PROGETTO STATISTICA COMPUTAZIONALE

di Millone A., Rossi S., Zanini V.

Il dataset scelto per questo progetto contiene 205 osservazioni analizzate su 26 variabili che descrivono determinate caratteristiche e fattori di numerose automobili attualmente in commercio; tutte queste variabili possono essere correlate al prezzo delle auto (variabile target); obiettivo dello studio è trovare quali variabili siano significative nella previsione del prezzo di un'auto e quanto bene queste variabili descrivono il prezzo di un'auto.

Il nostro dataset è composto dalle seguenti **26** variabili:

1. Car_ID: id unico di ogni osservazione;
2. Symboling (categoriale): è il rischio assicurativo assegnato ad ogni auto, +3 indica che l'auto è rischiosa, -3 che probabilmente è abbastanza sicura;
3. CarName (categoriale): nome del modello dell'auto;
4. **Fueltype** (categoriale): nome del tipo di carburante**;**
5. Aspiration (categoriale): tipo di aspirazione dell'auto;
6. **Doornumber** (categoriale): numero di portiere;
7. Carbody (categoriale): tipo di carrozzeria;
8. Drivewheel (categoriale): tipo di ruota mortice;
9. Enginelocation (categoriale): locazione del motore;
10. **Wheelbase** (numerica): passo della macchina (distanza tra ruota anteriore e posteriore);
11. Carlength (numerica): lunghezza dell'auto;
12. Carwidth (numerica): larghezza della macchina;
13. Carheight (numerica): altezza della macchina;
14. Curbweight (numerica): peso dell'auto senza passeggeri o bagagli;
15. Enginetype (categoriale): tipo di motore;
16. **Cylindernumber** (categoriale): numero di cilindri;
17. Enginesize (numerica): dimensione dell'auto;
18. Fuelsystem (categoriale): sistema di alimentazione;
19. Boreratio (numerica): rapporto tra il diametro dell'alesaggio del cilindro e la lunghezza della corsa del pistone;
20. Stroke (numerica): cilindrata dell'auto;
21. Compressionratio (numerica): rapporto di compressione, cioè rapporto tra il volume del cilindro e la camera di combustione;
22. Horsepower (numerica): cavalli della macchina;
23. Peakrpm (numerica): picco di giri per minute;
24. Citympg (numerica): numero di miglia percorsi in città;
25. **Highwaympg** (numerica): numero di miglia percorsi in autostrada**;**
26. Price (variabile target): prezzo del veicolo in €.

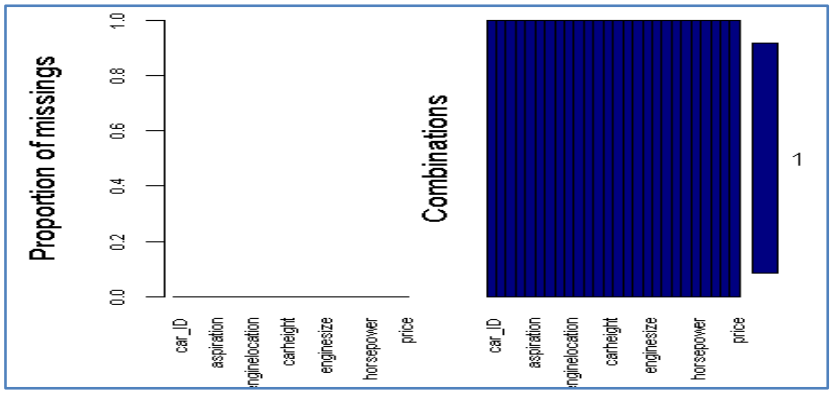
Iniziamo eliminando le seguenti **5** variabili che riteniamo non identificative per lo studio del modello: car_ID, carName, enginetype, enginelocation e doornumber. Ad esempio CarName viene eliminata perchè è una variabile categoriale con un grandissimo numero di livelli (vari modelli delle auto), altre variabili come enginelocation o doornumber riteniamo che siano parametri tecnici dell'auto poco utili ai fini dell'interpretazione o non discriminanti per il prezzo.

Per creare un modello robusto eseguiamo i seguenti passaggi in ordine:

1. Controllo dati mancanti ed eventuale imputazione;
2. Eliminare la collinearità tra i dati;
3. Trasformazione di Box-Cox per il target ed eventuali trasformazioni per le covariate;
4. Model Selection;
5. Outliers e punti influenti;
6. Eteroschedasticità;
7. Robust inference, Bootstrap;
8. Valutazioni finali, modello iniziale vs modello finale.

DATI MANCANTI e STATISTICHE DESCRITTIVE

Come si può osservare dal grafico qui di seguito nel dataset che abbiamo considerato non ci sono dati mancanti.



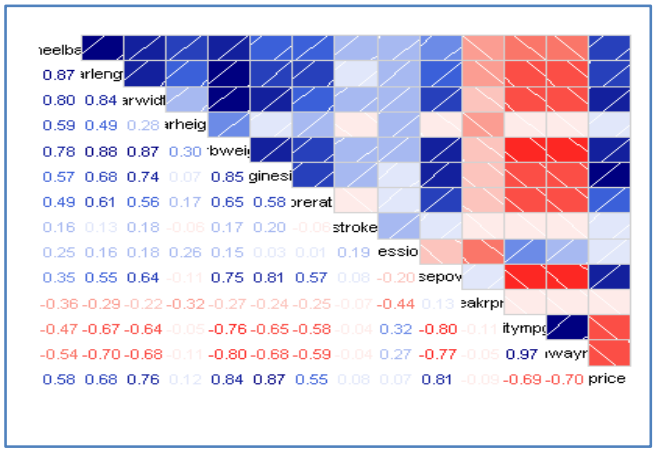
Inoltre da una prima analisi si nota che nessuna variabile factor ha un solo livello e nessuna variabile numerica ha varianza nulla.

Dalle statistiche descrittive si osserva che la lunghezza della macchina, il rapporto tra il diametro dell'alesaggio del cilindro e la lunghezza della corsa del pistone, il rapporto di compressione e la cilindrata e il numero di miglia percorsi in città e in autostrada hanno media e mediana coincidenti. Si vede che il peso dell'auto senza passeggeri o bagagli, il picco di giri per minuto e il prezzo hanno alta deviazione standard. Notiamo invece forte asimmetria per il prezzo, il rapporto di compressione e i cavalli delle auto.

Dalla stima del modello con tutte le **20** variabili incluse possiamo osservare l'R^2 è molto elevato (0.99 vicino a 1) ma ciò è dato proprio dal numero elevato di covariate. Tante variabili non sono significative, molto probabilmente questo è dovuto alla presenza di alta collinearità fra le singole covariate.

COLLINEARITÀ

La collinearità rappresenta un problema per la stima dei parametri e per gli standard error. Una prima idea sulla collinearità si può avere guardando il grafico delle correlazioni, anche se questo ultimo non ci dà una visione completa poichè alcune relazioni potrebbero essere nascoste.



Come da attese, nella matrice di correlazioni (riportata qui a fianco) si osservano alti valori positivi (cioè forte correlazione) tra le covariate che misurano le dimensioni del veicolo: tra "wheelbase" (distanza tra le due ruote) e "carlenght" (0.87), tra "carlength" e "curbweight" (peso veicolo senza passeggeri e bagagli) (0.88) e tra "carwidth" e "curbweight" (0.87). Si può notare anche alta correlazione (0.87) tra "enginesize" e la variabile target prezzo. Si osservano poi alti valori negativi fra la variabile "horsepower" e le variabili "citympg" (0.80) e "highwaympg" (0.77). Le miglia percorse in autostrada sono anche fortemente correlate negativamente (0.80) con il peso del veicolo senza bagagli e passeggeri.

