LINEAR 1 - Data set: CAR

INTRODUZIONE

Il dataset 'CAR' è composto da 60 unità statistiche per le quali è riportato il valore di 8 variabili. In particolare, viene considerato un insieme di 60 automobili per ogniuna delle quali viene misurato il valore delle seguenti variabili:

1. PRICE: prezzo di listino dell'autovettura (in particolare di un modello standard), espresso in dollari
2. COUNTRY: paese d'origine
3. RELIABILITY: grado di affidabilità (fattore codificato in livelli da 1 a 5)
4. MILIAGE: (consumo di carburante espresso in miglia / dollaro)
5. TYPE: tipologia di autovettura
6. WEIGHT: peso a vuoto misurato in libbre
7. DISP: capacità del motore (cilindrata), in litri
8. HP: potenza del veicolo

Variabile dipendente: PRICE. Le caratteristiche del veicolo sono variabili esplicative (o covariate).

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare semplice
3. Test di correlazione dei residui
4. Modello quadratico (con e senza outlier)
5. Modelli log lineari

```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"
```

```r
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\3.lin(5)\\1.linear\\car.test.txt"),sep=" "

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("Price","Mileage","Weight","Disp.","HP")

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

Table 1: Table continues below

model	Price	Country	Reliability	Mileage	Type	Weight
Eagle Summit 4	8,895	USA	4	33	Small	2,560
Ford Escort 4	7,402	USA	2	33	Small	2,345
Ford Festiva 4	6,319	Korea	4	37	Small	1,845
Honda Civic 4	6,635	Japan/USA	5	32	Small	2,260
Mazda Protege 4	6,599	Japan	5	32	Small	2,440
Mercury Tracer 4	8,672	Mexico	4	26	Small	2,285

Disp.	HP
97	113
114	90
81	63
91	92
113	103
97	82

STATISTICHE DESCRITTIVE

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

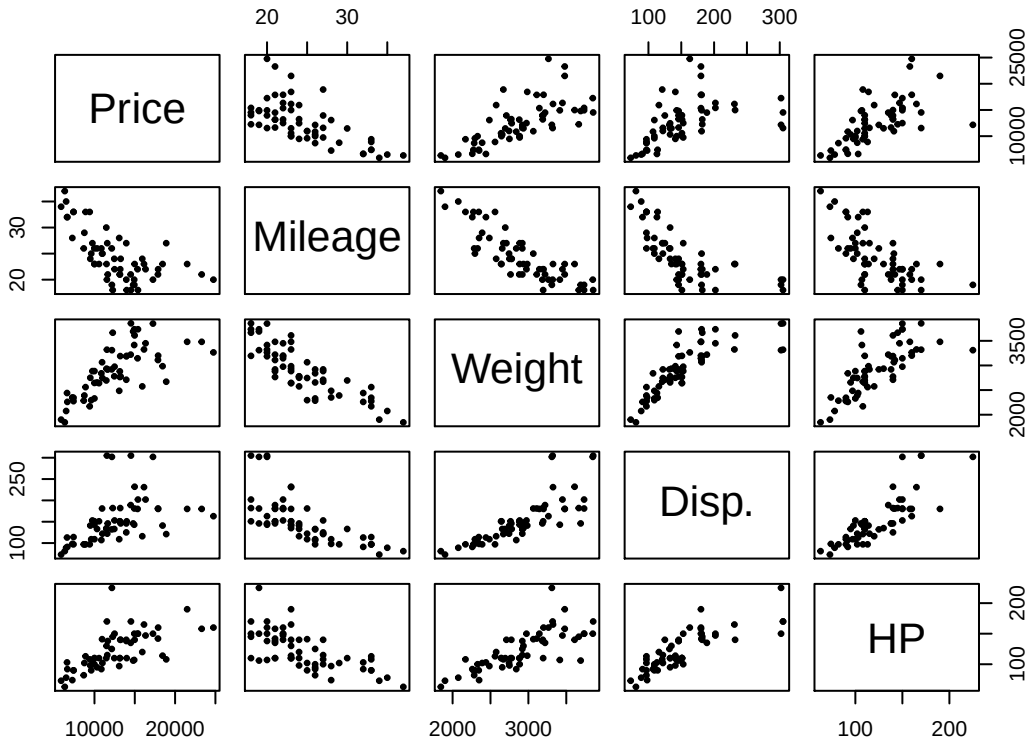
Price	Mileage	Weight	Disp.	HP
Min. : 5866	Min. :18.00	Min. :1845	Min. : 73.0	Min. : 63.0
1st Qu.: 9932	1st Qu.:21.00	1st Qu.:2571	1st Qu.:113.8	1st Qu.:101.5
Median :12216	Median :23.00	Median :2885	Median :144.5	Median :111.5
Mean :12616	Mean :24.58	Mean :2901	Mean :152.1	Mean :122.3
3rd Qu.:14933	3rd Qu.:27.00	3rd Qu.:3231	3rd Qu.:180.0	3rd Qu.:142.8
Max. :24760	Max. :37.00	Max. :3855	Max. :305.0	Max. :225.0

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

	Price	Mileage	Weight	Disp.	HP
Price	1	-0.6538	0.7018	0.4857	0.6536
Mileage	-0.6538	1	-0.8479	-0.6932	-0.6667

	Price	Mileage	Weight	Disp.	HP
Weight	0.7018	-0.8479	1	0.8033	0.7629
Disp.	0.4857	-0.6932	0.8033	1	0.8182
HP	0.6536	-0.6667	0.7629	0.8182	1

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



Non ci sono variabili collineari.

REGRESSIONE

Si regredisce il "prezzo" su "disp" dapprima in termini lineari.

