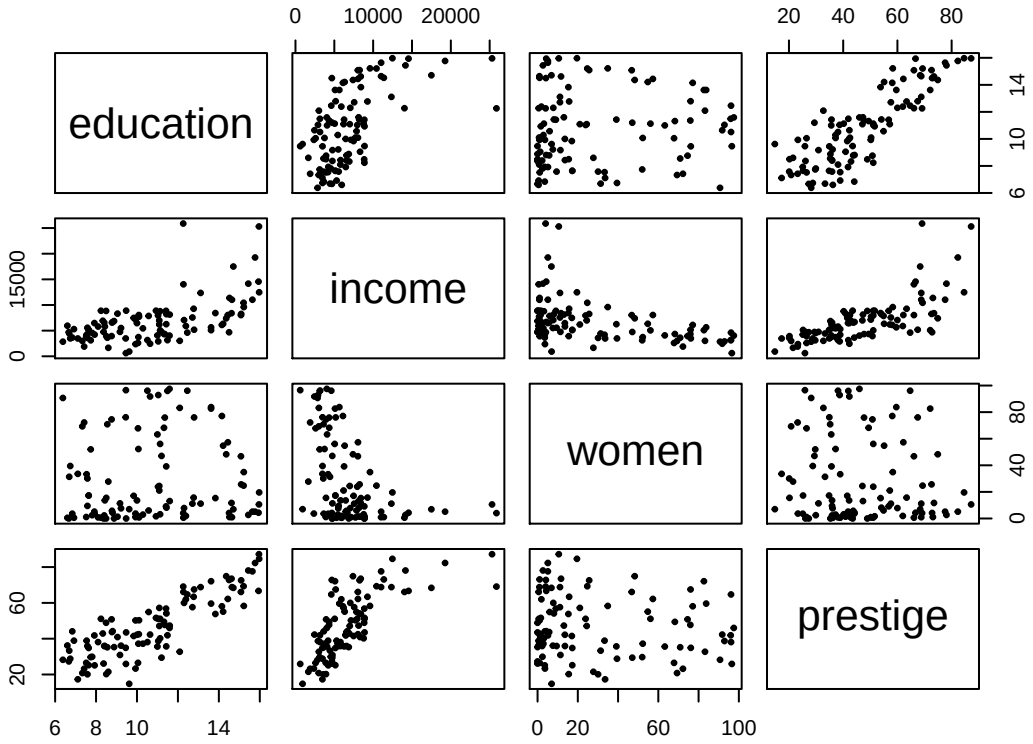
```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC])) #-- statistiche descrittive
```

education	income	women	prestige
Min. : 6.380	Min. : 611	Min. : 0.000	Min. :14.80
1st Qu.: 8.445	1st Qu.: 4106	1st Qu.: 3.592	1st Qu.:35.23
Median :10.540	Median : 5930	Median :13.600	Median :43.60
Mean :10.738	Mean : 6798	Mean :28.979	Mean :46.83
3rd Qu.:12.648	3rd Qu.: 8187	3rd Qu.:52.203	3rd Qu.:59.27
Max. :15.970	Max. :25879	Max. :97.510	Max. :87.20

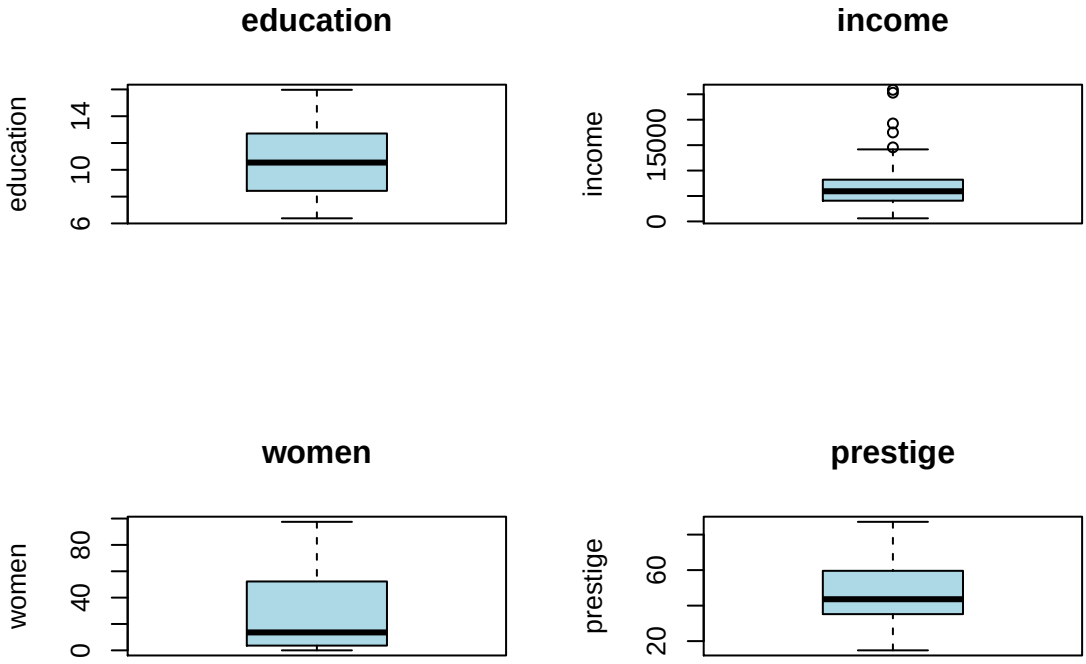
```r
pander(cor(d[,VAR_NUMERIC])) #-- matrice di correlazione
```

	education	income	women	prestige
education	1	0.5776	0.06185	0.8502
income	0.5776	1	-0.4411	0.7149
women	0.06185	-0.4411	1	-0.1183
prestige	0.8502	0.7149	-0.1183	1

