Esame Modelli Statistici

Nicola Zucchia

2023-02-15

Esplorazione dei dati

Questo lavoro consiste nell'analisi dei dati reperibili al link https://www.kaggle.com/datasets/hellbuoy/car-price-prediction?select=CarPrice_Assignment.csv e la costruzione di un modello di regressione lineare multipla.

Il primo passo consiste nel caricare i dati e darne una panoramica con le prime righe del dataset.

```
d <- read.csv("R Files/CarPrice_Assignment.csv", header = TRUE)
head(d)
```

```
##   car_ID symboling                  CarName fueltype aspiration doornumber
## 1      1         3       alfa-romero giulia      gas        std        two
## 2      2         3      alfa-romero stelvio      gas        std        two
## 3      3         1 alfa-romero Quadrifoglio      gas        std        two
## 4      4         2              audi 100 ls      gas        std       four
## 5      5         2               audi 100ls      gas        std       four
## 6      6         2                 audi fox      gas        std        two
##       carbody drivewheel enginelocation wheelbase carlength carwidth carheight
## 1 convertible        rwd           front      88.6     168.8     64.1      48.8
## 2 convertible        rwd           front      88.6     168.8     64.1      48.8
## 3   hatchback        rwd           front      94.5     171.2     65.5      52.4
## 4       sedan        fwd           front      99.8     176.6     66.2      54.3
## 5       sedan        4wd           front      99.4     176.6     66.4      54.3
## 6       sedan        fwd           front      99.8     177.3     66.3      53.1
##   curbweight enginetype cylindernumber enginesize fuelsystem boreratio stroke
## 1       2548       dohc           four        130       mpfi      3.47   2.68
## 2       2548       dohc           four        130       mpfi      3.47   2.68
## 3       2823       ohcv            six        152       mpfi      2.68   3.47
## 4       2337        ohc           four        109       mpfi      3.19   3.40
## 5       2824        ohc           five        136       mpfi      3.19   3.40
## 6       2507        ohc           five        136       mpfi      3.19   3.40
##   compressionratio horsepower peakrpm citympg highwaympg price
## 1              9.0        111    5000      21         27 13495
## 2              9.0        111    5000      21         27 16500
## 3              9.0        154    5000      19         26 16500
## 4             10.0        102    5500      24         30 13950
## 5              8.0        115    5500      18         22 17450
## 6              8.5        110    5500      19         25 15250
```

Il dataset in questione riporta varie informazioni riguardo lo stato, le caratteristiche fisiche e meccaniche e il prezzo di varie vetture. Lo scopo di questa analisi è quella di fornire un modello di regressione lineare multipla che veda il prezzo come variabile risposta: si cerca dunque di esprimere al meglio come le altre variabili influenzino il prezzo.

Si descrivono in sintesi le variabili contenute nel dataset:

- car_ID : variabile contatore per l'identificabilità delle righe all'interno del database
- symboling : variabile segnata come numerica per indicare il livello di sicurezza della vettura
- CarName : variabile qualitativa per il nome del modello di auto
- fueltype : variabile qualitativa per il tipo di combustibile
- aspiration : variabile qualitative per il tipo di aspirazione del motore
- doornumber : variabile qualitativa per il numero di portiere dell'auto
- carbody : variabile qualitativa per la struttura della vettura
- drivewheel : variabile qualitativa per il tipo di ruote motrici
- enginelocation : variabile qualitativa per la locazione del motore nell'auto
- wheelbase : variabile quantitativa per il passo (def: distanza tra gli assi)
- carlength : variabile quantitativa per la lunghezza della vettura
- carwidth : variabile quantitativa per la larghezza della vettura
- carheight : variabile quantitativa per l'altezza della vettura
- curbweight : variabiel quantitativa per il peso a vuoto della vettura
- enginetype : variabile qualitativa per il tipo del motore
- cylindernumber : variabile qualitativa per il numero dei cilindri del motore
- enginesize : variabile quantitativa per la grandezza del motore
- fuelsystem : variabile qualitativa per la modalità di alimentazione del motore
- boreratio : variabile quantitativa per il rapporto alesaggio (riguarda la dimensione del motore)
- stroke : variabile quantitativa per il movimento del pistone
- compressionratio : variabile quantitativa per il rapporto tra il volume del cilindro col pistone alla base rispetto a quello alla sommità
- horsepower : variabile quantitativa per il numero di cavalli
- peakrpm : variabile quantitativa per il numero massimo di giri del motore
- citympg : variabile quantitativa per il consumo in città (miles-per-gallon)
- highwaympg : variabile quantitativa per il consumo in autostrada (miles-per-gallon)
- price : variabile quantitativa per il prezzo

Ai fini dell'analisi, car_ID risulta essere inutile, di conseguenza viene eliminata.

```
d <- d[,-c(1)]
```

Un primo accertamento che viene fatto è la verifica dell'assenza di NA.

```
sum(is.na(d))
```

```
## [1] 0
```

Un altro passaggio preliminare consiste nel trasformare le variabili categoriali in fattori.

```
d$CarName <- as.factor(d$CarName)
d$fueltype <- as.factor(d$fueltype)
d$aspiration <- as.factor(d$aspiration)
d$doornumber <- as.factor(d$doornumber)
d$carbody <- as.factor(d$carbody)
d$drivewheel <- as.factor(d$drivewheel)
d$enginelocation <- as.factor(d$enginelocation)
d$enginetype <- as.factor(d$enginetype)
d$cylindernumber <- as.factor(d$cylindernumber)
d$fuelsystem <- as.factor(d$fuelsystem)
```

Il dataset così ottenuto è pronto per l'analisi e della forma seguente.

```
str(d)
```

```
## 'data.frame':    205 obs. of  25 variables:
##  $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ CarName         : Factor w/ 147 levels "alfa-romero giulia",..: 1 3 2 4 5 9 5 7 6 8 ...
##  $ fueltype        : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2 2 ...
##  $ aspiration      : Factor w/ 2 levels "std","turbo": 1 1 1 1 1 1 1 1 2 2 ...
##  $ doornumber      : Factor w/ 2 levels "four","two": 2 2 2 1 1 2 1 1 1 2 ...
##  $ carbody         : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4 4 5 4 3 ...
##  $ drivewheel      : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2 2 1 ...
##  $ enginelocation  : Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1 1 1 1 ...
##  $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength       : num  169 169 171 177 177 ...
##  $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
##  $ enginetype      : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4 4 4 4 4 ...
##  $ cylindernumber  : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2 2 2 2 2 ...
##  $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuelsystem      : Factor w/ 8 levels "1bbl","2bbl",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ boreratio       : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
##  $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
##  $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price           : num  13495 16500 16500 13950 17450 ...
```

Si nota come CarName abbia 147 livelli, un numero spropositato anche vista l'ampiezza del dataset. Viene dunque rimossa.

```
d <- d[,-c(2)]
```

La variabile symboling, secondo il dizionario fornito assieme ai dati, è una variabile categoriale assegnata per indicare il livello di di sicurezza dell'auto. +3 indica maggiore rischio, -3 più sicurezza. Viene dunque anch'essa convertita in fattore.

```
d$symboling <- as.factor(d$symboling)
```

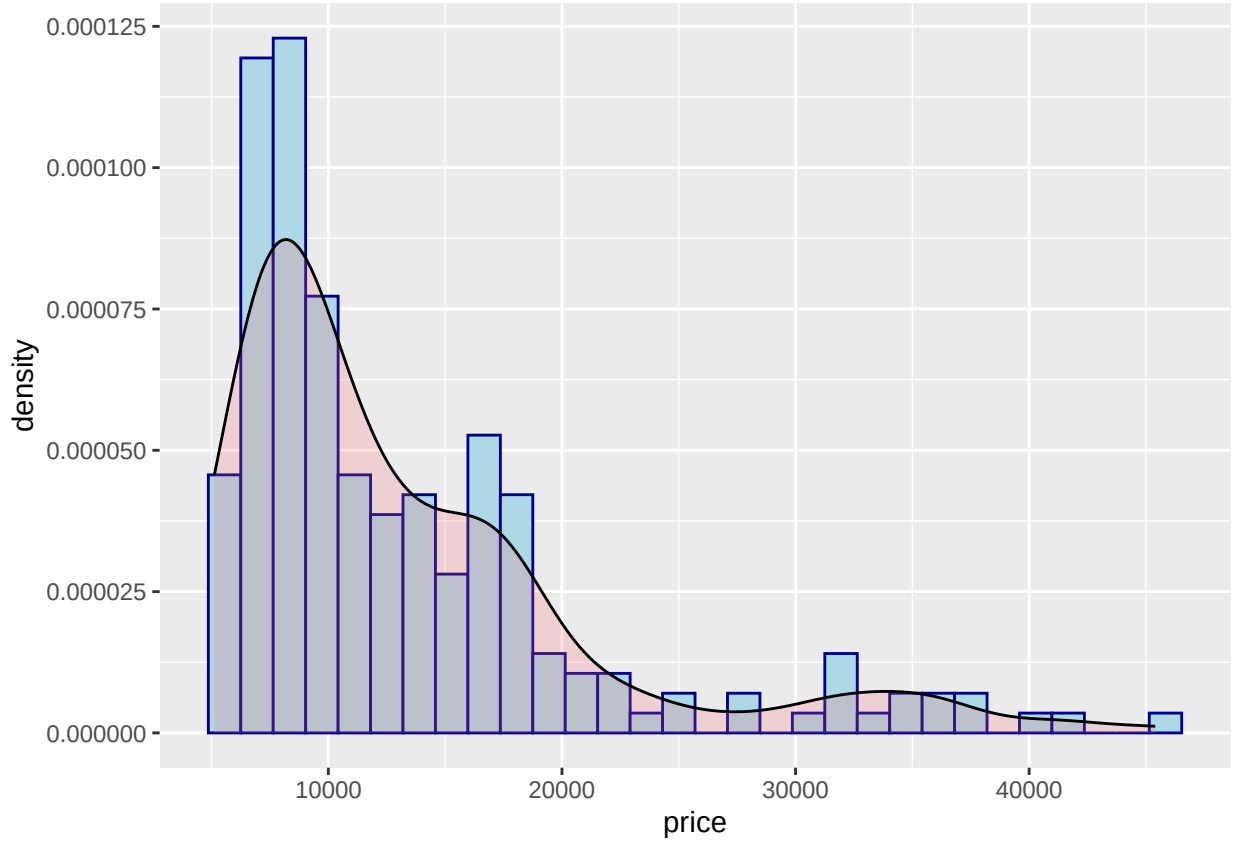
Analisi della variabile risposta

Si desidera verificare la distribuzione di price.

```
library(ggplot2)

ggplot(d, aes(x = price)) +
 geom_histogram(aes(y = after_stat(density)), colour = "darkblue", fill = "lightblue") +
 geom_density(alpha = .2, fill = "#FF6666")
```

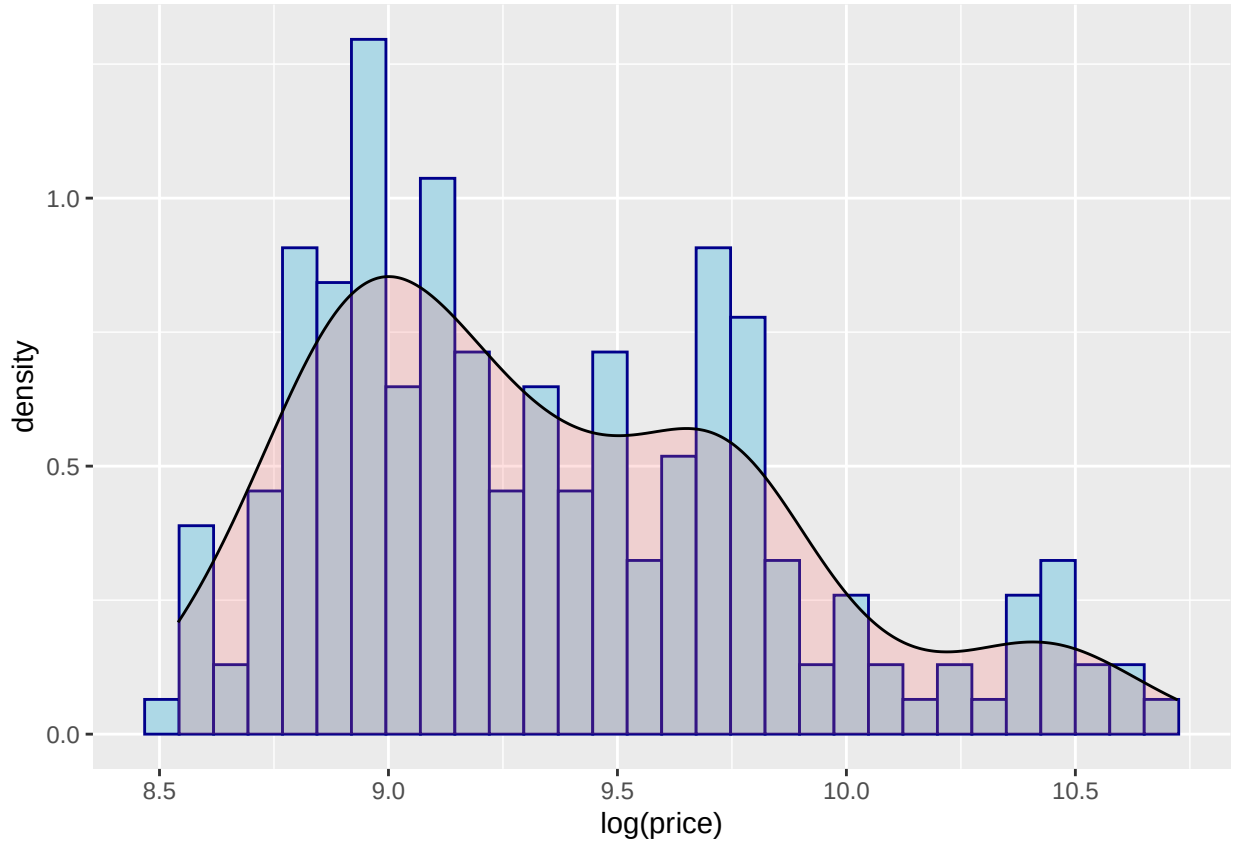
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Chiaramente non è normale. Si procede con una trasformazione logaritmica.

```
ggplot(d, aes(x = log(price))) +
 geom_histogram(aes(y = after_stat(density)), colour = "darkblue", fill = "lightblue") +
 geom_density(alpha = .2, fill = "#FF6666")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



La trasformazione logaritmica, sebbene non renda la variabile risposta normale, sembra perlomeno avere un comportamento più regolare e simmetrico. Si tenga anche presente che il dataset conta 205 osservazioni della variabile, e quindi attendersi una forma nitida è improbabile.

Altre variabili

Per prima cosa, si guarda alla matrice di correlazione per le variabili quantitative per avere un'idea di quali possano influenzare di più la variabile risposta e di quali possano causare multicollinearità.

```
corrs <- cor(d[,c(8,9,10,11,12,15,17,18,19,20,21,22,23,24)])
round(corrs, digits = 2)
```

```
##                   wheelbase carlength carwidth carheight curbweight enginesize
## wheelbase              1.00      0.87     0.80      0.59       0.78       0.57
## carlength              0.87      1.00     0.84      0.49       0.88       0.68
## carwidth               0.80      0.84     1.00      0.28       0.87       0.74
## carheight              0.59      0.49     0.28      1.00       0.30       0.07
## curbweight             0.78      0.88     0.87      0.30       1.00       0.85
## enginesize             0.57      0.68     0.74      0.07       0.85       1.00
## boreratio              0.49      0.61     0.56      0.17       0.65       0.58
## stroke                 0.16      0.13     0.18     -0.06       0.17       0.20
## compressionratio       0.25      0.16     0.18      0.26       0.15       0.03
## horsepower             0.35      0.55     0.64     -0.11       0.75       0.81
## peakrpm               -0.36     -0.29    -0.22     -0.32      -0.27      -0.24
## citympg               -0.47     -0.67    -0.64     -0.05      -0.76      -0.65
## highwaympg            -0.54     -0.70    -0.68     -0.11      -0.80      -0.68
## price                  0.58      0.68     0.76      0.12       0.84       0.87
```

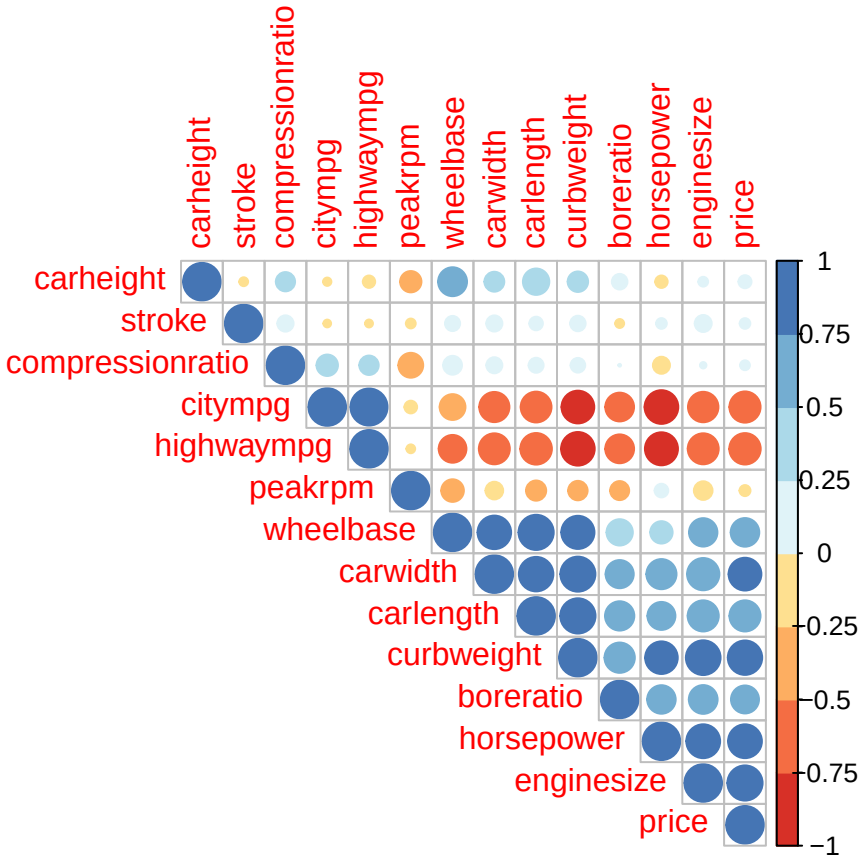
```
##                   boreratio stroke compressionratio horsepower peakrpm citympg
## wheelbase              0.49   0.16             0.25       0.35   -0.36   -0.47
## carlength              0.61   0.13             0.16       0.55   -0.29   -0.67
## carwidth               0.56   0.18             0.18       0.64   -0.22   -0.64
## carheight              0.17  -0.06             0.26      -0.11   -0.32   -0.05
## curbweight             0.65   0.17             0.15       0.75   -0.27   -0.76
## enginesize             0.58   0.20             0.03       0.81   -0.24   -0.65
## boreratio              1.00  -0.06             0.01       0.57   -0.25   -0.58
## stroke                -0.06   1.00             0.19       0.08   -0.07   -0.04
## compressionratio       0.01   0.19             1.00      -0.20   -0.44    0.32
## horsepower             0.57   0.08            -0.20       1.00    0.13   -0.80
## peakrpm               -0.25  -0.07            -0.44       0.13    1.00   -0.11
## citympg               -0.58  -0.04             0.32      -0.80   -0.11    1.00
## highwaympg            -0.59  -0.04             0.27      -0.77   -0.05    0.97
## price                  0.55   0.08             0.07       0.81   -0.09   -0.69
##                   highwaympg price
## wheelbase              -0.54  0.58
## carlength              -0.70  0.68
## carwidth               -0.68  0.76
## carheight              -0.11  0.12
## curbweight             -0.80  0.84
## enginesize             -0.68  0.87
## boreratio              -0.59  0.55
## stroke                 -0.04  0.08
## compressionratio        0.27  0.07
## horsepower             -0.77  0.81
## peakrpm                -0.05 -0.09
## citympg                 0.97 -0.69
## highwaympg              1.00 -0.70
## price                  -0.70  1.00
```

Si da un'alternativa grafica di tale matrice.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(RColorBrewer)
corrplot(corrs, type="upper", order="hclust", col=brewer.pal(n = 8, name = "RdYlBu"))
```



Da questa analisi si evince come le variabili quantitative maggiormente correlate con price sono: enginesize, curbweight, horsepower, carwidth, carlength, wheelbase, boreratio (positivamente); citympg, highwaympg (negativamente).