```
#-- R CODE
mod1 <- lm(Price~Disp.,d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7,049	1,395	5.052	4.658e-06
Disp.	36.61	8.653	4.231	8.369e-05

Table 6: Fitting linear model: Price ~ Disp.

Observations	Residual Std. Error	R^2	Adjusted R^2
60	3600	0.2359	0.2227

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

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Disp.	1	2.32e+08	2.32e+08	17.9	8.369e-05
Residuals	58	751,549,306	12,957,747	NA	NA

Il modello interpreta i dati e "Disp"" risulta essere significativa ma il fitting è molto basso ($R^2 = 0.2359$). Il test di White accetta l'ipotesi di omoschedaticità e il test di Durbin-Watson quella di non correlazione fra i residui.

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

Test.statistic	P.value
1.275	0.5285

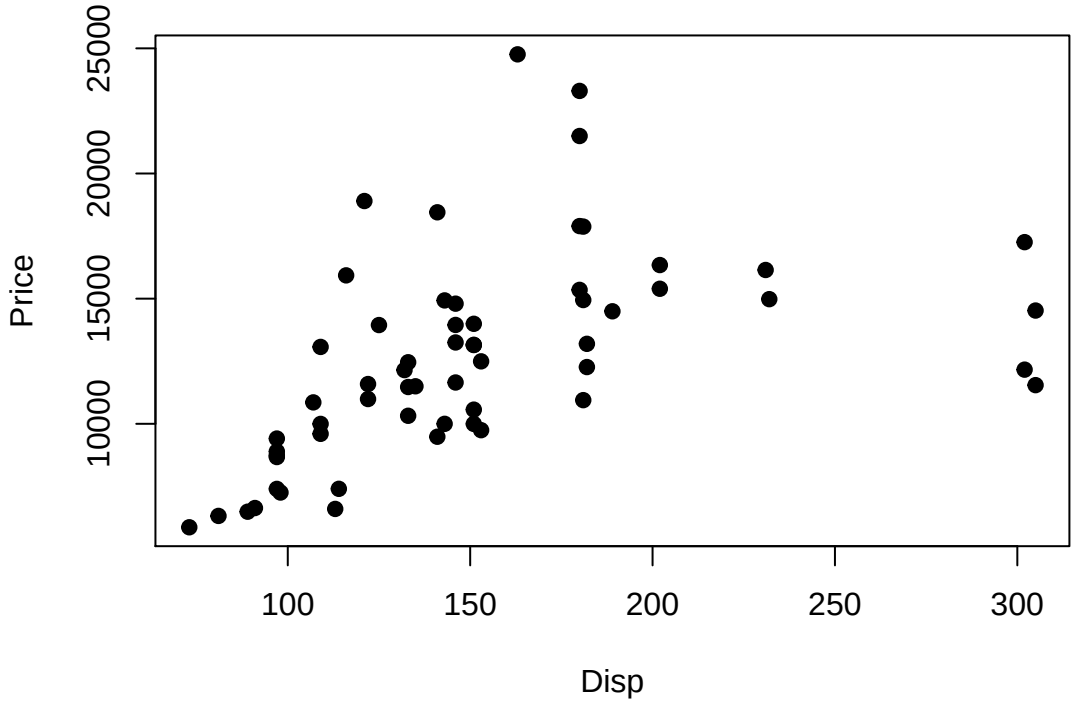
```
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

Table 9: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.264	0.00116 * *	true autocorrelation is greater than 0

Si consideri ora il grafico "Prezzo"" vs "Disp":