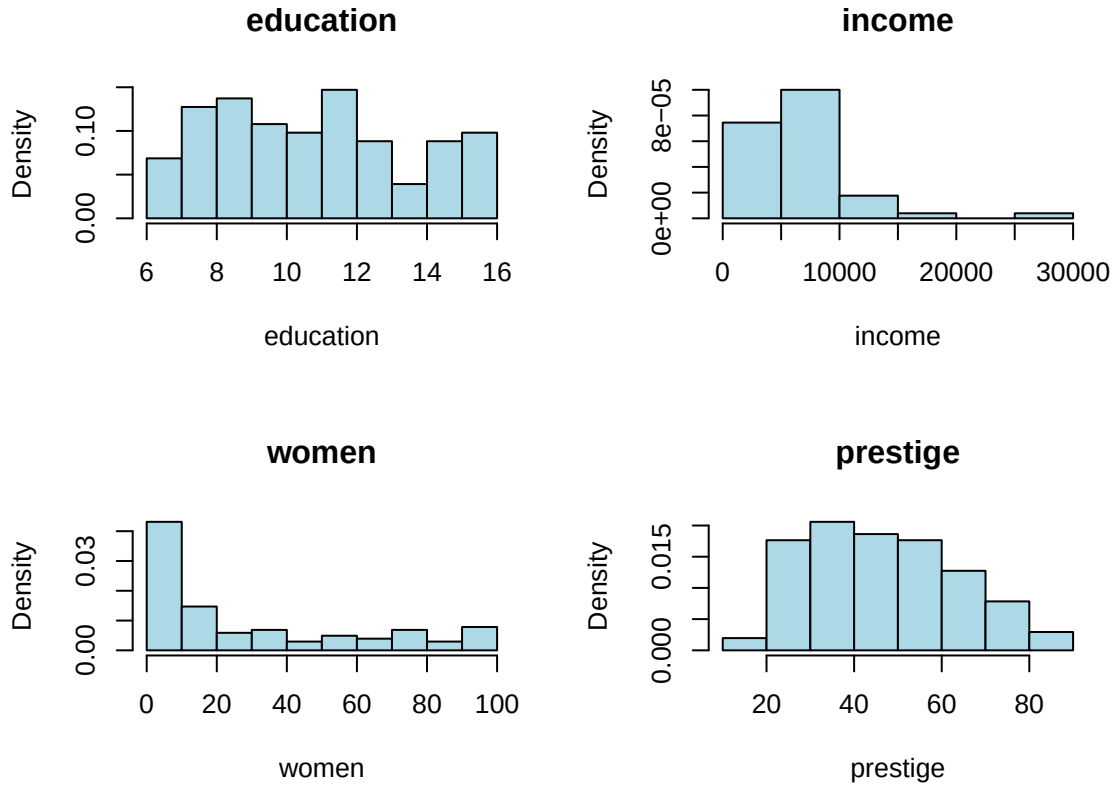
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




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




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



REGRESSIONE

Si analizza la dipendenza di "Prestige" da "Income" innanzitutto con una regressione lineare.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.14	2.268	11.97	5.135e-21
income	0.002897	0.0002833	10.22	3.192e-17

Table 5: Fitting linear model: prestige ~ income

Observations	Residual Std. Error	R^2	Adjusted R^2
102	12.09	0.5111	0.5062

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

Table 6: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	15,279	15,279	104.5	3.192e-17
Residuals	100	14,616	146.2	NA	NA