Si svolge un'attenta valutazione della multicollinearità.

```
fitC <- lm(price ~ citympg, data=d)
summary(fitC)
```

```
##
## Call:
## lm(formula = price ~ citympg, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9215  -3315  -1770   1974  22728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34395.44    1625.03   21.17   <2e-16 ***
## citympg      -837.40      62.38  -13.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5829 on 203 degrees of freedom
## Multiple R-squared:  0.4703, Adjusted R-squared:  0.4676
## F-statistic: 180.2 on 1 and 203 DF,  p-value: < 2.2e-16
```

```r
fitH <- lm(price ~ highwaympg, data=d)
summary(fitH)
```

```
##
## Call:
## lm(formula = price ~ highwaympg, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8698  -3519  -1302   1114  20956
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38162.86    1838.18   20.76   <2e-16 ***
## highwaympg   -809.27      58.34  -13.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5738 on 203 degrees of freedom
## Multiple R-squared:  0.4866, Adjusted R-squared:  0.4841
## F-statistic: 192.4 on 1 and 203 DF,  p-value: < 2.2e-16
```

```r
fitCH <- lm(price ~ citympg + highwaympg, data=d)
summary(fitCH)
```

```
##
## Call:
## lm(formula = price ~ citympg + highwaympg, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8792  -3558  -1368   1175  20755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37607.0     2013.6  18.676  < 2e-16 ***
## citympg       -176.1      258.7  -0.681  0.49685
## highwaympg    -646.8      245.7  -2.632  0.00914 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5746 on 202 degrees of freedom
## Multiple R-squared:  0.4878, Adjusted R-squared:  0.4827
## F-statistic:  96.2 on 2 and 202 DF,  p-value: < 2.2e-16
```

Si nota come l'inserimento esclusivo di una delle due variabili porti a un coefficiente con p-value < 0.05 e std. error nell'ordine di 10^-3. Tuttavia, l'inserimento di entrambe le variabili porta entrambi i coefficienti ad avere uno std. error di un ordine di grandezza superiorie e la significatività dei coefficienti è molto minore, essendo il p-value di d$citympg addirittura > 0.05, e l'altro anch'esso poco significativo. Questi sono i sintomi della multicollinearità per questa coppia di variabili: di conseguenza, si sceglie di escludere citympg in quanto leggermente meno correlata (anche se negativamente) con la variabile risposta e avendo essa coefficiente meno significativo nel modello con entrambe le esplicative.

```r
d <- d[,-c(22)]
```

Le altre combinazioni di variabili che possono portare a multicollinearità sulla base del coefficiente di

correlazione lineare sono:

- wheelbase, carlength, carwidth e curbweight tra di loro;
- enginesize, horsepower e curbweight tra di loro;
- citympg, highwaympg e horsepower tra di loro.

La presenza ridondante di curbweight e la sua alta correlazione con price potrebbe spiegare di per sè gran parte della variabilità senza però darci tante informazioni e rischiando multicollinearità con le altre variabili. Di conseguenza, si decide di tralasciarla. (in Italiano, il peso a vuoto della macchina).

La presenza ridondante di curbweight e la sua alta correlazione con price potrebbe spiegare di per sè gran parte della variabilità senza però darci tante informazioni e rischiando multicollinearità con le altre variabili. Di conseguenza, si decide di tralasciarla. (in Italiano, il peso a vuoto della macchina)

```
d <- d[,-c(12)]
```

Durante la costruzione dei modelli, si controllerà se verranno incluse più variabili che possano portare multicollinearità. In tal caso, si escluderà quella meno significativa.

Per quanto riguarda le variabili categoriali, si esegue un ANOVA test.

```
# definisco una funzione che, dato in input la variabile fattore, ritorna il p-value dell'ANOVA test
anova.p.calc <- function(v) {
  anova.result <- aov(d$price ~ v)
  a.summary <- summary(anova.result)
  p.value <- a.summary[[1]]$`Pr(>F)`[1]
  return (p.value)
}

# applico la funzione a ogni colonna del dataset corrispondente a fattori e stampo l'indice solo se il
fac.vars <- c(1,2,3,4,5,6,7,12,13,15)
sig.cat <- numeric()
not.sig.cat <- numeric()
for (i in fac.vars) {
  curr.p.val <- anova.p.calc(d[,c(i)])
  if (curr.p.val < 0.05) {
    sig.cat <- c(sig.cat,i)
  }
  else {
    not.sig.cat <- c(not.sig.cat,i)
  }
}
sig.cat
```

```
## [1]  1  3  5  6  7 12 13 15
```

```
not.sig.cat
```

```
## [1] 2 4
```

Elimino dal dataset le variabili categoriali non significative, onde evitare rumore nella costruzione del modello.

```
d <- d[,-c(not.sig.cat)]
```

Si propone un boxplot delle categoriali che influenzano significativamente il modello.

Si nota come nei boxplot delle ultime tre variabili risultino presenti dei livelli con un numero molto basso di osservazioni: questo nel modello potrebbe portare risultati inaffidabili. Si osservano le tabelle relative e si decide cosa farne.

```
table(d$enginetype)
```

```
##
##  dohc dohcv     l   ohc  ohcf  ohcv rotor
##    12     1    12   148    15    13     4
```

```
table(d$cylindernumber)
```

```
##
##  eight   five   four    six  three twelve    two
##      5     11    159     24      1      1      4
```

```
table(d$fuelsystem)
```

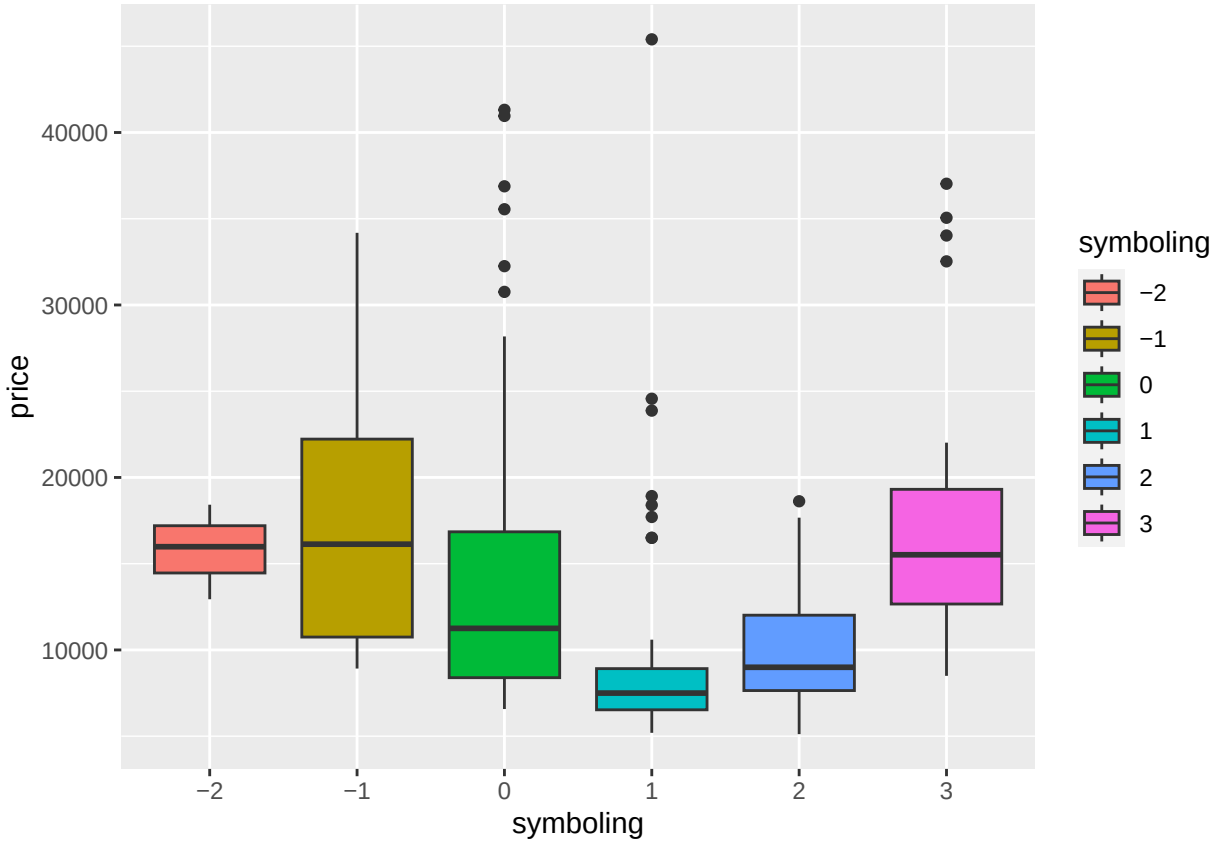
```
##
## 1bbl 2bbl 4bbl  idi  mfi mpfi spdi spfi
##   11   66    3   20    1   94    9    1
```

Si procede eliminando le righe dei livelli contenenti una sola osservazione. Si elimina anche "two" per cylindernumber, altrimenti i modelli lineari computano un NA come coefficiente per quel livello.

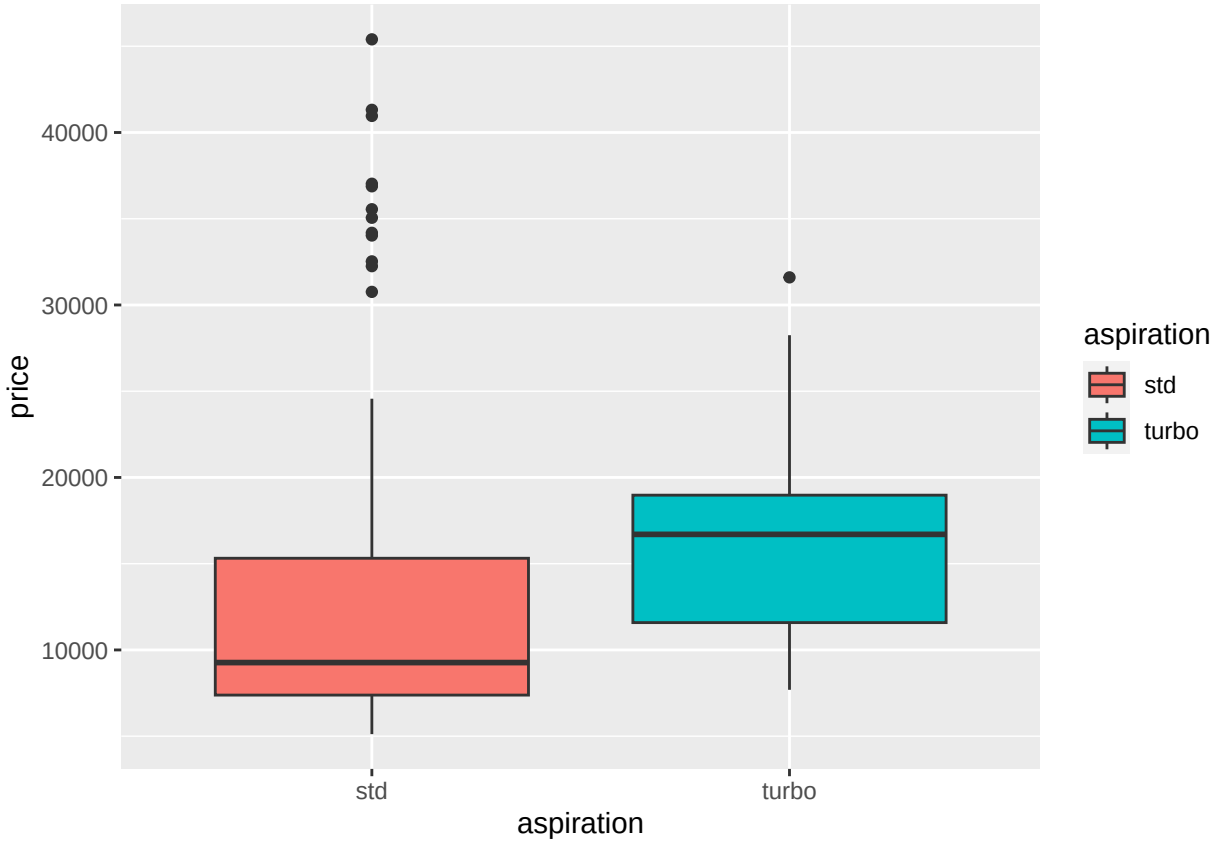
```
d <- d[d$enginetype != "dohcv", ]
d <- d[d$cylindernumber != "three", ]
d <- d[d$cylindernumber != "twelve", ]
d <- d[d$cylindernumber != "two", ]
d <- d[d$fuelsystem != "mfi", ]
d <- d[d$fuelsystem != "spfi", ]
d$enginetype <- droplevels(d$enginetype)
d$cylindernumber <- droplevels(d$cylindernumber)
d$fuelsystem <- droplevels(d$fuelsystem)
```

Vediamo come sono ora i boxplot.

