

IMPORTAZIONE DATI

Abbiamo importato un dataset contenente informazioni relative a 205 automobili e fittato un modello considerando il prezzo come variabile risposta. L'obiettivo è quello di ottenere un modello robusto.

Controllo importazione

```
## 'data.frame':    205 obs. of  25 variables:
## $ symboling      : int  3 3 1 2 2 2 1 1 1 0 ...
## $ normalized.losses: int  NA NA NA 164 164 NA 158 NA 158 NA ...
## $ make           : Factor w/ 22 levels "alfa-romero",...: 1 1 1 2 2 2 2 2 2 2 ...
## $ fuel.type      : 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 ...
## $ num.of.doors    : Factor w/ 2 levels "four","two": 2 2 2 1 1 2 1 1 1 2 ...
## $ body.style      : Factor w/ 5 levels "convertible",...: 1 1 3 4 4 4 4 5 4 3 ...
## $ drive.wheels    : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2 2 1 ...
## $ wheel.base      : num  88.6 88.6 94.5 99.8 99.4 ...
## $ length          : num  169 169 171 177 177 ...
## $ width           : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
## $ height          : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
## $ curb.weight     : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
## $ engine.type     : Factor w/ 7 levels "dohc","dohcv",...: 1 1 6 4 4 4 4 4 4 4 ...
## $ num.of.cylinders : Factor w/ 7 levels "eight","five",...: 3 3 4 3 2 2 2 2 2 2 ...
## $ engine.size     : int  130 130 152 109 136 136 136 136 131 131 ...
## $ fuel.system      : Factor w/ 8 levels "1bbl","2bbl",...: 6 6 6 6 6 6 6 6 6 6 ...
## $ bore            : 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 ...
## $ compression.ratio: 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 ...
## $ peak.rpm        : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
## $ city.mpg        : int  21 21 19 24 18 19 19 19 17 16 ...
## $ highway.mpg     : int  27 27 26 30 22 25 25 25 20 22 ...
## $ price           : int  13495 16500 16500 13950 17450 15250 17710 18920 23875 NA
```

Osserviamo che il fattore symboling non è importato correttamente (è un fattore, ma viene memorizzato come un intero). Procediamo correggendolo.

DATI MANCANTI (CONTEGGIO E IMPUTAZIONE)

Conteggio dati mancanti

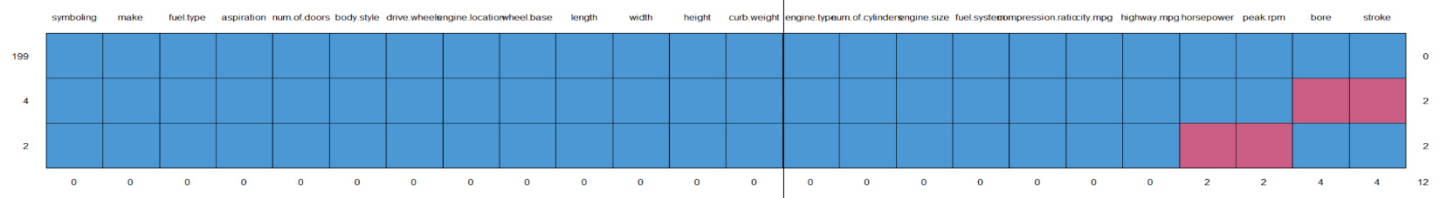
```
##      symboling normalized.losses      make      fuel.type
##      0          41                0                0
##      aspiration num.of.doors      body.style      drive.wheels
##      0          2                0                0
##      wheel.base      length      width      height
##      0          0                0                0
##      curb.weight      engine.type num.of.cylinders      engine.size
##      0          0                0                0
##      fuel.system      bore      stroke      compression.ratio
##      0          4                4                0
##      horsepower      peak.rpm      city.mpg      highway.mpg
##      2          2                0                0
##      price
##      4
##
##      variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
## 1 normalized.losses      0      0  41 20.00      0      0 integer    51
## 2 price                0      0   4  1.95      0      0 integer   186
## 3 bore                  0      0   4  1.95      0      0 numeric    38
## 4 stroke                0      0   4  1.95      0      0 numeric    36
## 5 num.of.doors          0      0   2  0.98      0      0 factor     2
## 6 horsepower            0      0   2  0.98      0      0 integer    59
```

Osserviamo che sono presenti valori mancanti e che per normalized.losses sono il 20%, quindi decidiamo di eliminarla. Ipotizziamo che i restanti dati mancanti siano di tipo MAR (Missing At Random) e procediamo applicando la procedura dell'imputazione.

Imputazione fattori (imputazione singola)

Applicando questa tipologia di imputazione sostituiamo i valori mancanti della variabile qualitativa considerata (num.of.doors) con la moda.

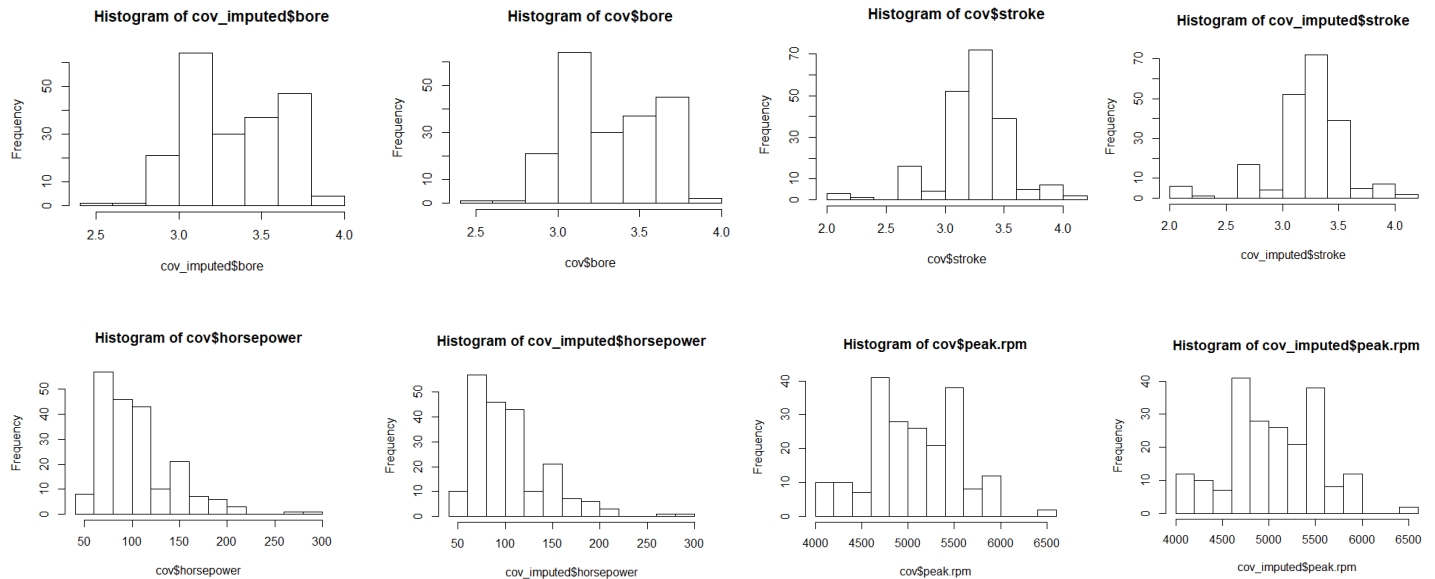
Imputazione covariate quantitative (imputazione multipla)



Notiamo che ci sono 199 osservazioni senza valori mancanti (12 dati mancanti in totale).

Procediamo con l'imputazione (ovviamente dopo l'imputazione nessuna variabile presenta valori mancanti).

Confrontiamo gli istogrammi delle variabili prima e dopo l'imputazione:



Vediamo che i grafici prima e dopo l'imputazione sono molto simili l'uno con l'altro e dunque non abbiamo particolari problemi.

COLLINEARITÀ

Modello 1

```
## Call:
## lm(formula = price ~ ., data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3676.7  -970.9    0.0   846.3  7935.4
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -10949.3672  17841.1182  -0.614  0.540421
## symboling-1    1120.0145   1286.0053   0.871  0.385319
## symboling0     1123.2539   1519.7514   0.739  0.461108
## symboling1      780.1666   1580.4515   0.494  0.622354
## symboling2    -106.8700   1649.7120  -0.065  0.948443
## symboling3     -634.3035   1735.7545  -0.365  0.715351
## makeaudi       2115.1270   2411.8400   0.877  0.382035
## makebmw        5121.6848   2307.7734   2.219  0.028110 *
## makechevrolet  -5049.4116   2292.9719  -2.202  0.029329 *
## makedodge     -5537.3322   1893.3228  -2.925  0.004037 **
## makehonda     -2599.3258   2200.0995  -1.181  0.239468
## makeisuzu     -4166.8230   2471.2188  -1.686  0.094045 .
## makejaguar    -1357.0827   2822.7979  -0.481  0.631455
## makemazda     -2640.6385   1703.9376  -1.550  0.123514
## makemercedes-benz 2412.3127   2553.8530   0.945  0.346537
## makemercury   -5018.4135   2966.5293  -1.692  0.092981 .
```

```

## makemitsubishi      -5566.4892   1908.2648  -2.917 0.004131 **
## makenissan           -2893.7552   1699.6384  -1.703 0.090916 .
## makepeugot          -8482.1475   4505.0283  -1.883 0.061845 .
## makeplymouth        -5401.3630   1873.6509  -2.883 0.004578 **
## makeporsche         3148.0056   2969.1841   1.060 0.290907
## makerenault         -867.1352   2489.3717  -0.348 0.728123
## makesaab            1730.5220   2119.4070   0.817 0.415625
## makesubaru          -12423.6371  3443.5386  -3.608 0.000432 ***
## maketoyota          -3692.0012   1569.5479  -2.352 0.020083 *
## makevolkswagen      -1481.1521   1860.7564  -0.796 0.427412
## makevolvo           -1955.0014   2301.6698  -0.849 0.397150
## fuel.typegas        -12243.0724  6857.4530  -1.785 0.076414 .
## aspirationturbo     1522.1112   860.9427   1.768 0.079294 .
## num.of.doorstwo      80.6449    517.2739   0.156 0.876338
## body.stylehardtop   -2418.1719  1220.9950  -1.980 0.049651 *
## body.stylehatchback -3142.0776  1108.0256  -2.836 0.005265 **
## body.styleседan     -2721.7352  1208.4198  -2.252 0.025893 *
## body.stylewagon     -3051.1078  1304.3065  -2.339 0.020768 *
## drive.wheelsfwd     -664.4283   928.5343  -0.716 0.475477
## drive.wheelsrwd    -110.4546  1238.4847  -0.089 0.929065
## wheel.base          227.3305    96.3686   2.359 0.019738 *
## length              -133.4508    51.7314  -2.580 0.010942 *
## width                590.0957   238.7799   2.471 0.014689 *
## height              -309.8589   144.7972  -2.140 0.034130 *
## curb.weight          6.2569     1.6862   3.711 0.000300 ***
## engine.type1        3457.8356  4194.2916   0.824 0.411137
## engine.typeohc      1014.0363  1243.4966   0.815 0.416218
## engine.typeohcf     9856.9098  2670.1634   3.692 0.000321 ***
## engine.typeohcv     -2892.1127  1296.3926  -2.231 0.027314 *
## engine.typerotor    4414.9017  5008.7459   0.881 0.379624
## num.of.cylindersfive -6373.7435  2831.8924  -2.251 0.025998 *
## num.of.cylindersfour -3298.2408  3513.5094  -0.939 0.349521
## num.of.cylinderssix -3654.0929  2694.5225  -1.356 0.177292
## num.of.cylindersthree NA          NA          NA          NA
## num.of.cylinders twelve -5635.0569  5255.9036  -1.072 0.285543
## num.of.cylinders two NA          NA          NA          NA
## engine.size          83.8520    26.0013   3.225 0.001576 **
## fuel.system2bbl      2298.5876  1504.6750   1.528 0.128910
## fuel.system4bbl      -762.2842  2597.7061  -0.293 0.769626
## fuel.systemidi       NA          NA          NA          NA
## fuel.systemmfi       139.9766  2656.3821   0.053 0.958052
## fuel.systemmpfi      1142.0301  1575.5432   0.725 0.469781
## fuel.systemspdi      232.1234  1843.4649   0.126 0.899982
## fuel.systemspfi      2491.6247  3104.4336   0.803 0.423595
## bore                 -2682.1766  1846.2480  -1.453 0.148574
## stroke               -1470.0187  1004.9503  -1.463 0.145819
## compression.ratio    -813.2263   506.4095  -1.606 0.110606
## horsepower           16.3822    25.3453   0.646 0.519127
## peak.rpm             1.9478     0.7168   2.718 0.007427 **
## city.mpg             -11.5778   134.7483  -0.086 0.931654
## highway.mpg          122.4225   117.6524   1.041 0.299920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1736 on 137 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.9673, Adjusted R-squared:  0.9523
## F-statistic: 64.36 on 63 and 137 DF, p-value: < 0.0000000000000022
##
## Single term deletions
##
## Model:
## price ~ symboling + make + fuel.type + aspiration + num.of.doors +
## body.style + drive.wheels + wheel.base + length + width +
## height + curb.weight + engine.type + num.of.cylinders + engine.size +
## fuel.system + bore + stroke + compression.ratio + horsepower +
## peak.rpm + city.mpg + highway.mpg
##          Df Sum of Sq  RSS      AIC F value    Pr(>F)
## <none>                    412815672 3049.6
## symboling          5  16481009 429296681 3047.4   1.0939    0.3666383
## make              20 378367038 791182710 3140.3   6.2784 0.00000000001157 ***
## fuel.type          0           0 412815672 3049.6
## aspiration          1   9418458 422234130 3052.1   3.1257    0.0792938 .
## num.of.doors        1    73240 412888912 3047.6   0.0243    0.8763385
## body.style          4  27843044 440658716 3054.7   2.3100    0.0609347 .
## drive.wheels        2   2949334 415765005 3047.0   0.4894    0.6140655