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

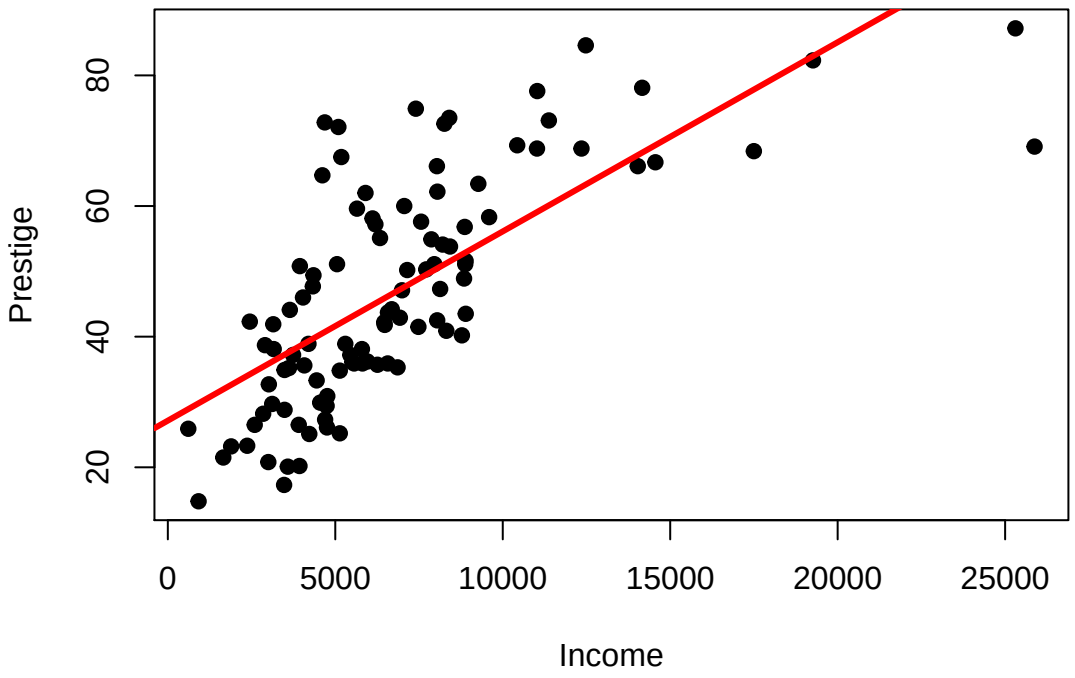
Test.statistic	P.value
8.096	0.01746

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

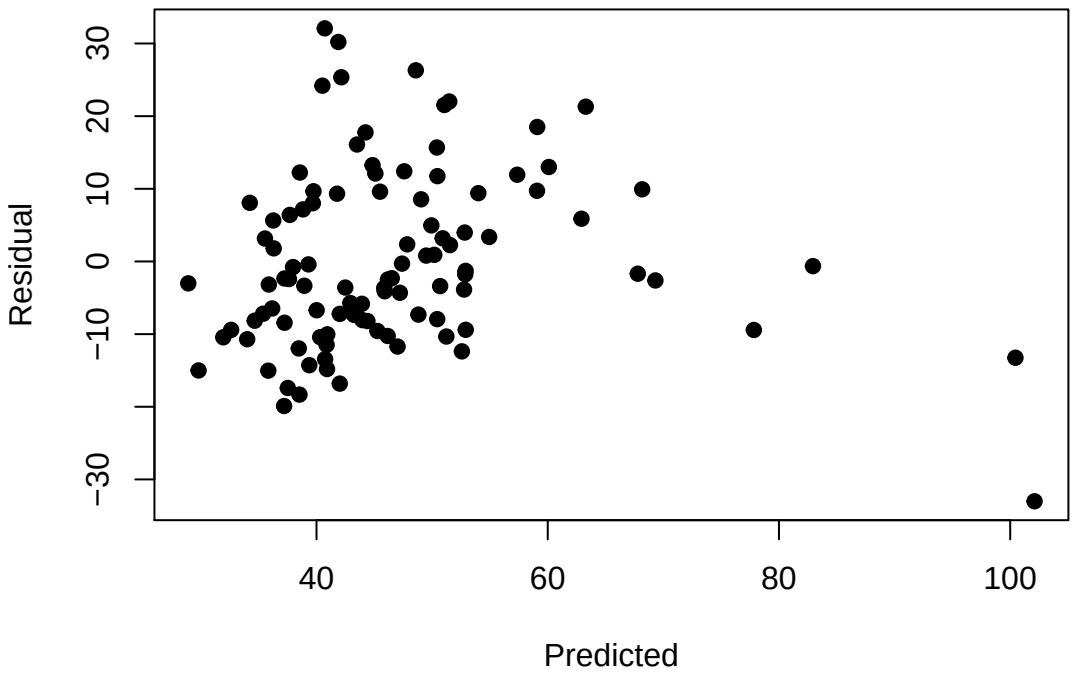
Table 8: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.126	3.278e-06 * * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(d$income,d$prestige,pch=19,xlab="Income",ylab="Prestige")
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
```

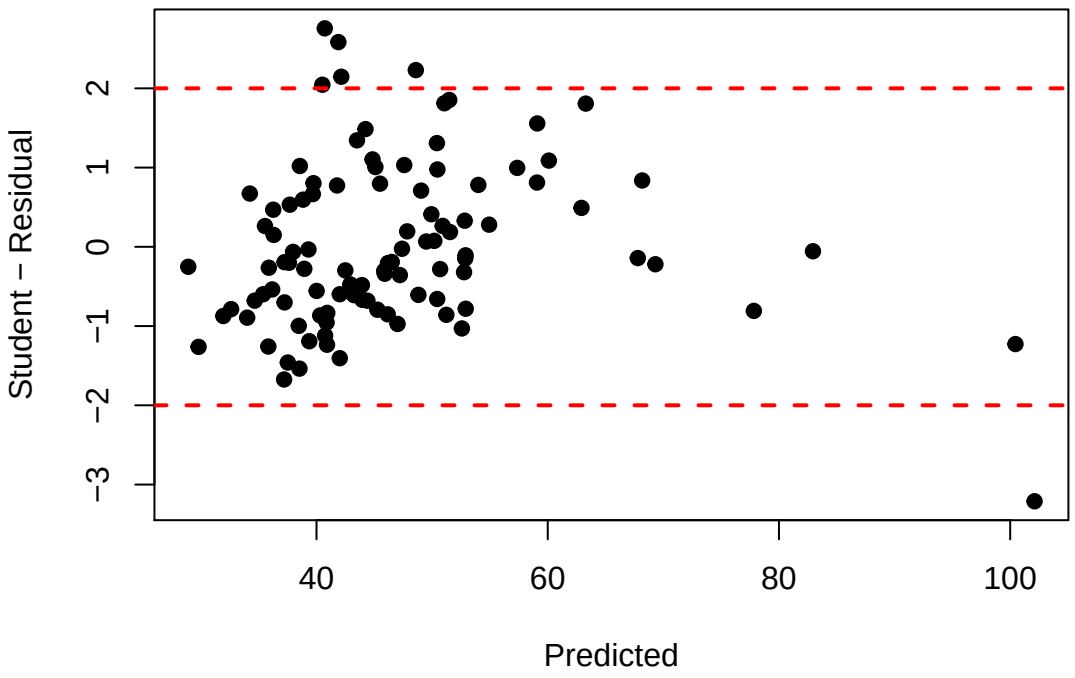