La cosa più corretta è calcolare le metriche VIF e TOL (VIF è una misura dell'inflazione della varianza, TOL misura la tolleranza è la parte di varianza di una covariata non spiegata dalle altre covariate) ed eliminare le covariate con valori fuori soglia, VIF (<5) e TOL (>0.3), cioè quelle che risultano collineari.
Nel caso in cui le covariate con valori fuori soglia siano troppo numerose, come si registra nel nostro caso, si può utilizzare la model selection, che tramite lo step AIC, propone il miglior modello che elimina la multicollinearità. Successivamente si valuterà nuovamente se è presente ancora multicollinearità e quindi si potrà procedere ad escludere ulteriori variabili, utilizzando VIF e TOL nel caso di variabili numeriche e il chi quadro normalizzato per le variabili categoriali.

MODEL SELECTION

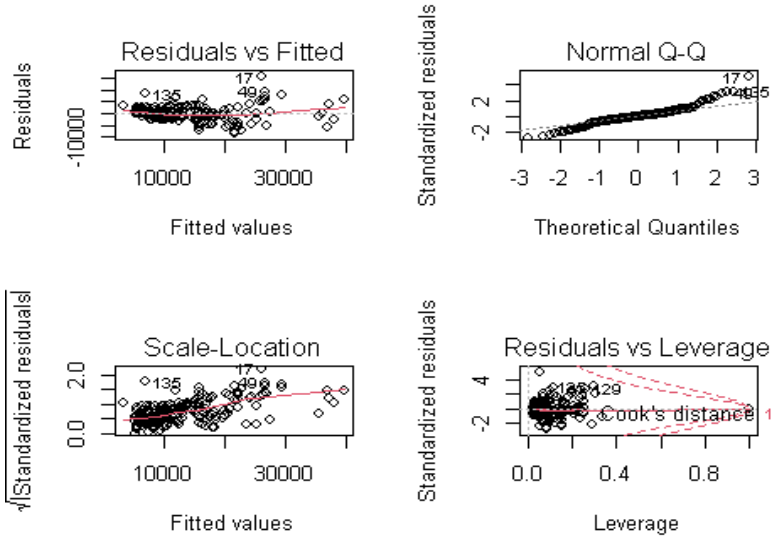
Nella procedura Model Selection vengono selezionate le migliori covariate da inserire con la procedura iterativa STEPAIC che sfrutta il criterio di informazione di Akaike (misura robusta).
Applicando questa procedura al nostro dataset si ottiene:

```
## price ~ carbody + drivewheel + cylindernumber + aspiration +
##     carwidth + carheight + enginesize + boreratio + stroke +
##     peakrpm + citympg + highwaympg
##                Df Sum of Sq        RSS    AIC F value              Pr(>F)
## <none>                      1345206247 3261.8
## carbody         4 151531525 1496737772 3275.7  5.1535           0.0005884 ***
## drivewheel      2  77648752 1422855000 3269.3  5.2816           0.0058882 **
## cylindernumber  6 361259918 1706466165 3298.6  8.1909 0.00000007405049682 ***
## aspiration      1 170951169 1516157417 3284.4 23.2560 0.00000297366289257 ***
## carwidth        1  48621137 1393827384 3267.1  6.6144           0.0109098 *
## carheight       1  65679501 1410885749 3269.6  8.9349           0.0031822 **
## enginesize      1 499170388 1844376635 3324.5 67.9065 0.00000000000003212 ***
## boreratio       1  14098981 1359305228 3262.0  1.9180           0.1677627
## stroke          1 133631931 1478838178 3279.3 18.1791 0.00003217321044547 ***
## peakrpm         1 203130536 1548336783 3288.7 27.6336 0.00000040578574554 ***
## citympg         1  36253025 1381459272 3265.3  4.9318           0.0275937 *
## highwaympg      1  73022231 1418228479 3270.7  9.9338           0.0018963 **
```

Dopo 10 iterazioni si è individuato il modello con minore AIC (AIC=3261.84 dall'iniziale 3287.07 del modello con tutte le covariate). Il modello all'ultimo step include le seguenti **12** covariate: carbody, drivewheel, cylindernumber, aspiration, carwidth, carheight, enginesize, boreratio, stroke, peakrpm, citympg e highwaympg.
Quindi si procede valutando di nuovo TOL e VIF delle numeriche; si può notare che il numero delle covariate da eliminare sia diminuito notevolmente. Infatti solo le variabili **highwaympg ,citympg** e **enginesize** hanno valori di TOL e VIF che superano i valori soglia. **Highwaympg e Citympg** sono logicamente collineari poichè indicano il numero di km rispettivamente percorsi in autostrada e in città. Dopo l'eliminazione di **highwaympg**, si ricalcolano i valori di VIF e TOL; si nota nuovamente che le metriche migliorano e quindi si procede ad eliminare solo **enginesize.** A questo punto si può notare che tutte le covariate rispettano le soglie previste sia per VIF (<5) sia per il TOL (>0.3).
Procedendo a valutare la statistica chi quadro normalizzato per i nostri fattori, si osserva che nessun valore è maggiore di 0.8, all'opposto siamo in presenza di valori tutti piuttosto piccoli. Per questo motivo possiamo affermare che tra nessuno dei nostri fattori c'è forte associazione e possiamo quindi decidere di includerli tutti nel nostro modello.

Fittiamo ora il modello con le **10** covariate che abbiamo mantenuto:

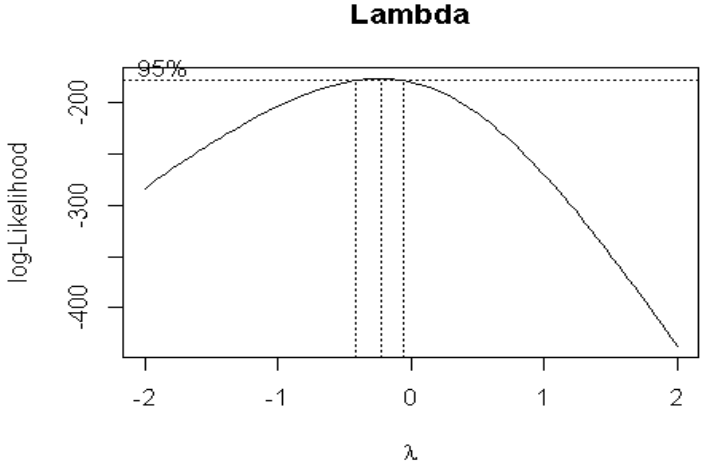
```
starting_model3 <- lm(price ~ carbody + drivewheel + cylindernumber + aspiration +
                      carwidth + carheight + boreratio + stroke +
                      peakrpm + citympg ,data=b_ms3)
```



Dalle diagnostiche si osserva: 1. una non linearità tra residui e valori fittati e i punti sono distribuiti sulla linea in modo non casuale; 2. il grafico Q-Qplot mostra una scarsa normalità (non allineamento sulla retta tra quantili teorici ed osservati) ,3. eteroschedasticita elevata,4. presenza di punti influenti

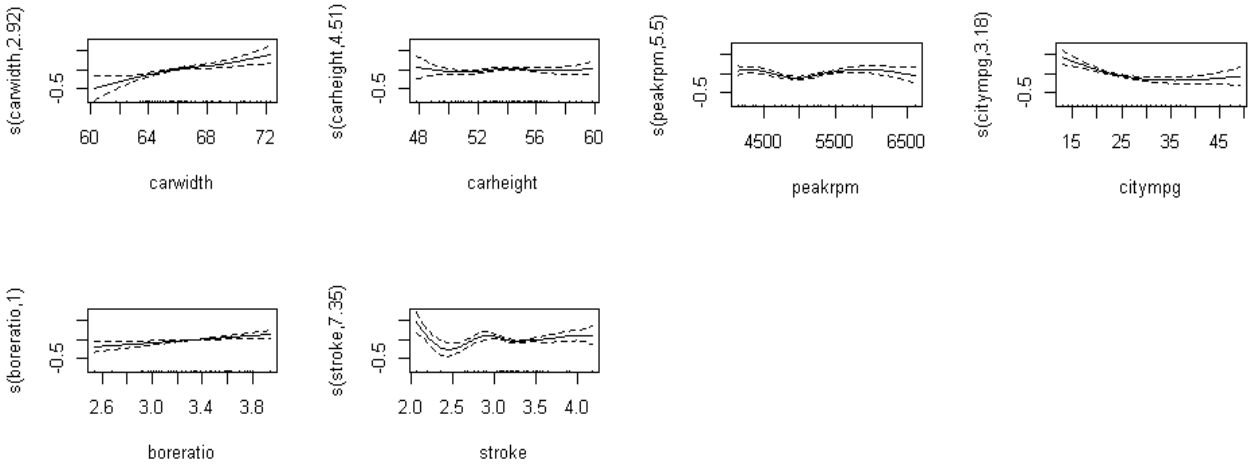
TRASFORMAZIONE BOX-COX FOR Y AND GAM FOR X