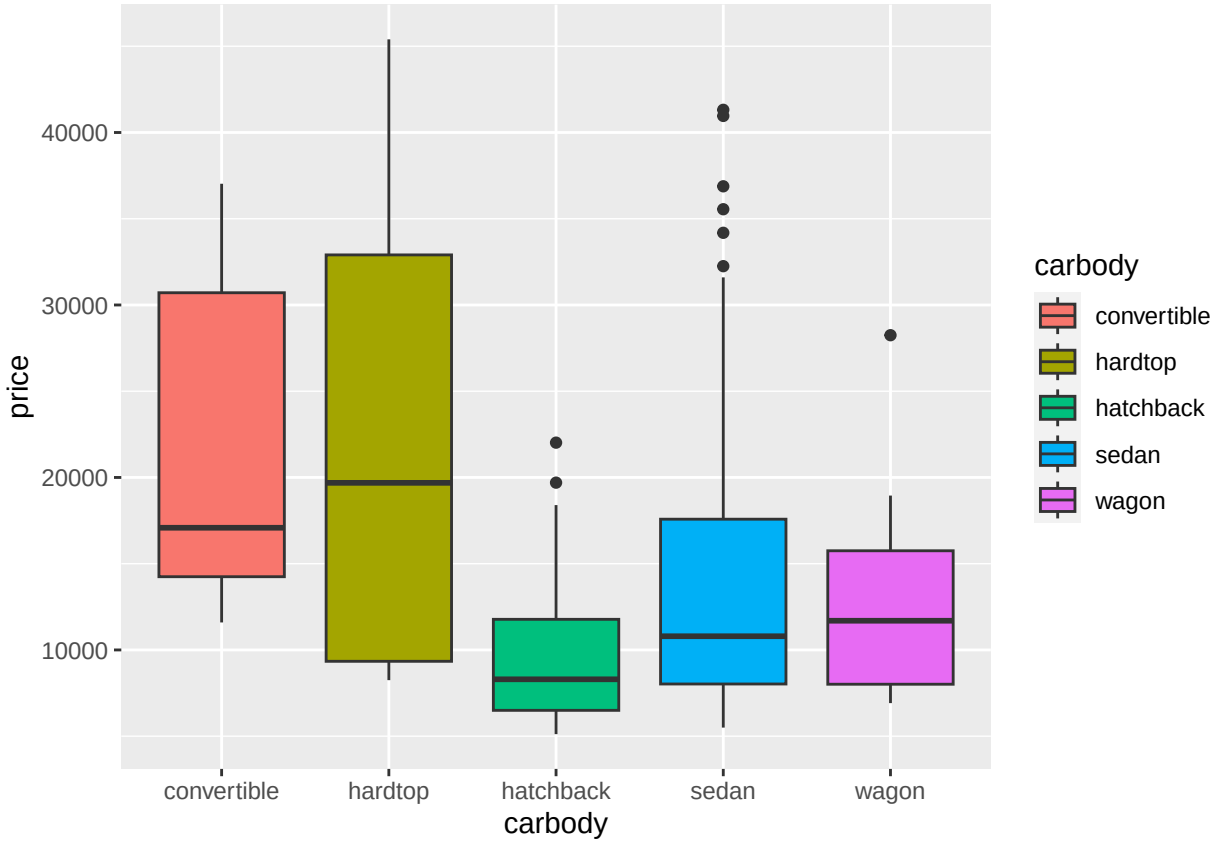
```
par(mfrow=c(3,3))
ggplot(d, aes(x = symboling, y = price, fill = symboling)) +
       geom_boxplot( )
```



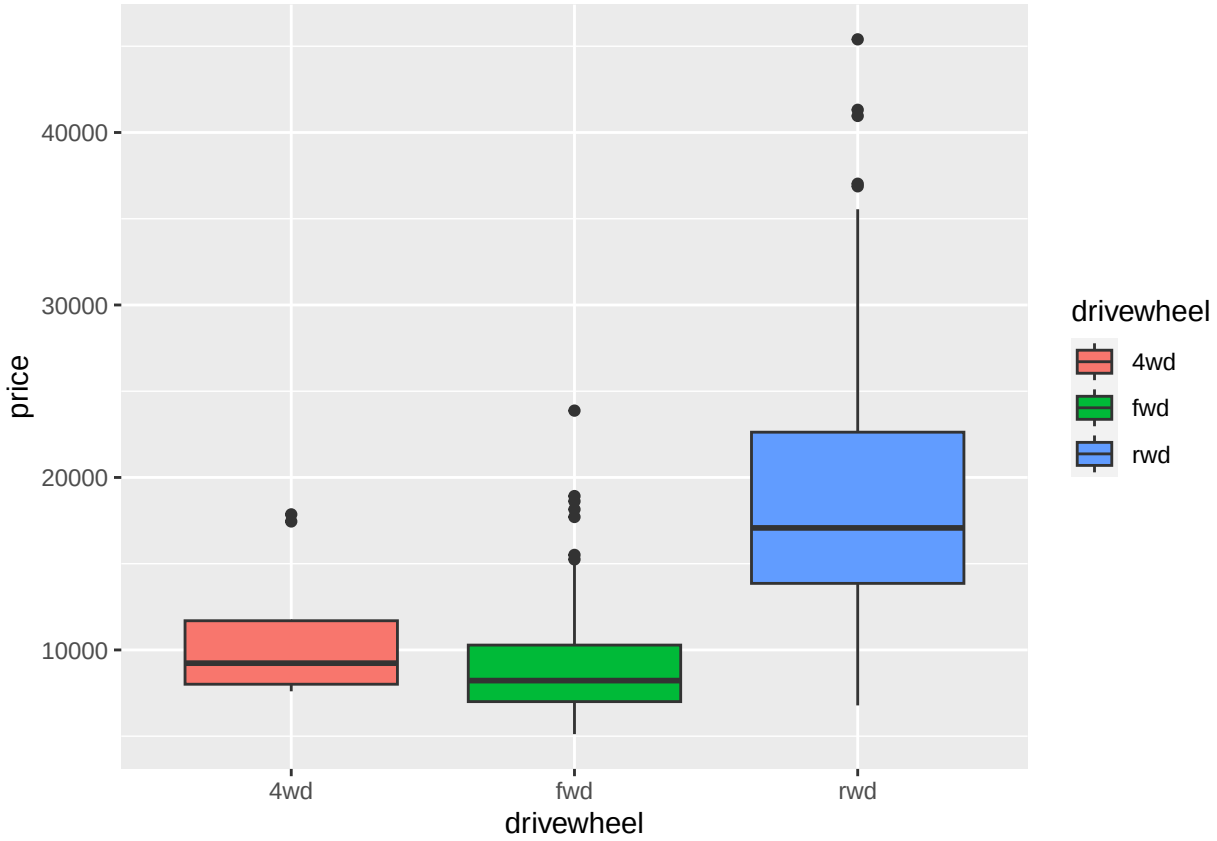

```
ggplot(d, aes(x = aspiration, y = price, fill = aspiration)) +
       geom_boxplot( )
```



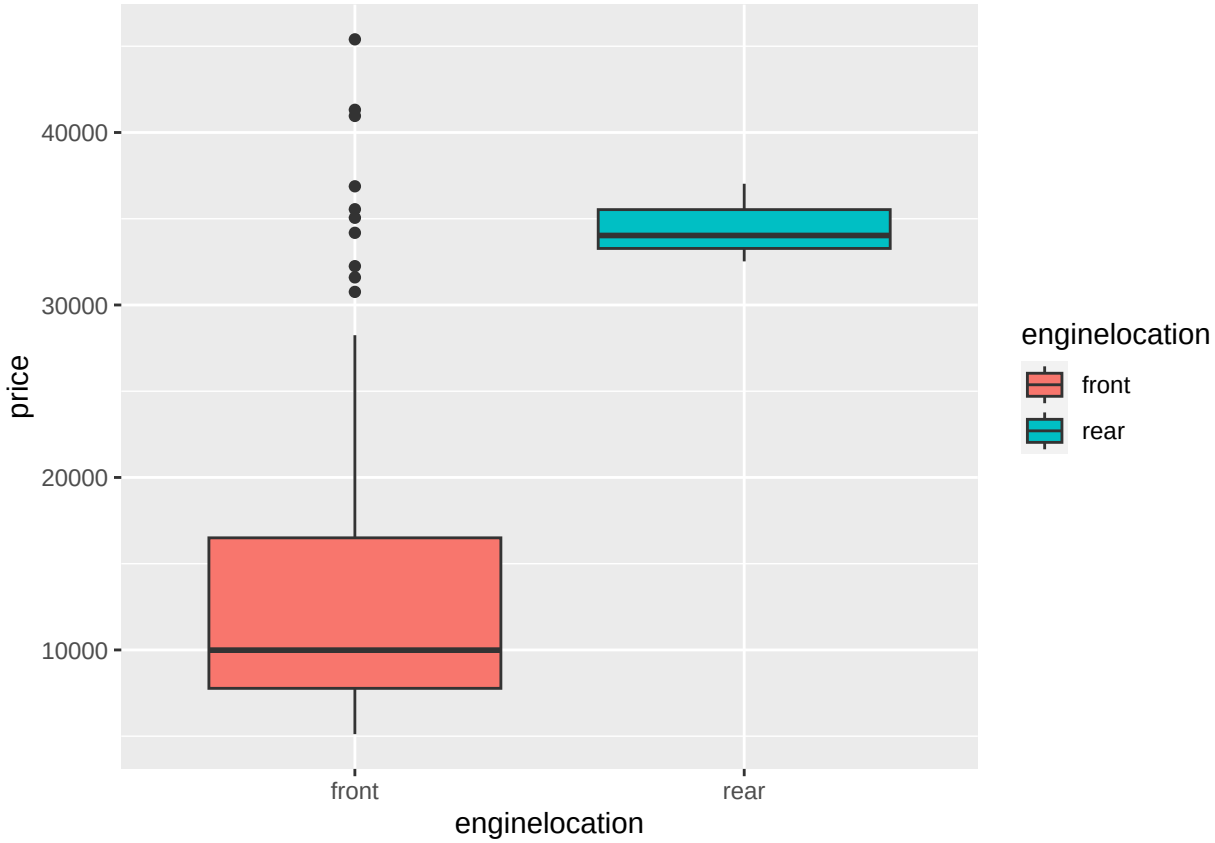

```
ggplot(d, aes(x = carbody, y = price, fill = carbody)) +
       geom_boxplot( )
```



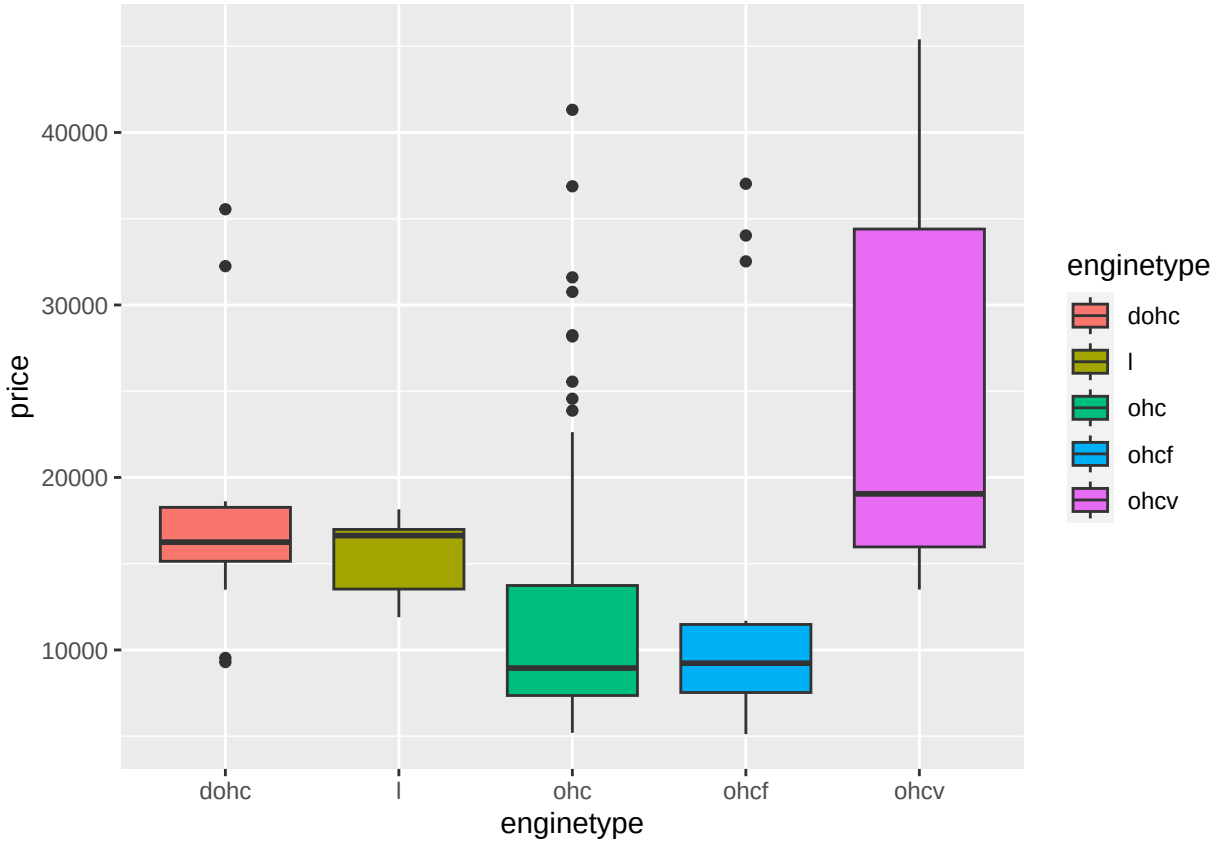

```
ggplot(d, aes(x = drivewheel, y = price, fill = drivewheel)) +
       geom_boxplot( )
```



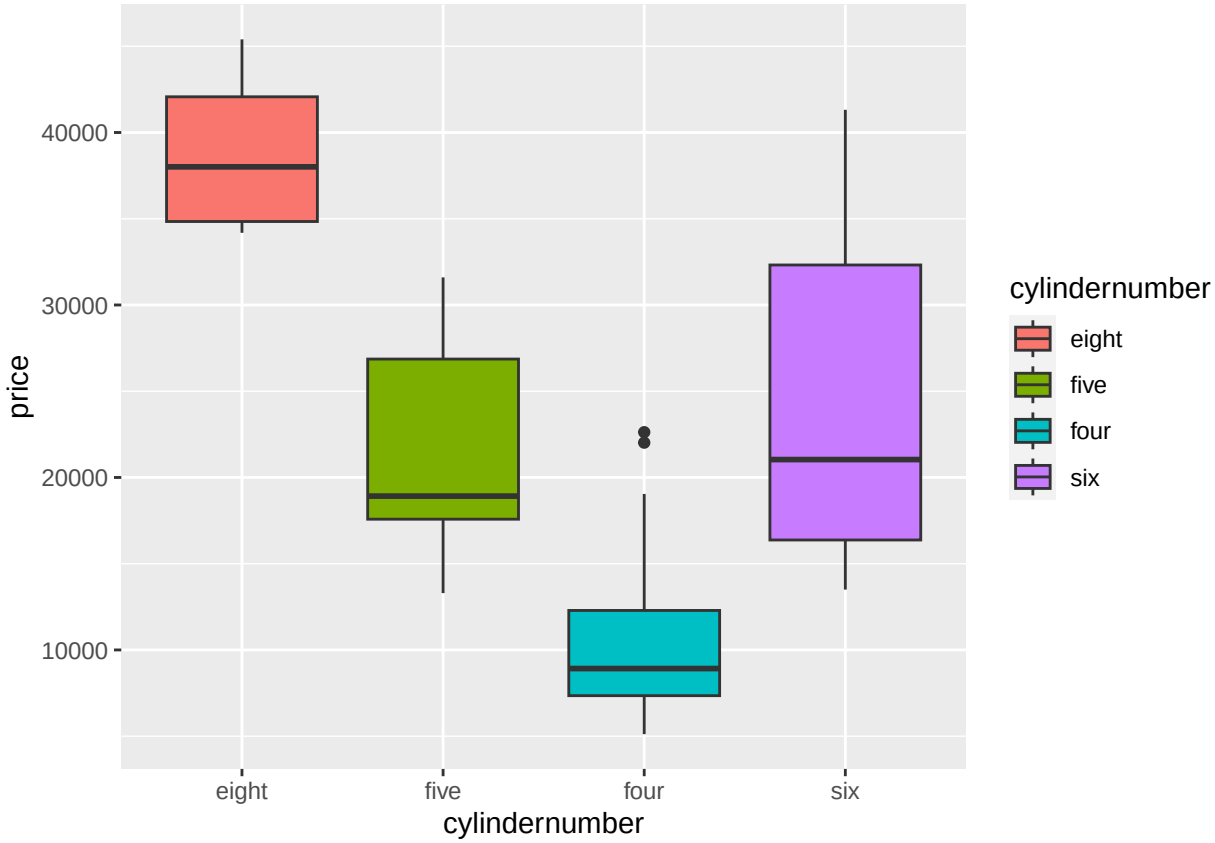

```
ggplot(d, aes(x = enginelocation, y = price, fill = enginelocation)) +
       geom_boxplot( )
```



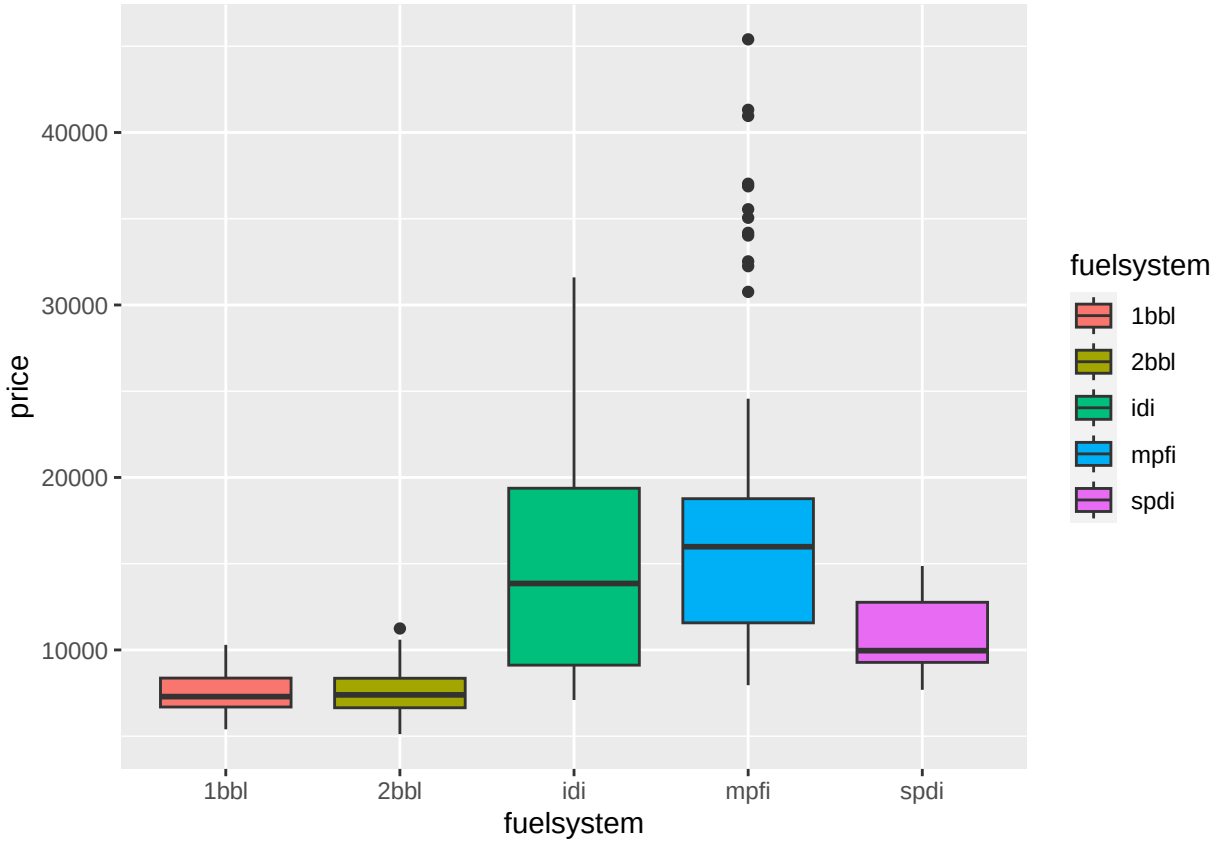

```
ggplot(d, aes(x = enginetype, y = price, fill = enginetype)) +
       geom_boxplot( )
```




```
ggplot(d, aes(x = cylindernumber, y = price, fill = cylindernumber)) +
       geom_boxplot( )
```




```
ggplot(d, aes(x = fuelsystem, y = price, fill = fuelsystem)) +
       geom_boxplot( )
```