```r
#-- R CODE
plot(fitted(mod1),resid(mod1),pch=19,xlab="Predicted",ylab="Residual")
```



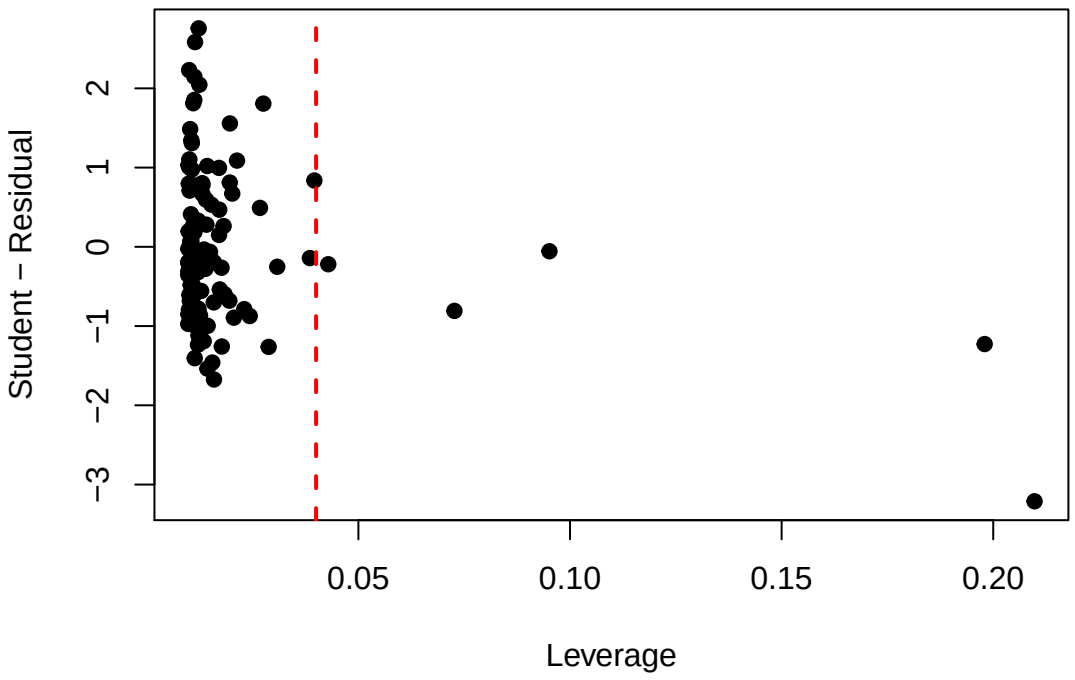
Il modello ha un discreto fitting ($R^2 = 0.5111$), "income" è significativa e gli errori sono sferici. Si nota piuttosto la presenza di outlier confermata dai grafici seguenti. Inoltre dai grafici prestige-income e residui-income traspare un legame non lineare non interpretato dal modello lineare semplice.

```
#-- R CODE
plot(fitted(mod1),rstudent(mod1),pch=19,xlab="Predicted",ylab="Student - Residual")
abline(h=-2,col=2,lty=2,lwd=2)
abline(h=2,col=2,lty=2,lwd=2)
```

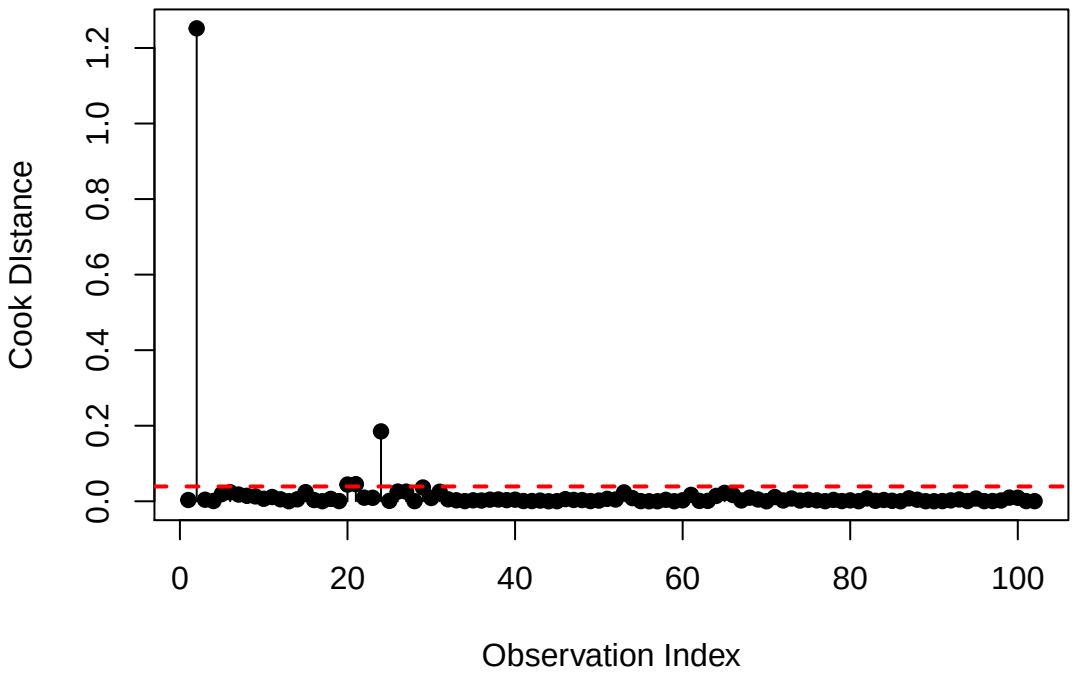