Per migliorare la non linearità ed eteroschedasticità del nostro modello si possono effettuare delle trasformazioni sul target oppure sulle covariate (anche su entrambe). La trasformazione Box-Cox consiste in una trasformazione della target, che permette ai dati di seguire una distribuzione normale, così da ridurre la non linearità. Il metodo consiste nel trovare un *lambda* (un parametro) adatto come esponente della target, viene scelto la trasformazione con Lambda più adatto, cioè che minimizza MSE.



In questo caso si ottiene un lambda pari a -0.222 , che non risulta adeguato e adatto allo studio; quindi si può optare per una trasformazione log-lineare. Inoltre si può anche provare ad applicare le trasformazioni analitiche sulle covariate, che permettono di visualizzare e verificare i loro andamenti a livello grafico.

Le covariate d'interesse sono 6 :1. larghezza auto; 2. altezza auto ; 3.rapporto tra le dimensioni del diametro dell'alesaggio del cilindro del motore e la lunghezza della corsa del pistone (BoreRatio); 4.volume del motore ; 5.giri di picco dell'auto ; 6.miglia in città



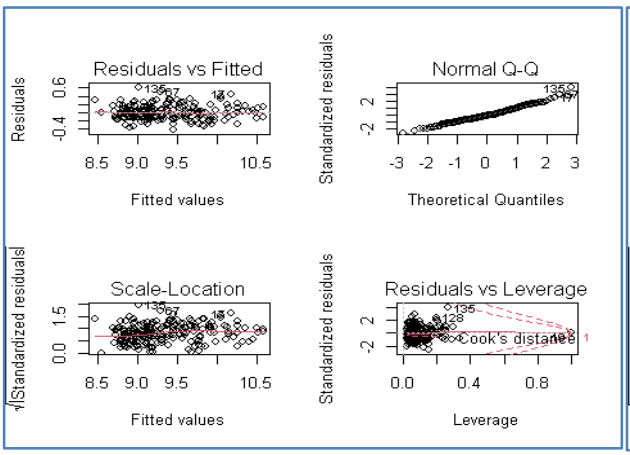
Dai grafici si nota che la variabile BoreRatio ha un andamento lineare, invece le altre variabili (larghezza auto , altezza auto , volume del motore, giri di picco dell'auto e miglia in citta) presentano andamenti non lineari. Inoltre dal confronto d'ipotesi si nota:

```
anova(starting_model_log, gam_modlog, test="LRT")
```

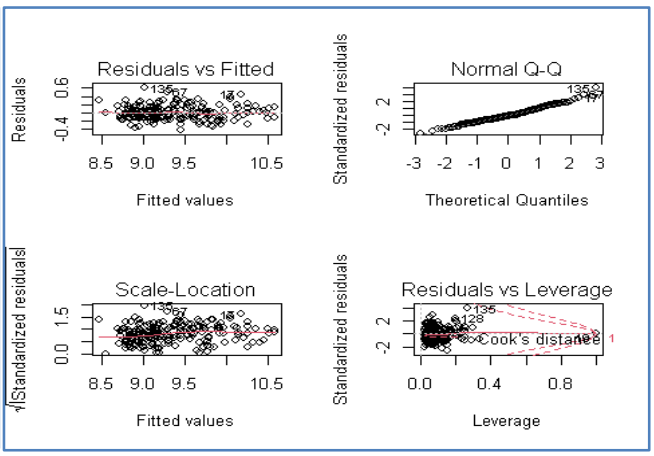
```
## Analysis of Variance Table
## Model 1: log(price + 1) ~ carbody + drivewheel + cylindernumber + aspiration +
##     carwidth + carheight + boreratio + stroke + peakrpm + citympg
## Model 2: log(price + 1) ~ carbody + drivewheel + cylindernumber + aspiration +
##     s(carwidth) + s(carheight) + s(boreratio) + s(stroke) + s(peakrpm) +
##     s(citympg)
##   Res.Df    RSS     Df Sum of Sq              Pr(>Chi)
## 1 185.00 5.7741
## 2 166.55 3.3786 18.449    2.3955 < 0.00000000000000022 ***
```

Il test dell'anova risulta significativo , questo significa che è corretto fare una trasformazione per le covariate. Per le 6 covariate d'interesse si è optato per una trasformazione con effetti quadratici.
Qui di seguito si possono osservare le diagnostiche relative al modello log-lineare ed al modello log-effetti quadratici.

Modello log-lineare



Modello log-effetti quadratici



Dal confronto tra le diagnostiche dei due modelli (modello log-lineare e modello log-effetti quadratici) non si notano differenze rilevanti, perché già nella trasformazione logaritmica sulla target si verifica un forte miglioramento dei principali aspetti del modello robusto (linearità ,eteroschedasticità , normalità e presenza dei punti influenti) . Infatti per entrambi i modelli si osserva: 1. una ottima linearità tra residui e valori fittati e punti distribuiti sulla linea (la funzione interpolante rossa è molto appiattita); 2. il grafico Q-Qplot mostra una buona normalità (allineamento/coincidenza sulla retta tra quantili teorici ed osservati) ,3. eteroschedasticità tende a ridursi sensibilmente (ma le varianze non sono ancora tutte costanti),4. poca presenza di punti influenti.

i modelli mostrano lievi differenze nei valori R^2_{adj} e Residual std.error

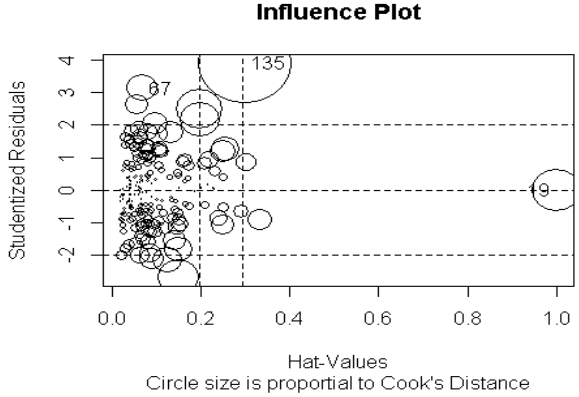
per il modello log-lineare	modello log-effetti quadratico
R^2_{adj}è 0.877, cioè	R^2_{adj}0.8739, cioè
spiega 87.70% della varianza di Prezzo	spiega 87.39% della varianza di Prezzo
Residual std. error 0.1767	Residual std.error 0.1789
cioè che la previsione del prezzo	cioè che la previsione del prezzo
risulta errata di 0.1767€	risulta errata di 0.1789€

Alcune considerazioni sugli effetti del modello log-lineare:

- Una variazione unitaria del rapporto tra le dimensioni del diametro dell'alesaggio del cilindro del motore e la lunghezza della corsa del pistone (bore ratio) determina una variazione del prezzo dell'auto pari al 22.95%
- Per un'auto con una cilindrata pari a 4000cc (stroke=4.170) si ottiene un valore di elasticità di 11.92,che significa che all'incremento dell'1% della cilindrata del motore si stima un incremento del 11.92% del prezzo dell'auto
- Considerando una macchina con prezzo 50000€, all'aumentare della cilindrata dell'auto (stroke), si stima un incremento del prezzo di 1490€.
- Passando da aspirazione 'std' ad aspirazione 'turbo' si determina una variazione del prezzo di 16.09% ,le aspirazione turbo hanno un prezzo più alto del 16.09% rispetto alle aspirazioni std.