```
#-- R CODE
plot(d$Disp.,d$Price,pch=19,col=1,xlab="Disp",ylab="Price") #-- scatter plot
```



L'andamento del grafico suggerisce che il legame non sia lineare. Si propone quindi un modello quadratico $Price = f(Disp, Disp^2)$.

REGRESSIONE CON MODELLO QUADRATICO

```
#-- R CODE
mod2 <- lm(Price~Disp.+I(Disp.^2),d) #-- stima del modello quadratico
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8,099	3,223	-2.513	0.01484
Disp.	219.3	36.94	5.936	1.828e-07
I(Disp.^2)	-0.4858	0.09632	-5.044	4.955e-06

Table 11: Fitting linear model: Price ~ Disp. + I(Disp.^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
60	3019	0.4717	0.4531

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

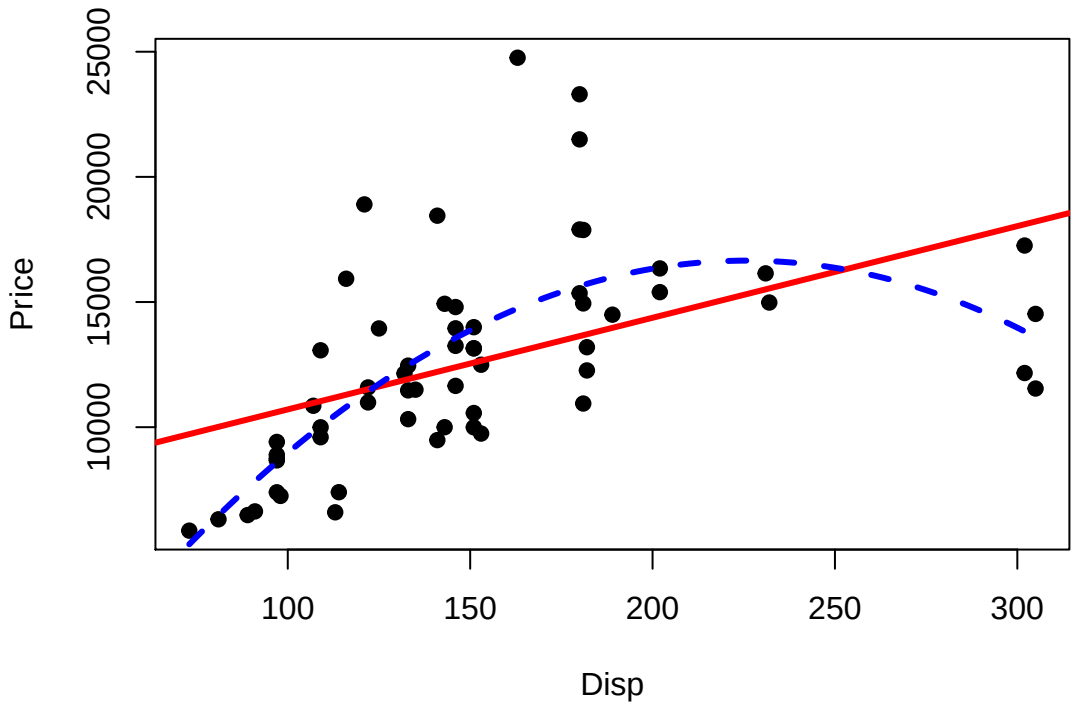
Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Disp.	1	2.32e+08	2.32e+08	25.45	4.936e-06
I(Disp.^2)	1	231,904,504	231,904,504	25.44	4.955e-06
Residuals	57	519,644,802	9,116,575	NA	NA