```

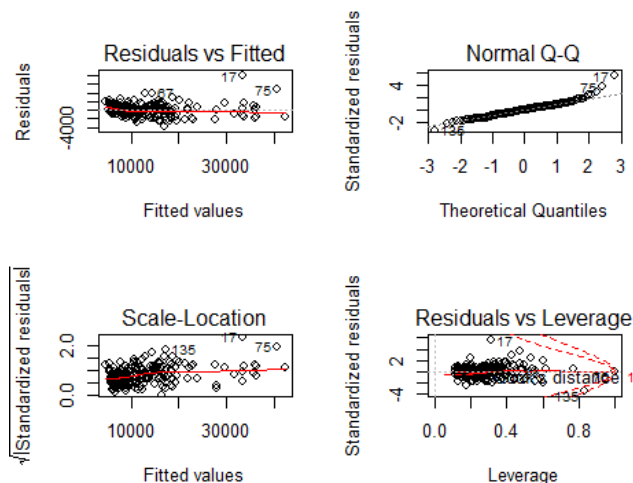
```
## wheel.base      1 16767952 429583624 3055.6 5.5647      0.0197384 *
## length          1 20052573 432868245 3057.1 6.6548      0.0109425 *
## width           1 18402848 431218520 3056.3 6.1073      0.0146889 *
## height          1 13798866 426614538 3054.2 4.5794      0.0341305 *
## curb.weight     1 41489930 454305602 3066.8 13.7691      0.0002996 ***
## engine.type     3 70308636 483124307 3075.2 7.7777 0.00007811166848 ***
## num.of.cylinders 4 49375683 462191355 3064.3 4.0965      0.0036225 **
## engine.size     1 31337952 444153624 3062.3 10.4000      0.0015757 **
## fuel.system     6 27904546 440720218 3050.7 1.5434      0.1685509
## bore           1 6359607 419175279 3050.7 2.1105      0.1485738
## stroke         1 6447512 419263184 3050.7 2.1397      0.1458186
## compression.ratio 1 7770604 420586276 3051.3 2.5788      0.1106057
## horsepower     1 1258883 414074555 3048.2 0.4178      0.5191272
## peak.rpm       1 22252700 435068371 3058.1 7.3849      0.0074270 **
## city.mpg       1 22246 412837917 3047.6 0.0074      0.9316539
## highway.mpg    1 3262547 416078219 3049.2 1.0827      0.2999199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fittiamo il nostro primo modello (ANCOVA).

Il valore dell' R^2 aggiustato è 0.9523 e quindi il nostro modello spiega il 95% circa della variabilità della risposta (osserviamo che si tratta già di un valore molto elevato).

Inoltre, notiamo che per fuel.type le stime non vengono calcolate. Questo potrebbe essere causato dal fatto che è presente il problema della collinearità e/o il problema della zero variance.

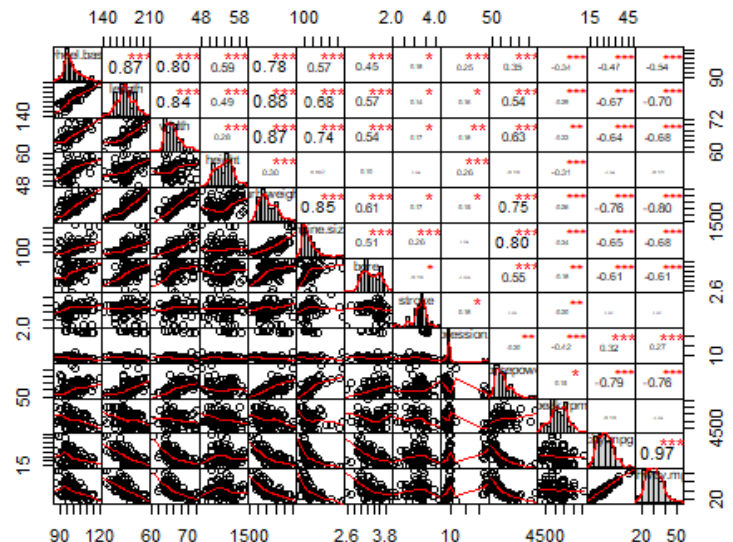
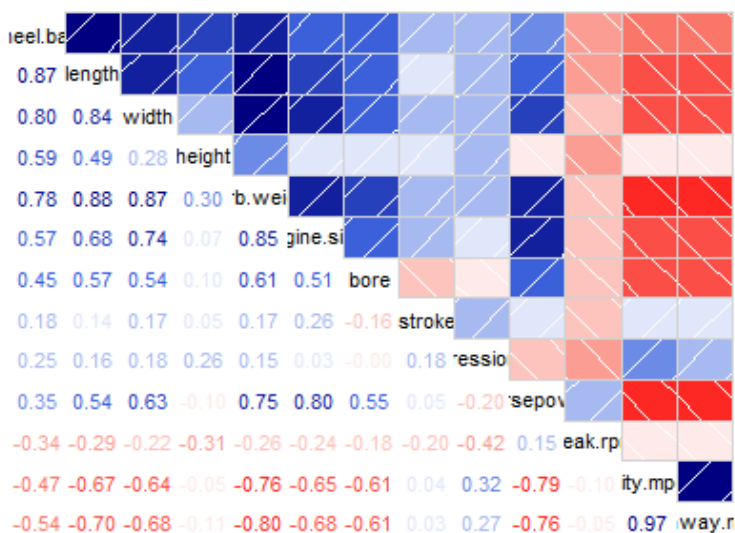
Diagnostiche fit1



Osservando i grafici relativi alle diagnostiche del nostro modello notiamo che:

- Grafico residui vs valori interpolati: l'assunzione di linearità sembra essere rispettata
- QQ_plot residui standardizzati: la distribuzione è leggermente asimmetrica a destra
- Scale-Location: il modello potrebbe soffrire di eteroschedasticità
- Grafico residui standardizzati vs leverage: sicuramente sono presenti degli outliers e dei punti di leva, tuttavia non siamo certe della presenza di punti influenti, poichè essendoci osservazioni con leva pari a 1 (18, 29, 43, 44, 47, 56, 123) queste non vengono rappresentate.

Correlazioni bivariate tra covariate numeriche



Osserviamo che ci sono correlazioni bivariate elevate, per esempio tra wheel.base e length (0.87), tra length e curb.weight (0.88), tra width e curb.weight (0.87), tra curb.weight e engine.size (0.85) e soprattutto tra city.mpg e highway.mpg (0.97).

TOL e VIF

Ora fittiamo il modello di regressione lineare multiplo della risposta su tutte le covariate numeriche e ne calcoliamo le diagnostiche TOL e VIF:

```
Call:
lm<diag>(mod = fit_numeric)

All Individual Multicollinearity Diagnostics Result

wheel.base 7.2022 0.1388 97.1676 106.5649 0.3726 -0.2785 1 0.0089 1.1723
length 9.6300 0.1038 135.2034 148.2792 0.3222 -0.3724 1 0.0066 1.2199
width 5.7427 0.1741 74.3023 81.4883 0.4173 -0.2220 0 0.0111 1.1242
height 2.1756 0.4596 18.4182 20.1994 0.6780 -0.0841 0 0.0293 0.7356
curb.weight 16.0639 0.0623 236.0015 258.8257 0.2495 -0.6211 1 0.0040 1.2766
engine.size 7.3299 0.1364 99.1685 108.7593 0.3694 -0.2834 1 0.0087 1.1756
bore 2.0940 0.4776 17.1386 18.7961 0.6911 -0.0810 0 0.0305 0.7112
stroke 1.3354 0.7488 5.2349 5.7632 0.8653 -0.0516 0 0.0478 0.3419
compression.ratio 2.1630 0.4623 18.2207 19.9828 0.6799 -0.0836 0 0.0295 0.7319
horsepower 8.0021 0.1250 109.6988 120.3080 0.3535 -0.3094 1 0.0080 1.1912
peak.rpm 2.0738 0.4822 16.8228 18.4497 0.6944 -0.0802 0 0.0308 0.7049
city.mpg 25.7950 0.0388 388.4557 426.0240 0.1969 -0.9974 1 0.0025 1.3085
highway.mpg 24.7024 0.0405 371.3371 407.2498 0.2012 -0.9551 1 0.0026 1.3062

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

wheel.base , length , height , curb.weight , bore , horsepower , highway.mpg , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.8542

* use method argument to check which regressors may be the reason of collinearity
=====
```

Vediamo che ci sono valori di VIF maggiori di 5 e dunque anche valori di TOL minori di 0.3. Decidiamo quindi di non mantenere tutte le covariate numeriche all’interno del nostro modello e anzi, di toglierle una ad una.

Andiamo prima però a vedere i chi quadri normalizzati e dunque le associazioni tra i fattori.

Associazioni tra fattori (chi quadri normalizzati)

X1 <fctr>	Row <fctr>	Column <dbl>	Chi.Square <dbl>	df <int>	p.value <dbl>	n <int>	u1 <dbl>	u2 <dbl>	nMinu1u2 <dbl>	Chi.Square.norm <dbl>
1	symboling	make	301.467	105	0.000	205	5	21	1025	0.294114066
2	symboling	fuel.type	14.662	5	0.012	205	5	1	205	0.071521884
3	symboling	aspiration	11.988	5	0.035	205	5	1	205	0.058476560
4	symboling	num.of.doors	99.258	5	0.000	205	5	1	205	0.484184138
5	symboling	body.style	109.795	20	0.000	205	5	4	820	0.133896546
6	symboling	drive.wheels	38.701	10	0.000	205	5	2	410	0.094393178
7	symboling	engine.type	79.527	30	0.000	205	5	6	1025	0.077587572
8	symboling	num.of.cylinders	55.856	30	0.003	205	5	6	1025	0.054493418
9	symboling	fuel.system	105.805	35	0.000	205	5	7	1025	0.103223975
10	make	fuel.type	49.043	21	0.000	205	21	1	205	0.239232222
11	make	aspiration	55.436	21	0.000	205	21	1	205	0.270419028
12	make	num.of.doors	38.831	21	0.010	205	21	1	205	0.189421435
13	make	body.style	165.361	84	0.000	205	21	4	820	0.201660203
14	make	drive.wheels	189.684	42	0.000	205	21	2	410	0.462645073
15	make	engine.type	598.195	126	0.000	205	21	6	1230	0.486337471
16	make	num.of.cylinders	479.586	126	0.000	205	21	6	1230	0.389907223
17	make	fuel.system	508.231	147	0.000	205	21	7	1435	0.354168166
18	fuel.type	aspiration	29.606	1	0.000	205	1	1	205	0.144418338
19	fuel.type	num.of.doors	6.058	1	0.014	205	1	1	205	0.029553448
20	fuel.type	body.style	10.129	4	0.038	205	1	4	205	0.049411438
21	fuel.type	drive.wheels	3.588	2	0.166	205	1	2	205	0.017501778
22	fuel.type	engine.type	18.768	6	0.005	205	1	6	205	0.091551010
23	fuel.type	num.of.cylinders	10.905	6	0.091	205	1	6	205	0.053194007
24	fuel.type	fuel.system	205.000	7	0.000	205	1	7	205	1.000000000
25	aspiration	num.of.doors	0.328	1	0.567	205	1	1	205	0.001600653
26	aspiration	body.style	1.597	4	0.809	205	1	4	205	0.007790954
27	aspiration	drive.wheels	4.857	2	0.088	205	1	2	205	0.023693807
28	aspiration	engine.type	10.590	6	0.102	205	1	6	205	0.051660325
29	aspiration	num.of.cylinders	13.864	6	0.031	205	1	6	205	0.067629333
30	aspiration	fuel.system	83.000	7	0.000	205	1	7	205	0.404877617
31	num.of.doors	body.style	118.590	4	0.000	205	1	4	205	0.578489529
32	num.of.doors	drive.wheels	2.654	2	0.265	205	1	2	205	0.012944771
33	num.of.doors	engine.type	14.306	6	0.026	205	1	6	205	0.069784999
34	num.of.doors	num.of.cylinders	9.839	6	0.132	205	1	6	205	0.047994513
35	num.of.doors	fuel.system	19.415	7	0.007	205	1	7	205	0.094708341
36	body.style	drive.wheels	26.590	8	0.001	205	4	2	410	0.064854545
37	body.style	engine.type	38.079	24	0.034	205	4	6	820	0.046437400
38	body.style	num.of.cylinders	27.784	24	0.269	205	4	6	820	0.033882858
39	body.style	fuel.system	44.702	28	0.024	205	4	7	820	0.054514129
40	drive.wheels	engine.type	85.535	12	0.000	205	2	6	410	0.208622823
41	drive.wheels	num.of.cylinders	57.892	12	0.000	205	2	6	410	0.141200930
42	drive.wheels	fuel.system	74.944	14	0.000	205	2	7	410	0.182789841
43	engine.type	num.of.cylinders	391.651	36	0.000	205	6	6	1230	0.318415218
44	engine.type	fuel.system	211.740	42	0.000	205	6	7	1230	0.172146639
45	num.of.cylinders	fuel.system	208.575	42	0.000	205	6	7	1230	0.169573195

C'è un chi quadro normalizzato maggiore di 0.9 e quindi decidiamo di eliminare o fuel.type o fuel.system, ovvero degli elementi della coppia di covariate che restituisce un chi quadro normalizzato pari a 1.

Modello 2 (Senza fuel.system)

Chi quadri-normalizzati

Non ci sono più chi quadri normalizzati maggiori di 0.9. Ora procediamo eliminando le covariate quantitative che avevano valori al di là della soglia per quanto riguarda TOL e VIF.

Modello 3 (Senza city.mpg)

Modello 4 (Senza curb.weight)

Modello 5 (Senza lenght)

Modello 6 (Senza horsepower)

Modello 7 (Senza width)

Modello 8 (Senza highway.mpg)

```
call:
lmcdiag(mod = (fit_numeric))
```

All Individual Multicollinearity Diagnostics Result

	VIF	TOL	wi	Fi	Leamer	CVIF	Klein	IND1	IND2
wheel.base	2.9755	0.3361	63.8733	77.0431	0.5797	-1.2030	0	0.0104	1.6445
height	1.9071	0.5243	29.3305	35.3781	0.7241	-0.7711	0	0.0162	1.1781
engine.size	2.1455	0.4661	37.0377	44.6744	0.6827	-0.8675	0	0.0144	1.3224
bore	1.6812	0.5948	22.0249	26.5661	0.7712	-0.6797	0	0.0184	1.0036
stroke	1.3002	0.7691	9.7078	11.7094	0.8770	-0.5257	0	0.0238	0.5719
compression.ratio	1.3076	0.7648	9.9449	11.9953	0.8745	-0.5287	0	0.0237	0.5826
peak.rpm	1.3914	0.7187	12.6568	15.2665	0.8477	-0.5626	0	0.0222	0.6968

```
1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

height , bore , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.833

* use method argument to check which regressors may be the reason of collinearity

Abbiamo infine ottenuto un modello senza valori di TOL minori di 0.3, nè valori di VIF maggiori di 5, procediamo perciò facendo il summary del nostro ultimo modello fittato (fit8).

```
## Call:
## lm(formula = price ~ symboling + make + fuel.type + aspiration +
##   num.of.doors + body.style + drive.wheels + wheel.base + height +
##   engine.type + num.of.cylinders + engine.size + bore + stroke +
##   compression.ratio + peak.rpm, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3798.1 -1023.5    0.0   835.4  9157.8
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10282.087   15111.360   0.680   0.497294
## symboling-1    1233.457    1361.688   0.906   0.366489
## symboling0     1189.337    1565.854   0.760   0.448726
## symboling1      881.279    1611.390   0.547   0.585261
## symboling2     144.647    1693.932   0.085   0.932065
## symboling3    -359.551    1800.976  -0.200   0.842033
## makeaudi      3421.733    2556.800   1.338   0.182841
## makebmw       5661.066    2441.014   2.319   0.021746 *
## makechevrolet -3273.061    2253.734  -1.452   0.148525
## makedodge     -4043.746    1923.842  -2.102   0.037243 *
## makehonda     -2469.630    1963.448  -1.258   0.210431
## makeisuzu     -2848.778    2214.303  -1.287   0.200252
## makejaguar     2462.318    2479.210   0.993   0.322229
## makemazda     -1054.765    1810.727  -0.583   0.561104
## makemercedes-benz 5910.873    2678.222   2.207   0.028843 *
## makemercury   -1520.316    2977.064  -0.511   0.610334
## makemitsubishi -4144.260    1954.317  -2.121   0.035614 *
## makenissan     -2044.741    1799.496  -1.136   0.257661
## makepeugot    -5632.750    4542.688  -1.240   0.216940
## makeplymouth  -4238.663    1915.901  -2.212   0.028463 *
```



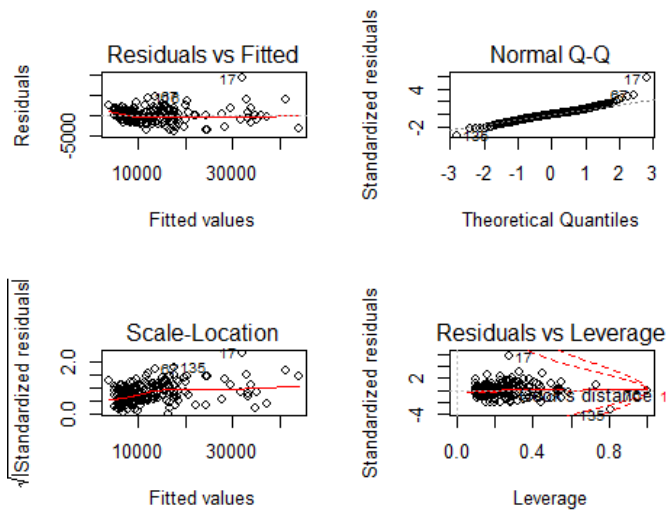
```

## makeporsche          6711.192    3072.849    2.184    0.030524 *
## makerenault          -910.572    2264.583   -0.402    0.688192
## makesaab             3213.042    2145.788    1.497    0.136412
## makesubaru          -7586.602    3397.202   -2.233    0.027027 *
## maketoyota          -2462.699    1650.629   -1.492    0.137820
## makevolkswagen       -598.445    1973.395   -0.303    0.762117
## makevolvo           171.007    2393.631    0.071    0.943142
## fuel.typegas        -5440.397    6594.437   -0.825    0.410692
## aspirationturbo      2458.082    685.971    3.583    0.000459 ***
## num.of.doorstwo     -275.535    556.703   -0.495    0.621371
## body.stylehardtop   -4038.088    1264.702   -3.193    0.001719 **
## body.stylehatchback -3931.236    1146.867   -3.428    0.000787 ***
## body.stylededan     -4024.866    1222.710   -3.292    0.001243 **
## body.stylewagon     -3946.638    1342.104   -2.941    0.003799 **
## drive.wheelsfwd     -1665.292    910.547   -1.829    0.069415 .
## drive.wheelsrwd     -692.208    1252.020   -0.553    0.581179
## wheel.base          216.031    80.670    2.678    0.008238 **
## height              -303.330    143.869   -2.108    0.036672 *
## engine.type1        4688.328    4484.227    1.046    0.297478
## engine.typeohc      1426.559    1256.362    1.135    0.258002
## engine.typeohcf     6466.065    2705.799    2.390    0.018112 *
## engine.typeohcv     -1645.690    1319.635   -1.247    0.214325
## engine.typerotor    5385.176    5028.479    1.071    0.285930
## num.of.cylindersfive -3404.690    2996.659   -1.136    0.257713
## num.of.cylindersfour -1870.329    3693.844   -0.506    0.613369
## num.of.cylinderssix -1130.928    2722.829   -0.415    0.678483
## num.of.cylindersthree NA          NA          NA          NA
## num.of.cylinderstwelve -6326.919    4756.794   -1.330    0.185524
## num.of.cylinderstwo NA          NA          NA          NA
## engine.size          126.201    22.150    5.698    0.0000000629 ***
## bore                -2589.230    1916.249   -1.351    0.178680
## stroke              -1466.281    1041.760   -1.408    0.161361
## compression.ratio    -373.654    475.417   -0.786    0.433145
## peak.rpm             1.861    0.640    2.908    0.004199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1921 on 149 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.9564, Adjusted R-squared:  0.9415
## F-statistic: 64.16 on 51 and 149 DF, p-value: < 0.00000000000000022
##
## Single term deletions
##
## Model:
## price ~ symboling + make + fuel.type + aspiration + num.of.doors +
## body.style + drive.wheels + wheel.base + height + engine.type +
## num.of.cylinders + engine.size + bore + stroke + compression.ratio +
## peak.rpm
##
##          Df Sum of Sq      RSS      AIC F value        Pr(>F)
## <none>                 550099625 3083.3
## symboling           5  14013159 564112785 3078.3   0.7591      0.5806825
## make                20  447305508 997405133 3162.9   6.0579 0.0000000001665 ***
## fuel.type           1   2512817 552612442 3082.2   0.6806      0.4106916
## aspiration           1  47406251 597505876 3097.9  12.8405      0.0004588 ***
## num.of.doors         1    904403 551004028 3081.6   0.2450      0.6213709
## body.style           4  48646390 598746015 3092.3   3.2941      0.0127842 *
## drive.wheels         2  18927292 569026917 3086.1   2.5633      0.0804426 .
## wheel.base           1  26476977 576576602 3090.7   7.1716      0.0082382 **
## height              1  16411511 566511136 3087.2   4.4452      0.0366723 *
## engine.type          3  41296974 591396599 3091.8   3.7286      0.0127307 *
## num.of.cylinders     4  20772858 570872483 3082.7   1.4066      0.2345287
## engine.size          1 119848919 669948544 3120.9  32.4623 0.00000006291886 ***
## bore                1   6740508 556840133 3083.7   1.8257      0.1786803
## stroke              1   7313989 557413614 3083.9   1.9811      0.1613608
## compression.ratio    1  2280579 552380205 3082.1   0.6177      0.4331447
## peak.rpm            1  31213224 581312849 3092.4   8.4544      0.0041987 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Osserviamo che l' R^2 aggiustato di questo modello è leggermente più basso rispetto al precedente e pari a 0.9415. Il modello presenta ancora NA nel summary in corrispondenza di num.of.cylindersthree e num.of.cylinderstwo.

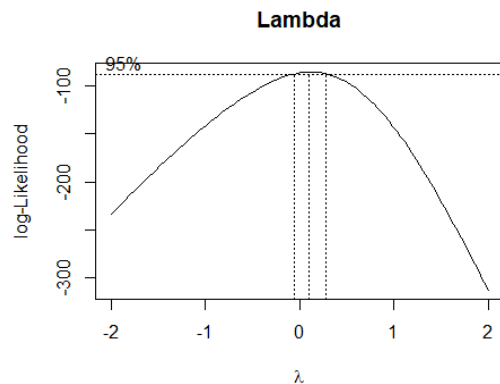
Diagnostiche fit8



Le diagnostiche del nostro modello sono molto simili a quelle del modello iniziale, ma ora sono tre le osservazioni che hanno leva pari a 1 (18, 47, 123).

LINEARITÀ

Trasformazione di Box-Cox



Vediamo che il valore che massimizza la funzione di log-verosimiglianza (e che minimizza l'errore nel calcolo dell'MSE) è 0.1010101. Procediamo con la trasformazione di Box-Cox approssimando lambda con 0 e dunque utilizzando il logaritmo della nostra variabile risposta (anche se dalle diagnostiche dei residui non si nota un'eccessiva non linearità).

Modello 9