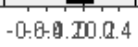
PUNTI INFLUENTI

Per irrobustire il modello si deve procedere all'eliminazione dei punti influenti nel modello log-lineare (i punti influenti possono determinare eteroschedasticità in quanto hanno alto residuo, cioè molta distanza tra valore osservato e previsto e un valore di x molto lontano dalla media), ovvero di quelle osservazioni che presentano un valore della distanza di Cook superiore alla soglia stabilita dall'UCLA, ovvero 4/n-p. La rimozione di queste osservazioni cambierà drasticamente la pendenza del nostro modello.



Qui a fianco si riporta il grafico delle distanze di Cook. Come si può osservare nel grafico, l'osservazione 19 mostra valori elevati nelle principali caratteristiche ma con residuo attorno allo zero, questo vuol dire vuol dire che pur avendo caratteristiche generali dell'auto buone non riesce a sfruttarle appieno, in quanto la stima del prezzo non rispecchia le sue potenzialità. Invece le osservazioni 67 e 135, pur non avendo un alto valore nelle principali caratteristiche dell'auto, mostrano un alto valore del residuo; cioè presentano un prezzo di previsione maggiore rispetto a quello che si prevederebbe sulla base delle caratteristiche dell'auto.

Variable		N	Estimate	p
carbody	convertible	5	Reference	
	hardtop	4	-0.26 (-0.47, -0.05)	0.02
	hatchback	64	-0.40 (-0.55, -0.25)	<0.001
	sedan	94	-0.30 (-0.46, -0.15)	<0.001
	wagon	25	-0.41 (-0.58, -0.23)	<0.001
drivewheel	4wd	9	Reference	
	fwd	116	-0.03 (-0.15, 0.10)	0.68
	rwd	67	0.15 (0.02, 0.27)	0.02
cylindernumber	eight	5	Reference	
	five	11	-0.24 (-0.42, -0.06)	0.01
	four	152	-0.51 (-0.70, -0.31)	<0.001
	six	21	-0.11 (-0.29, 0.08)	0.25
	two	3	-0.47 (-0.73, -0.21)	<0.001
aspiration	std	155	Reference	
	turbo	37	0.15 (0.08, 0.21)	<0.001
carwidth		192	0.06 (0.04, 0.09)	<0.001
carheight		192	0.02 (0.00, 0.03)	0.01
boreratio		192	0.28 (0.12, 0.44)	<0.001
stroke		192	0.09 (0.00, 0.18)	0.04
peakrpm		192	0.00 (-0.00, 0.00)	0.07
citympg		192	-0.01 (-0.02, -0.01)	<0.001



Come riportato sopra, il modello log-lineare, filtrato dai punti influenti (presenti in 13 osservazioni), presenta un adattamento pari a 0.8979, cioè spiega l' 89,79% della varianza relativa al Prezzo; tutte le variabili risultano significative. Il nostro residual standar erroe è pari a 0.1538 , che significa che la previsione del prezzo potrà risultare errata di 0.1538€ (un valore di errore più contenuto rispetto al modello iniziale in cui era pari a 0.1787).

ETEROSCHEDASTICITÀ

È uno dei punti più importanti da controllare, nel caso in cui la varianza dei residui non sia costante (ossia non vi è omoschedasticità), gli stimatori non sono efficienti (vengono sovrastimati gli s.e.).
Al fine di correggere la stima errata degli s.e., si utilizza la correzione proposta da White, il quale usa come stimatore di σ_i^2, e_i^2 il quale non è altro che l'i-esimo residuo della regressione di y su tutte le covariate x.

Per prima cosa, svolgiamo il test di Breusch-Pagan per vedere se vi è eteroschedasticità o meno.

```
## BP = 34.175, df = 17, p-value = 0.007971
```

Abbiamo ottenuto un valore del p-value non ancora sufficientemente elevato per rifiutare l'ipotesi di non eteroschedasticità .
Svolgiamo ora il Test di White, anch'esso verifica la presenza di eteroschedasticità come il test effettuato sopra, con la differenza che esso è più conservativo e meno severo.

```
## Chisquare = 2.045399, Df = 1, p = 0.15267
```

In questo caso vi è sufficiente evidenza empirica per rifiutare l'ipotesi nulla (presenza di eteroschedasticità). Nonostante i buoni risultati ottenuti con il Test di White, procediamo a correggere gli standard error con la formula degli s.e. corretti di White, in modo da ottenere un risultato migliore nel test di Breusch-Pagan.

Dopo aver calcolato gli s.e. corretti di White, osserviamo dei valori di s.e. maggiori rispetto ai valori precedenti alla correzione.

BOOTSTRAP

Svolgiamo ora l'ultimo step della procedura per ottenere un modello robusto. In questo passaggio si verificano i parametri per la stima puntuale e la significatività delle covariate del modello.

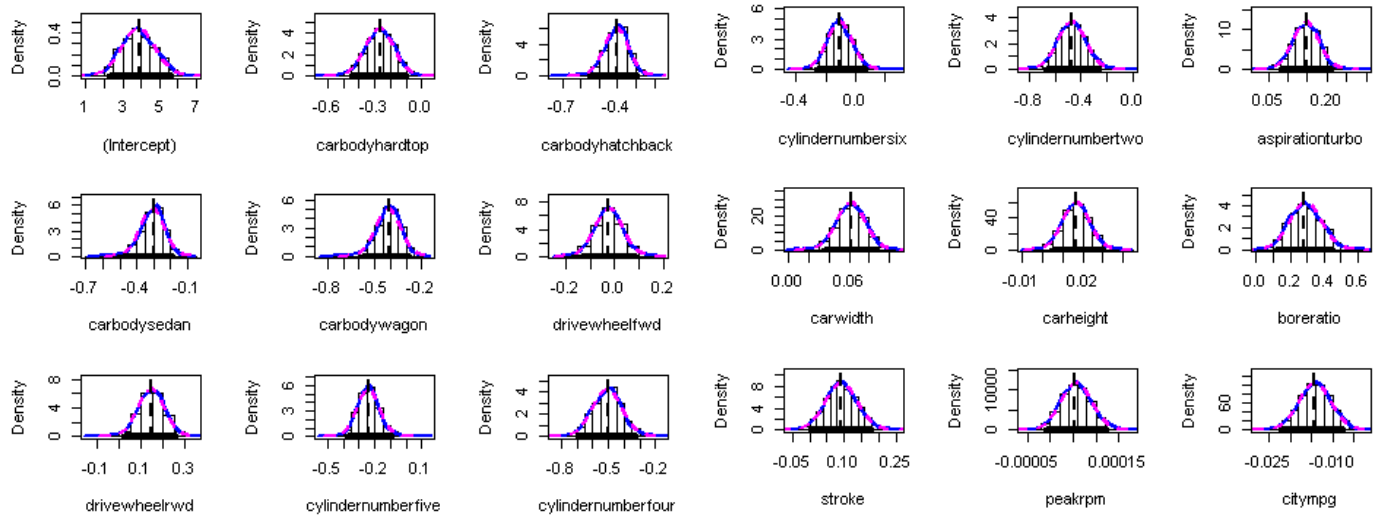
Questa procedura consiste nello stimare r volte (noi abbiamo svolto r=2000 ripetizioni) il modello usando come campione la popolazione stessa; si ricorda che ogni campione è estratto con reinserimento.