Il fitting raddoppia (osservare l'R^2) e anche la variabile $Disp^2$ risulta essere significativa: il modello quindi è realmente quadratico come si vede dalla seguente rappresentazione grafica:

```r
#-- R CODE
f_mod2 <- function(x) coefficients(mod2)[1]+coefficients(mod2)[2]*x+coefficients(mod2)[3]*x^2

plot(d$Disp.,d$Price,pch=19,xlab="Disp",ylab="Price")
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
curve(f_mod2,add=T,col="blue",lwd=3,lty=2) #-- abline del modello quadratico
```



E' confermata omoschedasticità e non correlazione degli errori.

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

Test.statistic	P.value
2.612	0.2709

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

Table 14: Durbin-Watson test: `mod2`

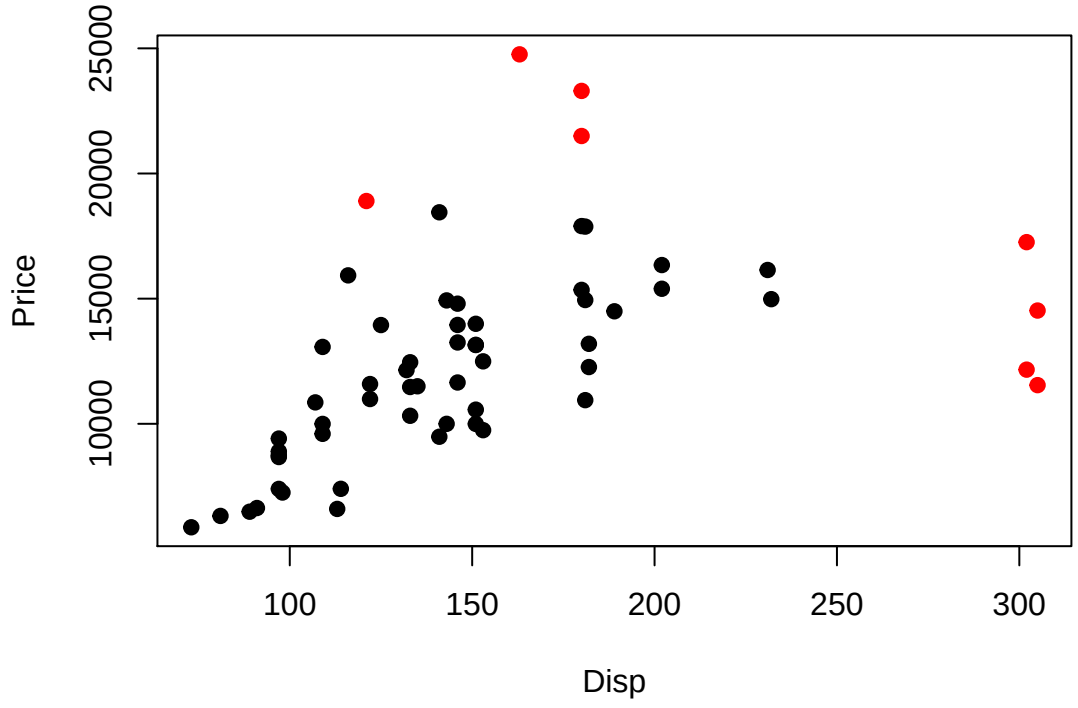
Test statistic	P value	Alternative hypothesis
1.68	0.08223	true autocorrelation is greater than 0

La rappresentazione grafica suggerisce però la presenza di outlier. Si analizza perciò la distribuzione dei valori estremi di "Price"" e "Disp".

```r
#-- R CODE
d$ESTREME <- 1 #-- inserisco una nuova colonna del dataset con obs. estreme

#-- ora applico la funzione FIND_EXTREME_OBSERVARION(variabile di interesse,fattore).
#-- si include anche l'osservazione 23 come outlier.
d$ESTREME[c(FIND_EXTREME_OBSERVARION(d$Price,2),FIND_EXTREME_OBSERVARION(d$Disp.,2),23)] <- 2

plot(d$Disp.,d$Price,pch=19,col=d$ESTREME,xlab="Disp",ylab="Price")
```




```
#-- d_noout è un nuovo data frame senza le osservazioni outlier
d_noout <- d[-c(FIND_EXTREME_OBSERVARION(d$Price,2),FIND_EXTREME_OBSERVARION(d$Disp.,2),23),]
```

