Elaborato di Data Mining

Melchioretto Nicolò (mat. 893145)
Finzi Micol Rebecca (mat. 882523)
Anderloni Hanan Francesca (mat. 889079)

Data set

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=car+details+v4.csv

Il Dataset che è stato usato contiene informazioni riguardanti la vendita di auto usate in India.

Le variabili del dataset sono:

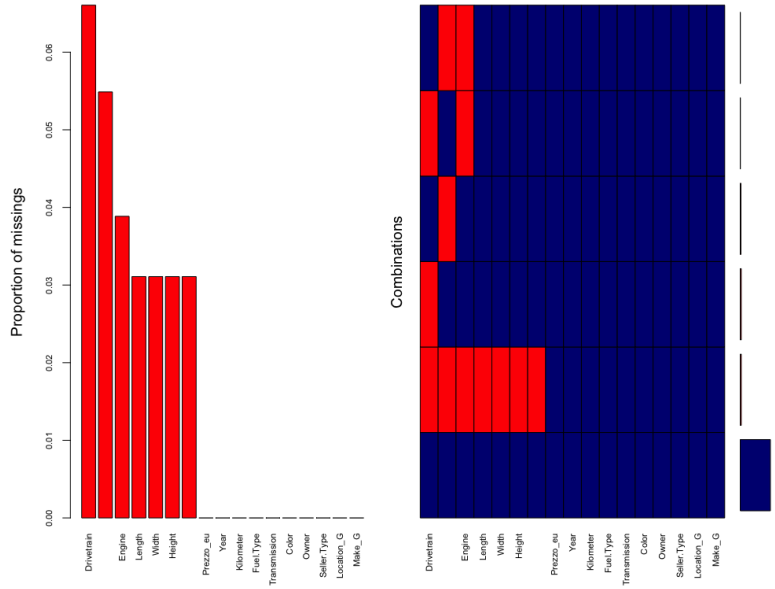
Make è il produttore dell'automobile
Model è il modello dell'automobile
Price è il prezzo di vendita dell'automobile (in euro: Price*0.011)
Year è l'anno di produzione dell'auto
Kilometer sono i chilometri totali percorsi
Fuel.Type è il tipo di carburante dell'auto
Transmission definisce se l'automobile è automatica o manuale
Location è la città in cui è stata venduta l'auto
Color è il colore dell'automobile
Owner è il numero di proprietari precedenti
Seller.Type indica se l'auto è venduta da privato o concessionario
Engine è la cilindrata della vettura in cc
Max.Power è la potenza massima in bhp@rpm
Max.Torque è la coppia massima in nm@rpm
Drivetrain è il tipo di trazione dell'automobile (totale AWD, posteriore RWD, anteriore FWD)
Length è la lunghezza dell'automobile in mm
Width è la larghezza dell'automobile in mm
Height è l' altezza dell'automobile in mm
Seating.Capacity è il numero massimo di persone che possono salire in macchina
Fuel.Tank.Capacity è la capacità massima di carburante dell'automobile in litri

Per il modello lineare abbiamo usato il prezzo in euro come variabile target, mentre per il modello logistico come variabile target abbiamo usato una variabile binaria, creata dalla variabile target precedente, usando come soglia per discriminare i due gruppi la mediana, cioè 9185 euro, dove 0 indica un prezzo dell'automobile sotto la mediana e 1 indica un prezzo dell'automobile al di sopra della mediana.

Modello robusto con target quantitativo

Dati mancanti

Dopo aver analizzato e sistemato le variabili del dataset sono stati gestiti i dati mancanti.

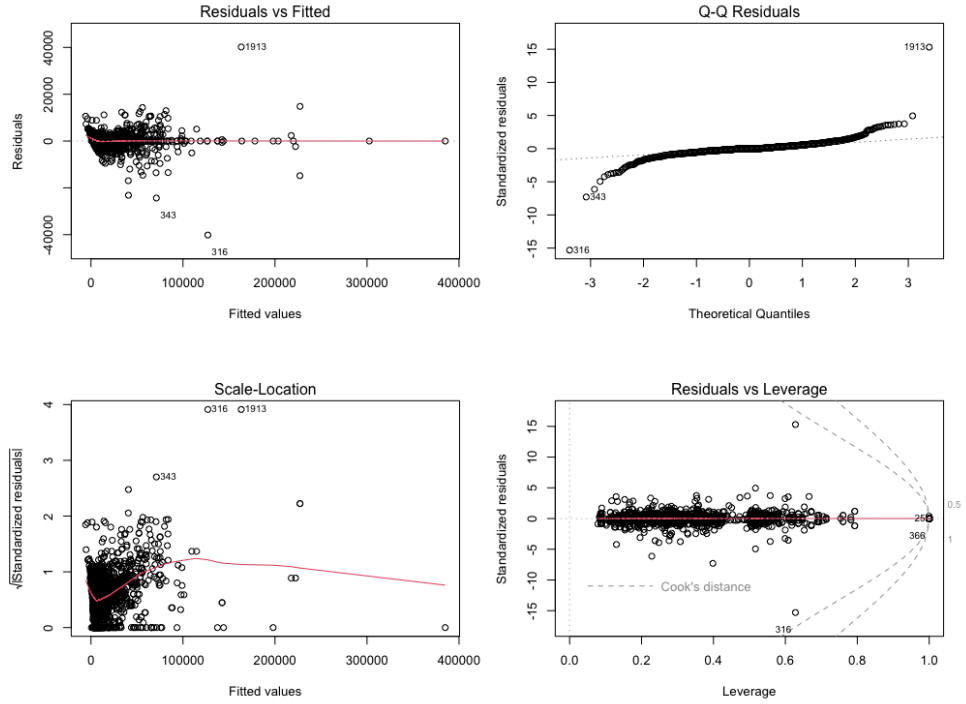


I dati mancanti sono MCAR, cioè sono *missing completely at random*, quindi sono stati risolti selezionando nel dataset solo le osservazioni complete (*cc analysis*). Le osservazioni iniziali del dataset sono 2059 di cui 1874 complete.