Questa procedura ci assicura una stima robusta del modello anche in caso in cui non tutte le assunzioni del modello siano rispettate.

```
BOOT.MOD_lm5_b2=Boot(lm5_boot2, R=1999)
summary(BOOT.MOD_lm5_b2, high.moments=TRUE)
```

Dopo aver usato la procedura Bootstrap, creiamo i grafici per verificare la significatività delle variabili. Si ricorda che la significatività degli intervalli di confidenza è da controllare prevalentemente per le variabili numeriche, mentre non è affidabile per le variabili fattoriali.

```
ci_perc <- Confint(BOOT.MOD_lm5_b2, level=c(.95), type="perc")
hist(BOOT.MOD_lm5_b2, legend="separate")
```



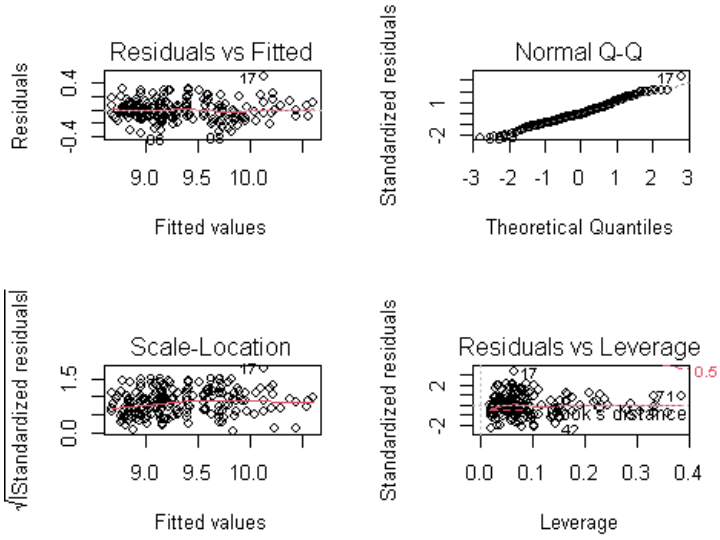
Si osserva che vi sono due variabili non significative per Bootstrap: **"stroke"** e "**peakrpm**".
Proviamo a stimare tre modelli: il primo senza la variabile "stroke", il secondo senza la variabile "peakrpm" e il terzo senza entrambe le variabili.

Poi mettiamo a confronto i grafici di diagnostica e il valore di R^2_{adj}.

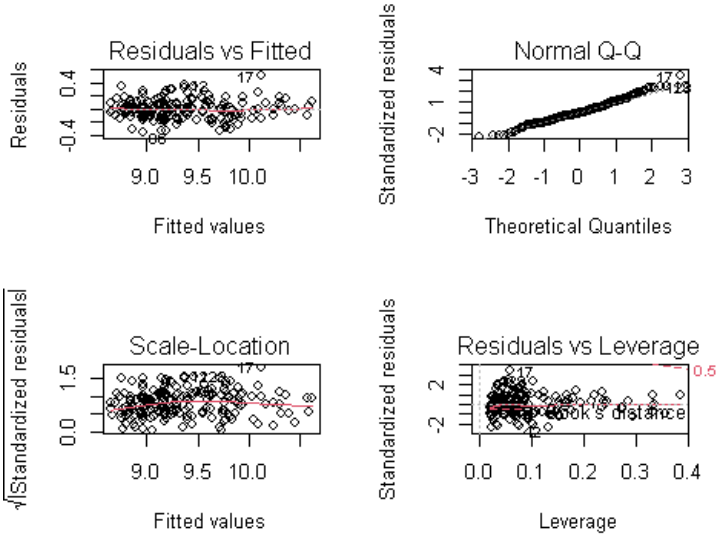
```
## Multiple R-squared:  0.9047, Adjusted R-squared:  0.896   #no stroke
## Multiple R-squared:  0.9052, Adjusted R-squared:  0.8966  #no peakrpm
## Multiple R-squared:  0.9033, Adjusted R-squared:  0.8951  #no stroke e peakrpm
```

Osserviamo che non vi sono sostanziali differenze tra i vari modelli, quello con R^2_{adj} maggiore è il modello che mantiene le due variabili non significative per Bootstrap ($R^2_{adj} = 0.898$).
Osserviamo ora i plot di diagnostica dei diversi modelli per vedere se vi sono delle differenze sostanziali tra i modelli.

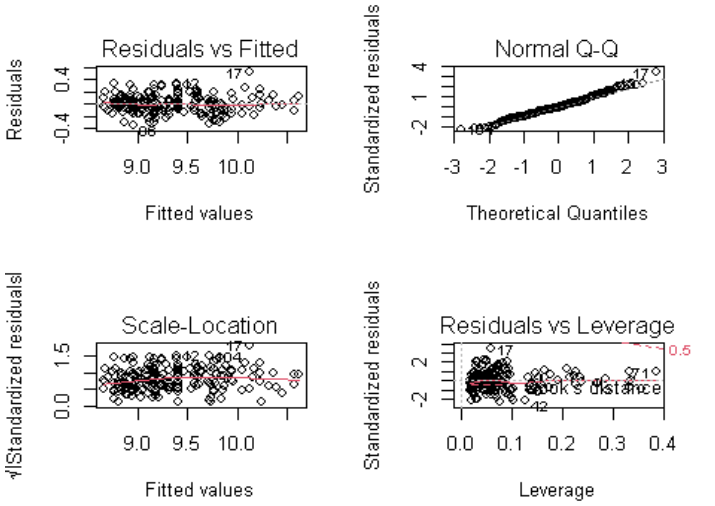
No Stroke



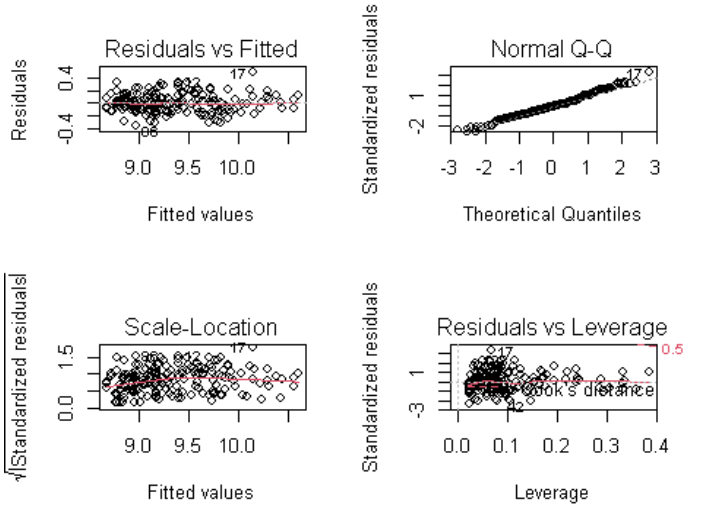
No PeakRpm



No Stroke e no PeakRpm



Modello con le due variabili



Non osserviamo nessuna differenza sostanziale nei grafici di diagnostica, il modello che sembra essere migliore è quello che mantiene le due variabili, quindi scegliamo quello come modello migliore.

Svolgiamo l'ultima analisi circa tutte le assunzioni principali del modello.