I valori estremi si discostano di molto dagli altri valori, come si vede dal grafico. Pertanto si procede ad eliminare gli outlier e a stimare nuovamente il modello. Si ripropone il modello lineare senza outlier.

```
#-- R CODE
mod_noout <- lm(Price~Disp.,d_noout) #-- modello senza outlier
pander(summary(mod_noout),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2,775	1,207	2.298	0.02575
Disp.	64.52	8.354	7.722	4.48e-10

Table 16: Fitting linear model: Price ~ Disp.

Observations	Residual Std. Error	R^2	Adjusted R^2
52	2222	0.5439	0.5348

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

Table 17: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Disp.	1	294,483,168	294,483,168	59.63	4.48e-10
Residuals	50	246,905,125	4,938,102	NA	NA

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

Test.statistic	P.value
0.8399	0.6571

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

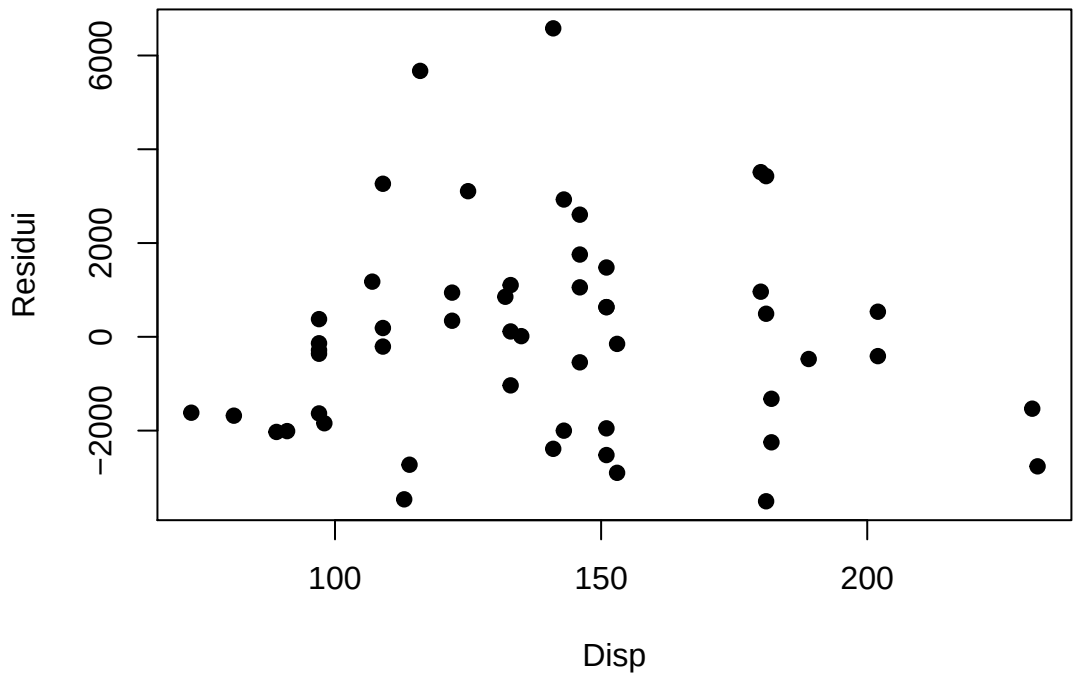
Table 19: Durbin-Watson test: `mod_noout`

Test statistic	P value	Alternative hypothesis
1.6	0.05863	true autocorrelation is greater than 0

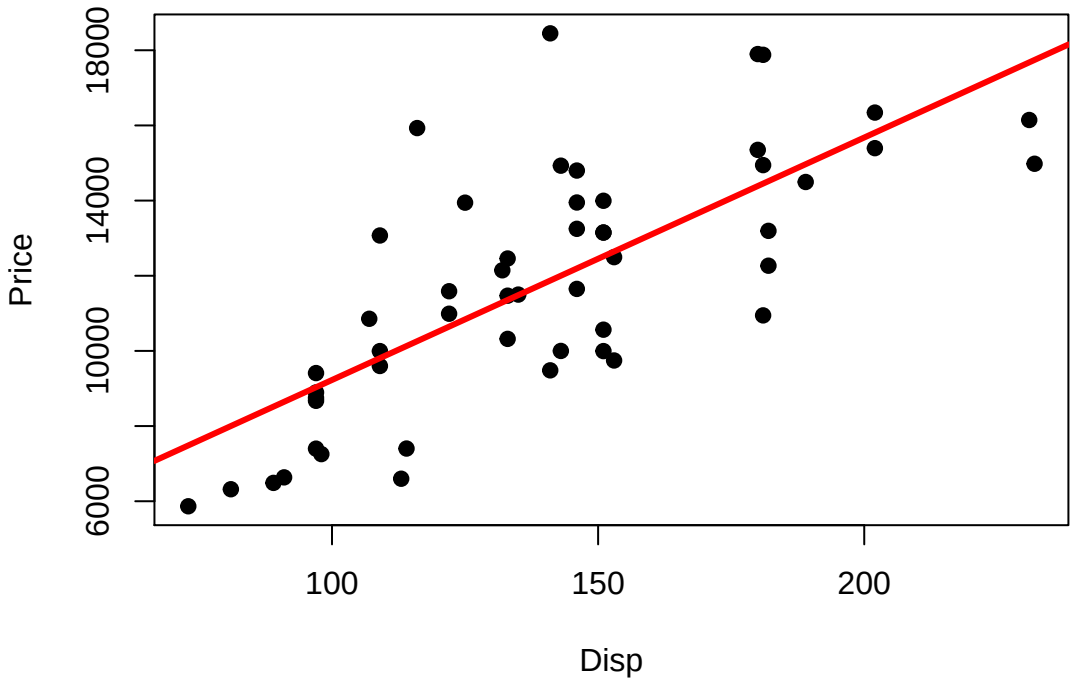
Il fitting migliora moltissimo e "Disp" rimane significativo. Gli errori sono ancora sferici anche se la non correlazione non è nettissima.

Tuttavia osservando il grafico residui-prezzi e il grafico della regressione lineare si nota ancora la presenza di una non linearità nella relazione non interpretata dall'interpolante lineare stessa.