```
## Call:
## lm(formula = log(price) ~ symboling + make + fuel.type + aspiration +
##   num.of.doors + body.style + drive.wheels + wheel.base + height +
##   engine.type + num.of.cylinders + engine.size + bore + stroke +
##   compression.ratio + peak.rpm, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29996 -0.06628 -0.00032  0.07343  0.32352
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.15207043  0.99010801   9.244 0.000000000000000231 ***
## symboling-1     0.02769637  0.08921887   0.310  0.756666
## symboling0    -0.01985216  0.10259595  -0.193  0.846832
## symboling1    -0.11176145  0.10557951  -1.059  0.291517
## symboling2    -0.10877441  0.11098774  -0.980  0.328646
## symboling3    -0.02684944  0.11800138  -0.228  0.820320
## makeaudi       0.35183105  0.16752352   2.100  0.037397 *
## makebmw       0.35750895  0.15993710   2.235  0.026884 *
## makechevrolet -0.27268685  0.14766641  -1.847  0.066783 .
## makedodge     -0.27097897  0.12605164  -2.150  0.033189 *
## makehonda     -0.10828454  0.12864667  -0.842  0.401293
## makeisuzu     -0.26099279  0.14508284  -1.799  0.074055 .
## makejaguar    -0.18077988  0.16243975  -1.113  0.267542
## makemazda    -0.01442205  0.11864022  -0.122  0.903410
## makemercedes-benz 0.09110508  0.17547920   0.519  0.604406
## makemercury   0.05655220  0.19505953   0.290  0.772278
## makemitsubishi -0.27767117  0.12804836  -2.168  0.031708 *
## makenissan    -0.09333660  0.11790435  -0.792  0.429835
## makepeugot   -0.59379184  0.29764044  -1.995  0.047867 *
## makeplymouth -0.30220811  0.12553132  -2.407  0.017288 *
## makeporsche   0.44206554  0.20133545   2.196  0.029662 *
```



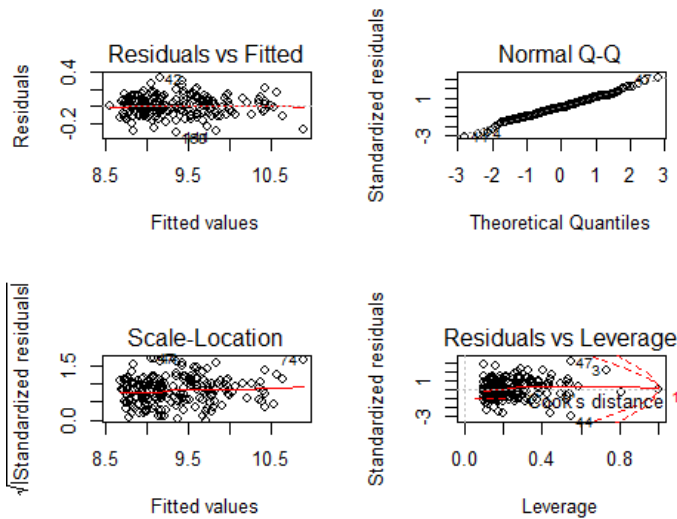
```

## makerenault      -0.04264704  0.14837726 -0.287      0.774188
## makesaab         0.40205773  0.14059367  2.860      0.004849 **
## makesubaru       -0.36947563  0.22258730 -1.660      0.099035 .
## maketoyota       -0.16258003  0.10815046 -1.503      0.134884
## makevolkswagen   0.04625885  0.12929838  0.358      0.721023
## makevolvo        -0.03201207  0.15683256 -0.204      0.838541
## fuel.typegas     -0.41279102  0.43207261 -0.955      0.340936
## aspirationturbo   0.18590720  0.04494539  4.136 0.000058762110774111 ***
## num.of.doorstwo  -0.05421642  0.03647561 -1.486      0.139293
## body.stylehardtop -0.27188790  0.08286424 -3.281      0.001288 **
## body.stylehatchback -0.30468676  0.07514360 -4.055 0.000080623270883889 ***
## body.styleseadan  -0.28680358  0.08011294 -3.580      0.000464 ***
## body.stylewagon   -0.25645202  0.08793569 -2.916      0.004089 **
## drive.wheelsfwd   -0.17687896  0.05965976 -2.965      0.003528 **
## drive.wheelsrwd    0.01014142  0.08203331  0.124      0.901778
## wheel.base        0.02302053  0.00528553  4.355 0.000024586756700195 ***
## height           -0.03465487  0.00942643 -3.676      0.000330 ***
## engine.type1      0.42901174  0.29381001  1.460      0.146349
## engine.typeohc     0.00259521  0.08231780  0.032      0.974892
## engine.typeohcf    0.21591658  0.17728602  1.218      0.225187
## engine.typeohcv    0.06886397  0.08646354  0.796      0.427037
## engine.typerotor   0.52019846  0.32946983  1.579      0.116481
## num.of.cylindersfive 0.18729135  0.19634340  0.954      0.341681
## num.of.cylindersfour 0.21211695  0.24202353  0.876      0.382207
## num.of.cylinderssix 0.18454116  0.17840184  1.034      0.302619
## num.of.cylindersthree NA NA NA NA
## num.of.cylinderstwelve 0.00556995  0.31166882  0.018      0.985765
## num.of.cylinderstwo NA NA NA NA
## engine.size       0.00588911  0.00145128  4.058 0.000079650481943306 ***
## bore             -0.11676684  0.12555410 -0.930      0.353869
## stroke            0.00969450  0.06825693  0.142      0.887248
## compression.ratio -0.03219315  0.03114971 -1.033      0.303046
## peak.rpm          0.00007089  0.00004193  1.691      0.093010 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1259 on 149 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.9532, Adjusted R-squared:  0.9372
## F-statistic: 59.49 on 51 and 149 DF, p-value: < 0.0000000000000022
##
## Single term deletions
##
## Model:
## log(price) ~ symboling + make + fuel.type + aspiration + num.of.doors +
## body.style + drive.wheels + wheel.base + height + engine.type +
## num.of.cylinders + engine.size + bore + stroke + compression.ratio +
## peak.rpm
##
##      Df Sum of Sq    RSS      AIC F value      Pr(>F)
## <none>                 2.3616 -789.24
## symboling             5  0.17619 2.5377 -784.78  2.2233      0.0548977 .
## make                  20  2.36993 4.7315 -689.56  7.4764 0.000000000000002565 ***
## fuel.type             1  0.01447 2.3760 -790.01  0.9127      0.3409360
## aspiration            1  0.27117 2.6327 -769.39 17.1089 0.000058762110774111 ***
## num.of.doors          1  0.03502 2.3966 -788.28  2.2093      0.1392930
## body.style            4  0.27869 2.6402 -774.82  4.3958      0.0021799 **
## drive.wheels          2  0.36328 2.7248 -764.48 11.4605 0.00002346264926101 ***
## wheel.base            1  0.30065 2.6622 -767.15 18.9694 0.00002458675670019 ***
## height                1  0.21421 2.5758 -773.79 13.5156      0.0003296 ***
## engine.type           3  0.02835 2.3899 -792.84  0.5963      0.6183651
## num.of.cylinders      4  0.02791 2.3895 -794.88  0.4402      0.7793816
## engine.size           1  0.26098 2.6225 -770.17 16.4664 0.00007965048194330 ***
## bore                  1  0.01371 2.3753 -790.08  0.8649      0.3538691
## stroke                1  0.00032 2.3619 -791.21  0.0202      0.8872485
## compression.ratio     1  0.01693 2.3785 -789.80  1.0681      0.3030457
## peak.rpm              1  0.04530 2.4069 -787.42  2.8580      0.0930100 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

L'R² aggiustato del modello trasformato è pari a 0.9372, praticamente identico al precedente.

Diagnostische fit9



Le diagnostiche del nuovo modello sono migliorate. Il modello è più lineare, soffre meno di eteroschedasticità e i residui sembrano meno asimmetrici. Ci sono ancora però dei valori con leva pari a 1 (18, 47 e 123).

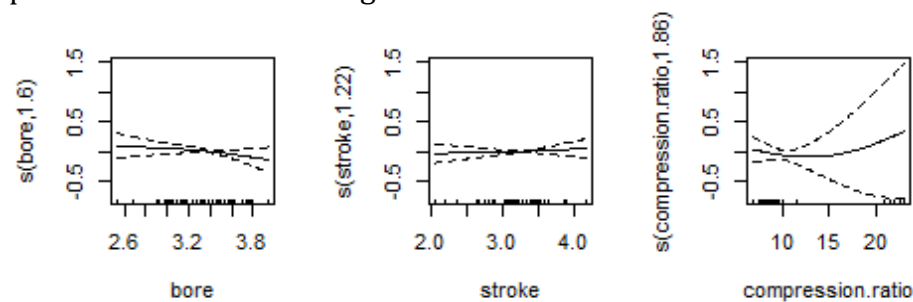
Trasformazione gam

Fittiamo il modello `fit_gam` per vedere se le variabili non significative possono diventarlo se trasformate come ci suggerisce la funzione `gam`:

```
## Family: gaussian
## Link function: identity
##
## Formula:
## log(price) ~ symboling + make + fuel.type + aspiration + num.of.doors +
## body.style + drive.wheels + wheel.base + height + engine.type +
## num.of.cylinders + engine.size + s(bore) + s(stroke) + s(compression.ratio) +
## peak.rpm
##
## Parametric coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   7.66899397 0.88999767    8.617 0.000000000000000000 ***
## symboling-1    0.02310844 0.08707032    0.265   0.791073
## symboling0   -0.01763372 0.10021160   -0.176   0.860563
## symboling1   -0.09899389 0.10326052   -0.959   0.339290
## symboling2   -0.10614094 0.10842819   -0.979   0.329232
## symboling3   -0.03693085 0.11538535   -0.320   0.749372
## makeaudi     -0.35399358 0.16601456   -2.132   0.034640 *
## makebmw       0.40513575 0.15847179    2.557   0.011585 *
## makechevrolet -0.24075832 0.14551632   -1.655   0.100153
## makedodge    -0.25235997 0.12398252   -2.035   0.043598 *
## makehonda    -0.09299240 0.12738188   -0.730   0.466532
## makeisuzu    -0.25054073 0.14345052   -1.747   0.082803 .
## makejaguar   -0.22052614 0.16375247   -1.347   0.180144
## makemazda    -0.00175073 0.11659959   -0.015   0.988041
## makemercedes-benz 0.15932468 0.17832092    0.893   0.373062
## makemercury  0.10324234 0.19139513    0.539   0.590412
## makemitsubishi -0.26590881 0.12600404   -2.110   0.036521 *
## makenissan    -0.08204821 0.11701308   -0.701   0.484292
## makepeugot   -0.25578965 0.12159680   -2.104   0.037114 *
## makeplymouth -0.28613764 0.12355696   -2.316   0.021947 *
## makeperssche 0.51735386 0.20096436    2.574   0.011029 *
## makerenault  -0.07178216 0.14777557   -0.486   0.627865
## makesaab     0.43740145 0.13904391    3.146   0.002004 **
## makesubaru   -0.31672177 0.22166277   -1.429   0.155166
## maketoyota   -0.14762328 0.10671120   -1.383   0.168638
## makevolkswagen 0.03708127 0.12781485    0.290   0.772135
## makevolvo    0.02451028 0.15446211    0.159   0.874137
## fuel.typegas  0.33538405 0.60319056    0.556   0.579043
## aspirationturbo 0.18521310 0.04517932    4.100 0.00006816516154209 ***
## num.of.doorsto -0.05475591 0.03555933   -1.540   0.125743
## num.stylehardtop -0.28074180 0.08110091   -3.462   0.000703 ***
## body.stylehatchback -0.30574895 0.07382836   -4.141 0.00005792742271285 ***
## body.stylesedan -0.29186573 0.07863633   -3.712   0.000291 ***
## body.stylewagon -0.25344796 0.08618938   -2.941   0.003805 **
## drive.wheelsfwd -0.17769521 0.05824449   -3.051   0.002707 **
## drive.wheelsrwd 0.01484228 0.08029142    0.185   0.853597
```

```
## wheel.base          0.02128003  0.00520311  4.090 0.00007075920491458 ***
## height             -0.03347828  0.00921689 -3.632  0.000387 ***
## engine.type1       0.16378364  0.12263757  1.336  0.183769
## engine.typeohc     -0.01857067  0.08287962 -0.224  0.823015
## engine.typeohcf     0.20163044  0.17610906  1.145  0.254100
## engine.typeohcv     0.07865636  0.08691309  0.905  0.366943
## engine.typerotor   0.39311683  0.18673618  2.105  0.036971 *
## num.of.cylindersfive 0.33964847  0.21410523  1.586  0.114801
## num.of.cylindersfour 0.36791500  0.25950901  1.418  0.158380
## num.of.cylinderssix 0.26861204  0.18403632  1.460  0.146540
## num.of.cylindersthree 0.41957329  0.20466906  2.050  0.042137 *
## num.of.cylinderstwelve 0.02335548  0.32777304  0.071  0.943291
## num.of.cylinderstwo 0.39311683  0.18673618  2.105  0.036971 *
## engine.size        0.00670854  0.00164787  4.071 0.00007609613709383 ***
## peak.rpm           0.00006562  0.00004187  1.567  0.119201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df   F p-value
## s(bore)      1.596  2.028 0.877  0.427
## s(stroke)    1.224  1.414 0.141  0.716
## s(compression.ratio) 1.865  2.171 1.031  0.230
##
## Rank: 76/78
## R-sq.(adj) =  0.94   Deviance explained = 95.6%
## GCV = 0.020524   Scale est. = 0.015043   n = 201
```

Osservando i p-value sembrerebbe che nessuna trasformazione possa essere significativa, ma procediamo osservando i grafici:



Effettivamente, per quanto riguarda le variabili bore e stroke notiamo un andamento molto lineare, mentre per quanto riguarda compression.ratio l'andamento non è molto lineare, proviamo quindi a inserire nel modello una componente quadratica per quanto riguarda compression.ratio.