```
#-- R CODE

plot(hatvalues(mod1),rstudent(mod1),pch=19,xlab="Leverage",ylab="Student - Residual")
abline(v=0.04,col=2,lty=2,lwd=2)
```



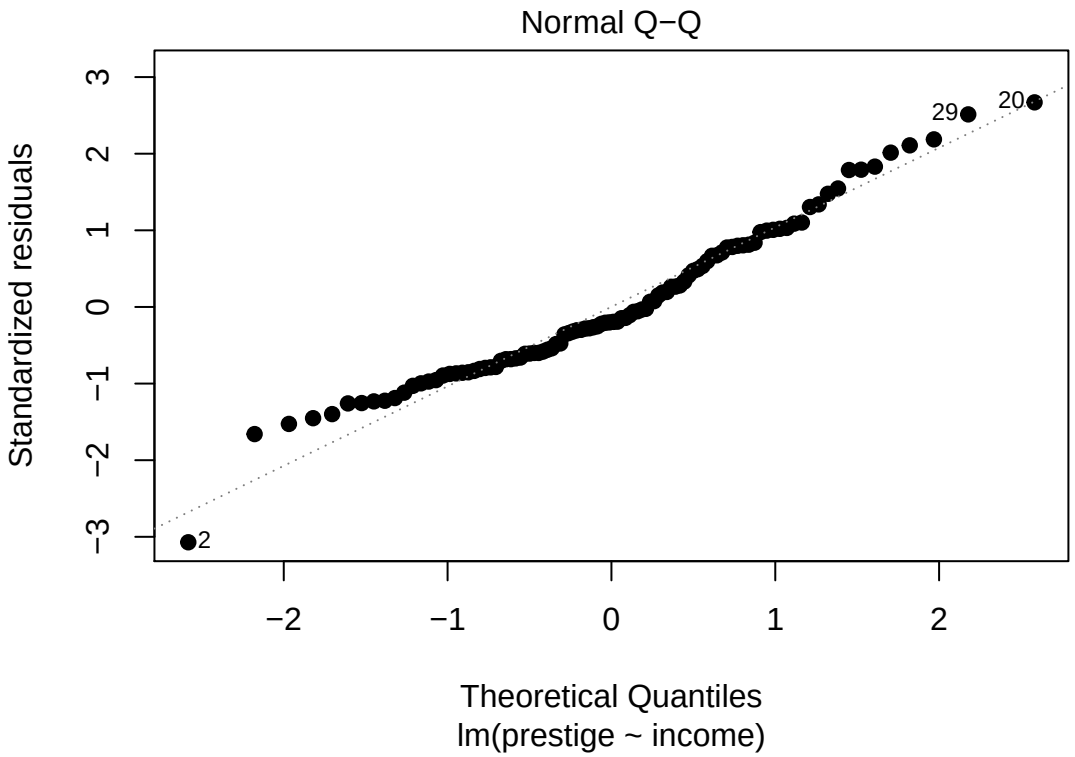

```
plot(cooks.distance(mod1),pch=19,xlab="Observation Index",ylab="Cook DIstance",type="h")
points(cooks.distance(mod1),pch=19)
abline(h=4/nrow(d),col=2,lty=2,lwd=2)
```



La distribuzione dei residui sembra normale eccetto per che una leggera asimmetria negativa ed emerge la presenza di outlier sulle code del Q-Q plot.

```
#-- R CODE

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



Pur dovendo eliminare gli outlier per avere risultati migliori ci si concentra sulla scelta di migliori interpolanti. Si verifica dapprima se e quali interpolanti di grado superiore al primo siano opportuni.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.18	3.515	4.035	0.0001078
income	0.006154	0.0007593	8.104	1.435e-12
I(income^2)	-1.433e-07	3.141e-08	-4.562	1.453e-05

Table 10: Fitting linear model: prestige ~ income + I(income^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
102	11.04	0.596	0.5879

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

Table 11: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	15,279	15,279	125.3	2.81e-19
I(income^2)	1	2,539	2,539	20.82	1.453e-05
Residuals	99	12,077	122	NA	NA