```
#-- R CODE
plot(d_noout$Disp.,resid(mod_noout),xlab="Disp",ylab="Residui",pch=19)
```




```
plot(d_noout$Disp.,d_noout$Price,xlab="Disp",ylab="Price",pch=19)
abline(mod_noout,col=2,lwd=3)
```



Si propone quindi un modello quadratico (come fatto precedentemente) ma senza outlier.

```
#-- R CODE
mod2_noout <- lm(Price~Disp.+I(Disp.^2),d_noout)
pander(summary(mod2_noout),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5,764	3,781	-1.525	0.1338
Disp.	187.4	52.44	3.574	0.0008003
I(Disp.^2)	-0.4135	0.1743	-2.372	0.02167

Table 21: Fitting linear model: Price ~ Disp. + I(Disp.^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
52	2126	0.5909	0.5742

```
pander(anova(mod2_noout),big.mark=",")
```

Table 22: Analysis of Variance Table

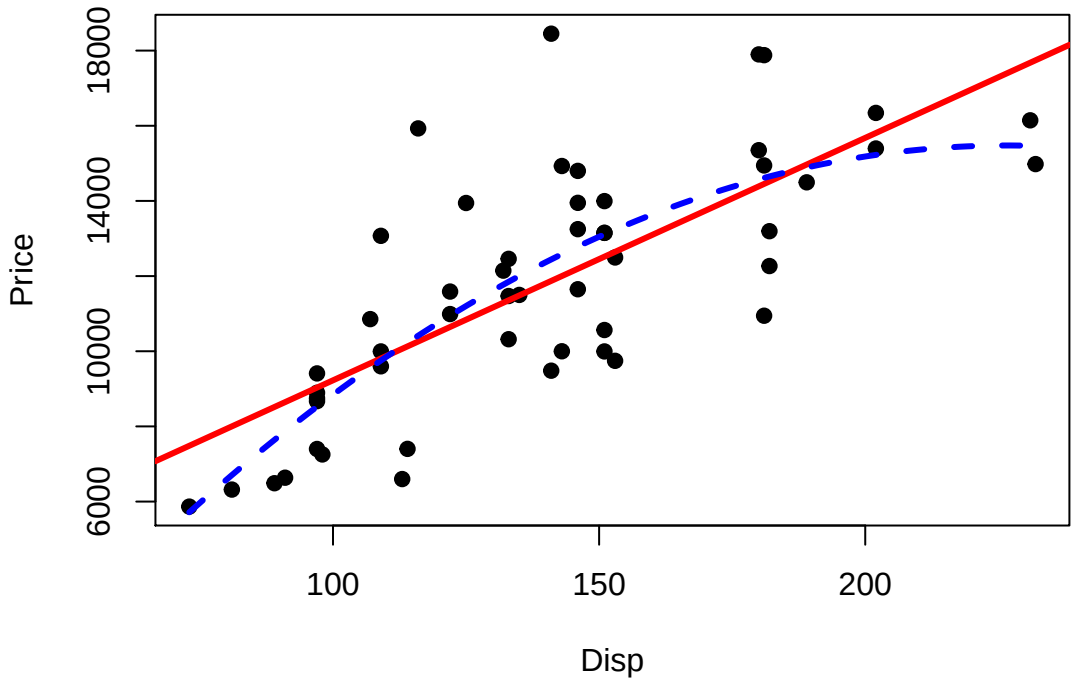
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Disp.	1	294,483,168	294,483,168	65.15	1.485e-10

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(Disp.^2)	1	25,427,064	25,427,064	5.626	0.02167
Residuals	49	221,478,061	4,519,960	NA	NA

Il modello migliora ancora il fitting e "Disp" risulta significativo come anche $Disp^2$ ma in modo non così netto come nel caso con outlier.

```
#-- R CODE
f_mod2_noout <- function(x) coefficients(mod2_noout)[1]+coefficients(mod2_noout)[2]*x+coefficients(mod2_

plot(d_noout$Disp.,d_noout$Price,pch=19,xlab="Disp",ylab="Price")
abline(mod_noout,col=2,lwd=3) #-- abline del modello lineare
curve(f_mod2_noout,add=T,col="blue",lwd=3,lty=2) #-- abline del modello quadratico
```



Il modello quadratico ha errori non solo omoschedastici ma con molta più chiarezza, anche incorrelati.

```
#-- R CODE
pander(white.test(mod2_noout),big.mark=",") #-- white test
```

Test.statistic	P.value
3.626	0.1631

```r
pander(dwtest(mod2_noout),big.mark=",") #-- Durbin-Whatson test
```

Table 24: Durbin-Watson test: `mod2_noout`

Test statistic	P value	Alternative hypothesis
1.922	0.3456	true autocorrelation is greater than 0

Proviamo ora se un modello cubico è più adatto ad interpretare i dati:

```r
#-- R CODE
mod3_noout <- lm(Price~Disp.+I(Disp.^2)+I(Disp.^3),d_noout) #-- modello cubico
pander(summary(mod3_noout),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12,525	12,834	-0.9759	0.334
Disp.	335.3	273.2	1.227	0.2257
I(Disp.^2)	-1.432	1.854	-0.7721	0.4438
I(Disp.^3)	0.002218	0.004021	0.5516	0.5838

Table 26: Fitting linear model: Price ~ Disp. + I(Disp.^2) + I(Disp.^3)

Observations	Residual Std. Error	R^2	Adjusted R^2
52	2141	0.5935	0.5681

```r
pander(anova(mod3_noout),big.mark=",")
```

Table 27: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Disp.	1	294,483,168	294,483,168	64.23	2.096e-10
I(Disp.^2)	1	25,427,064	25,427,064	5.546	0.02266
I(Disp.^3)	1	1,395,269	1,395,269	0.3043	0.5838
Residuals	48	220,082,792	4,585,058	NA	NA

Il modello sembra mantenere lo stesso fitting ma "disp", $dist^2$ e $disp^3$ non risultano significativi. Quindi si opta per ora sul modello quadratico. Si verifica ora un modello lineare-log in cui la variabile esplicativa $log(Disp)$.