Alcuni livelli potrebbero essere assimilabili a causa della loro media che sembra statisticamente uguale. Si fa un ANOVA per verificarlo. Questo potrebbe semplificare decisamente la formulazione del modello e ridurre il numero di coefficienti significativi.

```
d$symboling2 <- d$symboling
levels(d$symboling2) <- c("<0", "<0", "0", "<3", "<3", "3")
fit.symboling <- lm(price ~ symboling, data = d)
fit.symboling2 <- lm(price ~ symboling2, data = d)
anova(fit.symboling2,fit.symboling)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ symboling2
## Model 2: price ~ symboling
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    192 1.0178e+10
## 2    190 1.0164e+10  2  13426009 0.1255 0.8821
```

Il test suggerisce che non c'è evidenza per concludere che le medie dei gruppi siano diverse.

```
d$carbody2 <- d$carbody
levels(d$carbody2) <- c("ch", "ch", "h", "sw", "sw")
fit.carbody <- lm(price ~ carbody, data=d)
fit.carbody2 <- lm(price ~ carbody2, data=d)
anova(fit.carbody2,fit.carbody)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: price ~ carbody2
## Model 2: price ~ carbody
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    193 1.0273e+10
## 2    191 1.0212e+10  2  60568341 0.5664 0.5685
```

Anche qui il test porta a considerare la nuova variabile.

```
d$drivewheel2 <- d$drivewheel
levels(d$drivewheel2) <- c("4f", "4f", "r")
fit.drivewheel <- lm(price ~ drivewheel, data=d)
fit.drivewheel2 <- lm(price ~ drivewheel2, data=d)
anova(fit.drivewheel2,fit.drivewheel)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ drivewheel2
## Model 2: price ~ drivewheel
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    194 6999993345
## 2    193 6971525933  1  28467412 0.7881 0.3758
```

```
d$enginetype2 <- d$enginetype
levels(d$enginetype2) <- c("dl", "dl", "ohc", "ohc", "ohcv")
fit.enginetype <- lm(price ~ enginetype, data=d)
fit.enginetype2 <- lm(price ~ enginetype2, data=d)
anova(fit.enginetype2, fit.enginetype)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ enginetype2
## Model 2: price ~ enginetype
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    193 1.0001e+10
## 2    191 9.8970e+09  2 103696716 1.0006 0.3696
```

```
d$cylindernumber2 <- d$cylindernumber
levels(d$cylindernumber2) <- c("eight", "fivesix", "four", "fivesix")
fit.cylindernumber <- lm(price ~ cylindernumber, data=d)
fit.cylindernumber2 <- lm(price ~ cylindernumber2, data=d)
anova(fit.cylindernumber2,fit.cylindernumber)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ cylindernumber2
## Model 2: price ~ cylindernumber
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    193 4709523584
## 2    192 4678091251  1  31432333 1.2901 0.2575
```

```
d$fuelsystem2 <- d$fuelsystem
levels(d$fuelsystem2) <- c("12", "12", "im", "im", "s")
fit.fuelsystem <- lm(price ~ fuelsystem, data=d)
fit.fuelsystem2 <- lm(price ~ fuelsystem2, data=d)
anova(fit.fuelsystem2,fit.fuelsystem)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: price ~ fuelsystem2
## Model 2: price ~ fuelsystem
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    193 7867285019
## 2    191 7825859680  2  41425338 0.5055  0.604
```

Tutti i test ci hanno portato ad assumere un numero minore di livelli (che diventano dunque anche più popolati e quindi fornenti risultati sperabilmente più stabili).

L'ultimo passo preparatorio è dunque rimuovere le variabili originali.

```
d$symboling <- d$symboling2
d$carbody <- d$carbody2
d$drivewheel <- d$drivewheel2
d$enginetype <- d$enginetype2
d$cylindernumber <- d$cylindernumber2
d$ fuelsystem <- d$fuelsystem2
d <- d[,-c(21,22,23,24,25,26)]
```

Formulazione del modello

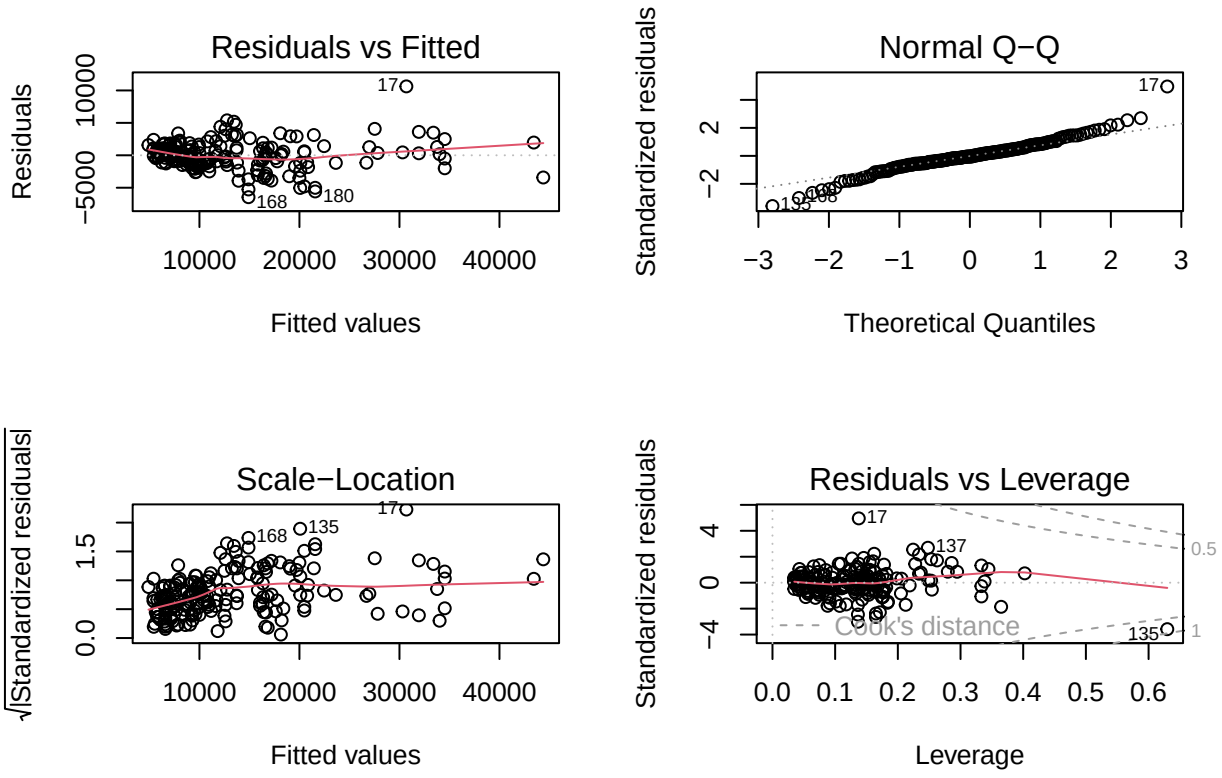
Si definisce il modello con tutte le variabili come esplicative, si userà per lo scope nella stepwise regression tramite metrica AIC.

```
fitAll <- lm(price ~ ., data=d)
summary(fitAll)
```

```
##
## Call:
## lm(formula = price ~ ., data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6455.5 -1136.1  -102.2  1094.9 10623.2
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -4.570e+04  1.311e+04  -3.487 0.000622 ***
## symboling0             1.317e+03  6.324e+02   2.083 0.038793 *
## symboling<3            8.200e+02  6.980e+02   1.175 0.241691
## symboling3             1.150e+03  1.087e+03   1.058 0.291670
## aspirationturbo        2.006e+03  8.288e+02   2.420 0.016565 *
## carbodyh              -9.896e+02  8.986e+02  -1.101 0.272319
## carbodysw              2.322e+01  9.316e+02   0.025 0.980140
## drivewheelr            2.121e+03  6.019e+02   3.524 0.000547 ***
## enginelocationrear     5.212e+03  2.045e+03   2.548 0.011708 *
## wheelbase              1.880e+01  9.792e+01   0.192 0.847931
## carlength             -3.082e+01  4.495e+01  -0.686 0.493957
## carwidth               7.933e+02  2.206e+02   3.595 0.000424 ***
## carheight              1.317e+02  1.183e+02   1.113 0.267170
## enginetypeohc          3.012e+03  6.833e+02   4.408 1.84e-05 ***
## enginetypeohcv        -5.022e+03  1.208e+03  -4.156 5.12e-05 ***
## cylindernumberfivesix -5.794e+03  2.146e+03  -2.700 0.007629 **
## cylindernumberfour    -6.732e+03  2.780e+03  -2.421 0.016515 *
## enginesize             1.504e+02  2.318e+01   6.488 9.12e-10 ***
```

```
## fuelsystemim           2.560e+02  6.035e+02   0.424 0.672024
## fuelsystems           -1.747e+03  1.121e+03  -1.558 0.121013
## boreratio             -4.486e+03  1.353e+03  -3.315 0.001122 **
## stroke                -3.768e+03  7.944e+02  -4.743 4.43e-06 ***
## compressionratio       8.162e+01  8.836e+01   0.924 0.356939
## horsepower             3.185e+01  2.115e+01   1.505 0.134052
## peakrpm                1.700e+00  5.638e-01   3.015 0.002966 **
## highwaympg             2.248e+01  6.647e+01   0.338 0.735656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2308 on 170 degrees of freedom
## Multiple R-squared:  0.9251, Adjusted R-squared:  0.914
## F-statistic: 83.94 on 25 and 170 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitAll)
```



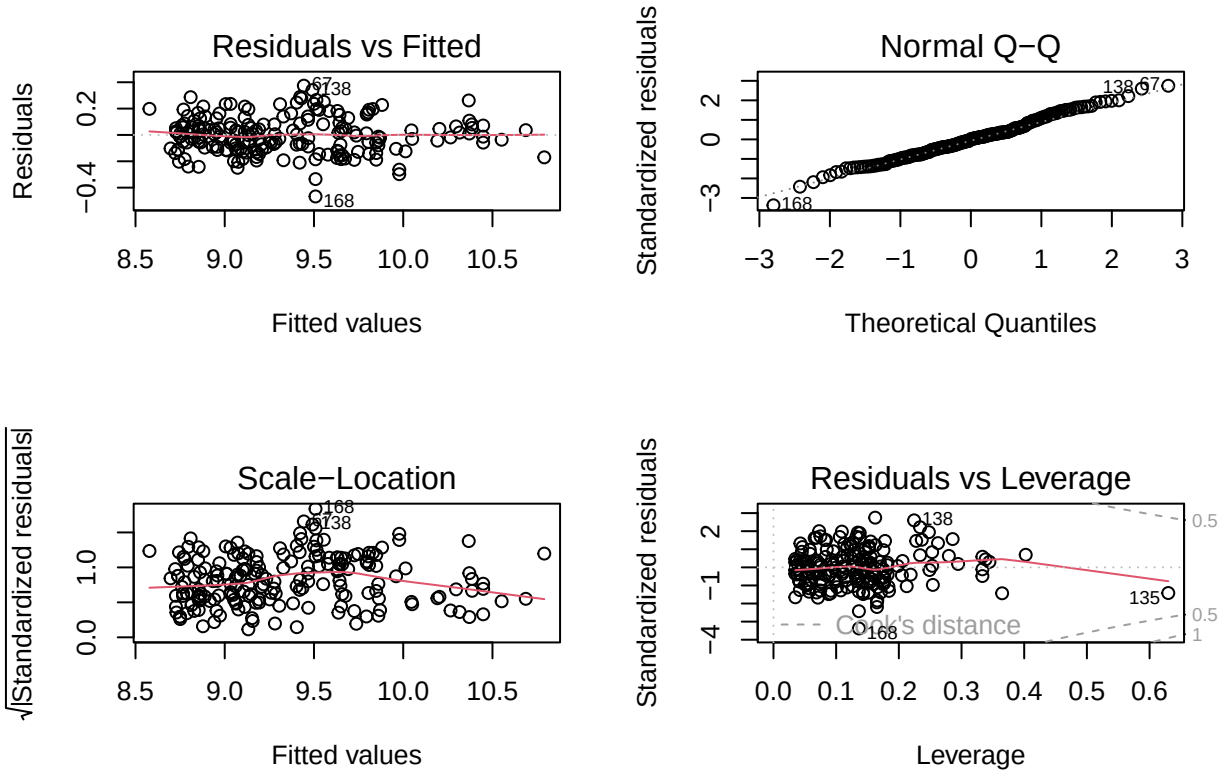
Risulta evidente nel grafico dei residui come siano violate le ipotesi di normalità, omoschedasticità e linearità. Si vede come sia necessaria una trasformazione.

```
fitAllLog <- lm(log(price) ~ ., data=d)
summary(fitAllLog)
```

```
##
## Call:
## lm(formula = log(price) ~ ., data = d)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46685 -0.10239  0.00059  0.07991  0.37399
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.080e+00  8.446e-01   6.015 1.07e-08 ***
## symboling0              8.517e-02  4.075e-02   2.090 0.038095 *
## symboling<3             1.426e-02  4.498e-02   0.317 0.751622
## symboling3              1.578e-01  7.007e-02   2.252 0.025590 *
## aspirationturbo         8.031e-02  5.341e-02   1.504 0.134492
## carbodyh               -8.890e-02  5.790e-02  -1.535 0.126580
## carbodysw              -9.181e-03  6.003e-02  -0.153 0.878629
## drivewheelr             1.626e-01  3.878e-02   4.192 4.44e-05 ***
## enginelocationrear      1.571e-01  1.318e-01   1.192 0.234839
## wheelbase               1.227e-03  6.310e-03   0.194 0.846019
## carlength               2.580e-03  2.897e-03   0.891 0.374332
## carwidth                4.889e-02  1.422e-02   3.438 0.000736 ***
## carheight               5.039e-03  7.624e-03   0.661 0.509550
## enginetypeohc           1.690e-01  4.403e-02   3.838 0.000175 ***
## enginetypeohcv         -1.374e-01  7.786e-02  -1.765 0.079353 .
## cylindernumberfivesix -1.124e-01  1.383e-01  -0.813 0.417594
## cylindernumberfour    -2.070e-01  1.792e-01  -1.155 0.249613
## enginesize              4.014e-03  1.494e-03   2.687 0.007932 **
## fuelsystemim            1.232e-01  3.889e-02   3.167 0.001826 **
## fuelsystems            -2.316e-02  7.225e-02  -0.321 0.748907
## boreratio              -1.649e-01  8.721e-02  -1.891 0.060353 .
## stroke                 -1.144e-01  5.119e-02  -2.236 0.026680 *
## compressionratio        5.554e-03  5.694e-03   0.976 0.330676
## horsepower              3.140e-03  1.363e-03   2.304 0.022454 *
## peakrpm                 5.077e-05  3.633e-05   1.397 0.164120
## highwaympg             -4.199e-03  4.283e-03  -0.980 0.328243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1487 on 170 degrees of freedom
## Multiple R-squared:  0.9227, Adjusted R-squared:  0.9114
## F-statistic: 81.21 on 25 and 170 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitAllLog)
```



I residui del modello con le stesse esplicative per la trasformata logaritmica del prezzo risultano invece fedeli alle ipotesi del modello.