Modello 10

```
## Call:
## lm(formula = log(price) ~ symboling + make + fuel.type + aspiration +
##   num.of.doors + body.style + drive.wheels + wheel.base + height +
##   engine.type + num.of.cylinders + engine.size + bore + stroke +
##   compression.ratio + I(compression.ratio^2) + peak.rpm, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30117 -0.06546 -0.00146  0.07629  0.32488
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.65384881  0.98472789  8.788 0.0000000000000353 ***
## symboling-1    0.02734084  0.08725959  0.313  0.754473
## symboling0   -0.01694471  0.10034822 -0.169  0.866138
## symboling1   -0.09931792  0.10335732 -0.961  0.338161
## symboling2   -0.10990514  0.10855104 -1.012  0.312964
## symboling3   -0.03518135  0.11544861 -0.305  0.760995
## makeaudi      0.38470232  0.16426847  2.342  0.020518 *
## makebmw       0.43481232  0.15886502  2.737  0.006960 **
## makechevrolet -0.22101131  0.14560892 -1.518  0.131187
## makedodge     -0.23679986  0.12389190 -1.911  0.057894 .
## makehonda     -0.06752090  0.12666876 -0.533  0.594798
## makeisuzu     -0.22383163  0.14252177 -1.571  0.118432
## makejaguar    -0.20222815  0.15905864 -1.271  0.205579
## makemazda     0.01438098  0.11649408  0.123  0.901920
## makemercedes-benz 0.21604771  0.17738450  1.218  0.225176
## makemercury   0.10189250  0.19146821  0.532  0.595410
## makemitsubishi -0.24793925  0.12568983 -1.973  0.050401 .
## makenissan    -0.05643816  0.11607261 -0.486  0.627522
## makepeugot   -0.66835587  0.29233080 -2.286  0.023656 *
```

```

## makeplymouth      -0.26988605  0.12332107 -2.188      0.030202 *
## makeporsche       0.50158858  0.19806878  2.532      0.012371 *
## makerenault       -0.04858744  0.14513434 -0.335      0.738268
## makesaab          0.45973141  0.13905461  3.306      0.001187 **
## makesubaru        -0.34848029  0.21782928 -1.600      0.111779
## maketoyota        -0.12936678  0.10644459 -1.215      0.226170
## makevolkswagen    0.05445672  0.12649301  0.431      0.667450
## makevolvo         0.02739623  0.15486250  0.177      0.859823
## fuel.typegas      0.49109088  0.53270198  0.922      0.358088
## aspirationturbo    0.17017954  0.04431911  3.840      0.000182 ***
## num.of.doorstwo   -0.05458276  0.03567479 -1.530      0.128149
## body.stylehardtop -0.28847397  0.08126266 -3.550      0.000517 ***
## body.stylehatchback -0.31454797  0.07357847 -4.275 0.00003409366507678 ***
## body.styleseadan  -0.29938315  0.07848346 -3.815      0.000200 ***
## body.stylewagon   -0.26050785  0.08601681 -3.029      0.002900 **
## drive.wheelsfwd   -0.17392492  0.05835917 -2.980      0.003368 **
## drive.wheelsrwd   0.01632103  0.08026238  0.203      0.839144
## wheel.base        0.02151469  0.00519762  4.139 0.00005825411653970 ***
## height            -0.03399878  0.00922241 -3.687      0.000318 ***
## engine.type1      0.58695625  0.29289309  2.004      0.046895 *
## engine.typeohc    -0.04352540  0.08219329 -0.530      0.597218
## engine.typeohcf   0.21908064  0.17339628  1.263      0.208408
## engine.typeohcv   0.07259020  0.08457524  0.858      0.392119
## engine.typerotor  0.78103808  0.33555197  2.328      0.021287 *
## num.of.cylindersfive 0.38982066  0.20532271  1.899      0.059567 .
## num.of.cylindersfour 0.40951710  0.24707909  1.657      0.099550 .
## num.of.cylinderssix 0.31344323  0.18051041  1.736      0.084569 .
## num.of.cylindersthree NA NA NA NA
## num.of.cylinderstwelve 0.12150833  0.30764989  0.395      0.693444
## num.of.cylinderstwo NA NA NA NA
## engine.size       0.00638319  0.00143043  4.462 0.00001595070321680 ***
## bore              -0.14236842  0.12313989 -1.156      0.249482
## stroke            0.04504961  0.06795264  0.663      0.508390
## compression.ratio -0.13479013  0.04778560 -2.821      0.005449 **
## I(compression.ratio^2) 0.00545409  0.00195707  2.787      0.006020 **
## peak.rpm          0.00005344  0.00004149  1.288      0.199705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1231 on 148 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.9555, Adjusted R-squared:  0.9399
## F-statistic: 61.15 on 52 and 148 DF, p-value: < 0.00000000000000022
##
## Single term deletions
##
## Model:
## log(price) ~ symboling + make + fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + height + engine.type +
##   num.of.cylinders + engine.size + bore + stroke + compression.ratio +
##   I(compression.ratio^2) + peak.rpm
##


|                        | Df | Sum of Sq | RSS    | AIC     | F value | Pr(>F)               |     |
|------------------------|----|-----------|--------|---------|---------|----------------------|-----|
| <none>                 |    |           | 2.2438 | -797.52 |         |                      |     |
| symboling              | 5  | 0.14296   | 2.3868 | -795.11 | 1.8859  | 0.1001153            |     |
| make                   | 20 | 2.41555   | 4.6594 | -690.65 | 7.9664  | 0.00000000000003284  | *** |
| fuel.type              | 1  | 0.01288   | 2.2567 | -798.37 | 0.8499  | 0.3580880            |     |
| aspiration             | 1  | 0.22354   | 2.4674 | -780.43 | 14.7446 | 0.0001820            | *** |
| num.of.doors           | 1  | 0.03549   | 2.2793 | -796.37 | 2.3409  | 0.1281491            |     |
| body.style             | 4  | 0.30376   | 2.5476 | -780.00 | 5.0089  | 0.0008164            | *** |
| drive.wheels           | 2  | 0.36565   | 2.6095 | -771.18 | 12.0591 | 0.000014066231893194 | *** |
| wheel.base             | 1  | 0.25977   | 2.5036 | -777.50 | 17.1341 | 0.000058254116539704 | *** |
| height                 | 1  | 0.20605   | 2.4499 | -781.86 | 13.5906 | 0.0003184            | *** |
| engine.type            | 3  | 0.03993   | 2.2837 | -799.98 | 0.8778  | 0.4541485            |     |
| num.of.cylinders       | 4  | 0.07262   | 2.3164 | -799.12 | 1.1974  | 0.3144215            |     |
| engine.size            | 1  | 0.30190   | 2.5457 | -774.15 | 19.9131 | 0.000015950703216804 |     |
| bore                   | 1  | 0.02027   | 2.2641 | -797.71 | 1.3367  | 0.2494819            |     |
| stroke                 | 1  | 0.00666   | 2.2505 | -798.92 | 0.4395  | 0.5083900            |     |
| compression.ratio      | 1  | 0.12063   | 2.3644 | -789.00 | 7.9565  | 0.0054492            | **  |
| I(compression.ratio^2) | 1  | 0.11775   | 2.3616 | -789.24 | 7.7666  | 0.0060202            | **  |
| peak.rpm               | 1  | 0.02516   | 2.2690 | -797.28 | 1.6593  | 0.1997054            |     |

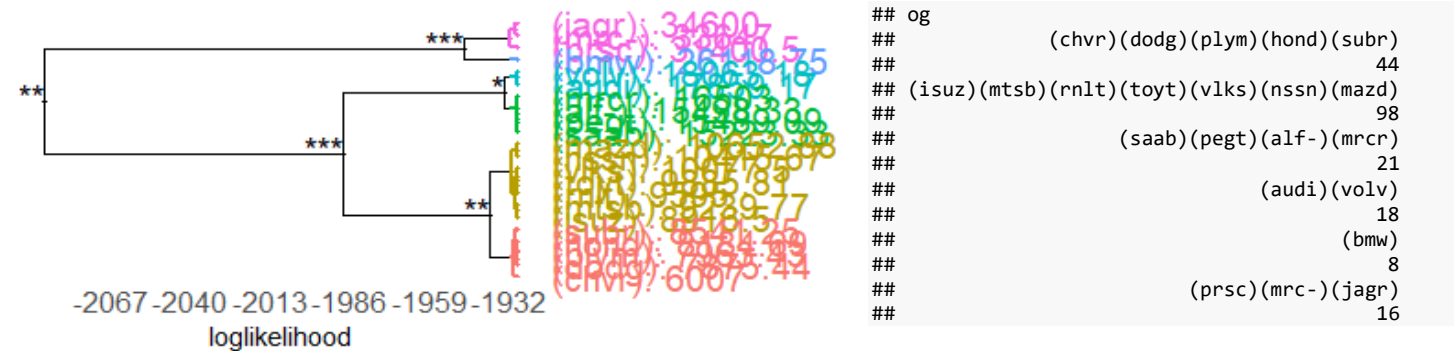

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Notiamo che la nuova variabile inserita è significativa e quindi la manteniamo nel modello.

Poichè make presenta un numero elevato di livelli, proviamo a vedere se aggregandoli il prezzo viene spiegato meglio.

Optimal grouping



```
## Call:
## lm(formula = log(price) ~ make, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65362 -0.18079 -0.00842  0.15492  0.71492
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.644102   0.161008  59.898 < 0.000000e+00 ***
## makeaudi       0.131417   0.197194   0.666   0.505993
## makebmw        0.473270   0.188799   2.507   0.013076 *
## makechevrolet  -0.948940   0.227700  -4.167   0.000479 ***
## makedodge      -0.702856   0.185916  -3.780   0.000213 ***
## makehonda      -0.661174   0.178623  -3.702   0.000285 ***
## makeisuzu      -0.577865   0.254576  -2.270   0.024406 *
## makejaguar      0.806312   0.227700   3.541   0.000508 ***
## makemazda      -0.435026   0.174638  -2.491   0.013648 *
## makemercedes-benz 0.762426   0.188799   4.038   0.0000798 ***
## makemercury     0.067196   0.322017   0.209   0.834942
## makemitsubishi -0.559876   0.178623  -3.134   0.002013 **
## makenissan      -0.470700   0.173909  -2.707   0.007456 **
## makepeugot     -0.006331   0.181642  -0.035   0.972236
## makeplymouth   -0.695446   0.192442  -3.614   0.000392 ***
## makeporsche     0.691858   0.212994   3.248   0.001387 **
## makerenault    -0.475594   0.254576  -1.868   0.063372 .
## makesaab       -0.028612   0.197194  -0.145   0.884798
## makesubaru     -0.615993   0.180013  -3.422   0.000770 ***
## maketoyota     -0.490471   0.168387  -2.913   0.004039 **
## makevolkswagen -0.446584   0.180013  -2.481   0.014030 *
## makevolvo      0.141590   0.181642   0.780   0.436713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2789 on 179 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.7241, Adjusted R-squared:  0.6917
## F-statistic: 22.37 on 21 and 179 DF, p-value: < 0.000000e+00
```

```
##
## Call:
## lm(formula = log(price) ~ og, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60628 -0.18533 -0.01521  0.15383  0.72659
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  8.96165    0.04095 218.863 < 0.000000000000002 ***
## og2          0.20008    0.04945   4.046    0.0000749 ***
## og3          0.67416    0.07204   9.358 < 0.000000000000002 ***
## og4          0.82045    0.07756  10.578 < 0.000000000000002 ***
## og5          1.15572    0.10439  11.071 < 0.000000000000002 ***
## og6          1.43484    0.08121  17.669 < 0.000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2716 on 195 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.7076
## F-statistic: 97.78 on 5 and 195 DF,  p-value: < 0.0000000000000022
```

Vediamo che l' R^2 aggiustato del modello che prevede la trasformazione di price in base alla variabile generata dall'optimal grouping è leggermente superiore a quello del modello che prevede la trasformazione di price in base a make. Decidiamo dunque di mantenere la variabile aggregata.

Modello 11

```
## Call:
## lm(formula = log(price) ~ symboling + og + fuel.type + aspiration +
##      num.of.doors + body.style + drive.wheels + wheel.base + height +
##      engine.type + num.of.cylinders + engine.size + bore + stroke +
##      compression.ratio + I(compression.ratio^2) + peak.rpm, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33660 -0.08997 -0.00791  0.07764  0.44975
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    5.38660657   0.95106336   5.664 0.0000000655 ***
## symboling-1     0.09929968   0.09842064   1.009   0.31450
## symboling0      0.12518002   0.10460090   1.197   0.23315
## symboling1      0.05287615   0.10635983   0.497   0.61976
## symboling2      0.06505960   0.10812897   0.602   0.54822
## symboling3      0.23155803   0.11758800   1.969   0.05062 .
## og2             0.09793474   0.03797526   2.579   0.01079 *
## og3             0.45475287   0.07621568   5.967 0.0000000146 ***
## og4             0.30632930   0.06992161   4.381 0.0000210710 ***
## og5             0.56220495   0.10166171   5.530 0.0000001249 ***
## og6             0.39556969   0.08721587   4.536 0.0000110890 ***
## fuel.typegas    1.11044839   0.57899778   1.918   0.05687 .
## aspirationturbo  0.16511633   0.04530833   3.644   0.00036 ***
## num.of.doorstwo -0.05938479   0.03932405  -1.510   0.13294
## body.stylehardtop -0.22160863   0.08833684  -2.509   0.01310 *
## body.stylehatchback -0.25890068   0.07552842  -3.428   0.00077 ***
## body.styleseadan -0.24155785   0.08207188  -2.943   0.00372 **
## body.stylewagon  -0.26353184   0.09032050  -2.918   0.00402 **
## drive.wheelsfwd  -0.10657315   0.06559521  -1.625   0.10616
## drive.wheelsrwd  0.02975888   0.08202271   0.363   0.71721
## wheel.base      0.01375384   0.00511459   2.689   0.00791 **
## height          -0.00052045   0.00827855  -0.063   0.94995
## engine.type1    -0.12323118   0.09965559  -1.237   0.21803
## engine.typeohc   0.09593627   0.06412361   1.496   0.13656
## engine.typeohcf  0.22178475   0.10361776   2.140   0.03381 *
## engine.typeohcv  0.08351517   0.08335548   1.002   0.31787
## engine.typeotor  0.39416065   0.29783788   1.323   0.18755
## num.of.cylindersfive 0.17872826   0.19124591   0.935   0.35140
## num.of.cylindersfour -0.04543107   0.21461545  -0.212   0.83262
## num.of.cylinderssix  0.06934944   0.13918635   0.498   0.61898
## num.of.cylindersthree 0.25089873   0.28308947   0.886   0.37677
## num.of.cylinders twelve -0.22092158   0.26494500  -0.834   0.40559
## num.of.cylinders two    NA              NA              NA              NA
## engine.size       0.00510838   0.00155275   3.290   0.00123 **
```



```
## bore                0.04883975  0.10864877  0.450    0.65365
## stroke              0.08357577  0.06624122  1.262    0.20886
## compression.ratio   -0.08063264  0.04843122 -1.665    0.09786 .
## I(compression.ratio^2) 0.00530325  0.00209811  2.528    0.01243 *
## peak.rpm            0.00013221  0.00003266  4.048 0.0000796982 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1429 on 163 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.934, Adjusted R-squared:  0.919
## F-statistic: 62.33 on 37 and 163 DF, p-value: < 0.0000000000000022
##
## Single term deletions
##
## Model:
## log(price) ~ symboling + og + fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + height + engine.type +
##   num.of.cylinders + engine.size + bore + stroke + compression.ratio +
##   I(compression.ratio^2) + peak.rpm
##
```

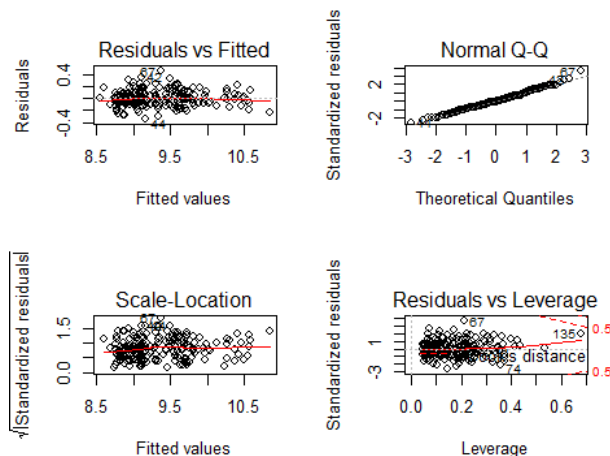
	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			3.3303	-748.15		
symboling	5	0.33593	3.6662	-738.83	3.2884	0.0074168 **
og	5	1.32908	4.6594	-690.65	13.0103	0.000000001172 ***
fuel.type	1	0.07515	3.4054	-745.66	3.6783	0.0568746 .
aspiration	1	0.27134	3.6016	-734.41	13.2808	0.0003601 ***
num.of.doors	1	0.04659	3.3769	-747.36	2.2805	0.1329439
body.style	4	0.24768	3.5780	-741.73	3.0307	0.0191937 *
drive.wheels	2	0.26705	3.5973	-736.65	6.5353	0.0018613 **
wheel.base	1	0.14775	3.4780	-741.42	7.2315	0.0079086 **
height	1	0.00008	3.3304	-750.15	0.0040	0.9499494
engine.type	4	0.15468	3.4850	-747.02	1.8926	0.1141539
num.of.cylinders	5	0.39653	3.7268	-735.54	3.8816	0.0023773 **
engine.size	1	0.22114	3.5514	-737.23	10.8235	0.0012283 **
bore	1	0.00413	3.3344	-749.90	0.2021	0.6536541
stroke	1	0.03252	3.3628	-748.20	1.5919	0.2088635
compression.ratio	1	0.05663	3.3869	-746.76	2.7719	0.0978558 .
I(compression.ratio^2)	1	0.13053	3.4608	-742.42	6.3890	0.0124348 *
peak.rpm	1	0.33477	3.6650	-730.90	16.3851	0.0000796981755 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Osserviamo un R^2 aggiustato (0.919) leggermente inferiore rispetto a quello precedente, ma un solo NA nel summary.

Diagnostiche fit11

Le diagnostiche del nostro modello sono molto simili a quelle del modello precedente. Le osservazioni che hanno leva pari ad 1 ora sono solo 2 (18, 47).



MODEL SELECTION (riportati solo primo e ultimo modello)