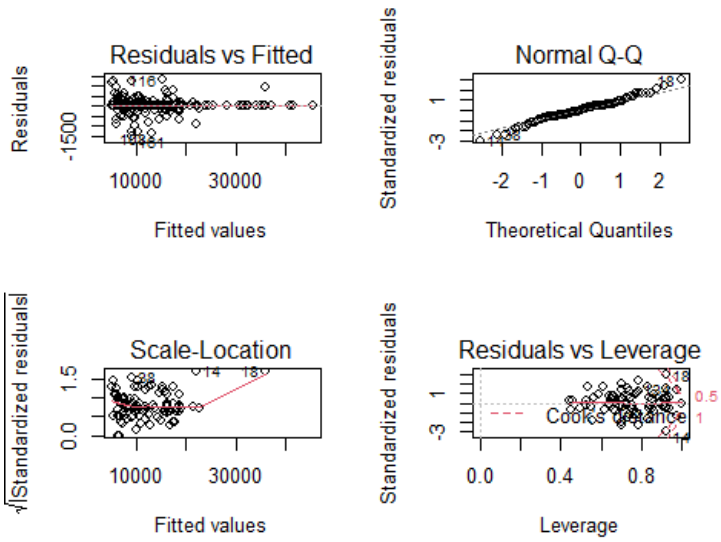
```
library(gvlma)
gvlma(lm5_boot2)
```

```
##                     Value p-value                Decision
## Global Stat        7.9332 0.09406 Assumptions acceptable.
## Skewness           3.1624 0.07535 Assumptions acceptable.
## Kurtosis           0.1642 0.68531 Assumptions acceptable.
## Link Function      1.5836 0.20824 Assumptions acceptable.
## Heteroscedasticity 3.0229 0.08209 Assumptions acceptable.
```

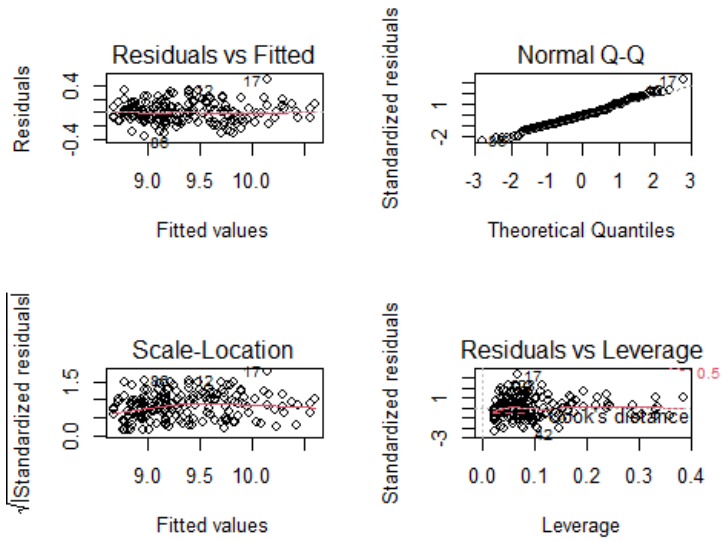
Tutte le assunzioni sono accettate, possiamo ritenere questo modello il migliore e, quindi, il modello definitivo.

CONFRONTO MODELLO INIZIALE VS MODELLO FINALE

MODELLO INIZIALE

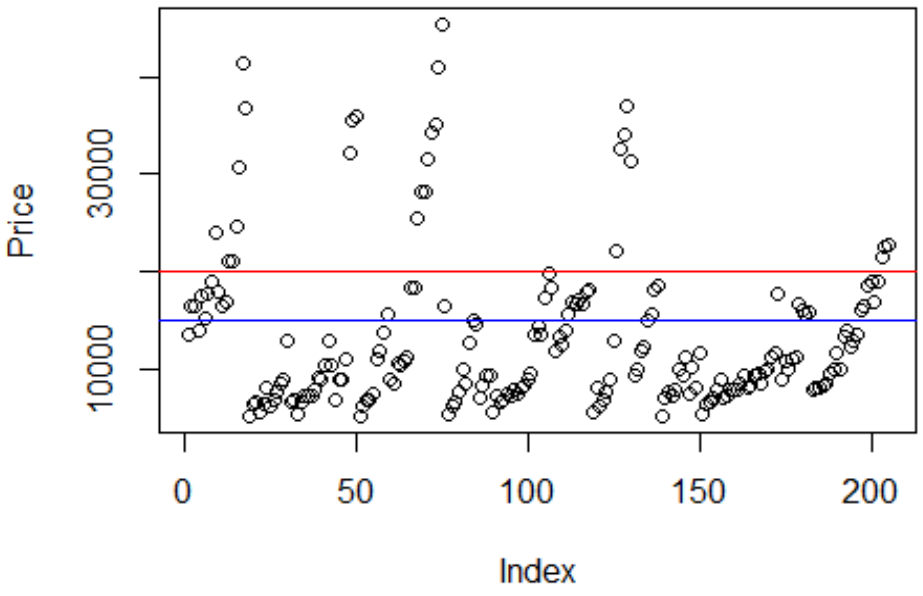


MODELLO FINALE



Osserviamo delle notevoli differenze in tutti e quattro i grafici di diagnostica, con le procedure utilizzate abbiamo soprattutto migliorato la linearità e l'eteroschedasticità dei residui.

MODELLO LOGISTICO



Abbiamo plottato tutte le unità per decidere quale soglia utilizzare per dividere le auto in "economiche" e "costose". Abbiamo scelto il valore 15,000$, così facendo otteniamo 141 unità con y=0 (economiche) e 64 unità con y=1 (costose). Creiamo la variabile target e stimiamo il primo modello.
Il modello non dà errore (l'algoritmo converge dopo 14 iterazioni), ma l'output non sembra essere molto affidabile poiché tutte le variabili risultano significative, hanno tutte lo stesso p-value (molto basso).

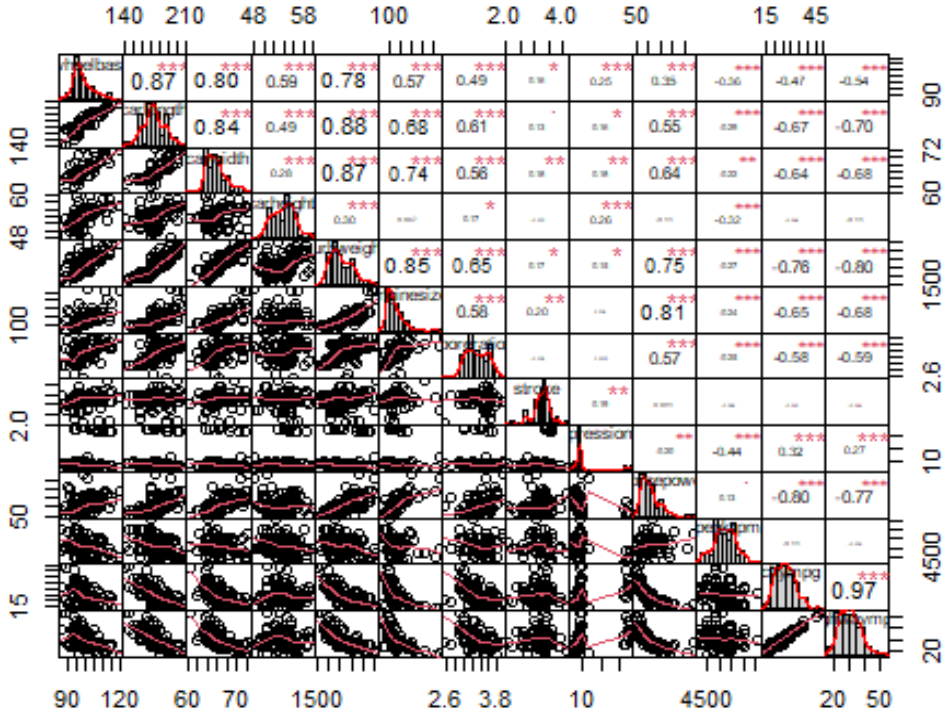
Iniziamo a controllare le varie problematiche, per capire quale sia la causa.
Prima di tutto verifichiamo che le variabili factor non abbiano un solo livello, poi calcoliamo la varianza delle variabili numeriche.

```
##      symboling       fueltype     aspiration        carbody     drivewheel
##              6              2              2              5              3
## cylindernumber     fuelsystem         target
##              7              8              2

##        wheelbase        carlength         carwidth        carheight
##      36.26178240    152.20868819      4.60189957      5.97079962
##       curbweight      enginesize       boreratio          stroke
##   271107.87431851  1734.11391679      0.07335631      0.09834309
## compressionratio      horsepower         peakrpm         citympg
##      15.77710432  1563.74112865 227515.30368245     42.79961741
##       highwaympg
##      47.42309900
```

Nessuna variabile presenta zero variance, il problema è da un'altra parte.