Si cerca con StepAIC un modello valido con meno esplicative.

```
require(MASS)
```

```
## Caricamento del pacchetto richiesto: MASS
```

```
fitEmpty <- lm(price ~ 1, data=d)
fitAIC <- stepAIC(fitEmpty, scope=formula(fitAll), direction="both")
```

```
## Start:  AIC=3517.69
## price ~ 1
##
##                    Df  Sum of Sq        RSS    AIC
## + enginesize        1 9572566027 2.5136e+09 3211.9
## + horsepower        1 7824378978 4.2618e+09 3315.4
## + cylindernumber    2 7376636697 4.7095e+09 3337.0
## + carwidth          1 6739435069 5.3467e+09 3359.8
## + highwaympg        1 6125442339 5.9607e+09 3381.1
## + carlength         1 5746787920 6.3394e+09 3393.2
## + drivewheel        1 5086166936 7.0000e+09 3412.6
## + wheelbase         1 4213513782 7.8726e+09 3435.7
## + fuelsystem        2 4218875263 7.8673e+09 3437.5
## + boreratio         1 3547402359 8.5388e+09 3451.6
## + enginetype        2 2085467438 1.0001e+10 3484.6
## + carbody           2 1813388415 1.0273e+10 3489.8
```

```
## + symboling         3 1908645539 1.0178e+10 3490.0
## + enginelocation    1 1395239528 1.0691e+10 3495.7
## + aspiration        1  469486100 1.1617e+10 3511.9
## + carheight         1  400819172 1.1685e+10 3513.1
## + stroke            1  144479491 1.1942e+10 3517.3
## + peakrpm           1  131093315 1.1955e+10 3517.6
## <none>                           1.2086e+10 3517.7
## + compressionratio  1   50884197 1.2035e+10 3518.9
##
## Step:  AIC=3211.91
## price ~ enginesize
##
##                    Df  Sum of Sq        RSS    AIC
## + enginelocation    1  271987309 2.2416e+09 3191.5
## + carwidth          1  218415594 2.2952e+09 3196.1
## + enginetype        2  240176465 2.2734e+09 3196.2
## + horsepower        1  211829996 2.3018e+09 3196.7
## + stroke            1  193313514 2.3203e+09 3198.2
## + fuelsystem        2  203402448 2.3102e+09 3199.4
## + drivewheel        1  149220971 2.3644e+09 3201.9
## + carbody           2  160491807 2.3531e+09 3203.0
## + peakrpm           1  136099743 2.3775e+09 3203.0
## + aspiration        1  116098778 2.3975e+09 3204.6
## + carlength         1  110327180 2.4033e+09 3205.1
## + cylindernumber    2  124721873 2.3889e+09 3205.9
## + highwaympg        1   99920848 2.4137e+09 3206.0
## + carheight         1   84609253 2.4290e+09 3207.2
## + wheelbase         1   74996350 2.4386e+09 3208.0
## + compressionratio  1   27211735 2.4864e+09 3211.8
## <none>                           2.5136e+09 3211.9
## + boreratio         1    1639684 2.5120e+09 3213.8
## + symboling         3   23605467 2.4900e+09 3216.1
## - enginesize        1 9572566027 1.2086e+10 3517.7
##
## Step:  AIC=3191.46
## price ~ enginesize + enginelocation
##
##                    Df  Sum of Sq        RSS    AIC
## + carwidth          1  451683863 1.7899e+09 3149.4
## + wheelbase         1  284979863 1.9566e+09 3166.8
## + carlength         1  258264517 1.9833e+09 3169.5
## + fuelsystem        2  204465926 2.0371e+09 3176.7
## + aspiration        1  149808295 2.0918e+09 3179.9
## + carheight         1  139427685 2.1022e+09 3180.9
## + enginetype        2  155184861 2.0864e+09 3181.4
## + drivewheel        1  132991318 2.1086e+09 3181.5
## + highwaympg        1  125525419 2.1161e+09 3182.2
## + stroke            1  113971919 2.1276e+09 3183.2
## + carbody           2  130875531 2.1107e+09 3183.7
## + horsepower        1   93364410 2.1482e+09 3185.1
## + cylindernumber    2   94294826 2.1473e+09 3187.0
## + peakrpm           1   54713231 2.1869e+09 3188.6
## + compressionratio  1   31836948 2.2098e+09 3190.7
## <none>                           2.2416e+09 3191.5
```

```
## + boreratio         1        148 2.2416e+09 3193.5
## + symboling         3   45037334 2.1966e+09 3193.5
## - enginelocation    1  271987309 2.5136e+09 3211.9
## - enginesize        1 8449313808 1.0691e+10 3495.7
##
## Step:  AIC=3149.36
## price ~ enginesize + enginelocation + carwidth
##
##                    Df  Sum of Sq        RSS    AIC
## + stroke            1   95519645 1694403437 3140.6
## + drivewheel        1   83188903 1706734180 3142.0
## + enginetype        2   92847856 1697075226 3142.9
## + peakrpm           1   69752611 1720170472 3143.6
## + cylindernumber    2   76300732 1713622351 3144.8
## + horsepower        1   51431178 1738491904 3145.6
## + carbody           2   63683356 1726239727 3146.3
## + fuelsystem        2   47566597 1742356485 3148.1
## + highwaympg        1   25841297 1764081786 3148.5
## + aspiration        1   25682995 1764240087 3148.5
## + boreratio         1   22116504 1767806579 3148.9
## <none>                           1789923082 3149.4
## + carheight         1   15265488 1774657594 3149.7
## + wheelbase         1   14190387 1775732695 3149.8
## + carlength         1    6237300 1783685782 3150.7
## + compressionratio  1      16261 1789906822 3151.4
## + symboling         3    5991655 1783931427 3154.7
## - carwidth          1  451683863 2241606945 3191.5
## - enginelocation    1  505255578 2295178660 3196.1
## - enginesize        1 1811070873 3600993956 3284.4
##
## Step:  AIC=3140.61
## price ~ enginesize + enginelocation + carwidth + stroke
##
##                    Df  Sum of Sq        RSS    AIC
## + enginetype        2  171996617 1522406820 3123.6
## + peakrpm           1   78568401 1615835035 3133.3
## + boreratio         1   58109198 1636294239 3135.8
## + drivewheel        1   51325877 1643077560 3136.6
## + aspiration        1   50381890 1644021547 3136.7
## + horsepower        1   39017758 1655385679 3138.0
## + cylindernumber    2   54291415 1640112022 3138.2
## + carbody           2   50295339 1644108098 3138.7
## <none>                           1694403437 3140.6
## + fuelsystem        2   27717034 1666686403 3141.4
## + wheelbase         1    9994066 1684409371 3141.4
## + highwaympg        1    7773250 1686630187 3141.7
## + carheight         1    6958551 1687444886 3141.8
## + compressionratio  1    5453452 1688949985 3142.0
## + carlength         1    1272326 1693131111 3142.5
## + symboling         3    6718553 1687684884 3145.8
## - stroke            1   95519645 1789923082 3149.4
## - enginelocation    1  395496573 2089900010 3179.7
## - carwidth          1  433231589 2127635026 3183.2
## - enginesize        1 1906535682 3600939119 3286.4
```

```
##
## Step:  AIC=3123.63
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype
##
##                    Df  Sum of Sq        RSS    AIC
## + boreratio         1  162351579 1360055241 3103.5
## + cylindernumber    2  158480548 1363926272 3106.1
## + peakrpm           1  130042203 1392364617 3108.1
## + horsepower        1   55088314 1467318507 3118.4
## + drivewheel        1   49120226 1473286594 3119.2
## + aspiration        1   46238065 1476168756 3119.6
## + fuelsystem        2   36928355 1485478465 3122.8
## + carbody           2   31749038 1490657782 3123.5
## <none>                           1522406820 3123.6
## + highwaympg        1    7001568 1515405252 3124.7
## + compressionratio  1    5402046 1517004775 3124.9
## + carlength         1    2388510 1520018311 3125.3
## + carheight         1    1732126 1520674695 3125.4
## + wheelbase         1     238302 1522168518 3125.6
## + symboling         3    9061063 1513345757 3128.5
## - enginetype        2  171996617 1694403437 3140.6
## - stroke            1  174668406 1697075226 3142.9
## - enginelocation    1  223026291 1745433112 3148.4
## - carwidth          1  353951621 1876358441 3162.6
## - enginesize        1 1897193409 3419600229 3280.2
##
## Step:  AIC=3103.52
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
##     boreratio
##
##                    Df  Sum of Sq        RSS    AIC
## + drivewheel        1  100754377 1259300864 3090.4
## + peakrpm           1   89664954 1270390287 3092.2
## + aspiration        1   85186493 1274868748 3092.8
## + horsepower        1   81099794 1278955447 3093.5
## + fuelsystem        2   53375130 1306680111 3099.7
## + carbody           2   46611818 1313443424 3100.7
## + highwaympg        1   23298453 1336756788 3102.1
## <none>                           1360055241 3103.5
## + cylindernumber    2   26272920 1333782321 3103.7
## + carheight         1    4814614 1355240627 3104.8
## + compressionratio  1    4257883 1355797358 3104.9
## + wheelbase         1    1590830 1358464411 3105.3
## + carlength         1    1213737 1358841504 3105.3
## + symboling         3    2994608 1357060633 3109.1
## - boreratio         1  162351579 1522406820 3123.6
## - enginelocation    1  217342537 1577397778 3130.6
## - enginetype        2  276238998 1636294239 3135.8
## - stroke            1  285395525 1645450766 3138.9
## - carwidth          1  422996307 1783051548 3154.6
## - enginesize        1 1975734676 3335789917 3277.4
##
## Step:  AIC=3090.44
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
```

```
##     boreratio + drivewheel
##
##                    Df  Sum of Sq        RSS    AIC
## + peakrpm           1   80596283 1178704581 3079.5
## + aspiration        1   78761241 1180539624 3079.8
## + horsepower        1   64634034 1194666830 3082.1
## + carbody           2   42341224 1216959640 3087.7
## + cylindernumber    2   35358689 1223942176 3088.9
## + highwaympg        1   21898239 1237402625 3089.0
## + fuelsystem        2   26182132 1233118733 3090.3
## <none>                           1259300864 3090.4
## + carheight         1    7657963 1251642901 3091.2
## + carlength         1     898628 1258402236 3092.3
## + compressionratio  1     487380 1258813484 3092.4
## + wheelbase         1      62417 1259238448 3092.4
## + symboling         3   13517656 1245783208 3094.3
## - drivewheel        1  100754377 1360055241 3103.5
## - enginelocation    1  213042695 1472343559 3119.1
## - boreratio         1  213985730 1473286594 3119.2
## - stroke            1  234718329 1494019194 3121.9
## - enginetype        2  285690052 1544990916 3126.5
## - carwidth          1  423966257 1683267122 3145.3
## - enginesize        1 1502206214 2761507079 3242.3
##
## Step:  AIC=3079.48
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
##     boreratio + drivewheel + peakrpm
##
##                    Df  Sum of Sq        RSS    AIC
## + aspiration        1   98635873 1080068707 3064.3
## + compressionratio  1   41178894 1137525687 3074.5
## + carbody           2   47782449 1130922132 3075.4
## + carheight         1   22472168 1156232413 3077.7
## + cylindernumber    2   32314584 1146389997 3078.0
## + horsepower        1   13643060 1165061521 3079.2
## + fuelsystem        2   23996819 1154707762 3079.4
## <none>                           1178704581 3079.5
## + wheelbase         1    2359593 1176344988 3081.1
## + carlength         1    1751746 1176952835 3081.2
## + highwaympg        1    1069883 1177634698 3081.3
## + symboling         3   21115802 1157588779 3081.9
## - peakrpm           1   80596283 1259300864 3090.4
## - drivewheel        1   91685706 1270390287 3092.2
## - enginelocation    1  126813329 1305517910 3097.5
## - boreratio         1  166574269 1345278850 3103.4
## - stroke            1  247103708 1425808289 3114.8
## - enginetype        2  318594790 1497299371 3122.4
## - carwidth          1  421804787 1600509368 3137.4
## - enginesize        1 1573133351 2751837932 3243.7
##
## Step:  AIC=3064.35
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
##     boreratio + drivewheel + peakrpm + aspiration
##
```

```
##                     Df  Sum of Sq        RSS    AIC
## + carbody            2   67571658 1012497050 3055.7
## + fuelsystem         2   59660854 1020407854 3057.2
## + carheight          1   35496919 1044571789 3059.8
## + cylindernumber     2   36362193 1043706515 3061.6
## + compressionratio   1   23374318 1056694390 3062.1
## + symboling          3   36940880 1043127827 3063.5
## <none>                            1080068707 3064.3
## + wheelbase          1    5648796 1074419911 3065.3
## + carlength          1    5644927 1074423780 3065.3
## + horsepower         1    3625856 1076442851 3065.7
## + highwaympg         1    3532000 1076536708 3065.7
## - drivewheel         1   83776864 1163845571 3077.0
## - aspiration         1   98635873 1178704581 3079.5
## - peakrpm            1  100470916 1180539624 3079.8
## - enginelocation     1  104679592 1184748299 3080.5
## - boreratio          1  202598125 1282666832 3096.0
## - carwidth           1  264270711 1344339419 3105.2
## - stroke             1  312229173 1392297881 3112.1
## - enginetype         2  338282199 1418350907 3113.8
## - enginesize         1 1671547110 2751615818 3245.6
##
## Step:  AIC=3055.68
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
##     boreratio + drivewheel + peakrpm + aspiration + carbody
##
##                     Df  Sum of Sq        RSS    AIC
## + fuelsystem         2   37805852  974691198 3052.2
## + cylindernumber     2   24704069  987792980 3054.8
## + carheight          1   10466414 1002030636 3055.6
## <none>                            1012497050 3055.7
## + compressionratio   1   10006697 1002490352 3055.7
## + highwaympg         1    3958122 1008538927 3056.9
## + carlength          1    1281472 1011215577 3057.4
## + horsepower         1     126363 1012370686 3057.7
## + wheelbase          1       1749 1012495300 3057.7
## + symboling          3   18112922  994384127 3058.1
## - carbody            2   67571658 1080068707 3064.3
## - drivewheel         1   76154917 1088651966 3067.9
## - enginelocation     1   80025346 1092522395 3068.6
## - peakrpm            1  110250886 1122747936 3073.9
## - aspiration         1  118425083 1130922132 3075.4
## - carwidth           1  217471092 1229968142 3091.8
## - boreratio          1  222289478 1234786528 3092.6
## - stroke             1  300133804 1312630853 3104.6
## - enginetype         2  319705237 1332202286 3105.5
## - enginesize         1 1644773161 2657270210 3242.8
##
## Step:  AIC=3052.23
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
##     boreratio + drivewheel + peakrpm + aspiration + carbody +
##     fuelsystem
##
##                     Df  Sum of Sq        RSS    AIC
```

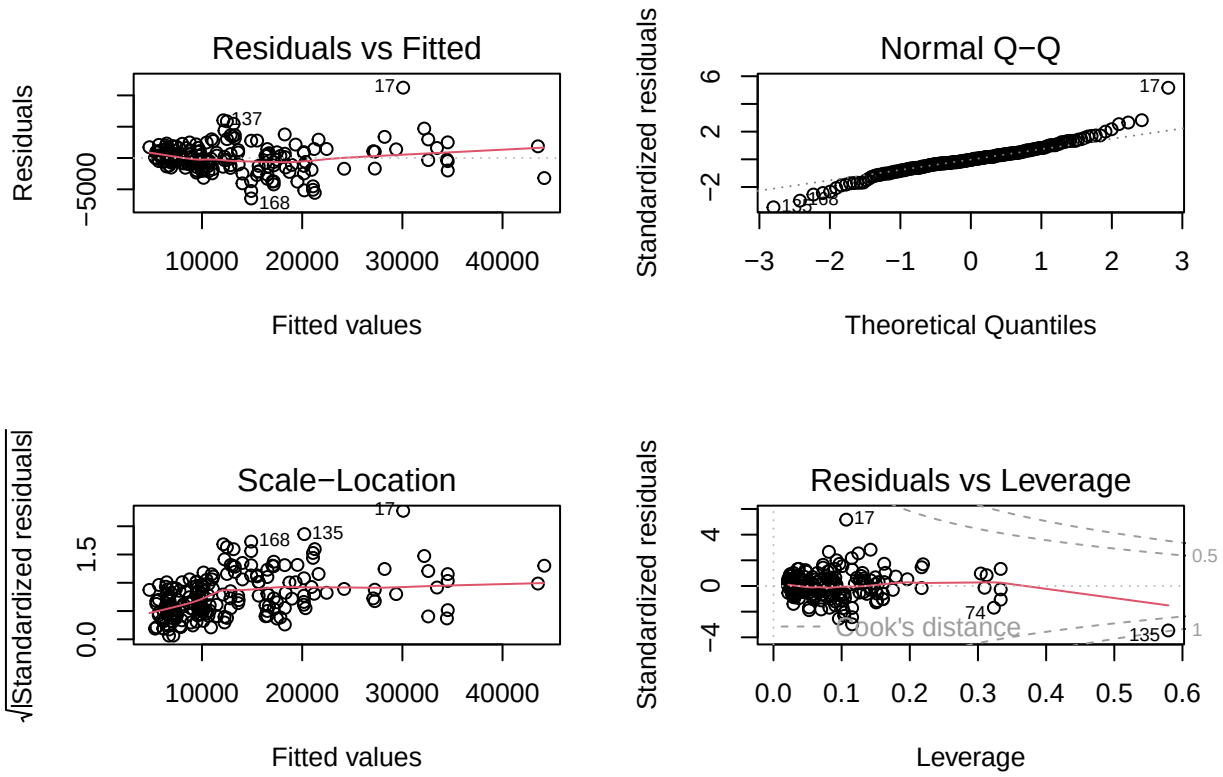
```
## + cylindernumber    2   27035787  947655411 3050.7
## <none>                            974691198 3052.2
## + carheight         1    3362057  971329141 3053.5
## + compressionratio  1    2248671  972442526 3053.8
## + carlength         1    1805842  972885356 3053.9
## + highwaympg        1    1515866  973175332 3053.9
## + horsepower        1     148038  974543160 3054.2
## + wheelbase         1     135800  974555398 3054.2
## + symboling         3   15388567  959302631 3055.1
## - fuelsystem        2   37805852 1012497050 3055.7
## - carbody           2   45716656 1020407854 3057.2
## - drivewheel        1   54549019 1029240217 3060.9
## - enginelocation    1   69132994 1043824192 3063.7
## - peakrpm           1  125222205 1099913403 3073.9
## - aspiration        1  143192051 1117883249 3077.1
## - carwidth          1  145498982 1120190180 3077.5
## - boreratio         1  208590190 1183281388 3088.2
## - stroke            1  271735754 1246426952 3098.4
## - enginetype        2  337445758 1312136956 3106.5
## - enginesize        1 1677393005 2652084203 3246.4
##
## Step:  AIC=3050.71
## price ~ enginesize + enginelocation + carwidth + stroke + enginetype +
##     boreratio + drivewheel + peakrpm + aspiration + carbody +
##     fuelsystem + cylindernumber
##
##                    Df Sum of Sq        RSS    AIC
## <none>                           947655411 3050.7
## + carheight         1   3049443  944605968 3052.1
## - cylindernumber    2  27035787  974691198 3052.2
## + compressionratio  1   1980159  945675253 3052.3
## + symboling         3  20721184  926934228 3052.4
## + horsepower        1   1570673  946084738 3052.4
## + highwaympg        1    962965  946692446 3052.5
## + wheelbase         1    638108  947017303 3052.6
## + carlength         1    374275  947281137 3052.6
## - carbody           2  36091474  983746885 3054.0
## - fuelsystem        2  40137569  987792980 3054.8
## - drivewheel        1  63235496 1010890907 3061.4
## - boreratio         1  71375257 1019030668 3062.9
## - enginelocation    1  76427964 1024083376 3063.9
## - carwidth          1 107620540 1055275951 3069.8
## - peakrpm           1 121141608 1068797019 3072.3
## - aspiration        1 140257877 1087913289 3075.8
## - stroke            1 167400137 1115055548 3080.6
## - enginetype        2 359077967 1306733379 3109.7
## - enginesize        1 577099278 1524754689 3141.9
```

```
summary(fitAIC)
```

```
##
## Call:
## lm(formula = price ~ enginesize + enginelocation + carwidth +
##     stroke + enginetype + boreratio + drivewheel + peakrpm +
##     aspiration + carbody + fuelsystem + cylindernumber, data = d)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6460.2 -1154.8   -84.6  1067.8 11234.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.415e+04  1.032e+04  -3.310 0.001127 **
## enginesize             1.730e+02  1.657e+01  10.441  < 2e-16 ***
## enginelocationrear     6.807e+03  1.791e+03   3.800 0.000199 ***
## carwidth               7.000e+02  1.553e+02   4.509 1.18e-05 ***
## stroke                -4.095e+03  7.283e+02  -5.623 7.10e-08 ***
## enginetypeohc          2.671e+03  6.161e+02   4.336 2.42e-05 ***
## enginetypeohcv        -5.047e+03  1.089e+03  -4.634 6.88e-06 ***
## boreratio             -4.642e+03  1.264e+03  -3.672 0.000318 ***
## drivewheelr            1.869e+03  5.407e+02   3.456 0.000685 ***
## peakrpm                1.929e+00  4.032e-01   4.784 3.59e-06 ***
## aspirationturbo        2.846e+03  5.529e+02   5.147 6.92e-07 ***
## carbodyh              -8.715e+02  8.472e+02  -1.029 0.305020
## carbodysw              1.484e+02  7.960e+02   0.186 0.852296
## fuelsystemim           5.455e+02  4.985e+02   1.094 0.275285
## fuelsystems           -2.128e+03  1.006e+03  -2.116 0.035755 *
## cylindernumberfivesix -4.420e+03  1.970e+03  -2.244 0.026043 *
## cylindernumberfour    -5.325e+03  2.606e+03  -2.043 0.042502 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2301 on 179 degrees of freedom
## Multiple R-squared:  0.9216, Adjusted R-squared:  0.9146
## F-statistic: 131.5 on 16 and 179 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitAIC)
```