```
## Start: AIC=-748.15
## log(price) ~ symboling + og + fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + height + engine.type +
##   num.of.cylinders + engine.size + bore + stroke + compression.ratio +
##   I(compression.ratio^2) + peak.rpm
##
```

	Df	Sum of Sq	RSS	AIC
- height	1	0.00008	3.3304	-750.15
- bore	1	0.00413	3.3344	-749.90
- stroke	1	0.03252	3.3628	-748.20
<none>			3.3303	-748.15
- num.of.doors	1	0.04659	3.3769	-747.36
- engine.type	4	0.15468	3.4850	-747.02

```
## - compression.ratio      1  0.05663 3.3869 -746.76
## - fuel.type              1  0.07515 3.4054 -745.66
## - I(compression.ratio^2) 1  0.13053 3.4608 -742.42
## - body.style             4  0.24768 3.5780 -741.73
## - wheel.base            1  0.14775 3.4780 -741.42
## - symboling             5  0.33593 3.6662 -738.83
## - engine.size           1  0.22114 3.5514 -737.23
## - drive.wheels          2  0.26705 3.5973 -736.65
## - num.of.cylinders       5  0.39653 3.7268 -735.54
## - aspiration            1  0.27134 3.6016 -734.41
## - peak.rpm              1  0.33477 3.6650 -730.90
## - og                    5  1.32908 4.6594 -690.65
##
## Step: AIC=-751.96
## log(price) ~ symboling + og + fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + engine.type + num.of.cylinders +
##   engine.size + compression.ratio + I(compression.ratio^2) +
##   peak.rpm
##
##              Df Sum of Sq    RSS    AIC
## <none>                        3.3667 -751.96
## + stroke                    1  0.03229 3.3344 -751.90
## - num.of.doors              1  0.04625 3.4130 -751.22
## - engine.type              4  0.15625 3.5230 -750.84
## - compression.ratio        1  0.05601 3.4227 -750.65
## + bore                     1  0.00367 3.3631 -750.18
## - fuel.type                1  0.06529 3.4320 -750.10
## + height                   1  0.00015 3.3666 -749.97
## - I(compression.ratio^2)    1  0.11971 3.4864 -746.94
## - body.style               4  0.24656 3.6133 -745.76
## - wheel.base               1  0.14677 3.5135 -745.39
## - symboling                 5  0.32198 3.6887 -743.60
## - drive.wheels             2  0.26929 3.6360 -740.50
## - aspiration                1  0.30881 3.6755 -736.32
## - num.of.cylinders          5  0.47471 3.8414 -735.45
## - peak.rpm                 1  0.34219 3.7089 -734.51
## - engine.size              1  0.74115 4.1079 -713.97
## - og                       5  1.76879 5.1355 -677.09
```

Modello 12

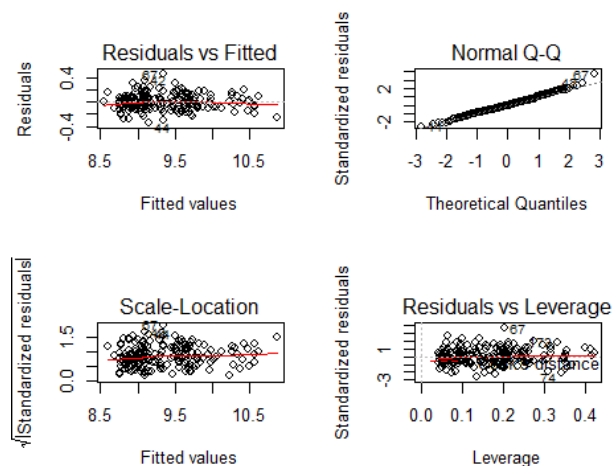
```
## Call:
## lm(formula = log(price) ~ symboling + og + fuel.type + aspiration +
##   num.of.doors + body.style + drive.wheels + wheel.base + engine.type +
##   num.of.cylinders + engine.size + compression.ratio + I(compression.ratio^2) +
##   peak.rpm, data = data_used)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33618 -0.08615 -0.01110  0.07631  0.46452
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.74923035   0.86445142   6.651 0.000000000403 ***
## symboling-1     0.10581815   0.09704546   1.090   0.277118
## symboling0      0.12801125   0.10255748   1.248   0.213719
## symboling1      0.05698275   0.10288448   0.554   0.580426
## symboling2      0.07570094   0.10495315   0.721   0.471750
## symboling3      0.22819749   0.11367317   2.007   0.046320 *
## og2             0.09251424   0.03417802   2.707   0.007503 **
## og3             0.41402589   0.06319462   6.552 0.000000000685 ***
## og4             0.29474567   0.06585685   4.476 0.000014104786 ***
## og5             0.53120439   0.08768798   6.058 0.000000008949 ***
## og6             0.37152383   0.08415537   4.415 0.000018159505 ***
## fuel.typegas    1.01446212   0.56541749   1.794   0.074604 .
## aspirationturbo  0.17306493   0.04435169   3.902   0.000138 ***
## num.of.doorstwo -0.05868270   0.03886012  -1.510   0.132919
## body.stylehardtop -0.22347188   0.08800035  -2.539   0.012021 *
## body.stylehatchback -0.25724690   0.07511366  -3.425   0.000775 ***
## body.styleседan -0.23684260   0.08166745  -2.900   0.004236 **
## body.stylewagon -0.25905241   0.08831585  -2.933   0.003829 **
## drive.wheelsfwd -0.09783427   0.06304690  -1.552   0.122622
## drive.wheelsrwd  0.02911269   0.07537529   0.386   0.699816
## wheel.base      0.01259850   0.00468328   2.690   0.007873 **
## engine.type1    -0.10400913   0.09822264  -1.059   0.291178
## engine.typeohc  0.09892902   0.06199680   1.596   0.112455
## engine.typeohcf  0.18376831   0.07877792   2.333   0.020862 *
## engine.typeohcv  0.06020953   0.07760990   0.776   0.438973
```

```
## engine.typerotor      0.50759404  0.20536530  2.472      0.014457 *
## num.of.cylindersfive  0.26482386  0.15892222  1.666      0.097525 .
## num.of.cylindersfour  0.06308169  0.16495580  0.382      0.702642
## num.of.cylinderssix   0.13823453  0.12468957  1.109      0.269194
## num.of.cylindertextree 0.34292336  0.24465200  1.402      0.162878
## num.of.cylindertwelve -0.33056783  0.23902394 -1.383      0.168525
## num.of.cylindertwo     NA          NA          NA          NA
## engine.size           0.00630318  0.00104269  6.045 0.000000009550 ***
## compression.ratio     -0.07859802  0.04729554 -1.662      0.098430 .
## I(compression.ratio^2) 0.00502870  0.00206982  2.430      0.016184 *
## peak.rpm              0.00013275  0.00003232  4.108 0.000062653858 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1424 on 166 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.9333, Adjusted R-squared:  0.9196
## F-statistic: 68.28 on 34 and 166 DF,  p-value: < 0.0000000000000022
##
## Single term deletions
##
## Model:
## log(price) ~ symboling + og + fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + engine.type + num.of.cylinders +
##   engine.size + compression.ratio + I(compression.ratio^2) +
##   peak.rpm
##
##              Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                3.3667 -751.96
## symboling             5  0.32198 3.6887 -743.60  3.1751    0.0091654 **
## og                    5  1.76879 5.1355 -677.09 17.4425 0.0000000000007324 ***
## fuel.type             1  0.06529 3.4320 -750.10  3.2191    0.0746040 .
## aspiration            1  0.30881 3.6755 -736.32 15.2264    0.0001383 ***
## num.of.doors          1  0.04625 3.4130 -751.22  2.2804    0.1329189
## body.style            4  0.24656 3.6133 -745.76  3.0392    0.0188834 *
## drive.wheels          2  0.26929 3.6360 -740.50  6.6388    0.0016839 **
## wheel.base            1  0.14677 3.5135 -745.39  7.2366    0.0078731 **
## engine.type           4  0.15625 3.5230 -750.84  1.9260    0.1084247
## num.of.cylinders      5  0.47471 3.8414 -735.45  4.6812    0.0005031 ***
## engine.size           1  0.74115 4.1079 -713.97 36.5433 0.00000000955038053 ***
## compression.ratio     1  0.05601 3.4227 -750.65  2.7617    0.0984305 .
## I(compression.ratio^2) 1  0.11971 3.4864 -746.94  5.9027    0.0161843 *
## peak.rpm              1  0.34219 3.7089 -734.51 16.8721 0.00006265385810925 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'R² aggiustato (0.9196) è praticamente uguale a quello del modello precedente ed è presente ancora un NA relativo alla variabile num.of.cylinders.

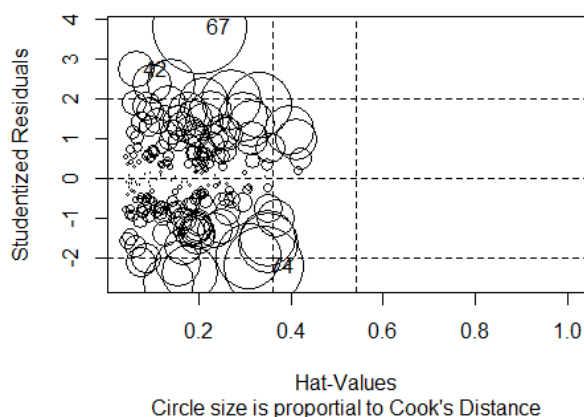
Diagnostiche fit12

I punti con valori di leva pari a 1 sono i 2 del caso precedente (18, 47).



PUNTI INFLUENTI

Influence Plot



```
##      StudRes      Hat      CookD
## 19      NaN 1.00000000      NaN
## 42  2.741740 0.06472589 0.01430209
## 50      NaN 1.00000000      NaN
## 67  3.801214 0.20404275 0.09789818
## 74 -2.190595 0.34136003 0.06946953
```

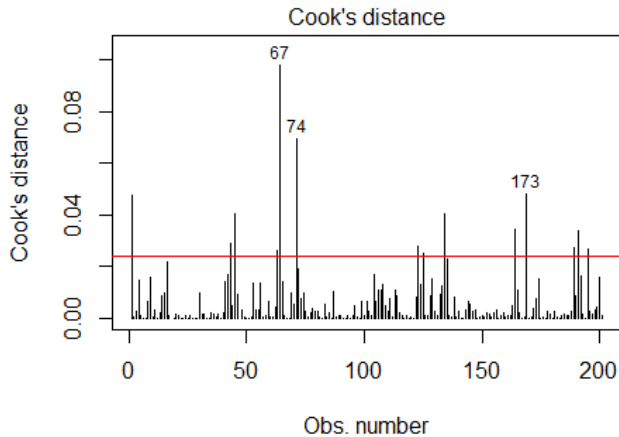
Procediamo calcolando le distanze di Cook relative ad ogni osservazione ed inserendole in un dataframe.

Soglia

Calcoliamo la soglia oltre la quale possiamo definire delle osservazioni influenti sul modello, considerando la loro distanza di Cook.

Abbiamo deciso di utilizzare la formula $4/(n-p)$ poichè stiamo lavorando con una numerosità non troppo elevata. Il valore della soglia è pari a 0.02424242.

Grafico punti influenti



Notiamo che la distanza di Cook di alcune osservazioni supera la soglia calcolata.

Dataset senza valori non influenti

Creiamo un dataset che contiene solo i valori non influenti.

Consideriamo come non influenti sul modello solo le osservazioni che hanno distanza di Cook minore della soglia. Abbiamo inoltre eliminato le osservazioni con distanza di leva pari a 1 (essendo la distanza di Cook in funzione della leverage, il software ci restituiva valore mancante).

Modello 13

```
## Call:
## lm(formula = log(price) ~ symboling + og + fuel.type + aspiration +
##   num.of.doors + body.style + drive.wheels + wheel.base + engine.type +
##   num.of.cylinders + engine.size + compression.ratio + I(compression.ratio^2) +
##   peak.rpm, data = data_noinflu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27110 -0.07548 -0.01011  0.07129  0.37324
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.07923775  0.85190069   7.136 0.00000000003675 ***
## symboling-1     0.08031017  0.13307883   0.603  0.547090
## symboling0     0.06634550  0.13658182   0.486  0.627839
## symboling1    -0.00576253  0.13688087  -0.042  0.966475
## symboling2     0.00444363  0.13723246   0.032  0.974211
## symboling3     0.13659500  0.14456403   0.945  0.346222
## og2            0.07831686  0.03034830   2.581  0.010809 *
## og3            0.40990201  0.05766118   7.109 0.00000000004263 ***
## og4            0.30539133  0.06011186   5.080 0.00000109018070 ***
## og5            0.56175720  0.08392324   6.694 0.000000000039429 ***
## og6            0.34869868  0.08875371   3.929  0.000129 ***
## fuel.typegas    1.23789968  0.55541696   2.229  0.027296 *
## aspirationturbo  0.17005238  0.04327391   3.930  0.000129 ***
## num.of.doorstwo -0.04899729  0.03400642  -1.441  0.151691
## body.stylehardtop -0.17998798  0.09472582  -1.900  0.059314 .
## body.stylehatchback -0.23128355  0.08304632  -2.785  0.006034 **
## body.styleседan -0.20864148  0.08799529  -2.371  0.018991 *
## body.stylewagon -0.22488354  0.09280333  -2.423  0.016558 *
## drive.wheelsfwd -0.10473272  0.05534179  -1.892  0.060330 .
## drive.wheelsrwd -0.00995707  0.07005730  -0.142  0.887167
## wheel.base      0.00730845  0.00449361   1.626  0.105934
```