```r
#-- R CODE
mod4_noout <- lm(Price~log(Disp.),d_noout) #-- stima modello log lineare
pander(summary(mod4_noout),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-33,695	5,449	-6.184	1.132e-07
log(Disp.)	9,273	1,109	8.36	4.643e-11

Table 29: Fitting linear model: Price ~ log(Disp.)

Observations	Residual Std. Error	R^2	Adjusted R^2
52	2125	0.5829	0.5746

```
pander(anova(mod4_noout),big.mark=",")
```

Table 30: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
log(Disp.)	1	315,593,437	315,593,437	69.88	4.643e-11
Residuals	50	225,794,856	4,515,897	NA	NA

$Log(Disp)$ risulta significativa ma il fitting peggiora leggermente ($R^2 = 0.5829$). Si verifica ora la sfericità dei residui:

```
#-- R CODE
pander(white.test(mod4_noout),big.mark=",")
```

Test.statistic	P.value
3.031	0.2197

```
pander(dwtest(mod4_noout),big.mark=",") #-- Durbin-Whatson test
```

Table 32: Durbin-Watson test: `mod4_noout`

Test statistic	P value	Alternative hypothesis
1.774	0.1778	true autocorrelation is greater than 0

Gli errori sono non correlati ma non sono più omoschedastici per $\alpha = 0.1$. Quindi si rimane sulla scelta del modello quadratico. Si verifica ora l'opportunità di un modello loglineare in cui come variabile dipendente si abbia $log(Price)$.

```
#-- R CODE
mod5_noout <- lm(I(log(Price))~Disp.,d_noout) #-- stima modello log lineare
pander(summary(mod5_noout),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.507	0.108	78.78	4.161e-54

	Estimate	Std. Error	t value	Pr(>|t|)
Disp.	0.005923	0.0007473	7.926	2.167e-10

Table 34: Fitting linear model: I(log(Price)) ~ Disp.

Observations	Residual Std. Error	R^2	Adjusted R^2
52	0.1988	0.5568	0.5479

```
pander(anova(mod5_noout),big.mark=",")
```

Table 35: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Disp.	1	2.482	2.482	62.82	2.167e-10
Residuals	50	1.976	0.03951	NA	NA

```
pander(white.test(mod5_noout),big.mark=",") #-- white test
```

Test.statistic	P.value
1.213	0.5452

```
pander(dwtest(mod5_noout),big.mark=",") #-- Durbin-Whatson test
```

Table 37: Durbin-Watson test: `mod5_noout`

Test statistic	P value	Alternative hypothesis
1.434	0.01387 *	true autocorrelation is greater than 0

"Disp" è significativo, gli errori sferici ma il fitting peggiora pertanto non si assume neanche questo modello. Si propone quindi un modello log-log in cui si abbia come variabile dipendente $log(Price)$ e come variabile esplicativa $log(Disp)$.

```
#-- R CODE
mod6_noout <- lm(I(log(Price))~I(log(Disp.)),d_noout) #-- stima modello log lineare
pander(summary(mod6_noout),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.088	0.4737	10.74	1.379e-14
I(log(Disp.))	0.8657	0.09645	8.976	5.345e-12

Table 39: Fitting linear model: I(log(Price)) ~ I(log(Disp.))

Observations	Residual Std. Error	R^2	Adjusted R^2
52	0.1848	0.6171	0.6094

```
pander(anova(mod6_noout),big.mark=",")
```

Table 40: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(Disp.))	1	2.751	2.751	80.57	5.345e-12
Residuals	50	1.707	0.03414	NA	NA

```
pander(white.test(mod6_noout),big.mark=",") #-- white test
```

Test.statistic	P.value
2.181	0.336

```
pander(dwtest(mod6_noout),big.mark=",") #-- Durbin-Whatson test
```

Table 42: Durbin-Watson test: `mod6_noout`

Test statistic	P value	Alternative hypothesis
1.6	0.05816	true autocorrelation is greater than 0

In definitiva, $log(Disp)$ è significativo, gli errori chiaramente sferici e il fitting migliora ($R^2 = 0.6171$): si sceglie quindi in definitiva il modello log-log. Si potrebbe ovviamente proseguire costruendo modelli più complessi che comprendano il logaritmo della variabile esplicativa e suoi termini quadratici.