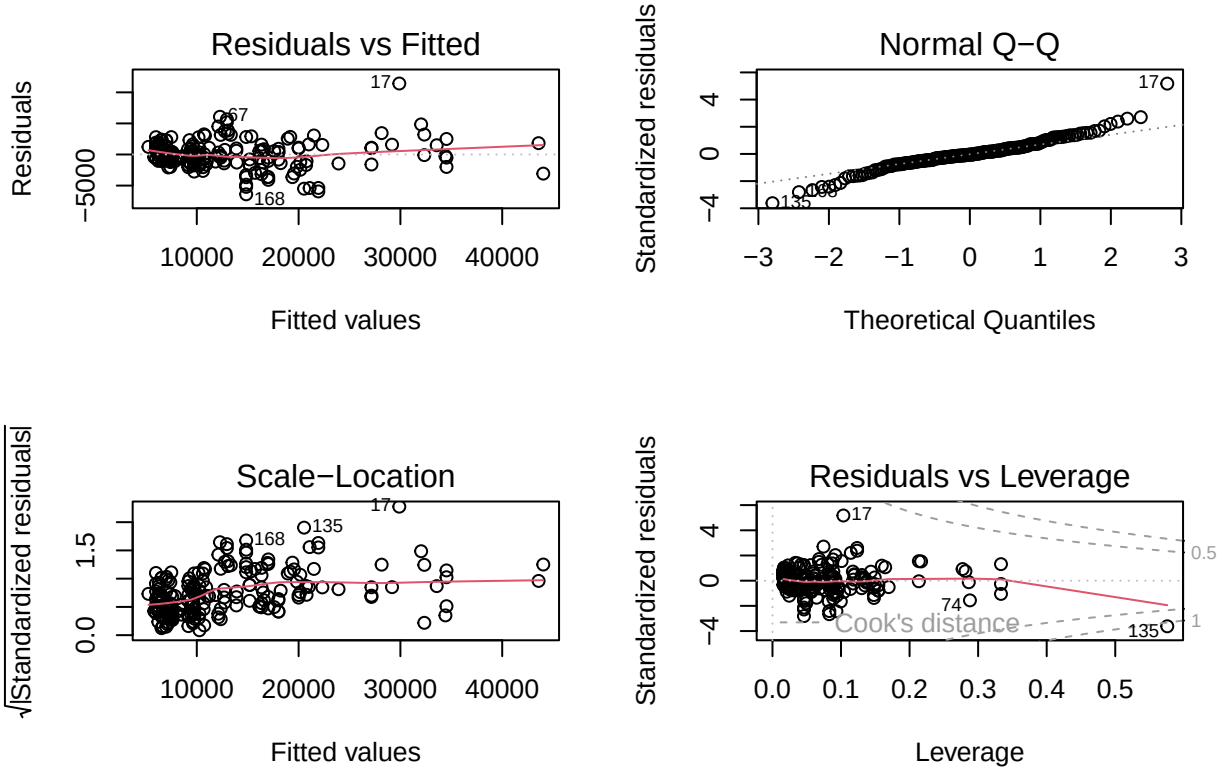
Si osserva che carbody non sembra significativa

```
fitNew <- lm(price ~ enginesize + enginelocation + carwidth +
    stroke + enginetype + boreratio + drivewheel + peakrpm +
    aspiration + fuelsystem + cylindernumber, data=d)
summary(fitNew)
```

```
##
## Call:
## lm(formula = price ~ enginesize + enginelocation + carwidth +
##     stroke + enginetype + boreratio + drivewheel + peakrpm +
##     aspiration + fuelsystem + cylindernumber, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6401.9 -1144.9  -120.1  1046.2 11422.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.558e+04  1.034e+04  -3.442 0.000717 ***
## enginesize             1.689e+02  1.670e+01  10.117  < 2e-16 ***
## enginelocationrear     6.825e+03  1.702e+03   4.010 8.87e-05 ***
## carwidth               7.145e+02  1.539e+02   4.643 6.58e-06 ***
## stroke                -4.027e+03  7.368e+02  -5.466 1.51e-07 ***
## enginetypeohc          2.630e+03  6.237e+02   4.216 3.92e-05 ***
## enginetypeohcv        -5.408e+03  1.094e+03  -4.945 1.73e-06 ***
## boreratio             -4.160e+03  1.267e+03  -3.284 0.001229 **
```

```
## drivewheelr             1.830e+03  5.420e+02   3.377 0.000898 ***
## peakrpm                 1.884e+00  4.076e-01   4.623 7.19e-06 ***
## aspirationturbo         2.761e+03  5.589e+02   4.940 1.77e-06 ***
## fuelsystemim            7.465e+02  4.937e+02   1.512 0.132239
## fuelsystems            -2.389e+03  1.014e+03  -2.356 0.019547 *
## cylindernumberfivesix -5.022e+03  1.939e+03  -2.590 0.010375 *
## cylindernumberfour    -6.204e+03  2.592e+03  -2.394 0.017701 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2331 on 181 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.9123
## F-statistic: 145.9 on 14 and 181 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitNew)
```



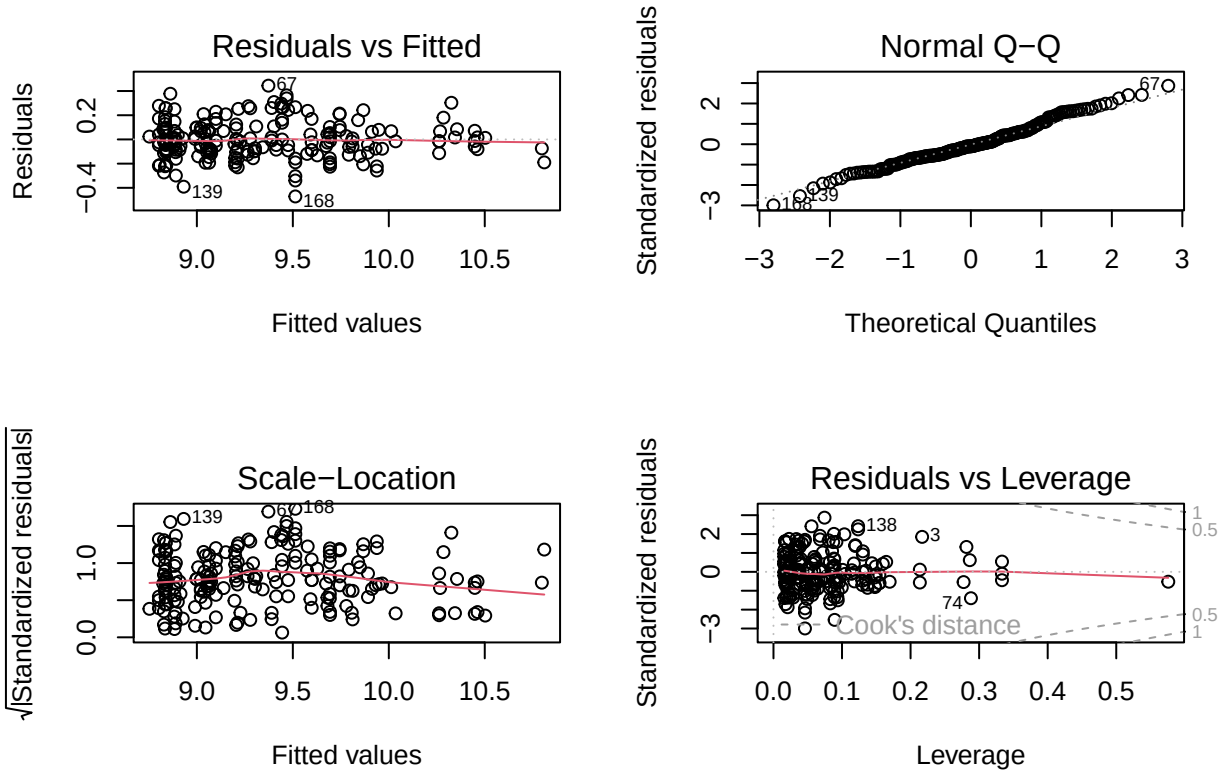
Si cerca una trasformazione.

```
fitNewLog <- lm(log(price) ~ enginesize + enginelocation + carwidth +
    stroke + enginetype + boreratio + drivewheel + peakrpm +
    aspiration + fuelsystem + cylindernumber, data=d)
summary(fitNewLog)
```

```
##
## Call:
## lm(formula = log(price) ~ enginesize + enginelocation + carwidth +
##     stroke + enginetype + boreratio + drivewheel + peakrpm +
```

```
##     aspiration + fuelsystem + cylindernumber, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47130 -0.09708 -0.01498  0.08921  0.44427
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.653e+00  7.139e-01   6.517 6.91e-10 ***
## enginesize              7.566e-03  1.153e-03   6.559 5.50e-10 ***
## enginelocationrear      2.607e-01  1.176e-01   2.218 0.027801 *
## carwidth                5.879e-02  1.063e-02   5.531 1.10e-07 ***
## stroke                 -2.048e-01  5.089e-02  -4.024 8.39e-05 ***
## enginetypeohc           1.033e-01  4.308e-02   2.399 0.017459 *
## enginetypeohcv         -2.174e-01  7.555e-02  -2.877 0.004493 **
## boreratio              -7.889e-02  8.750e-02  -0.902 0.368431
## drivewheelr             1.307e-01  3.744e-02   3.490 0.000607 ***
## peakrpm                 1.087e-04  2.816e-05   3.861 0.000157 ***
## aspirationturbo         1.542e-01  3.860e-02   3.994 9.45e-05 ***
## fuelsystemim            1.954e-01  3.410e-02   5.730 4.12e-08 ***
## fuelsystems             3.517e-02  7.004e-02   0.502 0.616210
## cylindernumberfivesix   6.381e-02  1.339e-01   0.476 0.634353
## cylindernumberfour     -4.954e-02  1.790e-01  -0.277 0.782285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.161 on 181 degrees of freedom
## Multiple R-squared:  0.9036, Adjusted R-squared:  0.8961
## F-statistic: 121.2 on 14 and 181 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitNewLog)
```



I risultati rispettano le ipotesi di normalità e linearità, mentre appare esserci eteroschedasticità e correlazione tra i residui.