Il primo modello è stato fittato scegliendo come variabile target il prezzo delle auto usate e come variabili indipendenti tutte le variabili del data set. I grafici diagnostici del primo modello sono riportati a seguito:

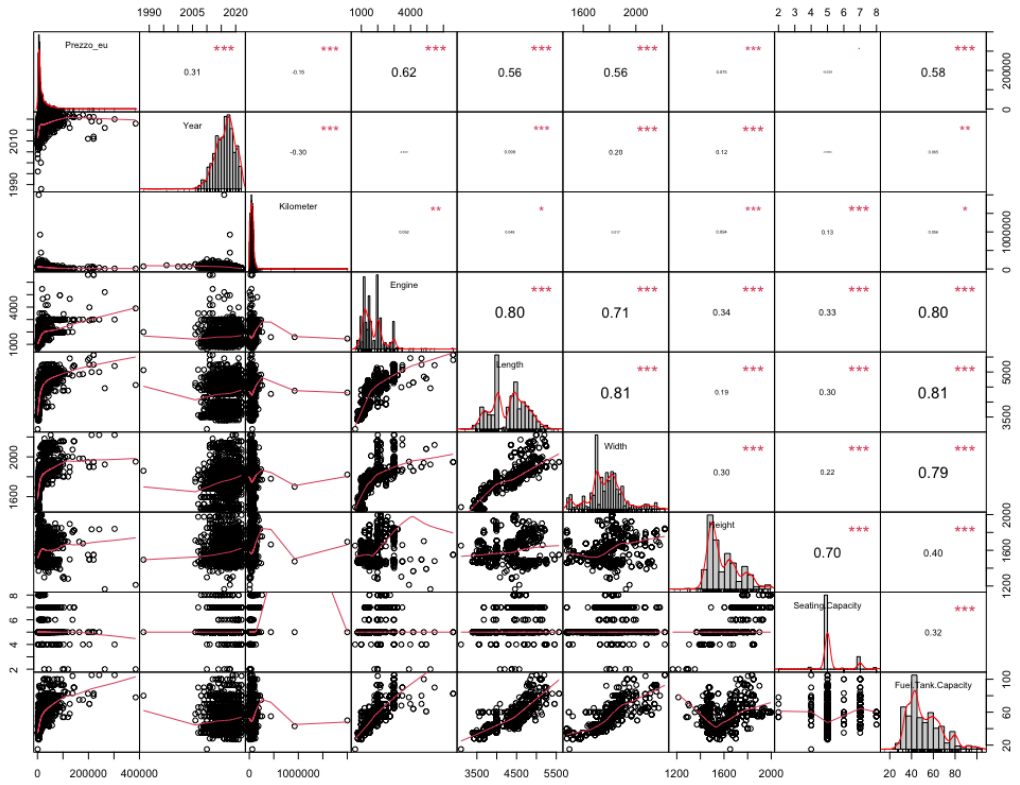
```
> drop1(fit_0, test="F")
Single term deletions

Model:
Prezzo_eu ~ Make + Model + Year + Kilometer + Fuel.Type + Transmission +
    Location + Color + Owner + Seller.Type + Engine + Drivetrain +
    Length + Width + Height + Seating.Capacity + Fuel.Tank.Capacity
                  Df    Sum of Sq          RSS   AIC  F value                Pr(>F)
<none>                              15259610513 31920
Make               0            0  15259610513 31920
Model            907 286891149948 302150760461 35702  17.0802 < 0.00000000000000022 ***
Year               1   4756665837  20016276350 32427 256.8540 < 0.00000000000000022 ***
Kilometer          1    177019307  15436629821 31940   9.5588             0.0020569 **
Fuel.Type          2     53160626  15312771139 31923   1.4353             0.2386375
Transmission       0            0  15259610513 31920
Location          65   2119918975  17379529488 32034   1.7611             0.0003155 ***
Color             15    109031048  15368641561 31904   0.3925             0.9811759
Owner              3    750982683  16010593196 32004  13.5174         0.00000001289 ***
Seller.Type        2     23134245  15282744759 31919   0.6246             0.5357234
Engine             1     26274588  15285885101 31922   1.4188             0.2339448
Drivetrain         0            0  15259610513 31920
Length             1     22263306  15281873819 31921   1.2022             0.2732046
Width              1     27125475  15286735988 31922   1.4647             0.2265242
Height             1        36403  15259646917 31918   0.0020             0.9646468
Seating.Capacity   1      2974726  15262585239 31919   0.1606             0.6886797
Fuel.Tank.Capacity 1     63033002  15322643515 31926   3.4037             0.0654089 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

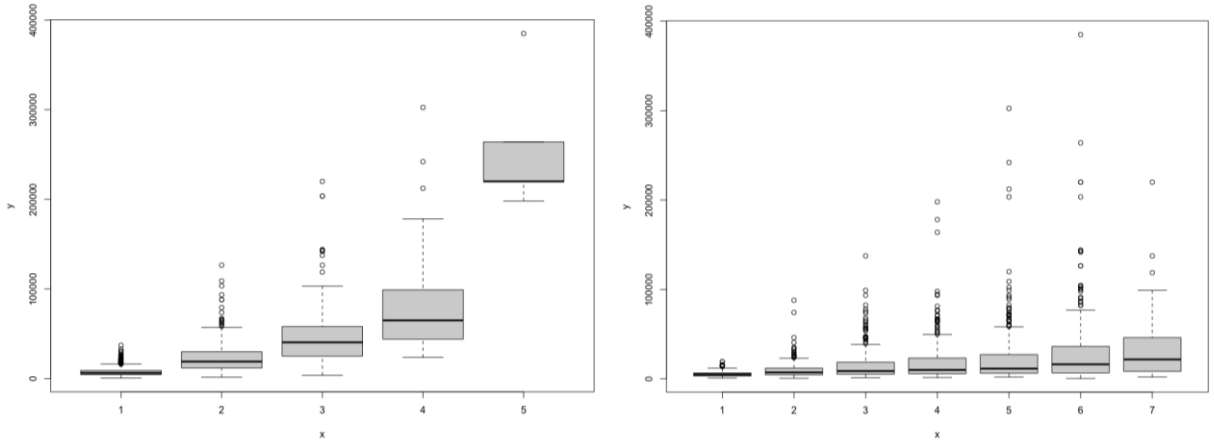


Analisi della collinearità

Dopo aver z si osserva il seguente grafico riferito alle variabili quantitative e successivamente si osservano i valori di TOL e VIF.



Per diminuire il numero di livelli delle variabili *Location* e *Make* si effettua quindi optimal grouping.



Mentre l'optimal grouping ha prodotto su *Make* una buona separazione tra i gruppi, lo stesso non si può dire di *Location*; ciò nonostante, per comodità la si mantiene comunque in questo modo.