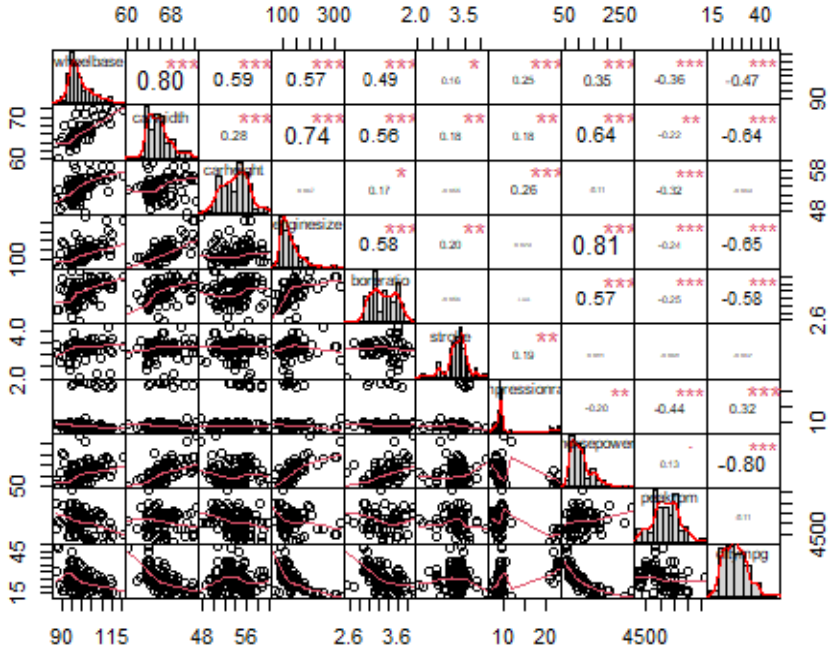
Analizziamo la correlazione fra le covariate.



Si osservano valori molto alti tra la variabile "ciytmpg" e "highwaympg", giustamente correlate poiché rappresentano rispettivamente il numero di km fatti in città e in autostrada; si decide di eliminare la variabile "highwaympg".

Altre variabili fortemente correlate sono quelle relative alle dimensioni della macchina; wheelbase (passo della macchina), carlength (lunghezza della macchina), carweight(peso della macchina), carwidth (larghezza della macchina). Si decide di eliminare la variabile peso e la variabile lunghezza.

Ricalcoliamo le correlazioni, dopo aver eliminato le variabili maggiormente correlate per vedere se la situazione è migliorata.



Si osservano ancora alcuni valori elevati di correlazione (tra 'horsepower' e 'citympg' e 'enginesize'). Quindi, si decide di eliminare le variabili 'horsepower' e 'carwidth' poichè fortemente correlate con altre variabili.

Verifichiamo il valore di VIF e TOL delle covariate numeriche rimaste.

```
##                      VIF    TOL      Wi      Fi Leamer    CVIF Klein   IND1
## Xwheelbase        3.2129 0.3112 62.2785 73.0271 0.5579 -5.3498     1 0.0111
## Xcarheight        1.9533 0.5119 26.8297 31.4601 0.7155 -3.2524     0 0.0182
## Xenginesize       2.6743 0.3739 47.1189 55.2511 0.6115 -4.4528     1 0.0133
## Xboreratio        1.9833 0.5042 27.6724 32.4484 0.7101 -3.3023     0 0.0179
## Xstroke           1.1729 0.8526  4.8645  5.7041 0.9234 -1.9529     0 0.0303
## Xcompressionratio 1.6212 0.6168 17.4832 20.5006 0.7854 -2.6995     0 0.0219
## Xpeakrpm          1.6127 0.6201 17.2443 20.2205 0.7874 -2.6853     0 0.0220
## Xcitympg          3.1441 0.3181 60.3402 70.7543 0.5640 -5.2351     1 0.0113
```

Tutte le covariate rispettano le soglie previste sia per VIF (<5) sia per il TOL (>0.3).

Stimiamo il modello dopo aver tolto le variabili fortemente correlate, testiamo la significatività delle variabili con il test LRT.

```
drop1(glm2, test="LRT")
```

```
##                  Df Deviance    AIC     LRT Pr(>Chi)
## <none>                42.039 110.04
## symboling         5   55.724 113.72 13.6853 0.017737 *
## fueltype          0   42.039 110.04  0.0000
## aspiration        1   42.041 108.04  0.0024 0.960967
## carbody           4   54.208 114.21 12.1698 0.016132 *
## drivewheel        2   53.892 117.89 11.8533 0.002667 **
## wheelbase         1   45.877 111.88  3.8385 0.050088 .
## carheight         1   42.105 108.11  0.0662 0.797009
## cylindernumber    6   49.285 105.28  7.2467 0.298635
## enginesize        1   42.219 108.22  0.1802 0.671164
## fuelsystem        6   47.775 103.78  5.7364 0.453360
## boreratio         1   42.574 108.57  0.5357 0.464231
## stroke            1   48.145 114.14  6.1069 0.013465 *
## compressionratio  1   42.429 108.43  0.3904 0.532079
## peakrpm           1   42.915 108.92  0.8766 0.349134
## citympg           1   42.183 108.18  0.1447 0.703620
```

Il modello converge, ma vi sono molte variabili non significative (per il test LRT), si procede con il calcolare il chi quadro normalizzato fra le variabili factor al fine di controllare quali siano i valori maggiormente elevati ed eliminare quelle variabili.

	X1	Row	Column	Chi.Square	df	p.value	n	u1	u2	nMinu1u2	Chi.Square.norm
1	1	symboling	fueltype	14.662	5	0.012	205	5	1	205	0.071521884
2	2	symboling	aspiration	11.988	5	0.035	205	5	1	205	0.058476560
3	3	symboling	carbody	109.795	20	0.000	205	5	4	820	0.133896546
4	4	symboling	drivewheel	38.701	10	0.000	205	5	2	410	0.094393178
5	5	symboling	cylindernumber	55.856	30	0.003	205	5	6	1025	0.054493418
6	6	symboling	fuelsystem	105.805	35	0.000	205	5	7	1025	0.103223975
7	7	symboling	target	20.961	5	0.001	205	5	1	205	0.102250395
8	8	fueltype	aspiration	29.606	1	0.000	205	1	1	205	0.144418338
9	9	fueltype	carbody	10.129	4	0.038	205	1	4	205	0.049411438
10	10	fueltype	drivewheel	3.588	2	0.166	205	1	2	205	0.017501778
11	11	fueltype	cylindernumber	10.905	6	0.091	205	1	6	205	0.053194007
12	12	fueltype	fuelsystem	205.000	7	0.000	205	1	7	205	1.000000000
13	13	fueltype	target	1.313	1	0.252	205	1	1	205	0.006406527
14	14	aspiration	carbody	1.597	4	0.809	205	1	4	205	0.007790954
15	15	aspiration	drivewheel	4.857	2	0.088	205	1	2	205	0.023693807
16	16	aspiration	cylindernumber	13.864	6	0.031	205	1	6	205	0.067629333
17	17	aspiration	fuelsystem	83.000	7	0.000	205	1	7	205	0.404877617
18	18	aspiration	target	7.416	1	0.006	205	1	1	205	0.036175523
19	19	carbody	drivewheel	26.590	8	0.001	205	4	2	410	0.064854545
20	20	carbody	cylindernumber	27.784	24	0.269	205	4	6	820	0.033882858
21	21	carbody	fuelsystem	44.702	28	0.024	205	4	7	820	0.054514129
22	22	carbody	target	11.924	4	0.018	205	4	1	205	0.058166582
23	23	drivewheel	cylindernumber	57.892	12	0.000	205	2	6	410	0.141200930
24	24	drivewheel	fuelsystem	74.944	14	0.000	205	2	7	410	0.182789841
25	25	drivewheel	target	90.187	2	0.000	205	2	1	205	0.439934703
26	26	cylindernumber	fuelsystem	208.575	42	0.000	205	6	7	1230	0.169573195
27	27	cylindernumber	target	83.766	6	0.000	205	6	1	205	0.408614106
28	28	fuelsystem	target	75.678	7	0.000	205	7	1	205	0.369162729

Si osserva che vi è chi-quadro normalizzato pari a 1 tra le variabili “fuelsystem” e “fueltype”, si decide di eliminare la variabile ‘fuelsystem’ poiché ai fini interpretativi la variabile “fueltype” (tipo di carburante della vettura: benzina, diesel o gas) è più interessante e utile.