Si vuole tentare una via più efficiente targetizzando da subito il log(price) invece che price.

```
fit0 <- lm(log(price) ~ ., data=d)
fit1 <- lm(log(price) ~ 1, data=d)
fit2 <- stepAIC(fit1, scope=formula(fit0), direction="both")
```

```
## Start:  AIC=-271
## log(price) ~ 1
##
##                    Df Sum of Sq    RSS     AIC
## + enginesize        1    35.686 12.994 -527.88
## + horsepower        1    33.847 14.832 -501.93
## + carwidth          1    30.600 18.080 -463.13
## + highwaympg        1    29.689 18.990 -453.50
## + carlength         1    29.148 19.531 -448.00
## + cylindernumber    2    25.151 23.528 -409.50
## + fuelsystem        2    24.735 23.944 -406.07
## + drivewheel        1    23.124 25.555 -395.31
## + wheelbase         1    20.003 28.677 -372.72
## + boreratio         1    17.417 31.263 -355.79
## + symboling         3    11.400 37.280 -317.30
## + enginetype        2     9.310 39.370 -308.60
## + carbody           2     7.016 41.663 -297.51
## + enginelocation    1     3.705 44.974 -284.52
```

```
## + aspiration        1     3.697 44.982 -284.48
## + carheight         1     2.582 46.097 -279.68
## + stroke            1     0.675 48.004 -271.74
## + peakrpm           1     0.644 48.036 -271.61
## <none>                          48.679 -271.00
## + compressionratio  1     0.351 48.328 -270.42
##
## Step:  AIC=-527.88
## log(price) ~ enginesize
##
##                    Df Sum of Sq    RSS     AIC
## + fuelsystem        2     4.006  8.988 -596.12
## + carlength         1     2.948 10.046 -576.31
## + carwidth          1     2.791 10.203 -573.27
## + horsepower        1     2.608 10.386 -569.78
## + highwaympg        1     2.531 10.462 -568.35
## + drivewheel        1     1.775 11.219 -554.66
## + aspiration        1     1.591 11.402 -551.48
## + wheelbase         1     1.347 11.647 -547.32
## + enginetype        2     1.172 11.822 -542.40
## + carbody           2     1.047 11.947 -540.34
## + carheight         1     0.900 12.094 -539.94
## + symboling         3     1.141 11.853 -539.88
## + cylindernumber    2     1.011 11.983 -539.75
## + stroke            1     0.575 12.419 -534.74
## + boreratio         1     0.560 12.434 -534.50
## + enginelocation    1     0.413 12.581 -532.20
## + peakrpm           1     0.368 12.626 -531.50
## + compressionratio  1     0.226 12.767 -529.32
## <none>                          12.994 -527.88
## - enginesize        1    35.686 48.679 -271.00
##
## Step:  AIC=-596.12
## log(price) ~ enginesize + fuelsystem
##
##                    Df Sum of Sq     RSS     AIC
## + horsepower        1    1.2160  7.7718 -622.61
## + highwaympg        1    1.1894  7.7984 -621.94
## + carwidth          1    0.9747  8.0132 -616.62
## + carlength         1    0.8730  8.1149 -614.15
## + aspiration        1    0.7616  8.2262 -611.48
## + enginetype        2    0.6659  8.3219 -607.21
## + drivewheel        1    0.5357  8.4522 -606.16
## + enginelocation    1    0.4524  8.5354 -604.24
## + stroke            1    0.4339  8.5540 -603.82
## + symboling         3    0.5942  8.3936 -603.53
## + wheelbase         1    0.3632  8.6246 -602.20
## + cylindernumber    2    0.4330  8.5548 -601.80
## + peakrpm           1    0.3390  8.6488 -601.66
## + carheight         1    0.3066  8.6812 -600.92
## + carbody           2    0.3767  8.6111 -600.51
## + boreratio         1    0.1064  8.8814 -596.45
## <none>                          8.9878 -596.12
## + compressionratio  1    0.0140  8.9738 -594.43
```

```
## - fuelsystem        2    4.0059 12.9937 -527.88
## - enginesize        1   14.9563 23.9442 -406.07
##
## Step:  AIC=-622.61
## log(price) ~ enginesize + fuelsystem + horsepower
##
##                    Df Sum of Sq     RSS     AIC
## + carwidth          1    1.2912  6.4806 -656.22
## + carlength         1    1.1365  6.6353 -651.60
## + wheelbase         1    1.0237  6.7481 -648.29
## + carheight         1    0.6640  7.1078 -638.12
## + enginetype        2    0.7270  7.0449 -637.86
## + carbody           2    0.6709  7.1009 -636.31
## + aspiration        1    0.4798  7.2920 -633.10
## + highwaympg        1    0.3986  7.3732 -630.93
## + symboling         3    0.5283  7.2435 -630.41
## + compressionratio  1    0.3768  7.3950 -630.35
## + drivewheel        1    0.3474  7.4244 -629.57
## + stroke            1    0.1591  7.6127 -624.67
## <none>                           7.7718 -622.61
## + enginelocation    1    0.0772  7.6946 -622.57
## + boreratio         1    0.0466  7.7252 -621.79
## + cylindernumber    2    0.1127  7.6591 -621.47
## + peakrpm           1    0.0043  7.7675 -620.72
## - horsepower        1    1.2160  8.9878 -596.12
## - fuelsystem        2    2.6143 10.3861 -569.78
## - enginesize        1    3.2110 10.9828 -556.83
##
## Step:  AIC=-656.22
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth
##
##                    Df Sum of Sq    RSS     AIC
## + carbody           2   0.64894 5.8316 -672.90
## + enginetype        2   0.64601 5.8346 -672.81
## + enginelocation    1   0.46439 6.0162 -668.80
## + drivewheel        1   0.36925 6.1113 -665.72
## + compressionratio  1   0.24582 6.2347 -661.80
## + symboling         3   0.36112 6.1194 -661.46
## + carlength         1   0.19753 6.2830 -660.29
## + carheight         1   0.19476 6.2858 -660.20
## + stroke            1   0.19405 6.2865 -660.18
## + wheelbase         1   0.11305 6.3675 -657.67
## + aspiration        1   0.08112 6.3994 -656.69
## + highwaympg        1   0.06585 6.4147 -656.23
## <none>                          6.4806 -656.22
## + cylindernumber    2   0.08798 6.3926 -654.90
## + boreratio         1   0.00603 6.4745 -654.41
## + peakrpm           1   0.00109 6.4795 -654.26
## - enginesize        1   0.88212 7.3627 -633.21
## - fuelsystem        2   1.11669 7.5973 -629.06
## - carwidth          1   1.29124 7.7718 -622.61
## - horsepower        1   1.53260 8.0132 -616.62
##
## Step:  AIC=-672.9
```

```
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody
##
##                    Df Sum of Sq    RSS     AIC
## + enginetype        2   0.47348 5.3581 -685.50
## + drivewheel        1   0.30667 5.5250 -681.49
## + enginelocation    1   0.28306 5.5486 -680.66
## + compressionratio  1   0.22152 5.6101 -678.49
## + symboling         3   0.29522 5.5364 -677.09
## + stroke            1   0.12650 5.7051 -675.20
## + aspiration        1   0.08149 5.7501 -673.66
## + wheelbase         1   0.06665 5.7650 -673.16
## + carheight         1   0.05973 5.7719 -672.92
## <none>                          5.8316 -672.90
## + cylindernumber    2   0.11545 5.7162 -672.82
## + carlength         1   0.05062 5.7810 -672.61
## + highwaympg        1   0.02611 5.8055 -671.78
## + boreratio         1   0.00466 5.8270 -671.06
## + peakrpm           1   0.00006 5.8316 -670.91
## - fuelsystem        2   0.59204 6.4237 -657.95
## - enginesize        1   0.56120 6.3928 -656.90
## - carbody           2   0.64894 6.4806 -656.22
## - carwidth          1   1.26926 7.1009 -636.31
## - horsepower        1   1.77482 7.6065 -622.83
##
## Step:  AIC=-685.5
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype
##
##                    Df Sum of Sq    RSS     AIC
## + drivewheel        1   0.30622 5.0519 -695.04
## + stroke            1   0.27811 5.0800 -693.95
## + symboling         3   0.29832 5.0598 -690.73
## + compressionratio  1   0.16609 5.1921 -689.67
## + cylindernumber    2   0.21359 5.1446 -689.47
## + enginelocation    1   0.13906 5.2191 -688.66
## + boreratio         1   0.06109 5.2971 -685.75
## + aspiration        1   0.05751 5.3006 -685.62
## <none>                          5.3581 -685.50
## + highwaympg        1   0.04286 5.3153 -685.08
## + carheight         1   0.03713 5.3210 -684.86
## + wheelbase         1   0.02260 5.3356 -684.33
## + carlength         1   0.01502 5.3431 -684.05
## + peakrpm           1   0.00380 5.3543 -683.64
## - enginetype        2   0.47348 5.8316 -672.90
## - carbody           2   0.47641 5.8346 -672.81
## - fuelsystem        2   0.64720 6.0053 -667.15
## - enginesize        1   0.88434 6.2425 -657.56
## - carwidth          1   1.16030 6.5184 -649.08
## - horsepower        1   1.80547 7.1636 -630.58
##
## Step:  AIC=-695.04
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel
```

```
##
##                     Df Sum of Sq    RSS     AIC
## + symboling          3   0.41828 4.6337 -705.97
## + cylindernumber     2   0.31475 4.7372 -703.64
## + stroke             1   0.19233 4.8596 -700.64
## + enginelocation     1   0.16396 4.8880 -699.50
## + boreratio          1   0.15523 4.8967 -699.15
## + compressionratio   1   0.12765 4.9243 -698.05
## <none>                           5.0519 -695.04
## + aspiration         1   0.04735 5.0046 -694.88
## + carheight          1   0.04053 5.0114 -694.61
## + highwaympg         1   0.03463 5.0173 -694.38
## + peakrpm            1   0.01157 5.0404 -693.49
## + carlength          1   0.00642 5.0455 -693.28
## + wheelbase          1   0.00006 5.0519 -693.04
## - drivewheel         1   0.30622 5.3581 -685.50
## - fuelsystem         2   0.43901 5.4909 -682.70
## - carbody            2   0.44266 5.4946 -682.57
## - enginetype         2   0.47304 5.5250 -681.49
## - enginesize         1   0.60159 5.6535 -674.98
## - carwidth           1   1.17530 6.2272 -656.04
## - horsepower         1   1.61330 6.6652 -642.72
##
## Step:  AIC=-705.97
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling
##
##                     Df Sum of Sq    RSS     AIC
## + cylindernumber     2   0.42008 4.2136 -720.60
## + boreratio          1   0.25457 4.3791 -715.05
## + enginelocation     1   0.10114 4.5325 -708.30
## + compressionratio   1   0.09247 4.5412 -707.93
## + stroke             1   0.09095 4.5427 -707.86
## + carheight          1   0.06306 4.5706 -706.66
## + aspiration         1   0.05437 4.5793 -706.29
## <none>                           4.6337 -705.97
## + wheelbase          1   0.03806 4.5956 -705.59
## + peakrpm            1   0.02408 4.6096 -705.00
## + highwaympg         1   0.02098 4.6127 -704.86
## + carlength          1   0.00256 4.6311 -704.08
## - symboling          3   0.41828 5.0519 -695.04
## - carbody            2   0.44049 5.0741 -692.18
## - drivewheel         1   0.42618 5.0598 -690.73
## - fuelsystem         2   0.48018 5.1138 -690.65
## - enginesize         1   0.46278 5.0964 -689.32
## - enginetype         2   0.57403 5.2077 -687.08
## - horsepower         1   1.25085 5.8845 -661.14
## - carwidth           1   1.29940 5.9330 -659.53
##
## Step:  AIC=-720.6
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + cylindernumber
##
##                     Df Sum of Sq    RSS     AIC
```

```
## + carheight         1   0.09416 4.1194 -723.03
## + compressionratio  1   0.07839 4.1352 -722.28
## + aspiration        1   0.07403 4.1395 -722.08
## + wheelbase         1   0.06827 4.1453 -721.80
## + enginelocation    1   0.05770 4.1559 -721.30
## + stroke            1   0.04579 4.1678 -720.74
## <none>                          4.2136 -720.60
## + carlength         1   0.04186 4.1717 -720.56
## + boreratio         1   0.02095 4.1926 -719.58
## - enginesize        1   0.06707 4.2806 -719.51
## + highwaympg        1   0.01009 4.2035 -719.07
## + peakrpm           1   0.00305 4.2105 -718.74
## - carbody           2   0.35551 4.5691 -708.73
## - cylindernumber    2   0.42008 4.6337 -705.97
## - symboling         3   0.52361 4.7372 -703.64
## - fuelsystem        2   0.47600 4.6896 -703.62
## - drivewheel        1   0.55919 4.7728 -698.18
## - enginetype        2   0.74129 4.9549 -692.84
## - horsepower        1   0.97736 5.1909 -681.72
## - carwidth          1   1.06923 5.2828 -678.28
##
## Step:  AIC=-723.03
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + cylindernumber +
##     carheight
##
##                     Df Sum of Sq    RSS     AIC
## + compressionratio  1   0.06764 4.0518 -724.28
## + aspiration        1   0.06366 4.0558 -724.08
## + enginelocation    1   0.04461 4.0748 -723.17
## <none>                          4.1194 -723.03
## + stroke            1   0.03101 4.0884 -722.51
## + boreratio         1   0.02401 4.0954 -722.18
## - enginesize        1   0.06497 4.1844 -721.96
## + wheelbase         1   0.01940 4.1000 -721.96
## + peakrpm           1   0.00942 4.1100 -721.48
## + carlength         1   0.00889 4.1105 -721.45
## + highwaympg        1   0.00106 4.1183 -721.08
## - carheight         1   0.09416 4.2136 -720.60
## - carbody           2   0.25770 4.3771 -715.14
## - fuelsystem        2   0.41137 4.5308 -708.37
## - cylindernumber    2   0.45118 4.5706 -706.66
## - symboling         3   0.54263 4.6620 -704.78
## - drivewheel        1   0.59691 4.7163 -698.51
## - enginetype        2   0.73578 4.8552 -694.82
## - carwidth          1   0.87836 4.9978 -687.15
## - horsepower        1   1.00286 5.1223 -682.33
##
## Step:  AIC=-724.28
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + cylindernumber +
##     carheight + compressionratio
##
##                     Df Sum of Sq    RSS     AIC
```

```
## + stroke            1   0.04998 4.0018 -724.71
## - enginesize        1   0.03433 4.0861 -724.62
## <none>                          4.0518 -724.28
## + highwaympg        1   0.03763 4.0141 -724.11
## + wheelbase         1   0.03259 4.0192 -723.86
## + carlength         1   0.02145 4.0303 -723.32
## + enginelocation    1   0.02113 4.0306 -723.30
## + peakrpm           1   0.02091 4.0309 -723.29
## + boreratio         1   0.01910 4.0327 -723.20
## - compressionratio  1   0.06764 4.1194 -723.03
## + aspiration        1   0.01509 4.0367 -723.01
## - carheight         1   0.08341 4.1352 -722.28
## - carbody           2   0.25218 4.3039 -716.44
## - fuelsystem        2   0.25955 4.3113 -716.11
## - cylindernumber    2   0.43155 4.4833 -708.44
## - symboling         3   0.50787 4.5596 -707.13
## - drivewheel        1   0.55161 4.6034 -701.26
## - enginetype        2   0.73161 4.7834 -695.74
## - carwidth          1   0.82075 4.8725 -690.12
## - horsepower        1   0.91376 4.9655 -686.42
##
## Step:  AIC=-724.71
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + cylindernumber +
##     carheight + compressionratio + stroke
##
##                     Df Sum of Sq    RSS     AIC
## + boreratio          1   0.07305 3.9287 -726.32
## + peakrpm            1   0.04121 3.9606 -724.74
## <none>                          4.0018 -724.71
## + highwaympg         1   0.03221 3.9696 -724.29
## - stroke             1   0.04998 4.0518 -724.28
## + wheelbase          1   0.03103 3.9708 -724.23
## + aspiration         1   0.02005 3.9817 -723.69
## - carheight          1   0.06472 4.0665 -723.56
## + carlength          1   0.01676 3.9850 -723.53
## - enginesize         1   0.06959 4.0714 -723.33
## + enginelocation     1   0.01033 3.9915 -723.22
## - compressionratio   1   0.08661 4.0884 -722.51
## - carbody            2   0.23491 4.2367 -717.53
## - fuelsystem         2   0.24591 4.2477 -717.02
## - symboling          3   0.38367 4.3855 -712.76
## - cylindernumber     2   0.36498 4.3668 -711.60
## - drivewheel         1   0.42773 4.4295 -706.81
## - enginetype         2   0.76059 4.7624 -694.60
## - carwidth           1   0.83329 4.8351 -689.63
## - horsepower         1   0.85048 4.8523 -688.94
##
## Step:  AIC=-726.32
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + cylindernumber +
##     carheight + compressionratio + stroke + boreratio
##
##                     Df Sum of Sq    RSS     AIC
```

```
## + highwaympg        1   0.05876 3.8700 -727.27
## - cylindernumber    2   0.06733 3.9961 -726.99
## <none>                          3.9287 -726.32
## + aspiration        1   0.02961 3.8991 -725.80
## + peakrpm           1   0.02608 3.9027 -725.63
## - carheight         1   0.06048 3.9892 -725.33
## + wheelbase         1   0.01972 3.9090 -725.31
## + carlength         1   0.01606 3.9127 -725.12
## + enginelocation    1   0.01253 3.9162 -724.95
## - boreratio         1   0.07305 4.0018 -724.71
## - compressionratio  1   0.08895 4.0177 -723.93
## - stroke            1   0.10393 4.0327 -723.20
## - enginesize        1   0.13779 4.0665 -721.56
## - fuelsystem        2   0.24504 4.1738 -718.46
## - carbody           2   0.26422 4.1930 -717.56
## - symboling         3   0.37858 4.3073 -714.29
## - drivewheel        1   0.45294 4.3817 -706.93
## - enginetype        2   0.82439 4.7531 -692.99
## - horsepower        1   0.90445 4.8332 -687.71
## - carwidth          1   0.90588 4.8346 -687.65
##
## Step:  AIC=-727.27
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + cylindernumber +
##     carheight + compressionratio + stroke + boreratio + highwaympg
##
##                     Df Sum of Sq    RSS     AIC
## - cylindernumber    2   0.04124 3.9112 -729.20
## - carheight         1   0.02672 3.8967 -727.93
## <none>                          3.8700 -727.27
## + enginelocation    1   0.02667 3.8433 -726.63
## + peakrpm           1   0.02505 3.8449 -726.55
## - highwaympg        1   0.05876 3.9287 -726.32
## + aspiration        1   0.01199 3.8580 -725.88
## + wheelbase         1   0.00912 3.8609 -725.74
## + carlength         1   0.00515 3.8648 -725.53
## - boreratio         1   0.09960 3.9696 -724.29
## - stroke            1   0.10987 3.9798 -723.79
## - compressionratio  1   0.14468 4.0146 -722.08
## - enginesize        1   0.15418 4.0242 -721.62
## - fuelsystem        2   0.21806 4.0880 -720.53
## - carbody           2   0.26211 4.1321 -718.43
## - symboling         3   0.32982 4.1998 -717.24
## - drivewheel        1   0.41511 4.2851 -709.30
## - carwidth          1   0.71877 4.5887 -695.88
## - horsepower        1   0.78130 4.6513 -693.23
## - enginetype        2   0.86294 4.7329 -691.82
##
## Step:  AIC=-729.2
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + carheight +
##     compressionratio + stroke + boreratio + highwaympg
##
##                     Df Sum of Sq    RSS     AIC
```

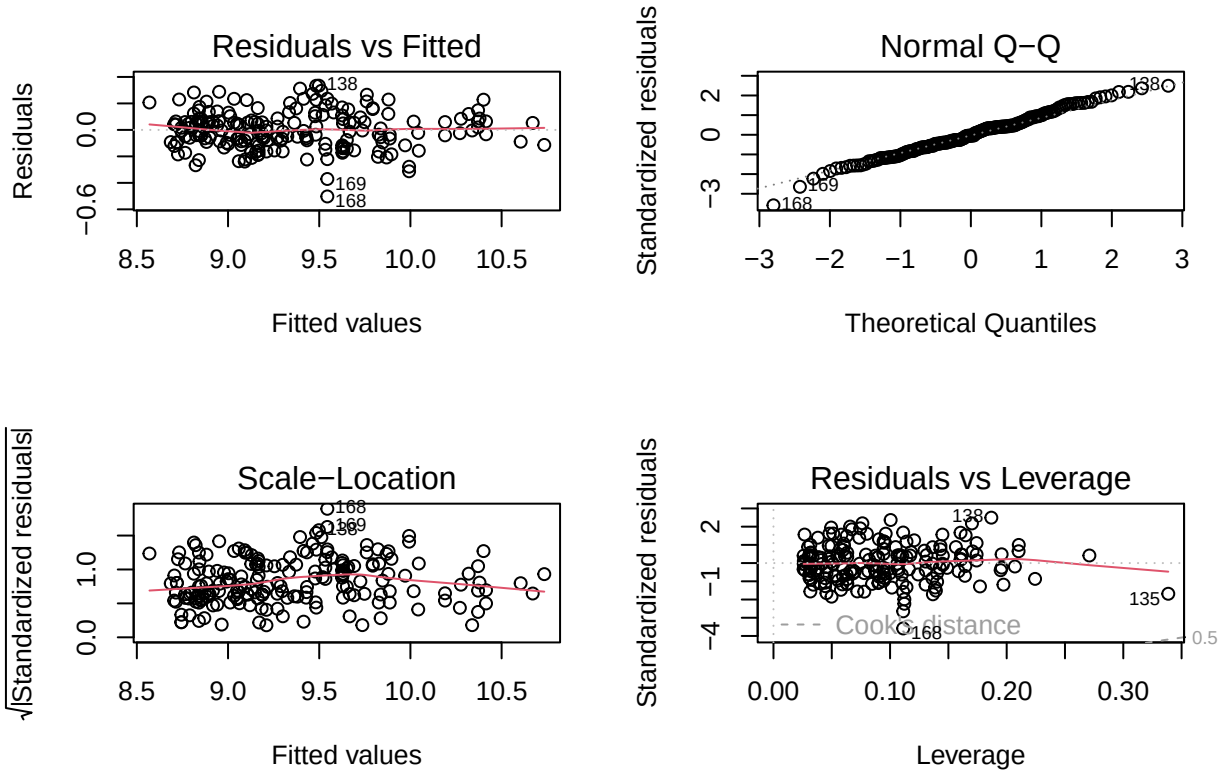
```
## - carheight         1   0.01981 3.9310 -730.21
## <none>                          3.9112 -729.20
## + enginelocation    1   0.03316 3.8781 -728.87
## + peakrpm           1   0.02965 3.8816 -728.69
## + aspiration        1   0.01224 3.8990 -727.81
## + wheelbase         1   0.00433 3.9069 -727.41
## + cylindernumber    2   0.04124 3.8700 -727.27
## + carlength         1   0.00062 3.9106 -727.23
## - highwaympg        1   0.08484 3.9961 -726.99
## - fuelsystem        2   0.20749 4.1187 -723.06
## - compressionratio  1   0.17682 4.0880 -722.53
## - symboling         3   0.29898 4.2102 -720.76
## - stroke            1   0.22551 4.1367 -720.21
## - carbody           2   0.30357 4.2148 -718.55
## - drivewheel        1   0.38803 4.2992 -712.66
## - boreratio         1   0.39102 4.3022 -712.52
## - enginesize        1   0.61045 4.5217 -702.77
## - enginetype        2   0.89077 4.8020 -692.98
## - carwidth          1   0.85057 4.7618 -692.63
## - horsepower        1   0.92796 4.8392 -689.47
##
## Step:  AIC=-730.21
## log(price) ~ enginesize + fuelsystem + horsepower + carwidth +
##     carbody + enginetype + drivewheel + symboling + compressionratio +
##     stroke + boreratio + highwaympg
##
##                    Df Sum of Sq    RSS     AIC
## <none>                          3.9310 -730.21
## + enginelocation    1   0.03737 3.8937 -730.08
## + peakrpm           1   0.02510 3.9059 -729.46
## + carheight         1   0.01981 3.9112 -729.20
## + wheelbase         1   0.01430 3.9167 -728.92
## + aspiration        1   0.01222 3.9188 -728.82
## + carlength         1   0.00525 3.9258 -728.47
## + cylindernumber    2   0.03434 3.8967 -727.93
## - highwaympg        1   0.11523 4.0463 -726.54
## - fuelsystem        2   0.22512 4.1562 -723.29
## - symboling         3   0.28518 4.2162 -722.48
## - compressionratio  1   0.20599 4.1370 -722.20
## - stroke            1   0.24215 4.1732 -720.49
## - carbody           2   0.34735 4.2784 -717.61
## - drivewheel        1   0.37107 4.3021 -714.53
## - boreratio         1   0.38472 4.3158 -713.91
## - enginesize        1   0.61102 4.5421 -703.89
## - enginetype        2   0.89929 4.8303 -693.83
## - carwidth          1   0.88924 4.8203 -692.24
## - horsepower        1   0.90826 4.8393 -691.46
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = log(price) ~ enginesize + fuelsystem + horsepower +
##     carwidth + carbody + enginetype + drivewheel + symboling +
##     compressionratio + stroke + boreratio + highwaympg, data = d)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50262 -0.09246 -0.00588  0.08259  0.33406
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.3966766  0.6920379   7.798 5.09e-13 ***
## enginesize        0.0040808  0.0007758   5.260 4.10e-07 ***
## fuelsystemim      0.0996731  0.0362503   2.750  0.00658 **
## fuelsystems      -0.0325984  0.0642455  -0.507  0.61250
## horsepower        0.0049103  0.0007657   6.413 1.24e-09 ***
## carwidth          0.0623850  0.0098314   6.345 1.78e-09 ***
## carbodyh         -0.1114226  0.0519136  -2.146  0.03320 *
## carbodysw        -0.0032561  0.0531230  -0.061  0.95119
## enginetypeohc     0.1748643  0.0406462   4.302 2.79e-05 ***
## enginetypeohcv   -0.1530921  0.0662281  -2.312  0.02195 *
## drivewheelr       0.1494563  0.0364613   4.099 6.30e-05 ***
## symboling0        0.0782913  0.0384688   2.035  0.04331 *
## symboling<3       0.0092163  0.0406044   0.227  0.82070
## symboling3        0.1311169  0.0575542   2.278  0.02391 *
## compressionratio  0.0121197  0.0039684   3.054  0.00260 **
## stroke           -0.1429638  0.0431748  -3.311  0.00112 **
## boreratio        -0.2508623  0.0601045  -4.174 4.68e-05 ***
## highwaympg       -0.0082509  0.0036121  -2.284  0.02354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1486 on 178 degrees of freedom
## Multiple R-squared:  0.9192, Adjusted R-squared:  0.9115
## F-statistic: 119.2 on 17 and 178 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit2)
```