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

Test.statistic	P.value
3.619	0.1637

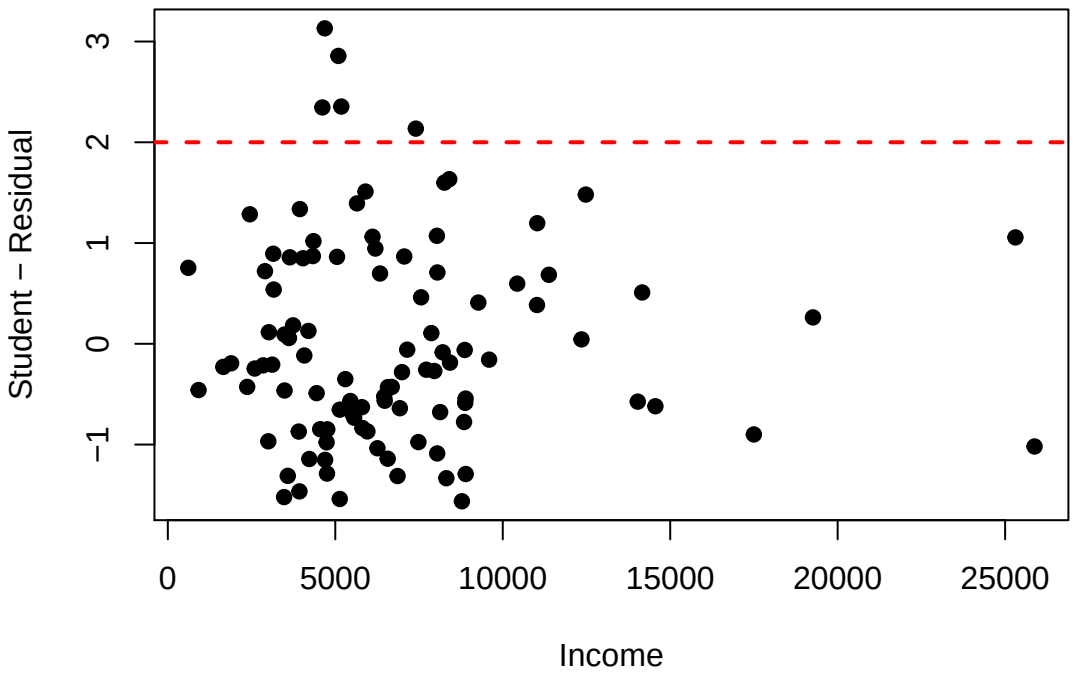
```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 13: Durbin-Watson test: `mod2`

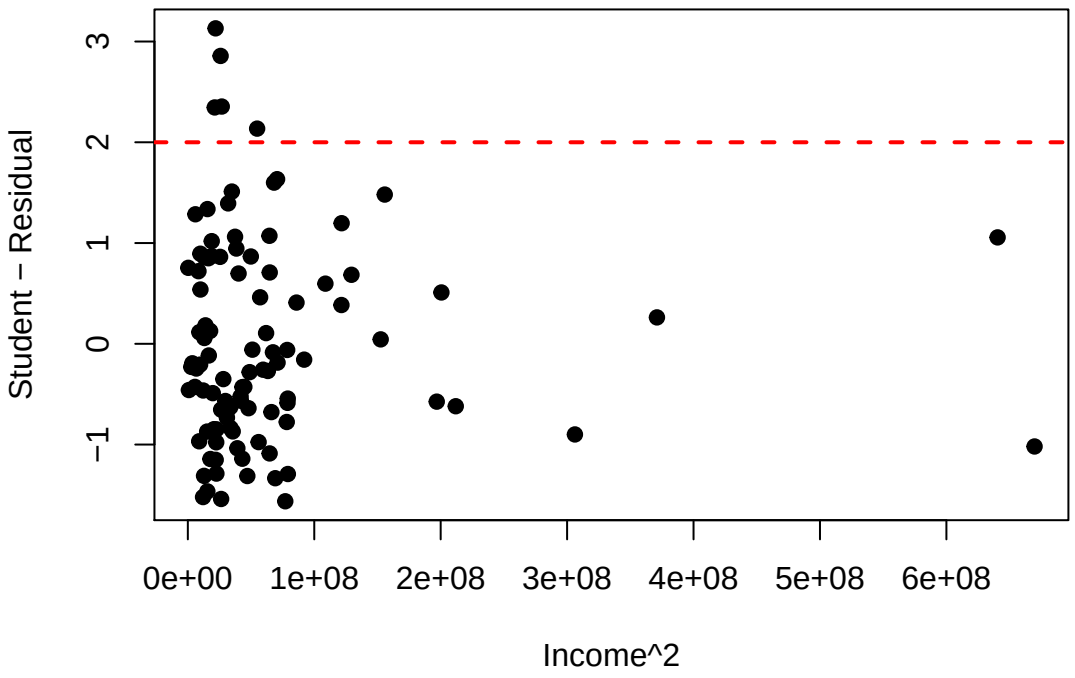
Test statistic	P value	Alternative hypothesis
1.198	1.717e-05 * * *	true autocorrelation is greater than 0

Il fitting migliora nettamente e risultano significativi sia il temine "income" lineare che quadratico. Lo si vede anche dai grafici residui-income residui – income^2 ove i residui sono compresi in intervalli di valori più contenuti.

```
#-- R CODE
plot(d$income,rstudent(mod2),pch=19,xlab="Income",ylab="Student - Residual")
abline(h=-2,col=2,lty=2,lwd=2)
abline(h=2,col=2,lty=2,lwd=2)
```