Ora ristimiamo il modello e analizziamo le significatività con il test LRT.

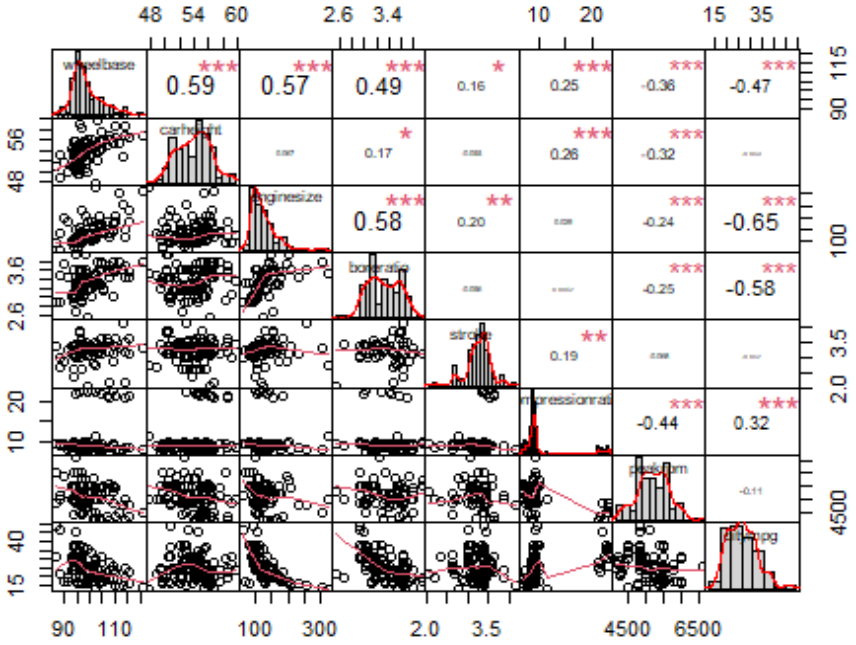
```
## <none>                47.775 103.775
## symboling         5   60.654 106.654 12.8788 0.0245411 *
## fueltype          1   47.808 101.808  0.0329 0.8560404
## aspiration        1   47.829 101.829  0.0545 0.8153976
## carbody           4   61.150 109.150 13.3750 0.0095818 **
## drivewheel        2   62.277 114.277 14.5025 0.0007093 ***
## wheelbase         1   51.437 105.437  3.6618 0.0556749 .
## carheight         1   47.913 101.913  0.1385 0.7098156
## cylindernumber    6   54.065  98.065  6.2900 0.3915005
## enginesize        1   47.853 101.853  0.0778 0.7802359
## boreratio         1   47.804 101.804  0.0290 0.8647626
## stroke            1   55.222 109.222  7.4467 0.0063554 **
## compressionratio  1   47.776 101.776  0.0016 0.9685059
## peakrpm           1   48.859 102.859  1.0840 0.2978059
## citympg           1   48.803 102.803  1.0282 0.3105731
```

La situazione migliora, ma ancora tante variabili non sono significative.

Svolgiamo la procedura StepAIC, poi stimiamo il modello che ci ha proposto la procedura AIC e osserviamo quali variabili sono significative per il test LRT.

```
## ##                   Df Deviance     AIC     LRT    Pr(>Chi)
## <none>                 57.602  91.602
## symboling          5   77.624 101.624 20.0223   0.0012377 **
## carbody            4   71.299  97.299 13.6970   0.0083275 **
## drivewheel         2   70.025 100.025 12.4231   0.0020062 **
## wheelbase          1   65.066  97.066  7.4643   0.0062935 **
## enginesize         1   74.286 106.286 16.6840 0.000044153 ***
## stroke             1   70.156 102.156 12.5542   0.0003953 ***
## compressionratio   1   64.954  96.954  7.3523   0.0066976 **
## citympg            1   77.730 109.730 20.1277 0.000007244 ***
```

Il modello converge dopo 9 iterazioni; tutte le variabili sono significative per il test LRT.
Controlliamo correlazione fra le variabili scelte dalla procedura step.



Non vi sono valori altissimi, controlliamo i valori di VIF e TOL.

```
##                      VIF    TOL      Wi      Fi Leamer    CVIF Klein   IND1
## Xwheelbase        3.2129 0.3112 62.2785 73.0271 0.5579 -5.3498     1 0.0111
## Xcarheight        1.9533 0.5119 26.8297 31.4601 0.7155 -3.2524     0 0.0182
## Xenginesize       2.6743 0.3739 47.1189 55.2511 0.6115 -4.4528     1 0.0133
## Xboreratio        1.9833 0.5042 27.6724 32.4484 0.7101 -3.3023     0 0.0179
## Xstroke           1.1729 0.8526  4.8645  5.7041 0.9234 -1.9529     0 0.0303
## Xcompressionratio 1.6212 0.6168 17.4832 20.5006 0.7854 -2.6995     0 0.0219
## Xpeakrpm          1.6127 0.6201 17.2443 20.2205 0.7874 -2.6853     0 0.0220
## Xcitympg          3.1441 0.3181 60.3402 70.7543 0.5640 -5.2351     1 0.0113
```

I valori rispettano tutte le soglie. Ora controlliamo quasi separation, ossia se le osservazioni divise tra 0 e 1 si distribuiscono in maniera uniforme rispetto ai livelli delle variabili factor.

```
table(c5$target, c5$symboling)
##
##     -2 -1  0  1  2  3
##   0  1  9 46 45 26 14
##   1  2 13 21  9  6 13

table(c5$target, c5$fueltype)
##
##     diesel gas
##   0     11 130
##   1      9  55

table (c5$target, c5$carbody)
##
##     convertible hardtop hatchback sedan wagon
##   0           2       4        57    60    18
##   1           4       4        13    36     7

table (c5$target, c5$aspiration)
##
##     std turbo
##   0 123    18
##   1  45    19
```

Non c'è separation fra la variabile target e le singole covariate factor.

```
R_2_glm4 <- 1 - (57.602/254.547); R_2_glm4
## [1] 0.7737078
```

Questo modello ha una stima del valore di R^2 pari a 0.77 .

Proviamo a fittare un nuovo modello, dopo aver eliminato le variabili "stroke" e "compression ratio", relative a due parametri tecnici del motore, poco utili ai fini dell'interpretazione. Analizziamo i risultati ottenuti.

```
##            Df Deviance    AIC    LRT   Pr(>Chi)
## <none>          74.197 104.20
## symboling   5   88.360 108.36 14.163  0.0146048 *
## carbody     4   92.440 114.44 18.243  0.0011063 **
## drivewheel  2   94.846 120.85 20.648 0.00003283 ***
## wheelbase   1   89.985 117.98 15.788 0.00007085 ***
## enginesize  1   87.295 115.30 13.097  0.0002957 ***
## citympg     1   91.038 119.04 16.841 0.00004065 ***
```

Otteniamo un modello con meno covariate rispetto al precedente, quindi, l'interpretazione dello stesso è più semplice. Le variabili, per il test LRT, sono tutte significative.

```
R_2_glm5 <- 1 - (74.197/254.547); R_2_glm5
```

```
## [1] 0.7085136
```

Questo modello ha un valore della stima di R^2 pari a 0.71, il valore è più basso rispetto al primo modello fittato, quindi, questo secondo modello spiega meno i nostri dati; però è da tenere in considerazione che questo modello è più snello ergo più facilmente interpretabile.