I risultati sembrano decisamente migliorati.

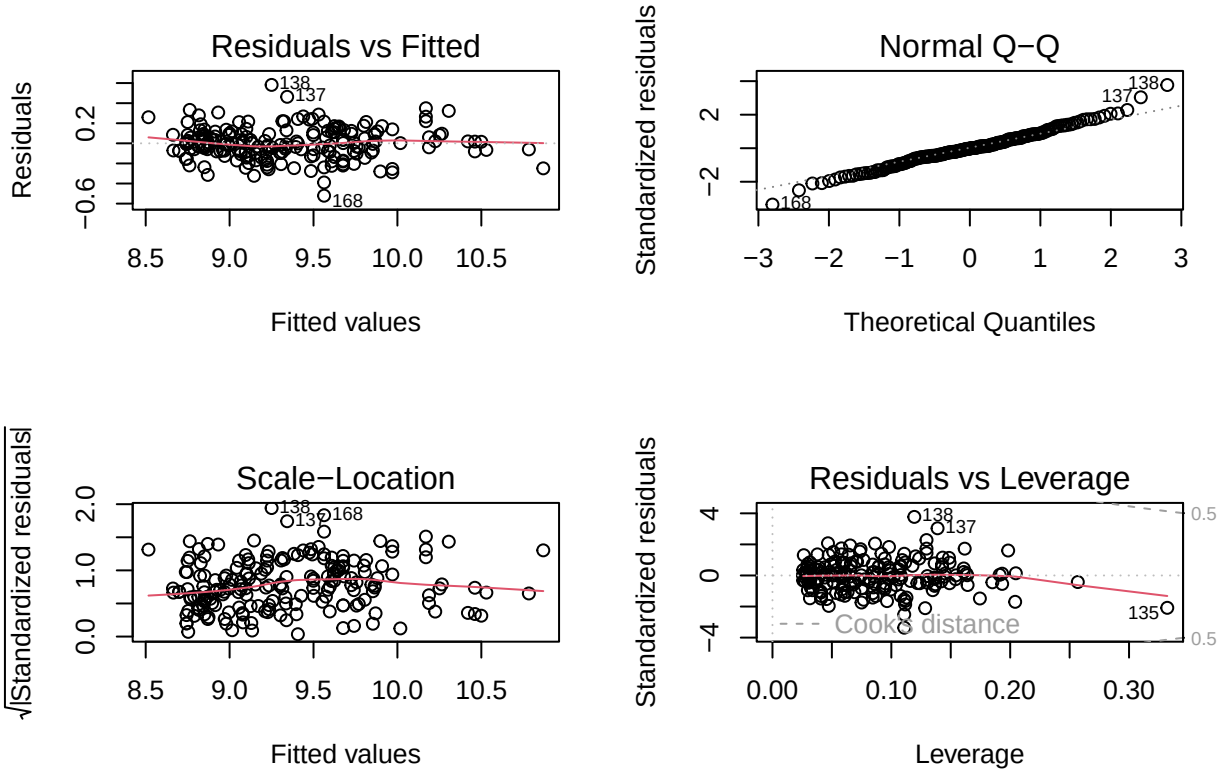
Possibili multicollinearità: horsepower e enginesize. Si rimuove horsepower in virtù del fatto che è meno correlata con price.

```
fit3 <- lm(log(price) ~ enginesize + fuelsystem +
    carwidth + carbody + enginetype + drivewheel + symboling +
    compressionratio + stroke + boreratio + highwaympg, data=d)
summary(fit3)
```

```
##
## Call:
## lm(formula = log(price) ~ enginesize + fuelsystem + carwidth +
##     carbody + enginetype + drivewheel + symboling + compressionratio +
##     stroke + boreratio + highwaympg, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52118 -0.08448 -0.00298  0.09274  0.58166
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.288976   0.750050   8.385 1.47e-14 ***
## enginesize        0.006541   0.000746   8.768 1.38e-15 ***
## fuelsystemim      0.185519   0.037273   4.977 1.51e-06 ***
## fuelsystems       0.046088   0.069774   0.661 0.509759
## carwidth          0.054837   0.010799   5.078 9.53e-07 ***
```

```
## carbodyh          -0.078197   0.057152  -1.368 0.172951
## carbodysw          0.021982   0.058615   0.375 0.708091
## enginetypeohc      0.184811   0.044939   4.112 5.96e-05 ***
## enginetypeohcv    -0.156839   0.073273  -2.140 0.033671 *
## drivewheelr        0.176824   0.040064   4.414 1.75e-05 ***
## symboling0         0.059076   0.042433   1.392 0.165587
## symboling<3        0.004918   0.044920   0.109 0.912941
## symboling3         0.197843   0.062630   3.159 0.001859 **
## compressionratio   0.003755   0.004147   0.905 0.366458
## stroke            -0.147865   0.047762  -3.096 0.002278 **
## boreratio         -0.250786   0.066501  -3.771 0.000221 ***
## highwaympg        -0.014667   0.003840  -3.819 0.000184 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1644 on 179 degrees of freedom
## Multiple R-squared:  0.9006, Adjusted R-squared:  0.8917
## F-statistic: 101.3 on 16 and 179 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit3)
```



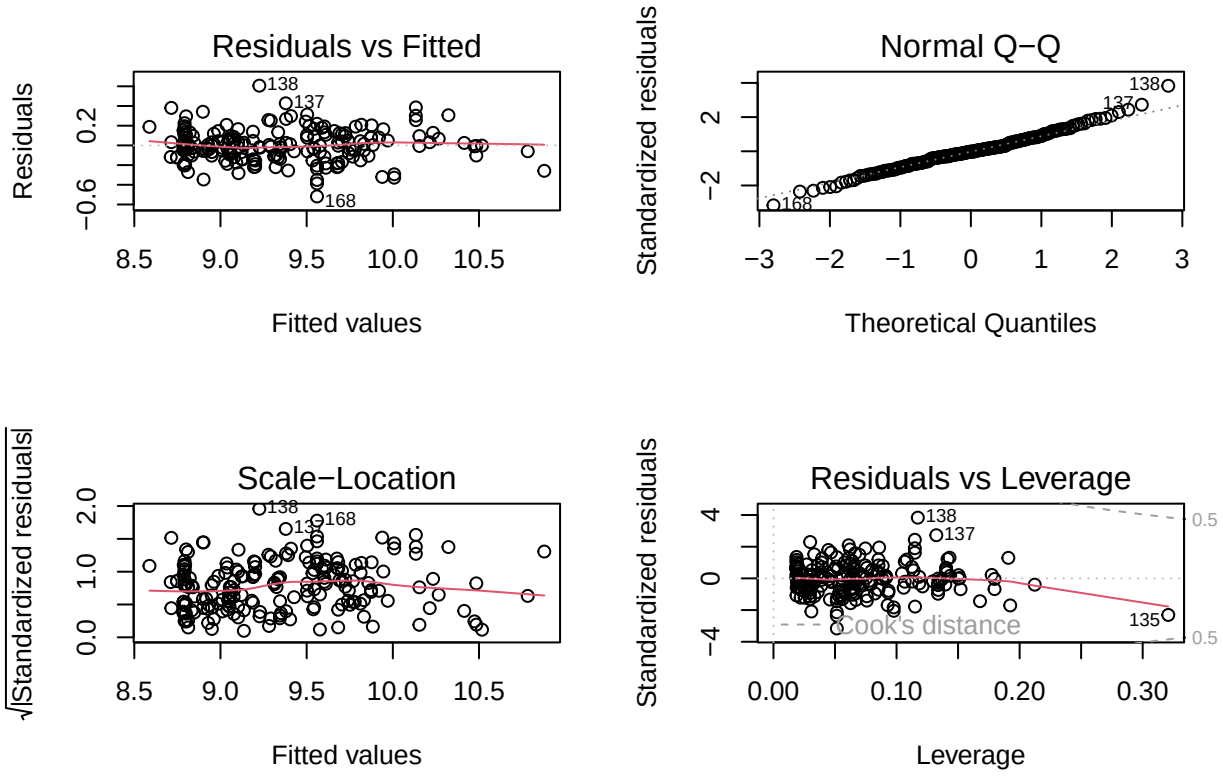
Ancora una volta carbody non è significativa nel modello.

```
fit4 <- lm(log(price) ~ enginesize + fuelsystem +
    carwidth + enginetype + drivewheel + symboling +
    stroke + boreratio + highwaympg, data=d)
```

```
summary(fit4)
```

```
##
## Call:
## lm(formula = log(price) ~ enginesize + fuelsystem + carwidth +
##     enginetype + drivewheel + symboling + stroke + boreratio +
##     highwaympg, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51832 -0.10345 -0.00372  0.09681  0.60641
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.1434610  0.7130589   8.616 3.28e-15 ***
## enginesize      0.0067925  0.0007429   9.143  < 2e-16 ***
## fuelsystemim    0.2207848  0.0352021   6.272 2.53e-09 ***
## fuelsystems     0.0480827  0.0706711   0.680 0.497133
## carwidth        0.0564719  0.0102368   5.517 1.17e-07 ***
## enginetypeohc   0.1805417  0.0455640   3.962 0.000106 ***
## enginetypeohcv -0.1885309  0.0744007  -2.534 0.012120 *
## drivewheelr     0.1731969  0.0402420   4.304 2.74e-05 ***
## symboling0      0.0597895  0.0433626   1.379 0.169640
## symboling<3    -0.0259203  0.0437753  -0.592 0.554505
## symboling3      0.1273661  0.0559233   2.278 0.023918 *
## stroke         -0.1568378  0.0482172  -3.253 0.001362 **
## boreratio      -0.2479127  0.0678301  -3.655 0.000336 ***
## highwaympg     -0.0127273  0.0031262  -4.071 6.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1684 on 182 degrees of freedom
## Multiple R-squared:  0.894,  Adjusted R-squared:  0.8864
## F-statistic: 118.1 on 13 and 182 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit4)
```



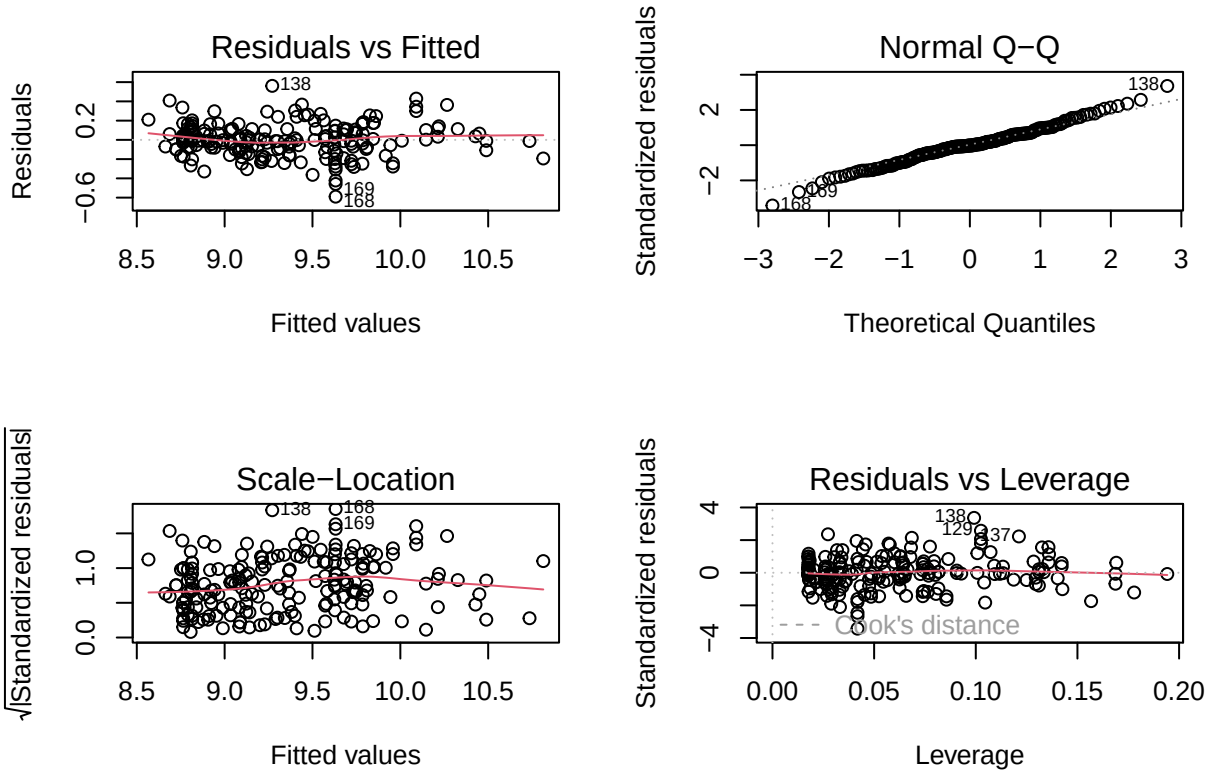
Le variabili stroke e boreratio sono poco correlate con price, sebbene la loro significatività nel modello si decide di toglierle per ridurre la complessità del modello.

```
fit5 <- lm(log(price) ~ enginesize + fuelsystem + carwidth +
    enginetype + drivewheel + symboling +
    highwaympg, data = d)
summary(fit5)
```

```
##
## Call:
## lm(formula = log(price) ~ enginesize + fuelsystem + carwidth +
##     enginetype + drivewheel + symboling + highwaympg, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58991 -0.09493 -0.00328  0.10267  0.56114
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.2912125  0.7124033   7.427 4.01e-12 ***
## enginesize       0.0053028  0.0006741   7.866 3.01e-13 ***
## fuelsystemim     0.2059591  0.0365869   5.629 6.66e-08 ***
## fuelsystems     -0.0289103  0.0696087  -0.415  0.67839
## carwidth         0.0524020  0.0106450   4.923 1.89e-06 ***
## enginetypeohc    0.1656197  0.0472304   3.507  0.00057 ***
## enginetypeohcv -0.0909356  0.0738408  -1.232  0.21970
## drivewheelr      0.1632804  0.0401893   4.063 7.17e-05 ***
```

```
## symboling0      0.0674015  0.0451037   1.494  0.13679
## symboling<3    -0.0053050  0.0453313  -0.117  0.90697
## symboling3      0.1636100  0.0573980   2.850  0.00486 **
## highwaympg     -0.0133609  0.0031302  -4.268 3.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1759 on 184 degrees of freedom
## Multiple R-squared:  0.8831, Adjusted R-squared:  0.8761
## F-statistic: 126.3 on 11 and 184 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit5)
```




```
shapiro.test(fit5$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit5$residuals
## W = 0.99015, p-value = 0.1995
```

Conclusioni

Il modello lineare ottenuto con fit5 appare soddisfacente: i residui rispettano le ipotesi, l'R^2 è dell'87% e gli standard error sono bassi. Si propone come modello interpretativo per il prezzo delle auto, tuttavia si consiglia di ampliare il dataset con nuove osservazioni per validare i risultati ottenuti.