```
plot(d$income^2,rstudent(mod2),pch=19,xlab="Income^2",ylab="Student - Residual")
abline(h=-2,col=2,lty=2,lwd=2)
abline(h=2,col=2,lty=2,lwd=2)
```



I modelli di grado 3 e 4 non sono adeguati perché i parametri non sono significativi.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.27	5.668	2.693	0.00832
income	0.005705	0.001986	2.872	0.004998
I(income^2)	-9.595e-08	1.962e-07	-0.4891	0.6259
I(income^3)	-1.271e-12	5.196e-12	-0.2446	0.8073

Table 15: Fitting linear model: prestige ~ income + I(income^2) + I(income^3)

Observations	Residual Std. Error	R^2	Adjusted R^2
102	11.1	0.5963	0.5839

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

Table 16: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	15,279	15,279	124.1	4.18e-19
I(income^2)	1	2,539	2,539	20.62	1.597e-05
I(income^3)	1	7.366	7.366	0.05981	0.8073
Residuals	98	12,070	123.2	NA	NA

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

Test.statistic	P.value
3.414	0.1814

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

Table 18: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
1.206	2.071e-05 * * *	true autocorrelation is greater than 0

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.78	8.504	1.973	0.05135
income	0.004778	0.004359	1.096	0.2756
I(income^2)	6.994e-08	7.212e-07	0.09698	0.9229
I(income^3)	-1.195e-11	4.497e-11	-0.2658	0.791
I(income^4)	2.164e-16	9.051e-16	0.2391	0.8115

Table 20: Fitting linear model: prestige ~ income + I(income^2) + I(income^3) + I(income^4)

Observations	Residual Std. Error	R^2	Adjusted R^2
102	11.15	0.5965	0.5799

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

Table 21: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	15,279	15,279	122.9	6.222e-19
I(income^2)	1	2,539	2,539	20.42	1.756e-05

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(income^3)	1	7.366	7.366	0.05923	0.8082
I(income^4)	1	7.111	7.111	0.05718	0.8115
Residuals	97	12,062	124.4	NA	NA

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

Test.statistic	P.value
3.473	0.1761

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

Table 23: Durbin-Watson test: `mod4`

Test statistic	P value	Alternative hypothesis
1.21	2.025e-05 * * *	true autocorrelation is greater than 0

Si propone ora un modello log-lin in cui la variabile $log(Prestige)$ viene regredita su "Income".

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.353	0.05466	61.34	3.664e-81
income	6.208e-05	6.829e-06	9.091	9.727e-15

Table 25: Fitting linear model: I(log(prestige)) ~ income

Observations	Residual Std. Error	R^2	Adjusted R^2
102	0.2914	0.4525	0.447

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

Table 26: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	7.018	7.018	82.64	9.727e-15
Residuals	100	8.492	0.08492	NA	NA

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

Test.statistic	P.value
19.7	5.263e-05

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

Table 28: Durbin-Watson test: `mod5`

Test statistic	P value	Alternative hypothesis
1.246	4.919e-05 * * *	true autocorrelation is greater than 0