L'analisi della collinearità per le variabili quantitative produce il seguente:

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
X_matrixPrezzo_eu	2.2896	0.4368	310.7825	355.3642	0.6609	0	0	0.0018	0.9478
X_matrixYear	1.4437	0.6927	106.9214	122.2592	0.8323	0	0	0.0029	0.5172
X_matrixKilometer	1.1374	0.8792	33.1248	37.8765	0.9376	0	0	0.0036	0.2034
X_matrixEngine	4.2133	0.2373	774.3983	885.4856	0.4872	0	0	0.0010	1.2834
X_matrixLength	5.6069	0.1784	1110.2513	1269.5167	0.4223	0	0	0.0007	1.3827
X_matrixWidth	3.8097	0.2625	677.1428	774.2789	0.5123	0	0	0.0011	1.2411
X_matrixHeight	2.8014	0.3570	434.1450	496.4230	0.5975	0	0	0.0015	1.0821
X_matrixSeating.Capacity	2.4825	0.4028	357.2715	408.5221	0.6347	0	0	0.0017	1.0049
X_matrixFuel.Tank.Capacity	4.8704	0.2053	932.7575	1066.5615	0.4531	0	0	0.0009	1.3373

Per risolvere la collinearità si effettua l'analisi delle componenti principali tra le variabili collineari (*Length, Width, Fuel.Tank.Capacity, Engine*). Decidiamo di estrarre le PC dalla matrice di correlazione poiché le unità di misura e gli ordini di grandezza sono differenti.

L'analisi delle componenti principali produce il seguente:

Importance of components:

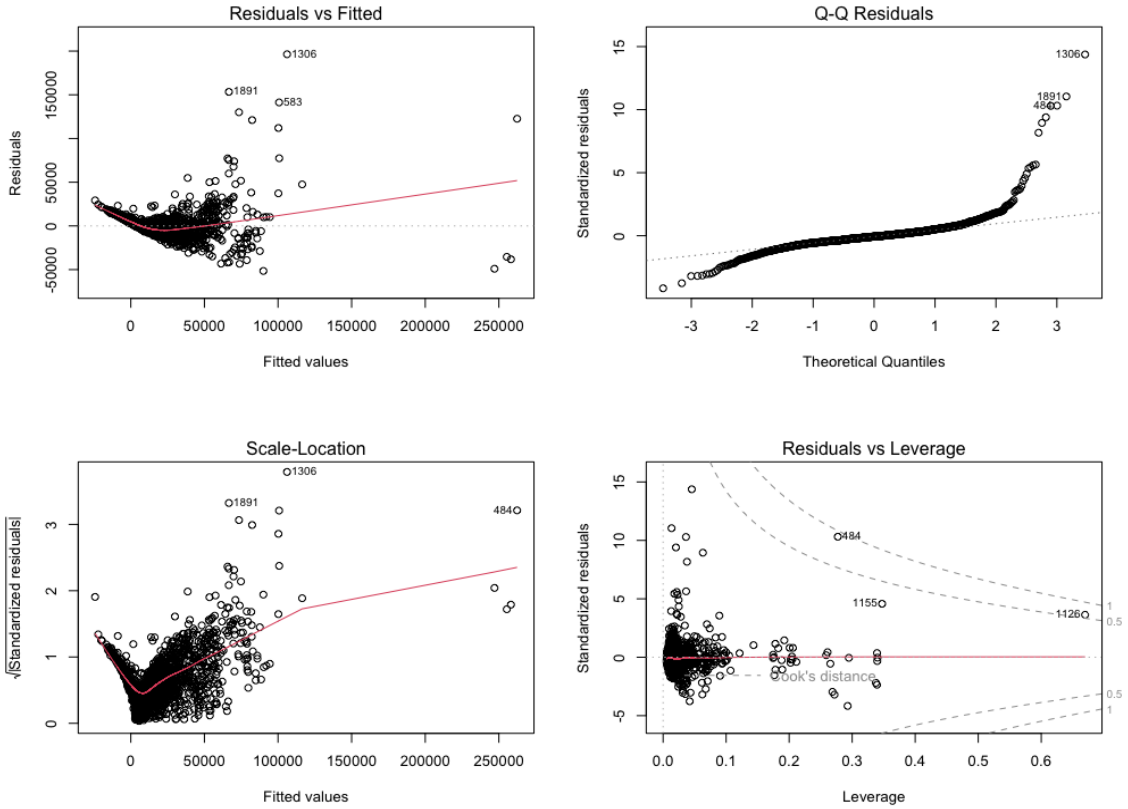
	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.8320556	0.53461904	0.43631306	0.40912801
Proportion of Variance	0.8391069	0.07145438	0.04759227	0.04184643
Cumulative Proportion	0.8391069	0.91056130	0.95815357	1.00000000

Decidiamo quindi di inserire nel modello solo la prima PC.

Ricalcolando l'analisi della collinearità per le variabili quantitative si produce il seguente:

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
X_matrix2Prezzo_eu	2.1644	0.4620	435.0090	544.0524	0.6797	0	0	0.0012	1.3173
X_matrix2Year	1.2533	0.7979	94.6470	118.3721	0.8932	0	0	0.0021	0.4949
X_matrix2Kilometer	1.1284	0.8862	47.9807	60.0079	0.9414	0	0	0.0024	0.2787
X_matrix2Height	2.0679	0.4836	398.9753	498.9862	0.6954	0	0	0.0013	1.2645
X_matrix2Seating.Capacity	2.1192	0.4719	418.1180	522.9273	0.6869	0	0	0.0013	1.2931
X_matrix2PC1	2.2320	0.4480	460.2804	575.6586	0.6693	0	0	0.0012	1.3515

La collinearità è stata risolta.
Si osservi ora il fit del modello post-collinearità, con i relativi grafici diagnostici.



Analisi della linearità

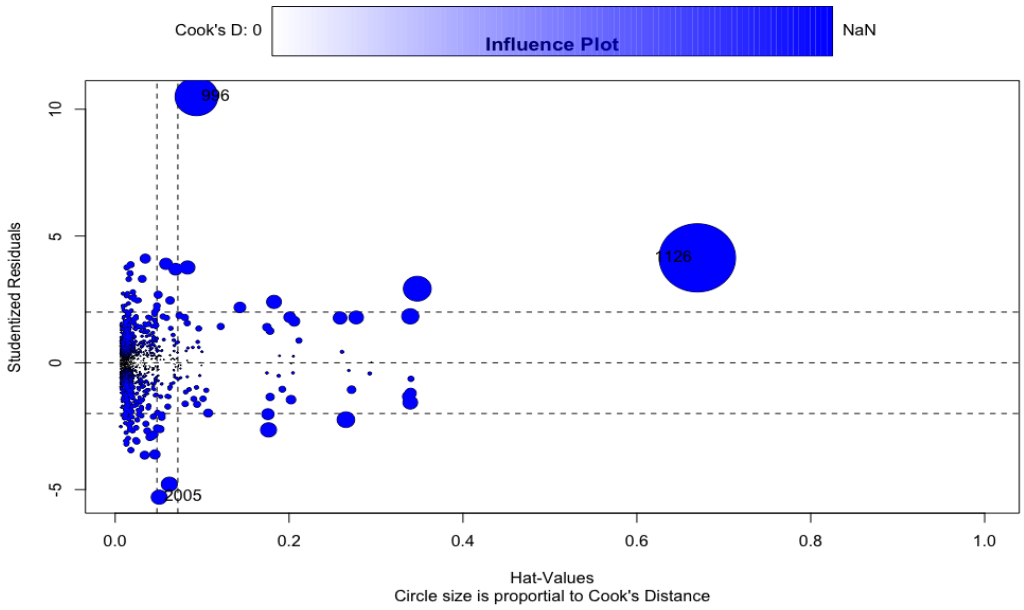
Per risolvere la linearità viene creata una trasformazione box cox della variabile risposta e siccome il valore della lambda risulta essere pari a -0,10 la variabile target *Prezzo_eu* viene trasformata tramite logaritmo in *ylog*. Da questo modello viene prodotto il seguente RESET che evidenzia un modello ancora non correttamente specificato.