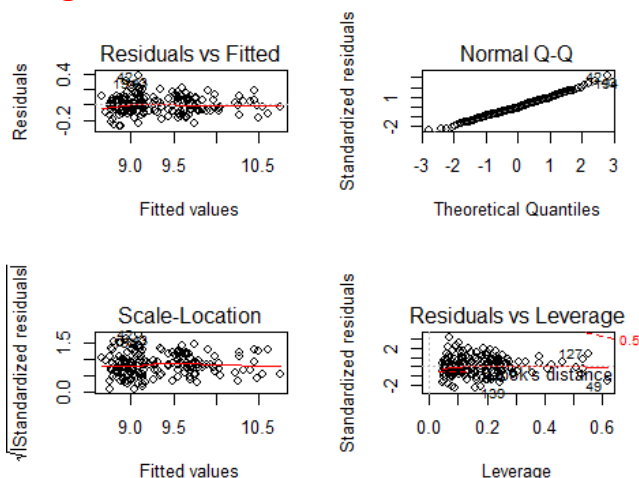
```
## engine.type1 -0.06050170 0.09786878 -0.618 0.537374
## engine.typeohc 0.03992108 0.06255085 0.638 0.524292
## engine.typeohcf 0.09897509 0.07860164 1.259 0.209889
## engine.typeohcv 0.00546028 0.07071805 0.077 0.938556
## engine.typerotor 0.51900788 0.19125383 2.714 0.007422 **
## num.of.cylindersfive 0.26223005 0.14994923 1.749 0.082346 .
## num.of.cylindersfour 0.03062491 0.15906051 0.193 0.847579
## num.of.cylinderssix 0.06689579 0.12252064 0.546 0.585869
## num.of.cylinderstwo NA NA NA NA
## engine.size 0.00731983 0.00096511 7.584 0.00000000000309 ***
## compression.ratio -0.07164817 0.04290618 -1.670 0.097001 .
## I(compression.ratio^2) 0.00521427 0.00189575 2.751 0.006673 **
## peak.rpm 0.00011493 0.00002919 3.937 0.000125 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1235 on 152 degrees of freedom
## Multiple R-squared:  0.9455, Adjusted R-squared:  0.934
## F-statistic: 82.33 on 32 and 152 DF, p-value: < 0.0000000000000022
##
## Single term deletions
##
## Model:
## log(price) ~ symboling + og + fuel.type + aspiration + num.of.doors +
## body.style + drive.wheels + wheel.base + engine.type + num.of.cylinders +
## engine.size + compression.ratio + I(compression.ratio^2) +
## peak.rpm
##
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			2.3178	-744.25		
symboling	5	0.22810	2.5459	-736.88	2.9916	0.0132243 *
og	5	1.49873	3.8166	-661.99	19.6568	0.00000000000004482 ***
fuel.type	1	0.07575	2.3936	-740.30	4.9674	0.0272960 *
aspiration	1	0.23548	2.5533	-728.35	15.4423	0.0001288 ***
num.of.doors	1	0.03166	2.3495	-743.74	2.0760	0.1516910 .
body.style	4	0.12659	2.4444	-742.41	2.0753	0.0867448 .
drive.wheels	2	0.14726	2.4651	-736.85	4.8284	0.0092677 **
wheel.base	1	0.04034	2.3582	-743.06	2.6452	0.1059345 .
engine.type	4	0.04027	2.3581	-749.06	0.6603	0.6205258 .
num.of.cylinders	3	0.27057	2.5884	-729.82	5.9146	0.0007635 ***
engine.size	1	0.87717	3.1950	-686.87	57.5237	0.000000000003091924 ***
compression.ratio	1	0.04252	2.3604	-742.89	2.7885	0.0970012 .
I(compression.ratio^2)	1	0.11536	2.4332	-737.26	7.5653	0.0066731 **
peak.rpm	1	0.23632	2.5542	-728.29	15.4976	0.0001254 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notiamo che l' R^2 aggiustato è uguale a 0.934, quindi è leggermente aumentato rispetto a quello del modello precedente. Rimangono valori mancanti per una delle stime relative a num.of.cylinders (fattore su più livelli).

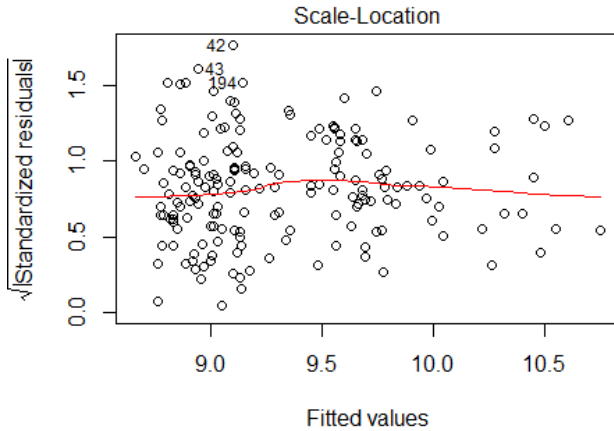
Diagnostiche fit13



Osserviamo che le diagnostiche in generale sembrano essere molto simili a quelle del modello precedente. Per quanto riguarda però il grafico residui standardizzati vs leverage notiamo che tutte le unità statistiche hanno leverage non superiore a 0.6.

ETEROSCHEDASTICITÀ

Scale-Location fit13



Controlliamo l'assunto di omoschedasticità utilizzando i test, anche se dal grafico sembrerebbe rispettato.

Breusch-Pagan Test

```
## studentized Breusch-Pagan test
##
## data: fit13
## BP = 23.12, df = 32, p-value = 0.8745
```

Proviamo ad effettuare anche il test di White (più preciso).

Test di White

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.547028, Df = 1, p = 0.21357
```

Siamo portati ad accettare l'ipotesi nulla di omoschedasticità.

Il nostro modello 13 è dunque il modello robusto.

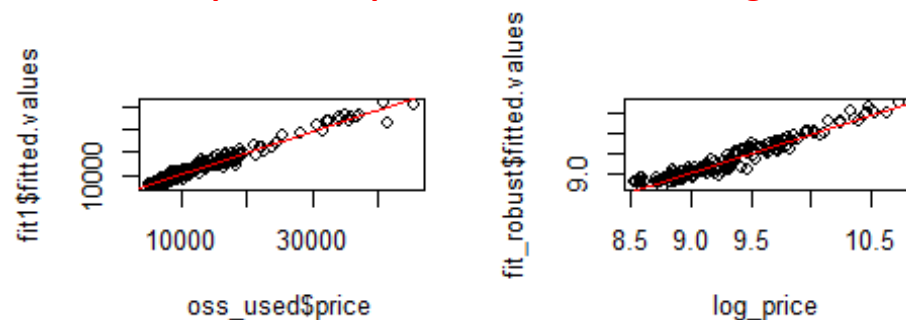
Confronto modello iniziale/modello finale

```
ncvTest(fit1)
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 56.17918, Df = 1, p = 0.000000000000066159
ncvTest(fit13)
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.547028, Df = 1, p = 0.21357
```

Osserviamo che i passaggi intermedi per la costruzione del modello robusto hanno portato alla risoluzione del problema di eteroschedasticità (che era presente nel nostro modello iniziale).

MODELLO ROBUSTO

Confronto risposta e risposta stimata tra starting model e robust model



Osserviamo che nel grafico relativo al modello finale i punti associati alle osservazioni con prezzo più alto si discostano meno dalla linea rossa.

BOOTSTRAP

Controllo ipotesi normalità

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: fit_robust$residuals
```

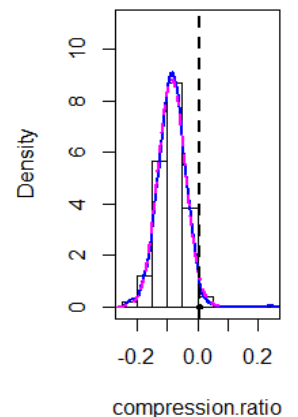
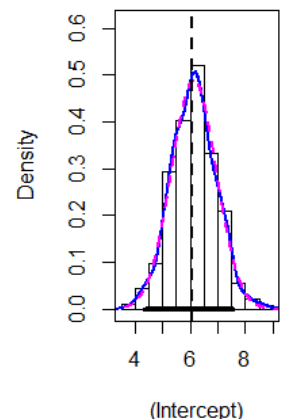
```
## W = 0.99353, p-value = 0.5941
```

La normalità del nostro modello è rispettata e la numerosità è elevata. Applichiamo la procedura bootstrap per vedere se l'inferenza fatta sulla base del nostro modello robusto è corretta.

```
Number of bootstrap replications R = 1159
      original      bootBias      bootSE      bootMed      bootskew      bootKurtosis
(Intercept)      6.07923775 0.055398641 0.816041279 6.12919013 0.066397 0.1049427
symboling-1      0.08031017 0.001027053 0.057988666 0.07660946 0.523969 0.7175085
symboling0      0.06634550 0.017025317 0.063568816 0.08219459 0.224272 0.2183152
symboling1      -0.00576253 0.014006801 0.065974408 0.00555728 0.283295 0.3609519
symboling2      0.00444363 0.017857726 0.067481996 0.02314761 0.109446 0.0311659
symboling3      0.13659500 0.006861261 0.057618667 0.14140763 0.144380 0.3787237
og2      0.07831686 -0.014058241 0.03527524 0.06593270 -0.454076 0.9090103
og3      0.40990201 -0.013609454 0.062631722 0.39673036 -0.220113 0.4792546
og4      0.30539133 -0.007714945 0.072345460 0.30234583 -0.317323 0.3578425
og5      0.56175720 -0.022491094 0.086406296 0.53930404 -0.238598 0.3807793
og6      0.34869868 -0.010617380 0.125587660 0.33942524 -0.090478 0.2905073
fuel.typegas      1.23789968 -0.031855093 0.518372017 1.22316421 -0.164382 0.3508028
aspirationturbo    0.17005238 -0.010110239 0.041460089 0.16294222 -0.290622 0.0094346
num.of.doorstwo   -0.04899729 -0.002844764 0.035015136 -0.05129954 -0.052275 -0.1888848
body.stylehardtop -0.17998798 -0.022632453 0.127071175 -0.19186864 -0.930647 8.3442006
body.stylehatchback -0.23128355 -0.015354560 0.102579410 -0.24104147 -1.794799 13.9425733
body.stylehatchback -0.20864148 -0.013399552 0.108349933 -0.21907030 -1.532543 11.1158425
body.stylewagon   -0.22488354 -0.014304153 0.110822811 -0.23720561 -1.487824 10.1116984
drive.wheelsfwd   -0.10473272 -0.002668820 0.049885380 0.10551748 0.067997 0.3283748
drive.wheelsrwd   -0.00995707 0.001299819 0.070857242 -0.00882202 -0.163653 0.6062458
wheel.base      0.00730845 0.000954620 0.005223259 0.00820730 0.244449 0.3951395
engine.type1      -0.06050170 -0.045393836 0.121140503 -0.10101547 -0.280298 0.1957152
engine.typeohc    0.03992108 -0.021294118 0.078302653 0.01843413 0.086451 0.2516005
engine.typeohcf   0.09897509 -0.044198755 0.108434276 0.06401819 -0.373582 0.0701195
engine.typeohcv   0.00546028 -0.010906309 0.081045571 -0.00128334 -0.354973 0.6249236
engine.typerotor  0.51900788 0.056787476 0.217639925 0.56308835 0.414816 0.3186472
num.of.cylindersfive 0.26223005 0.039645347 0.160766504 0.29478091 0.409456 0.8341524
num.of.cylindersfour 0.03062491 0.006303034 0.06267135 0.503177 0.5568007 0.4909235
num.of.cylinderssix 0.06689579 0.031644632 0.152215781 0.07635649 0.538756 0.4909235
engine.size      0.00731983 0.000127292 0.001020636 0.00746556 0.100294 0.0390518
compression.ratio -0.07164817 -0.012931913 0.045241817 -0.08375366 0.111512 2.0234589
I(compression.ratio^2) 0.00521427 0.000330138 0.001947381 0.00548381 -0.432720 4.9434546
peak.rpm      0.00011493 -0.000010662 0.000033278 0.00010535 -0.184227 0.0303988
```

Calcoliamo gli intervalli di confidenza percentuali boot e riportiamo alcuni grafici:

```
## Bootstrap bca confidence intervals
##
##              Estimate      2.5 %      97.5 %
## (Intercept)      6.0792377521 4.36951305690 7.5419785584
## symboling-1      0.0803101714 -0.02699263028 0.2127810827
## symboling0      0.0663455012 -0.06045891680 0.1746498573
## symboling1      -0.0057625253 -0.13699037843 0.1180483009
## symboling2      0.0044436289 -0.13496162983 0.1235657110
## symboling3      0.1365950020 -0.00443200832 0.2867242962
## og2      0.0783168600 0.02366206318 0.1491969961
## og3      0.4099020086 0.29641223533 0.5485802624
## og4      0.3053913257 0.15809084284 0.4368713747
## og5      0.5617572022 0.40435386615 0.7334789493
## og6      0.3486986767 0.08898002638 0.6123970733
## fuel.typegas      1.2378996792 0.21247220545 2.2662291025
## aspirationturbo    0.1700523813 0.09002320973 0.2492370181
## num.of.doorstwo   -0.0489972903 -0.11817512520 0.0214647370
## body.stylehardtop -0.1799879802 -0.40653172779 0.0933412619
## body.stylehatchback -0.2312835506 -0.37802508519 0.085217625
## body.stylehatchback -0.2086414803 -0.38839044618 0.0194316895
## body.stylewagon   -0.2248835391 -0.41123990788 0.0065207029
## drive.wheelsfwd   -0.1047327212 -0.20458362743 -0.0023723176
## drive.wheelsrwd   -0.0099570748 -0.15065120323 0.1215331636
## wheel.base      0.0073084451 -0.00251537400 0.0169599881
## engine.type1      -0.0605017030 -0.25539992242 0.1853133134
## engine.typeohc    0.0399210763 -0.08823133493 0.2363563590
## engine.typeohcf   0.0989750852 -0.08445997350 0.3358829951
## engine.typeohcv   0.0054602823 -0.15011454009 0.1670796138
## engine.typerotor  0.5190078751 0.10989866354 0.9209252145
## num.of.cylindersfive 0.2622300456 -0.05809238847 0.5543104900
## num.of.cylindersfour 0.0306249061 -0.33804975475 0.3932152357
## num.of.cylinderssix 0.0668957909 -0.20473598121 0.4015653876
## engine.size      0.0073198324 0.00525628466 0.0091974741
## compression.ratio -0.0716481654 -0.14611163092 0.0251912268
## I(compression.ratio^2) 0.0052142740 0.00176632579 0.0087487448
## peak.rpm      0.0001149319 0.00005743685 0.0001825645
```



Tutte le stime MLE sono contenute negli intervalli boot, quindi possiamo dire che l'inferenza è corretta.

MODELLO LOGISTICO

DATI

Consideriamo la costruzione del modello robusto e utilizziamo quanto ottenuto per lavorare su un modello logistico.

RISPOSTA BINARIA

Abbiamo reso la variabile risposta binaria in modo tale che venga attribuito valore 1 alle auto costose (price>15000) e valore 0 alle altre.

MODELLO LOGISTICO 1

ERRORE

Osserviamo che c'è qualche problema. Poiché nella costruzione del modello robusto abbiamo già lavorato su collinearità e su zero variance, proviamo a controllare se c'è separazione/quasi-separazione.

SEPARAZIONE/QUASI-SEPARAZIONE

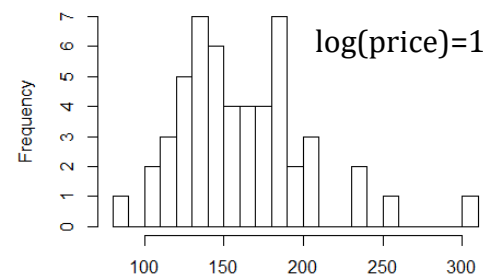
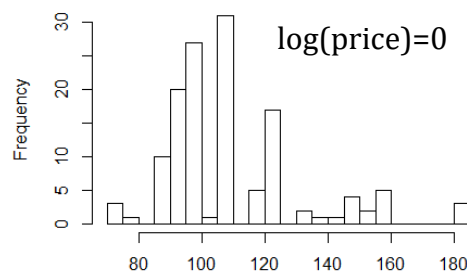
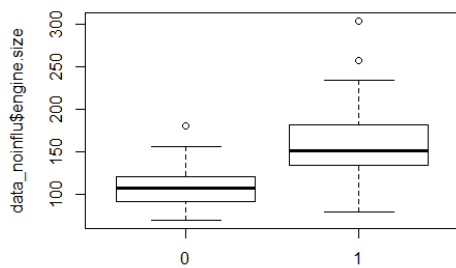
Controllo significatività covariate

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Single term deletions
##
## Model:
## logprice ~ symboling + og + fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + engine.type + num.of.cylinders +
##   engine.size + compression.ratio + I(compression.ratio^2) +
##   peak.rpm
##
##           Df Deviance      AIC      LRT      Pr(>Chi)
## <none>          648.79   714.79
## symboling        5   865.05   921.05  216.26 < 0.00000000000000022 ***
## og              5  1009.22  1065.22  360.44 < 0.00000000000000022 ***
## fuel.type       1   504.61   568.61   0.00      1
## aspiration      1   648.79   712.79   0.00      1
## num.of.doors    1   576.70   640.70   0.00      1
## body.style      4   576.70   634.70   0.00      1
## drive.wheels    2   648.79   710.79   0.00      1
## wheel.base     1   504.61   568.61   0.00      1
## engine.type     4   648.79   706.79   0.00      1
## num.of.cylinders 3   720.87   780.87   72.09  0.000000000000001525 ***
## engine.size     1   720.87   784.87   72.09 < 0.00000000000000022 ***
## compression.ratio 1   576.70   640.70   0.00      1
## I(compression.ratio^2) 1   648.79   712.79   0.00      1
## peak.rpm       1   504.61   568.61   0.00      1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Osserviamo che le covariate symboling, og ed engine.size hanno p-value<0.00000000000000022, che num.of cylinders ha p-value pari a 0.000000000000001525 e che tutte le altre non sono significative.

Controllo separazione/quasi-separazione

```
table(data_noinflu$logprice, data_noinflu$og)
##
##      1  2  3  4  5  6
## 0 43 82  6  2  0  0
## 1  0  8 13 13  8 10
table(data_noinflu$logprice, data_noinflu$symboling)
##
##     -2 -1  0  1  2  3
## 0  0  9 43 44 24 13
## 1  1 13 15  8  4 11
```



```
data_noinflu$logprice
## data_noinflu$logprice: 0
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.0   92.0   108.0   109.5  121.0   181.0
## -----
## data_noinflu$logprice: 1
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   80.0  135.5   152.0   160.5  181.5   304.0
```

Nelle tables relative a symboling e a og (fattori) sono presenti valori pari a 0 mentre, osservando le statistiche descrittive e gli istogrammi relativi a engine.size stratificati per logprice, possiamo dire che la maggior parte delle unità statistiche si trova tra il primo e il terzo quartile sia per logprice=1 che per logprice=0 e che il primo quartile della distribuzione di engine.size per logprice=1 è maggiore del terzo quartile per logprice=0. Possiamo allora affermare che symboling, og e engine.size sono causa del problema della quasi-separazione. Proviamo a eliminarle dal modello.

MODELLO LOGISTICO 2

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Single term deletions
##
## Model:
## logprice ~ fuel.type + aspiration + num.of.doors + body.style +
##   drive.wheels + wheel.base + engine.type + num.of.cylinders +
##   compression.ratio + I(compression.ratio^2) + peak.rpm
##
##          Df Deviance   AIC    LRT   Pr(>Chi)
## <none>          55.061  99.061
## fuel.type       1   55.071  97.071  0.0099   0.920749
## aspiration      1   62.899 104.899  7.8381   0.005116 **
## num.of.doors    1   55.402  97.402  0.3412   0.559134
## body.style      4   67.075 103.075 12.0141   0.017247 *
## drive.wheels    2   73.347 113.347 18.2859   0.000107 ***
## wheel.base      1   62.719 104.719  7.6575   0.005654 **
## engine.type     4   59.572  95.572  4.5103   0.341324
## num.of.cylinders 3   81.720 119.720 26.6585  0.000006942 ***
## compression.ratio 1   58.907 100.907  3.8462   0.049860 *
## I(compression.ratio^2) 1   56.307  98.307  1.2460   0.264320
## peak.rpm        1   55.459  97.459  0.3973   0.528470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
table(data_noinflu$logprice, data_noinflu$num.of.cylinders)
##
##      eight five four six three twelve two
## 0      0      1 126  3      0      0  3
## 1      3      9  20 19      0      0  1
```

Vediamo che ora è la covariata num.of.cylinders molto più significativa rispetto alle altre e che causa il problema della quasi-separazione.