Il parametro associato alla variabile "income" è significativo ma il fitting è peggiore e gli errori sono non correlati ma viene respinta l'ipotesi di omoschedasticità. Se si analizza quindi il modello lin-log in cui la variabile prestige è regredita rispetto a $log(Income)$; i parametri sono significativi ma il fitting è leggermente peggiore rispetto al caso quadratico e gli errori omoschedastici ma viene respinta l'ipotesi di loro non correlazione.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-139.9	16.95	-8.249	6.602e-13
I(log(income))	21.56	1.953	11.04	5.352e-19

Table 30: Fitting linear model: prestige ~ I(log(income))

Observations	Residual Std. Error	R^2	Adjusted R^2
102	11.61	0.5492	0.5447

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

Table 31: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(income))	1	16,417	16,417	121.8	5.352e-19
Residuals	100	13,478	134.8	NA	NA

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

Test.statistic	P.value
2.908	0.2336

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

Table 33: Durbin-Watson test: `mod6`

Test statistic	P value	Alternative hypothesis
1.108	2.121e-06 * * *	true autocorrelation is greater than 0

Il modello log-log in cui la variabile $log(Prestige)$ viene regredita su $log(Income)$ ha un fitting solo leggermente peggiore che il modello quadratico, i parametri sono significativi ma viene respinta sia l'ipotesi di omoschedasticità che di non correlazione dei residui.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-139.9	16.95	-8.249	6.602e-13
I(log(income))	21.56	1.953	11.04	5.352e-19

Table 35: Fitting linear model: prestige ~ I(log(income))

Observations	Residual Std. Error	R^2	Adjusted R^2
102	11.61	0.5492	0.5447

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

Table 36: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(income))	1	16,417	16,417	121.8	5.352e-19
Residuals	100	13,478	134.8	NA	NA

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

Test.statistic	P.value
2.908	0.2336

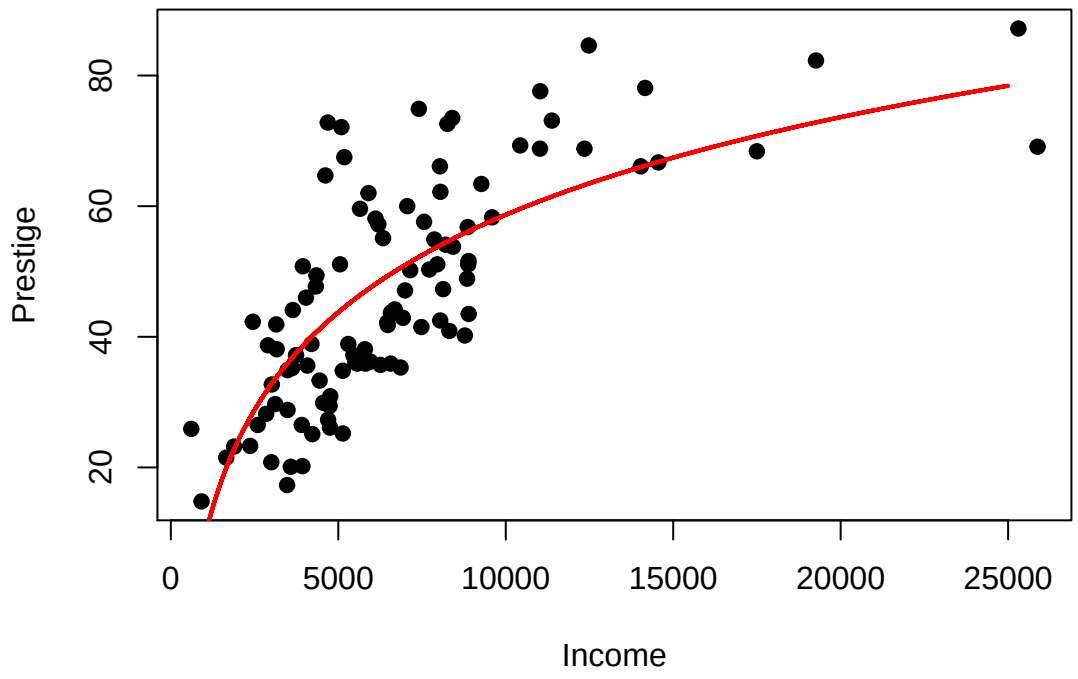
```r
pander(dwtest(mod7),big.mark=",") #-- Durbin-Whatson test
```

Table 38: Durbin-Watson test: `mod7`

Test statistic	P value	Alternative hypothesis
1.108	2.121e-06 * * *	true autocorrelation is greater than 0

Test statistic	P value	Alternative hypothesis

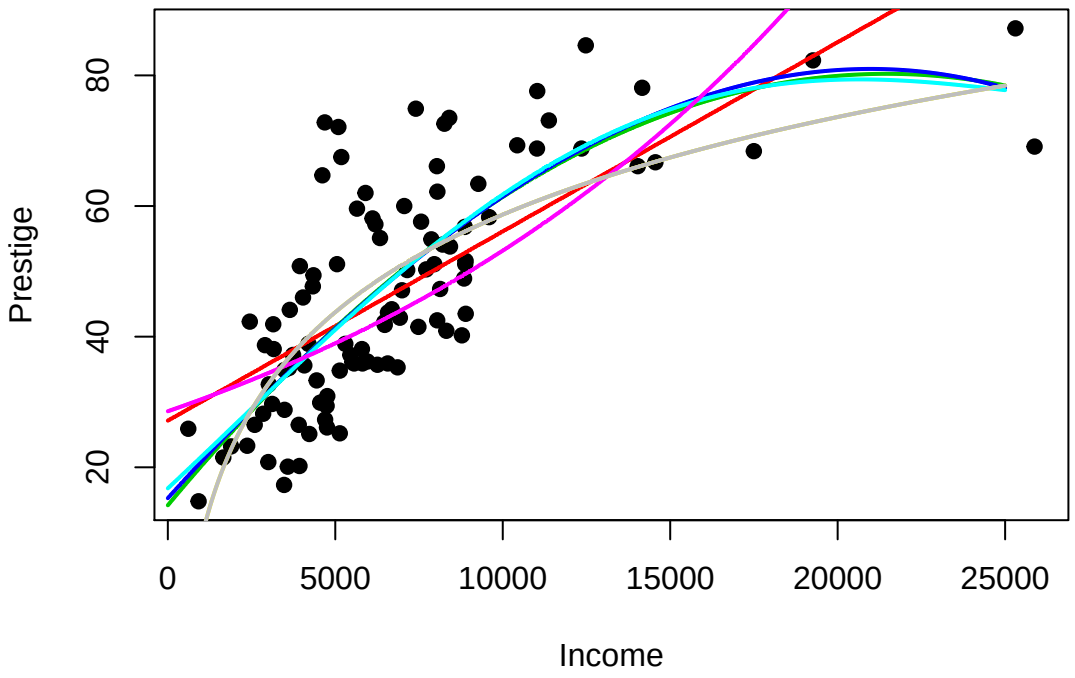
```
#-- R CODE
plot(d$income,d$prestige,pch=19,xlab="Income",ylab="Prestige")
lines(seq(0,25000,0.1),predict(mod7,data.frame(income=seq(0,25000,0.1))),col=2,lwd=2)
```



Il modello prescelto è quindi quello quadratico.

Si rappresentano congiuntamente i diversi modelli:

```
#-- R CODE
plot(d$income,d$prestige,pch=19,xlab="Income",ylab="Prestige")
lines(seq(0,25000,1),predict(mod1,data.frame(income=seq(0,25000,1))),col=2,lwd=2)
lines(seq(0,25000,1),predict(mod2,data.frame(income=seq(0,25000,1))),col=3,lwd=2)
lines(seq(0,25000,1),predict(mod3,data.frame(income=seq(0,25000,1))),col=4,lwd=2)
lines(seq(0,25000,1),predict(mod4,data.frame(income=seq(0,25000,1))),col=5,lwd=2)
lines(seq(0,25000,1),exp(predict(mod5,data.frame(income=seq(0,25000,1)))),col=6,lwd=2)
lines(seq(0,25000,1),predict(mod6,data.frame(income=seq(0,25000,1))),col=7,lwd=2)
lines(seq(0,25000,1),predict(mod7,data.frame(income=seq(0,25000,1))),col=8,lwd=2)
```


Prestige
20
40
60
80
0
5000
10000
15000
20000
25000
Income