```
data:  fit_l0
RESET = 134.32, df1 = 1, df2 = 1828, p-value < 0.00000000000000022
```

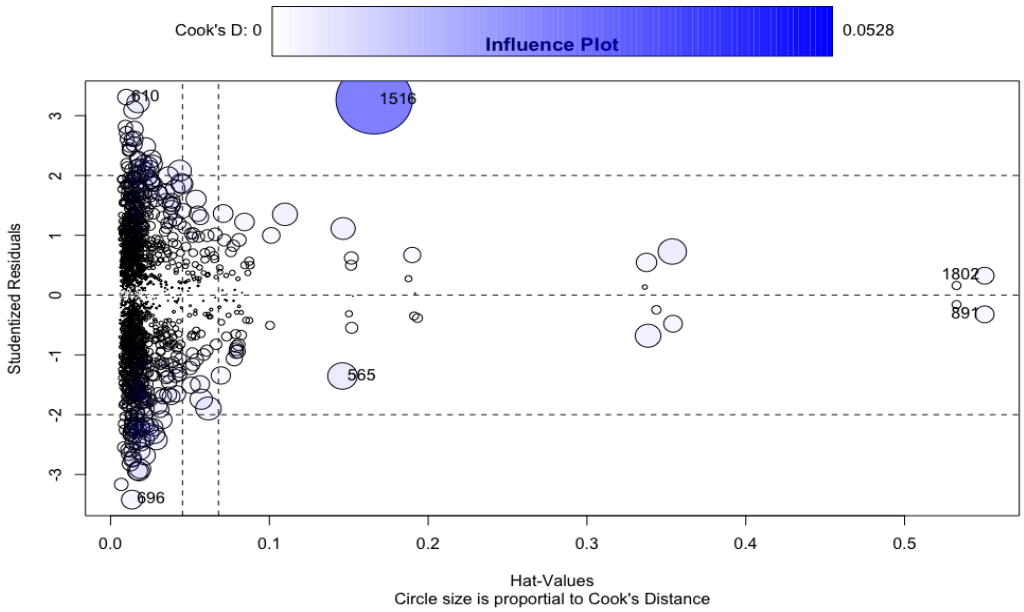
Come si vede dal codice, le gam risultano inizialmente problematiche (le trasformazioni non miglioravano il RESET) e si è ipotizzato, anche guardando alla distribuzione della

variabili tramite boxplot, che la causa fossero i valori influenti; si è quindi deciso di gestire prima questi ultimi.

Punti influenti

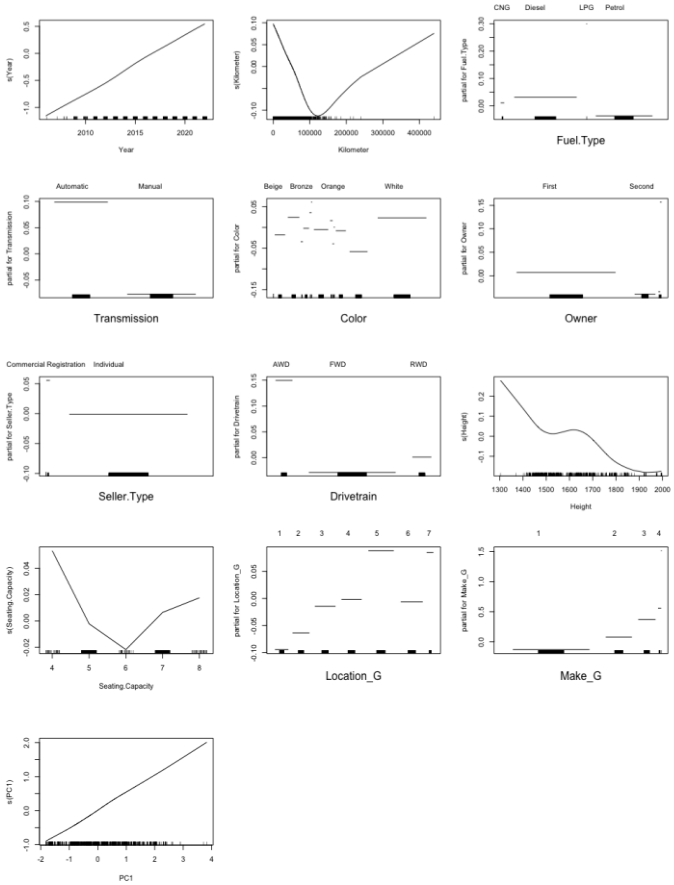


I punti influenti che verranno eliminati dal dataset sono corrispondenti a 108 osservazioni. Il medesimo grafico post-rimozione di queste è il seguente:



Risoluzione linearità

Dopo la rimozione dei punti influenti vengono ricalcolate le gam e le splines risultanti sono le seguenti:



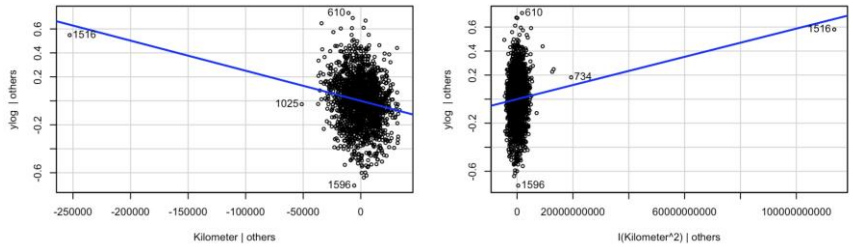
Guardando le gam è stato fatto il seguente modello:
fit_l2_noin ← lm(ylog ~ Year + Kilometer + I(Kilometer^2) + Fuel.Type + Transmission + Color + Owner + Seller.Type + Drivetrain + Height + I(Height^2) + I(Height^3) + Seating.Capacity + Location_G+Make_G+ PC1, data = r2_noinflu)

Osservando drop1 e RESET decidiamo di rimuovere dal modello *Height* e le sue relative potenze, *Color* e *Seller.Type* in quanto peggiora il RESET. Il modello dunque è il seguente:
fit_clo ← lm(ylog ~ Year + Kilometer + I(Kilometer^2) + Fuel.Type + Transmission + Owner + Drivetrain + Seating.Capacity + Location_G + Make_G + PC1, data = r2_noinflu)

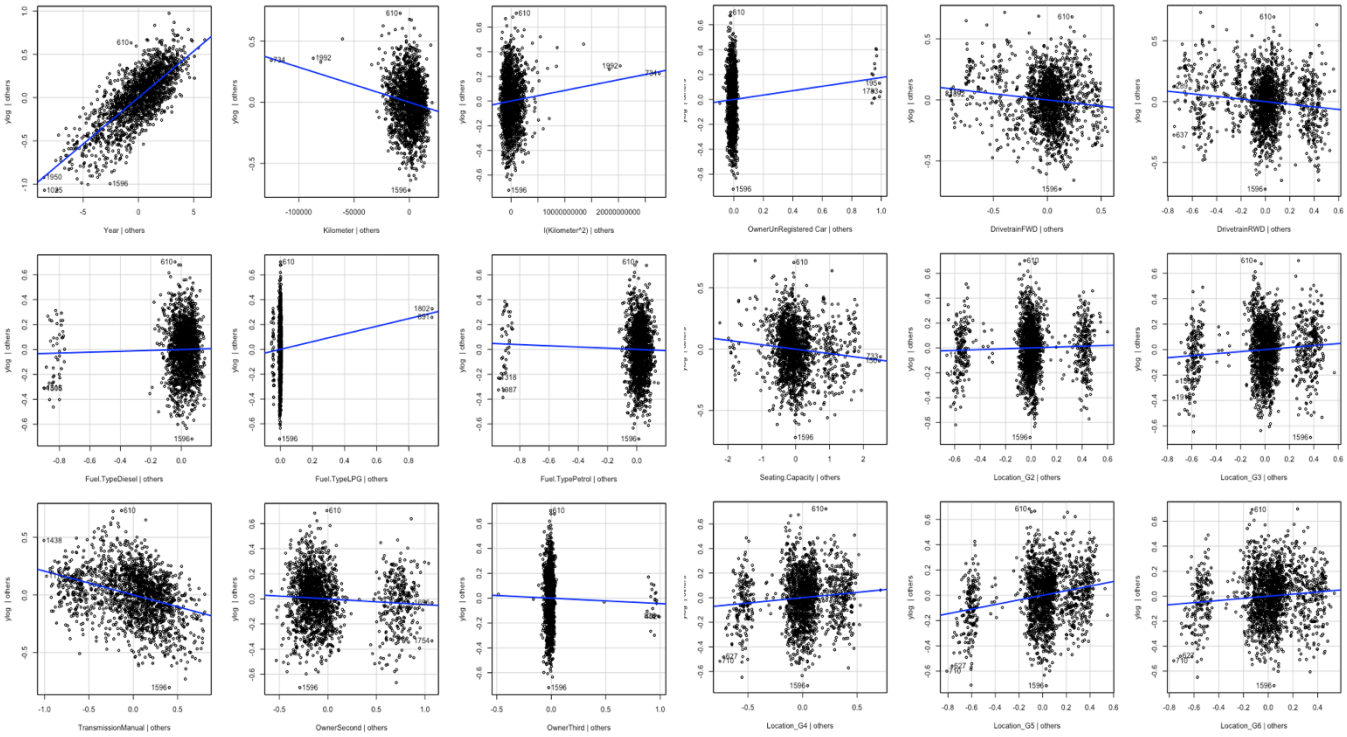
Il reset del modello risulta pari a:

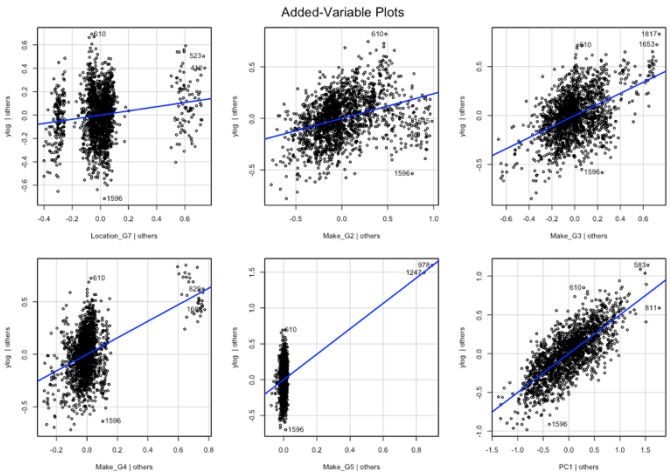
```
data:  fit_clo
RESET = 76.637, df1 = 1, df2 = 1740, p-value < 0.00000000000000022
```

Osservando i partial plot, si nota che l'effetto di *Kilometer* (e relativa potenza) è fortemente influenzato dall'osservazione 1516.



Si procede quindi alla rimozione manuale dell'osservazione. I partial plots risultanti vengono riportati in seguito.

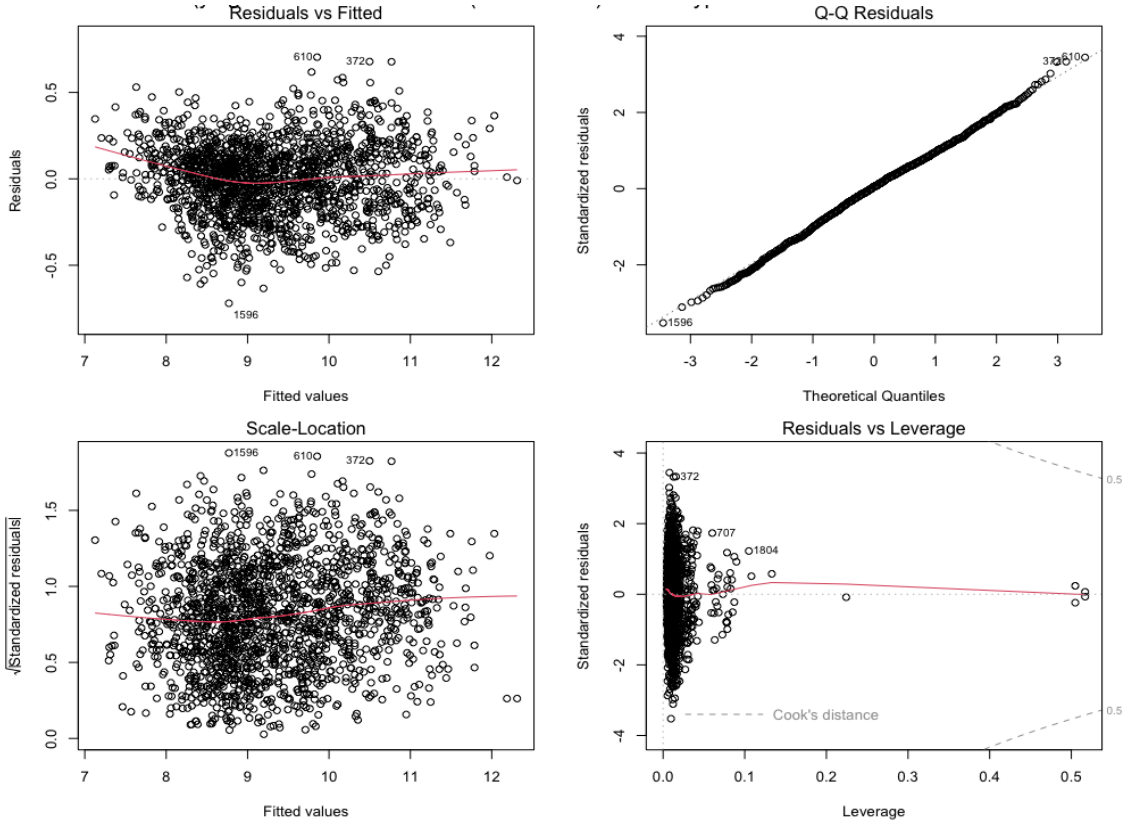




Il modello fittato con le trasformazioni suggerite da gam e dopo aver rimosso l'osservazione nr. 1516 è il seguente:

fit_cloz ← lm(ylog ~ Year + Kilometer + I(Kilometer^2) + Fuel.Type + Transmission + Owner + Drivetrain + Seating.Capacity + Location_G + Make_G+ PC1, data = r_z)

e produce i seguenti grafici diagnostici:



Model selection

ll modello *fit_clozm* che minimizza lAIC tramite procedura Stepwise è identico al precedente *fit_cloz*.

Eteroschedasticità

Dai grafici diagnostici precedenti, si nota una presenza di eteroschedasticità, che viene confermata sia dal Breusch-Pagan Test (in figura) sia dal While Test.