Proviamo allora a fittare il modello senza num.of.cylinders.

MODELLO LOGISTICO 3

```
## Call:
## glm(formula = logprice ~ fuel.type + aspiration + num.of.doors +
##   body.style + drive.wheels + wheel.base + engine.type + compression.ratio +
```

```
##      I(compression.ratio^2) + peak.rpm, family = "binomial", data = data_noinflu)
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -1.77843  -0.27362  -0.10068   0.03685   2.38141
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -47.7803470  23.5287333  -2.031  0.04228 *
## fuel.typegas    -13.1985198  16.2749401  -0.811  0.41738
## aspirationturbo   1.0218710   1.5451654   0.661  0.50840
## num.of.doorstwo  0.8970588   1.2485658   0.718  0.47247
## body.stylehardtop -4.6878841   2.1990408  -2.132  0.03302 *
## body.stylehatchback -4.7613084   1.7439236  -2.730  0.00633 **
## body.styleседan  -4.0120938   1.9475168  -2.060  0.03939 *
## body.stylewagon  -6.9383251   2.4904355  -2.786  0.00534 **
## drive.wheelsfwd  -1.1441974   1.7133081  -0.668  0.50424
## drive.wheelsrwd   2.1678199   1.8416266   1.177  0.23915
## wheel.base       0.5487414   0.1402453   3.913  0.0000913 ***
## engine.type1     -5.6920747   2.1938647  -2.595  0.00947 **
## engine.typeohc    -0.6255469   1.3246480  -0.472  0.63676
## engine.typeohcf   1.1382331   1.9232997   0.592  0.55398
## engine.typeohcv   2.1301749   1.5500468   1.374  0.16936
## engine.typerotor  -1.3171907   1.6333830  -0.806  0.42000
## compression.ratio 0.8561902   1.8777231   0.456  0.64841
## I(compression.ratio^2) -0.0641995   0.0829392  -0.774  0.43890
## peak.rpm         0.0013705   0.0007846   1.747  0.08068 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 219.77  on 184  degrees of freedom
## Residual deviance:  81.72  on 166  degrees of freedom
## AIC: 119.72
##
## Number of Fisher Scoring iterations: 7
##
## Single term deletions
##
## Model:
## logprice ~ fuel.type + aspiration + num.of.doors + body.style +
##      drive.wheels + wheel.base + engine.type + compression.ratio +
##      I(compression.ratio^2) + peak.rpm
##              Df Deviance      AIC      LRT      Pr(>Chi)
## <none>                81.720 119.72
## fuel.type             1   82.364 118.36  0.6443    0.422152
## aspiration             1   82.164 118.16  0.4446    0.504916
## num.of.doors          1   82.242 118.24  0.5223    0.469859
## body.style            4   96.981 126.98 15.2609    0.004190 **
## drive.wheels          2  102.897 136.90 21.1772  0.0000252017 ***
## wheel.base            1  109.318 145.32 27.5987  0.0000001493 ***
## engine.type           5  100.685 128.69 18.9654    0.001951 **
## compression.ratio     1   81.926 117.93  0.2066    0.649482
## I(compression.ratio^2) 1   82.338 118.34  0.6181    0.431765
## peak.rpm              1   85.006 121.01  3.2862    0.069865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ora il problema della quasi-separazione è stato risolto.

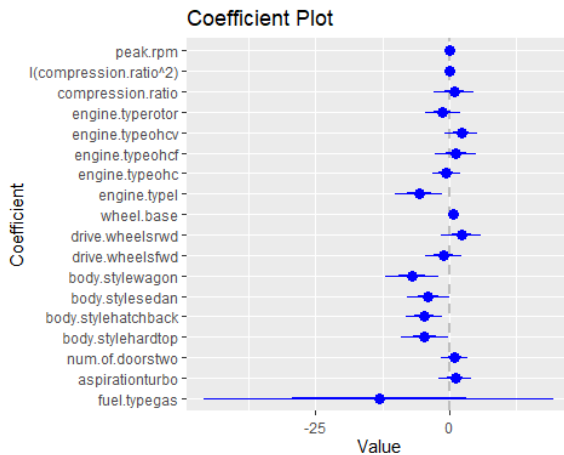
OR

	OR	2.5 %	97.5 %
## (Intercept)	0.00	0.00	0.01
## fuel.typegas	0.00	0.00	27396488.95
## aspirationturbo	2.78	0.14	65.56
## num.of.doorstwo	2.45	0.22	30.99
## body.stylehardtop	0.01	0.00	0.52
## body.stylehatchback	0.01	0.00	0.24
## body.styleседan	0.02	0.00	0.73
## body.stylewagon	0.00	0.00	0.09
## drive.wheelsfwd	0.32	0.01	13.15
## drive.wheelsrwd	8.74	0.24	478.36
## wheel.base	1.73	1.36	2.38
## engine.type1	0.00	0.00	0.17
## engine.typeohc	0.53	0.03	6.83

```
## engine.typeohcf      3.12  0.07    140.61
## engine.typeohcv      8.42  0.38    192.35
## engine.typeotor      0.27  0.01      5.44
## compression.ratio    2.35  0.08     81.56
## I(compression.ratio^2) 0.94  0.82      1.07
## peak.rpm             1.00  1.00      1.00
```

Possiamo confrontare i p-value nel summary e gli intervalli di confidenza relativi agli OR: se l'intervallo di confidenza contiene l'1, la covariata non è significativa.

Coefplot



Osserviamo le caratteristiche delle automobili che hanno prezzo elevato.

Previsioni

```
##          predicted
## observed      0      1
##          0 0.65945946 0.05945946
##          1 0.04864865 0.23243243
accuracy <- 0.65945946 + 0.23243243
```

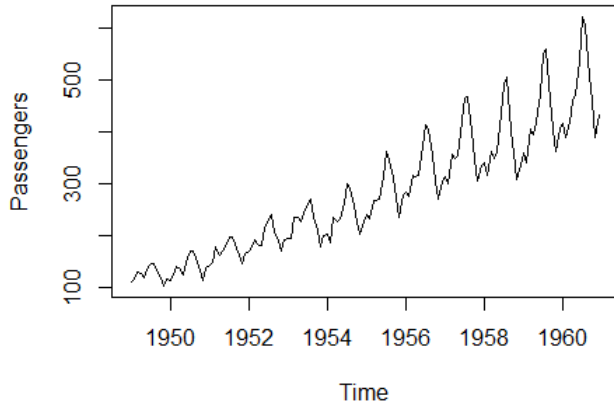
Otteniamo un'accuracy elevata (pari quasi al 90%), ma dobbiamo ricordare che non si tratta di una misura robusta (misura ottimistica).

SERIE STORICA

IMPORTAZIONE DATI

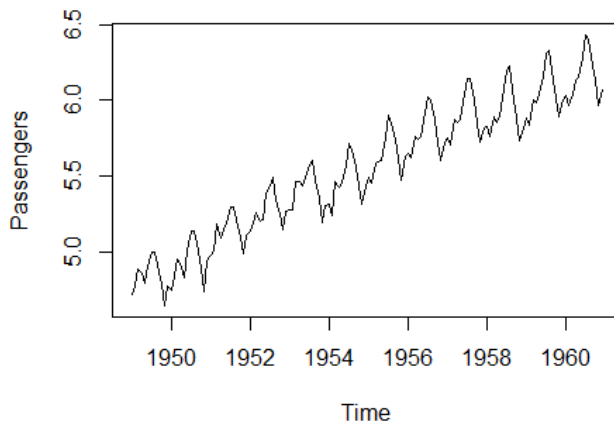
Abbiamo importato un dataset contenente informazioni relative al numero di passeggeri presenti sui voli aerei a partire da gennaio del 1949 fino a dicembre del 1961. L'obiettivo è quello di trovare il processo stocastico che ha generato la serie storica considerata e di farne una previsione.

GRAFICO DELLA SERIE



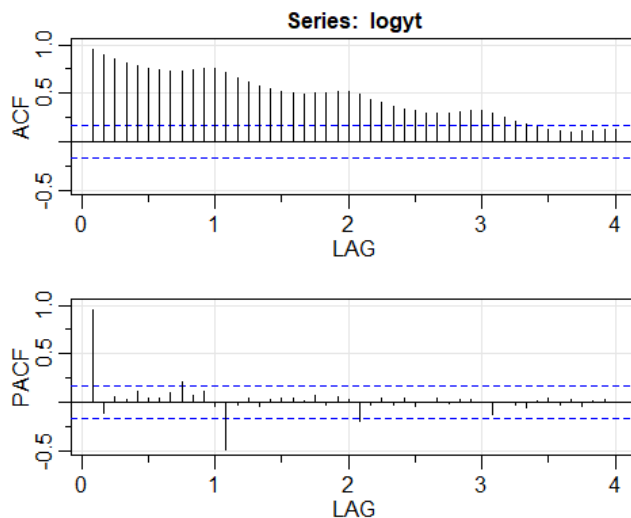
Osserviamo che la varianza non è costante e quindi applichiamo una trasformazione logaritmica per stabilizzarla.

TRASFORMAZIONE LOGARITMICA



La varianza ora è stabile.

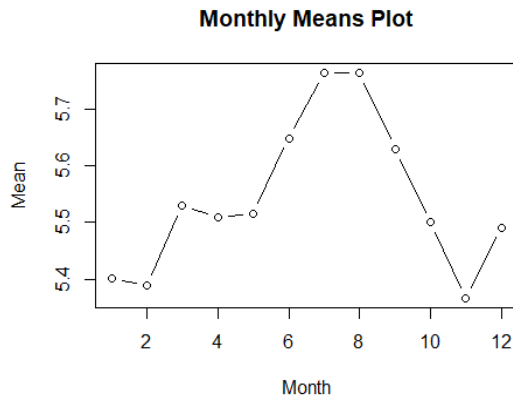
ACF E PACF



Componente stagionale: ipotizziamo $AR(0)$ e $MA(3)$.
Componente non stagionale: ipotizziamo $AR(1)$ mentre, per quanto riguarda la componente a media mobile, vediamo che tutte le autocorrelazioni sono molto elevate. Procediamo lavorando prima sulla componente stagionale.

COMPONENTE STAGIONALE

Controllo stagionalità

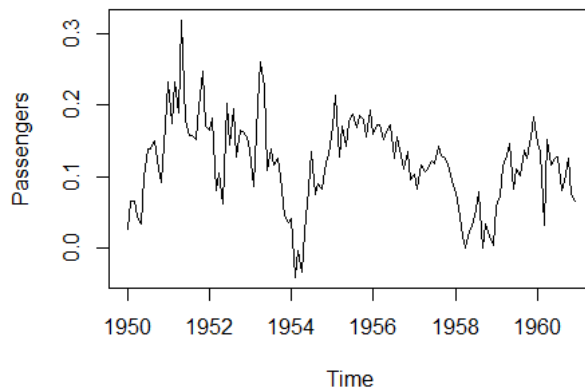


Poiché la media della risposta cambia nei diversi mesi, possiamo affermare che c'è chiara stagionalità.

Controllo differenziazioni

Vediamo che basta una sola differenziazione di ordine 12 per rendere la serie stazionaria nella sua componente stagionale.

Differenziazione

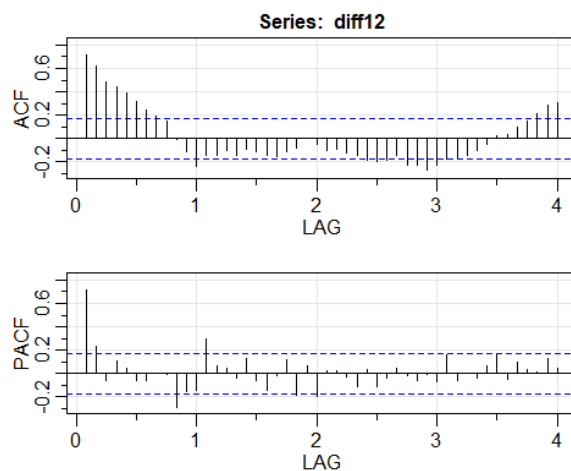


Ora la nostra serie è stazionaria per quanto riguarda la sua parte stagionale.

Adesso possiamo procedere lavorando sulla componente non stagionale.

COMPONENTE NON STAGIONALE

ACF e PACF



ADF test

```
## #####  
## # Augmented Dickey-Fuller Test Unit Root Test #  
## #####  
##  
## Test regression trend  
##
```

```
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.105416 -0.017336 -0.001595  0.018657  0.104638
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0548079  0.0200684   2.731  0.00742 **
## z.lag.1      -0.3469365  0.1105933  -3.137  0.00222 **
## tt           -0.0001897  0.0001212  -1.566  0.12045
## z.diff.lag1  -0.1773169  0.1102939  -1.608  0.11094
## z.diff.lag2   0.1327550  0.1055479   1.258  0.21129
## z.diff.lag3   0.0351851  0.1053101   0.334  0.73897
## z.diff.lag4   0.0410516  0.1054239   0.389  0.69778
## z.diff.lag5   0.1884087  0.1043706   1.805  0.07394 .
## z.diff.lag6   0.1939619  0.1042608   1.860  0.06566 .
## z.diff.lag7   0.0926435  0.1047215   0.885  0.37838
## z.diff.lag8   0.1842671  0.1023650   1.800  0.07474 .
## z.diff.lag9   0.3223770  0.1012341   3.184  0.00191 **
## z.diff.lag10  0.1736772  0.1035293   1.678  0.09643 .
## z.diff.lag11  0.0533492  0.1038694   0.514  0.60861
## z.diff.lag12 -0.2633223  0.0938585  -2.806  0.00600 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03745 on 104 degrees of freedom
## Multiple R-squared:  0.4236, Adjusted R-squared:  0.346
## F-statistic: 5.458 on 14 and 104 DF, p-value: 9.959e-08
##
## Value of test-statistic is: -3.137 3.3411 4.9474
##
## Critical values for test statistics:
##      1pct  5pct 10pct
## tau3 -3.99 -3.43 -3.13
## phi2  6.22  4.75  4.07
## phi3  8.43  6.49  5.47
```

Osserviamo che $\text{value}(-3.137) > \text{tau3}(-3.43)$ quindi non rifiutiamo $H_0: \rho=0$ e possiamo dire che la serie ha radici unitarie. Inoltre, $\text{value}(4.9474) < \text{phi3}(6.49)$ quindi non rifiutiamo $H_0: \beta_2=0$ e possiamo dire che la serie non ha un trend.

```
## #####
## # Augmented Dickey-Fuller Test Unit Root Test #
## #####
##
## Test regression drift
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.096405 -0.015333 -0.000812  0.018072  0.114315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.029280  0.011782   2.485  0.01453 *
## z.lag.1      -0.249907  0.092231  -2.710  0.00787 **
## z.diff.lag1  -0.239858  0.103515  -2.317  0.02244 *
## z.diff.lag2   0.085232  0.101786   0.837  0.40429
## z.diff.lag3  -0.011181  0.101757  -0.110  0.91272
## z.diff.lag4  -0.006046  0.101738  -0.059  0.95273
## z.diff.lag5   0.144126  0.101158   1.425  0.15719
## z.diff.lag6   0.150986  0.101276   1.491  0.13900
## z.diff.lag7   0.049665  0.101756   0.488  0.62651
## z.diff.lag8   0.145999  0.100089   1.459  0.14764
## z.diff.lag9   0.290161  0.099804   2.907  0.00445 **
## z.diff.lag10  0.135007  0.101233   1.334  0.18521
## z.diff.lag11  0.012921  0.101302   0.128  0.89875
## z.diff.lag12 -0.293537  0.092486  -3.174  0.00197 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0377 on 105 degrees of freedom
## Multiple R-squared:  0.41, Adjusted R-squared:  0.3369
## F-statistic: 5.612 on 13 and 105 DF, p-value: 1.128e-07
##
##
## Value of test-statistic is: -2.7096 3.7342
##
## Critical values for test statistics:
##      1pct 5pct 10pct
## tau2 -3.46 -2.88 -2.57
## phi1  6.52  4.63  3.81
```

Osserviamo che $\text{value}(-2.7096) > \text{tau2}(-2.88)$ quindi non rifiutiamo $H_0: \rho=0$ e possiamo dire che la serie ha radici unitarie. Inoltre, $\text{value}(3.7342) < \text{phi1}(4.63)$ quindi non rifiutiamo $H_0: \beta_1=0$ e possiamo dire che la serie non ha un drift (infatti, $\text{mean}=0.12$ prossima a 0).

```
## #####
## # Augmented Dickey-Fuller Test Unit Root Test #
## #####
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.085309 -0.022677  0.001226  0.024508  0.116005
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## z.lag.1      -0.03082    0.02777  -1.110   0.270
## z.diff.lag1  -0.38153    0.08849  -4.312 3.64e-05 ***
## z.diff.lag2  -0.02665    0.09349  -0.285   0.776
## z.diff.lag3  -0.12269    0.09353  -1.312   0.192
## z.diff.lag4  -0.11825    0.09337  -1.267   0.208
## z.diff.lag5   0.03932    0.09417   0.418   0.677
## z.diff.lag6   0.04553    0.09418   0.483   0.630
## z.diff.lag7  -0.05936    0.09403  -0.631   0.529
## z.diff.lag8   0.04121    0.09296   0.443   0.658
## z.diff.lag9   0.18814    0.09317   2.019   0.046 *
## z.diff.lag10  0.03173    0.09454   0.336   0.738
## z.diff.lag11 -0.08796    0.09505  -0.925   0.357
## z.diff.lag12 -0.36951    0.08939  -4.133 7.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03861 on 106 degrees of freedom
## Multiple R-squared:  0.3758, Adjusted R-squared:  0.2993
## F-statistic: 4.91 on 13 and 106 DF, p-value: 1.137e-06
##
##
## Value of test-statistic is: -1.1099
##
## Critical values for test statistics:
##      1pct 5pct 10pct
## tau1 -2.58 -1.95 -1.62
```

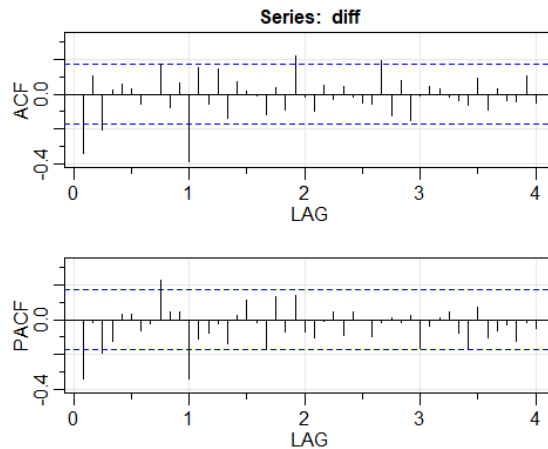
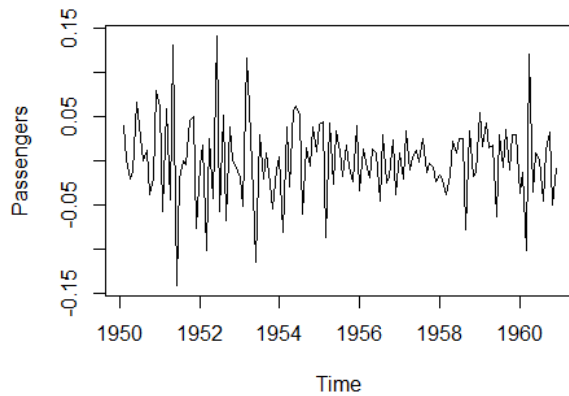
Osserviamo che $\text{value}(-1.1099) > \text{tau1}(-1.95)$ quindi non rifiutiamo $H_0: \rho=0$ e possiamo dire che la serie ha radici unitarie.

Controllo differenziazioni

Vediamo che è necessaria una differenziazione per rendere la nostra serie stazionaria.

Differenziazione

Grafici



Componente stagionale:
ipotizziamo AR(1)
e MA(1).
Componente non stagionale:
ipotizziamo AR(1)
e MA(1).

Funzione auto.arima

```
## Series: diff
## ARIMA(0,0,1)(0,0,1)[12] with zero mean
##
## Coefficients:
##          ma1      sma1
##      -0.4018  -0.5569
## s.e.   0.0896   0.0731
##
## sigma^2 estimated as 0.001369: log likelihood=244.7
## AIC=-483.39 AICc=-483.2 BIC=-474.77
```

La funzione auto.arima applicata alla serie resa stazionaria ci suggerisce un modello ARIMA(0,0,1)(0,0,1)[12] con media zero.

Proviamo a utilizzare la stessa funzione sul logaritmo della serie originale per vedere se vengono confermati i passaggi effettuati (differenziazione di ordine 12 prima e differenziazione di ordine 1 poi).

```
## Series: logyt
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##          ma1      sma1
##      -0.4018  -0.5569
## s.e.   0.0896   0.0731
##
## sigma^2 estimated as 0.001371: log likelihood=244.7
## AIC=-483.4 AICc=-483.21 BIC=-474.77
```

Ci viene suggerito un modello ARIMA(0,1,1)(0,1,1)[12].

MODELLO

Scelta del modello

ARIMA (1,1,1) (1,1,3) [12]

```
$fit
call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = 5), include.mean = !no.constant, transform.pars = trans, fixed = fixed,
optim.control = list(trace = trc, REPORT = 1, reltol = tol))

Coefficients:
          ar1      ma1      sar1      sma1      sma2      sma3
          0.1470 -0.5566 -0.9839  0.4109 -0.5091  0.0099
          s.e.  0.2344  0.2012  0.1303  0.2151  0.1486  0.1057

sigma^2 estimated as 0.001308: log likelihood = 245.64, aic = -477.28

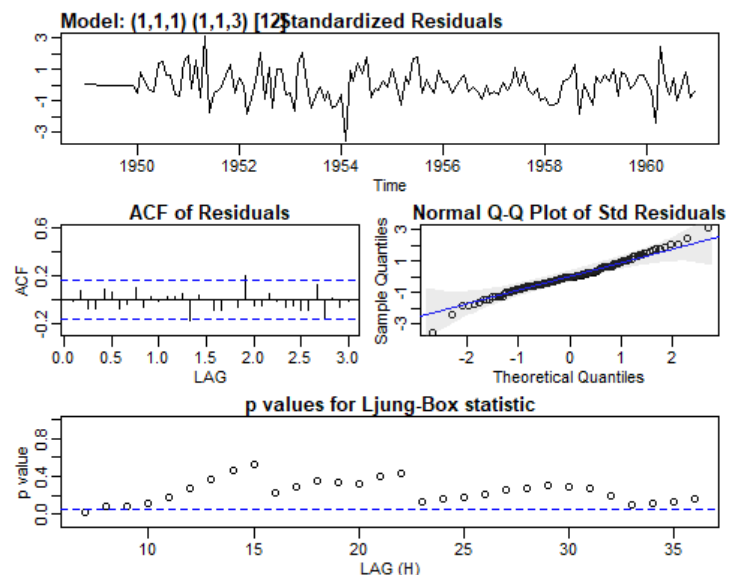
$degrees_of_freedom
[1] 125

$stable
      Estimate      SE t.value p.value
ar1    0.1470 0.2344  0.6271  0.5318
ma1   -0.5566 0.2012 -2.7672  0.0065
sar1   -0.9839 0.1303 -7.5529  0.0000
sma1    0.4109 0.2151  1.9104  0.0584
sma2   -0.5091 0.1486 -3.4269  0.0008
sma3    0.0099 0.1057  0.0938  0.9254

$AIC
[1] -3.36111

$AICC
[1] -3.356729

$BIC
[1] -3.219375
```



ARIMA (0,1,1) (1,1,2) [12]

```
$fit
Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
Q), period = S), include.mean = !no.constant, transform.pars = trans, fixed = fixed,
optim.control = list(trace = trc, REPORT = 1, reltol = tol))

Coefficients:
      ma1      sar1      sma1      sma2
-0.4288   -0.9736   0.4077  -0.4945
s.e.      0.0898   0.1736   0.2546   0.1729

sigma^2 estimated as 0.001315: log likelihood = 245.45, aic = -480.9

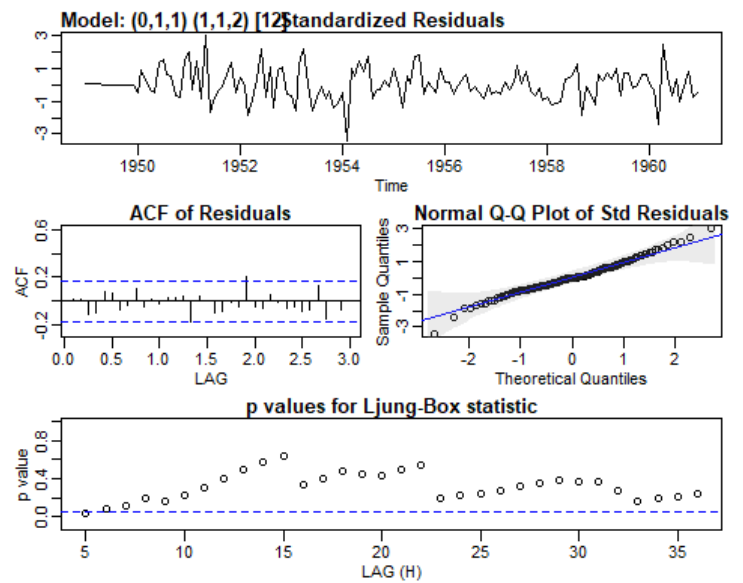
$degrees_of_freedom
[1] 127

$table
      Estimate      SE t.value p.value
ma1    -0.4288  0.0898  -4.7750  0.0000
sar1    -0.9736  0.1736  -5.6099  0.0000
sma1     0.4077  0.2546   1.6015  0.1118
sma2    -0.4945  0.1729  -2.8595  0.0050

$AIC
[1] -3.386609

$AICC
[1] -3.384552

$BIC
[1] -3.285369
```



ARIMA (0,1,1) (1,1,1) [12]

```
$fit
Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
Q), period = S), include.mean = !no.constant, transform.pars = trans, fixed = fixed,
optim.control = list(trace = trc, REPORT = 1, reltol = tol))

Coefficients:
      ma1      sar1      sma1
-0.4143   -0.1116   -0.4817
s.e.      0.0899   0.1548   0.1363

sigma^2 estimated as 0.001341: log likelihood = 244.96, aic = -481.91

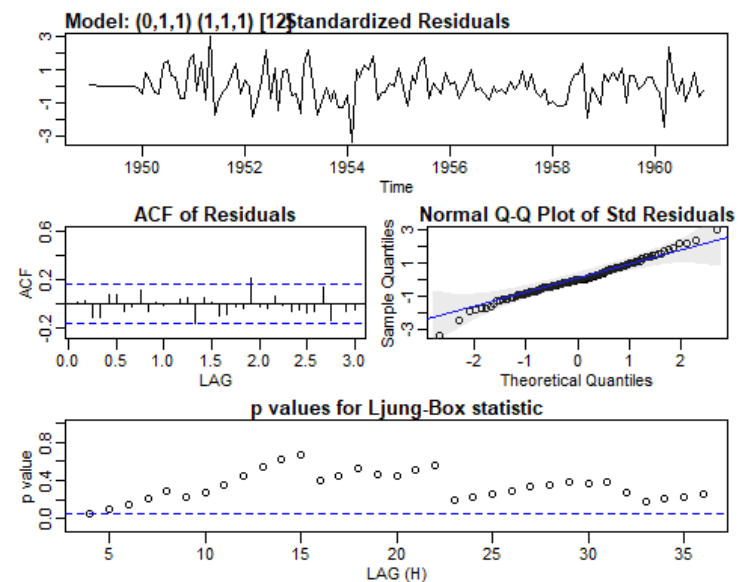
$degrees_of_freedom
[1] 128

$table
      Estimate      SE t.value p.value
ma1    -0.4143  0.0899  -4.6064  0.0000
sar1    -0.1116  0.1548  -0.7214  0.4720
sma1    -0.4817  0.1363  -3.5338  0.0006

$AIC
[1] -3.393754

$AICC
[1] -3.39253

$BIC
[1] -3.312763
```



ARIMA (0,1,1) (0,1,1) [12]

```
$fit
Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
Q), period = S), include.mean = !no.constant, transform.pars = trans, fixed = fixed,
optim.control = list(trace = trc, REPORT = 1, reltol = tol))

Coefficients:
      ma1      sma1
-0.4018   -0.5569
s.e.      0.0896   0.0731

sigma^2 estimated as 0.001348: log likelihood = 244.7, aic = -483.4

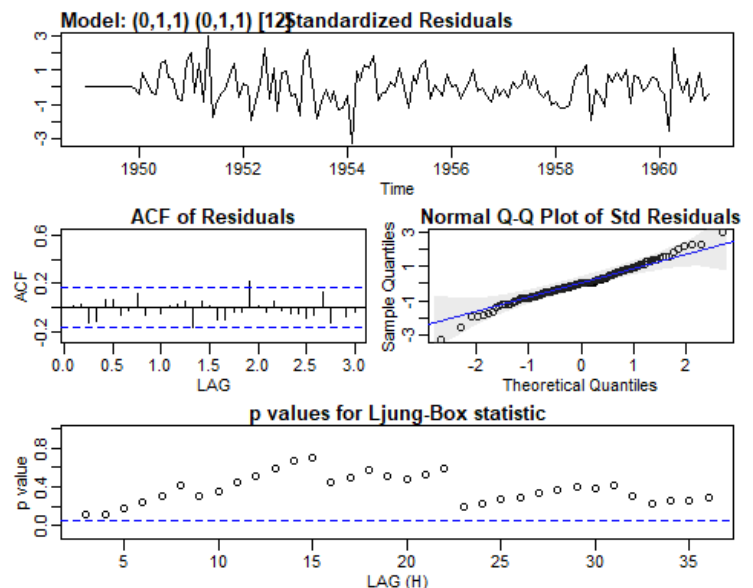
$degrees_of_freedom
[1] 129

$table
      Estimate      SE t.value p.value
ma1    -0.4018  0.0896  -4.4825  0.0000
sma1    -0.5569  0.0731  -7.6190  0.0000

$AIC
[1] -3.404219

$AICC
[1] -3.403611

$BIC
[1] -3.343475
```



Partendo dal modello ARIMA(1,1,1,1,3)[12] e considerando la significatività dei parametri, otteniamo lo stesso modello suggeritoci dalla funzione auto.arima: ARIMA(0,1,1,0,1,1)[12]. I residui sembrano essere random, non autocorrelati (ACF=0 e p-value>0.05), distribuiti normalmente e l'AIC è pari a -483.4.

Fit

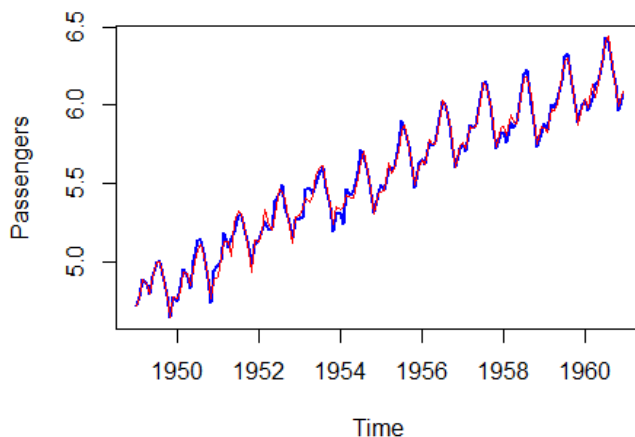
```
## Series: logyt
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##          ma1          sma1
##       -0.4018   -0.5569
## s.e.    0.0896    0.0731
##
## sigma^2 estimated as 0.001371: log likelihood=244.7
## AIC=-483.4   AICc=-483.21   BIC=-474.77
```

Analisi residui

```
## Series: mod$residuals
## ARIMA(0,0,0) with zero mean
##
## sigma^2 estimated as 0.001228: log likelihood=278.22
## AIC=-554.44   AICc=-554.41   BIC=-551.47
```

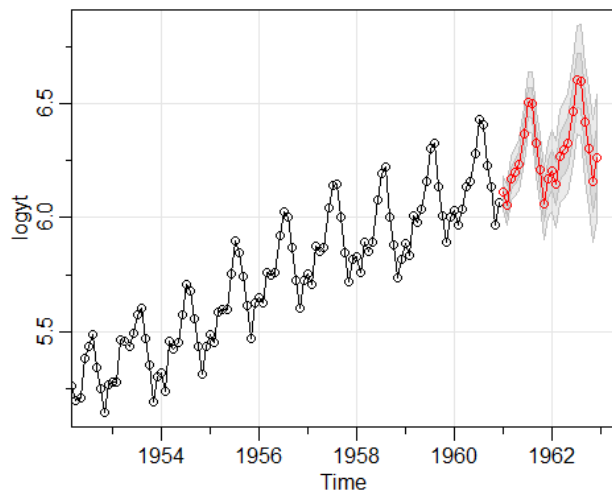
I residui del nostro modello sono white noise (infatti, da quanto visto prima, non risultavano problematici).

Plot (in blu logyt e in rosso i valori previsti)



Osserviamo che il modello fittato cattura molto bene l'andamento del logaritmo della nostra serie originale.

Forecast



Abbiamo fatto una previsione di due anni (24 mesi).

RIASSUNTO