```
data:  fit_clozm
BP = 123.42, df = 24, p-value = 0.000000000000002373
```

Ciò significa che l'inferenza semplice è fuorviante; si ricorre all'impiego della correzione di White degli errori standard; viene generata la seguente inferenza corretta per l'eteroschedasticità.

Significatività dei coefficienti (non robusta) - prime righe

Variable		N	Estimate	p
Year		1765	0.11 (0.10, 0.11)	<0.001
Kilometer		1765	-0.00 (-0.00, -0.00)	<0.001
I(Kilometer^2)		1765	0.00 (0.00, 0.00)	0.006
Fuel.Type	CNG	42	Reference	
	Diesel	902	0.04 (-0.03, 0.10)	0.306
	LPG	2	0.31 (0.02, 0.60)	0.038
	Petrol	819	-0.05 (-0.11, 0.02)	0.148
Transmission	Automatic	772	Reference	
	Manual	993	-0.21 (-0.23, -0.18)	<0.001
Owner	First	1438	Reference	
	Second	294	-0.05 (-0.07, -0.02)	0.001

Inferenza con errori corretti - prime righe

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-206.8056154273635912	4.4022186762495874	-46.9776	< 0.00000000000000022	***
Year	0.1073045994119981	0.0021793751308992	49.2364	< 0.00000000000000022	***
Kilometer	-0.0000028641270804	0.0000004573375893	-6.2626	0.0000000004759	***
I(Kilometer^2)	0.0000000000086286	0.0000000000029481	2.9269	0.003469	**
Fuel.TypeDiesel	0.0357257342037648	0.0381953373360687	0.9353	0.349741	
Fuel.TypeLPG	0.3093346595616249	0.0616201623827059	5.0200	0.0000005695374	***
Fuel.TypePetrol	-0.0484220014676491	0.0364557467551697	-1.3282	0.184273	
TransmissionManual	-0.2056340261687352	0.0132949549358721	-15.4671	< 0.00000000000000022	***
OwnerSecond	-0.0457219197160628	0.0148420105352773	-3.0806	0.002098	**

Notiamo che l'inferenza corretta conferma in generale la significatività precedente.

Proviamo a fittare un modello senza *Fuel.Type* poiché ha solo un livello significativo e notiamo che il RESET migliora notevolmente sia come valore che come p-value, quindi la specificazione ne beneficia; togliendo $Kilometer^2$ la specificazione invece ne risente.

Il modello post-eteroschedasticità risulta essere il seguente:

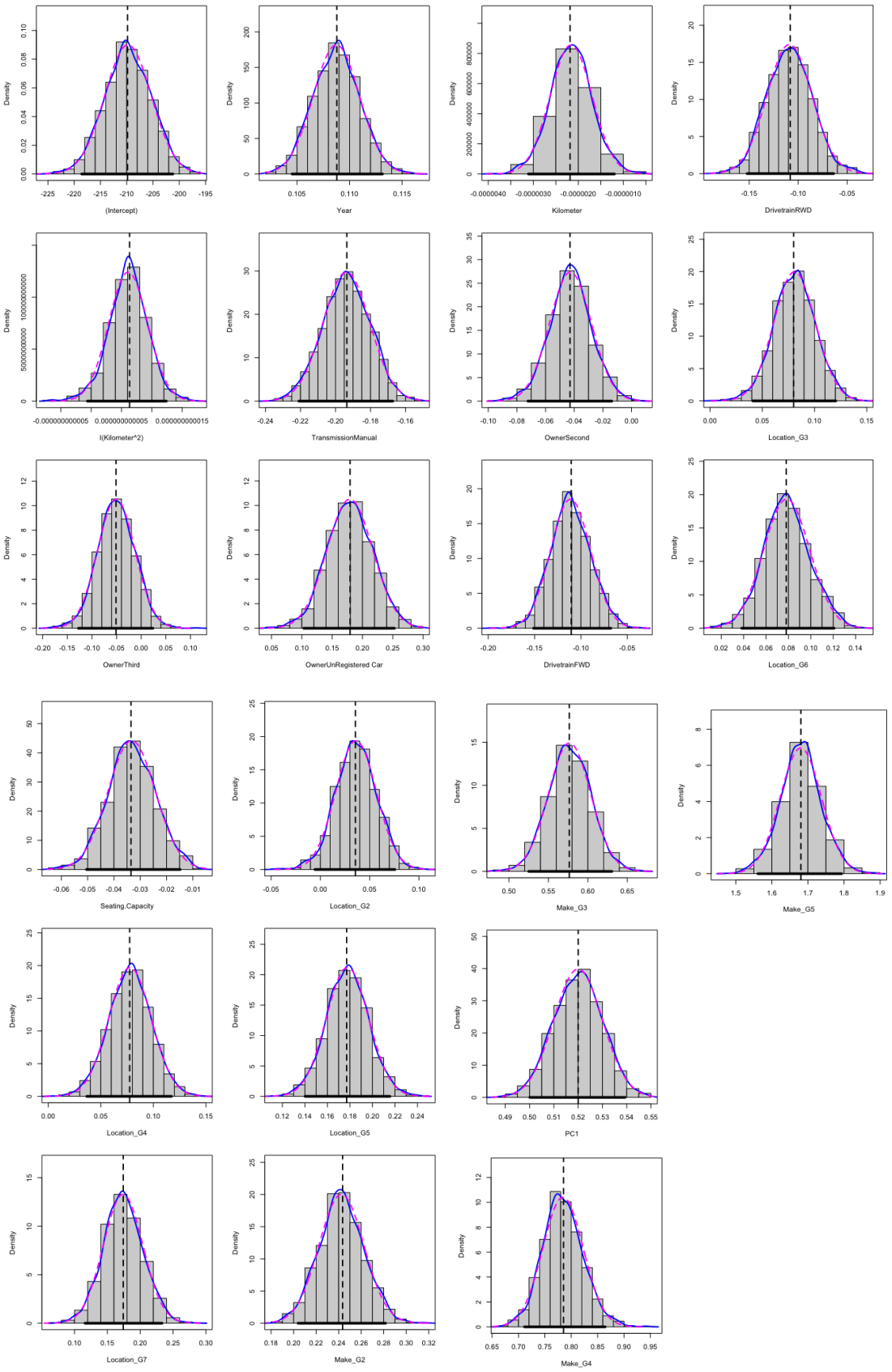
```
fit_clozme ← lm(ylog ~ Year + Kilometer + I(Kilometer^2) + Transmission +
                Owner + Drivetrain + Seating.Capacity + Location_G + Make_G + PC1,
data = r_z)
```

con:

```
data:  fit_clozme
RESET = 65.219, df1 = 1, df2 = 1742, p-value = 0.000000000000001238
```

Bootstrap

Gli intervalli di confidenza bootstrap sono i seguenti:



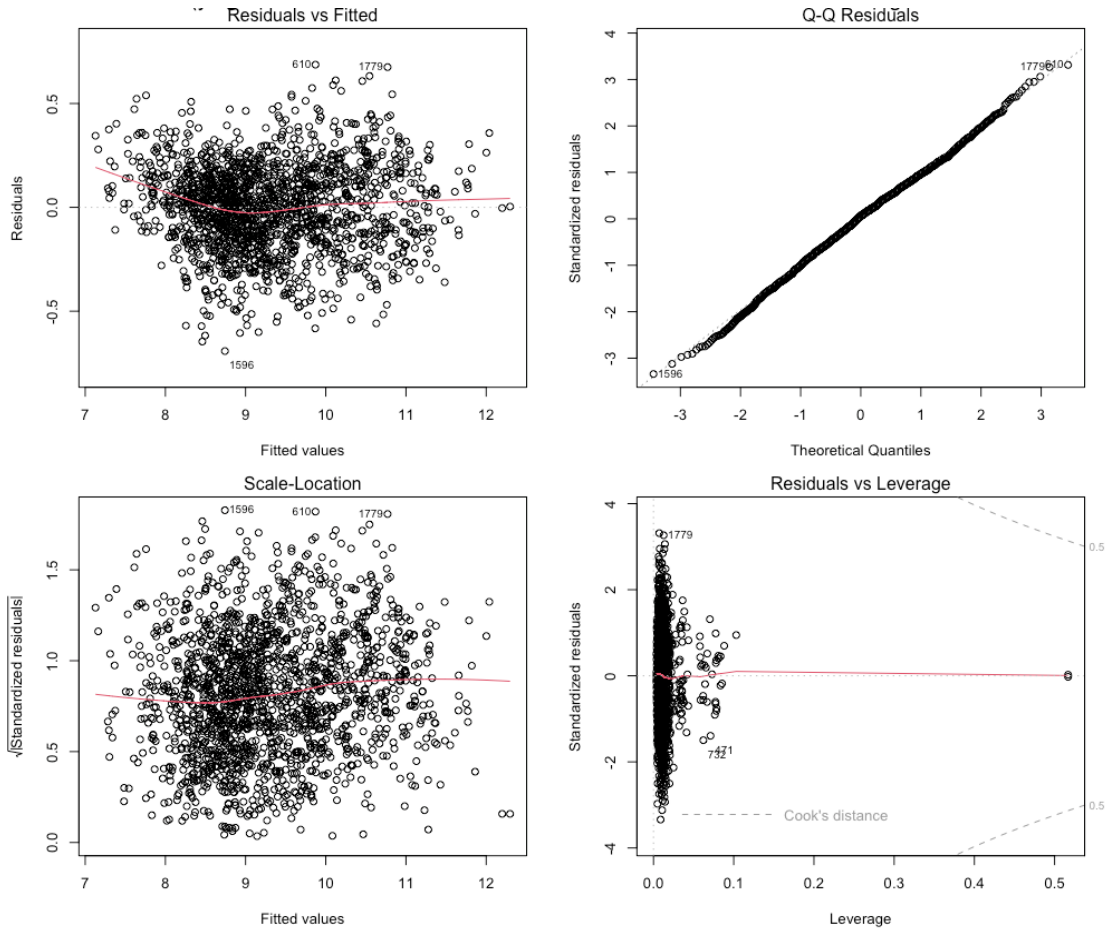
L'inferenza prodotta dai coefficienti risulta robusta dal bootstrap, tranne per un livello di *Owner* (trascurabile) e $Kilometer^2$.
Nonostante il RESET migliori leggermente includendo $Kilometer^2$, il bootstrap lo boccia, quindi lo rimuoviamo e creiamo il seguente modello definitivo:

```
fit_finale ← lm(ylog ~ Year + Kilometer + Transmission +
                Owner + Drivetrain + Seating.Capacity + Location_G + Make_G + PC1,
data = r_z2)
```

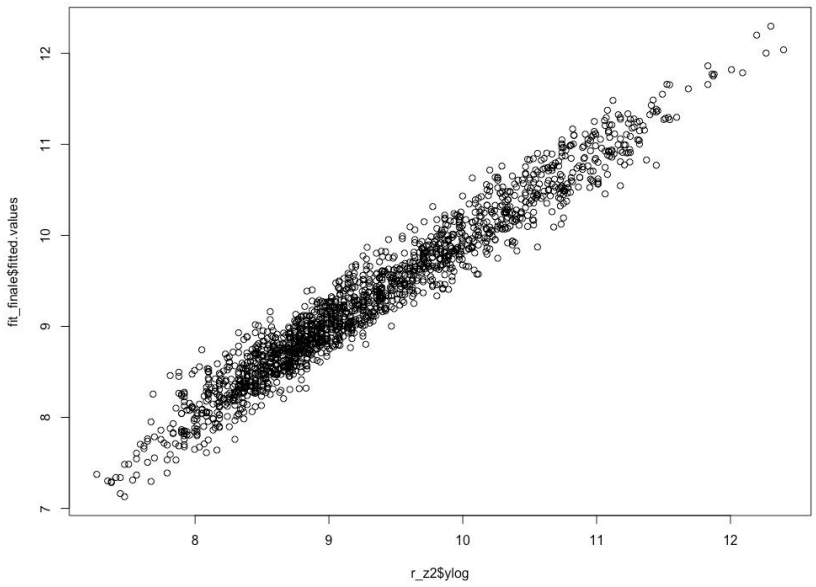
La stime dei coefficienti sono:

	Estimate		
(Intercept)	-211.9221252960	Location_G2	0.0369900320
Year	0.1098096642	Location_G3	0.0810045264
Kilometer	-0.0000013003	Location_G4	0.0791588466
TransmissionManual	-0.1937582782	Location_G5	0.1785272400
OwnerSecond	-0.0442437047	Location_G6	0.0796333028
OwnerThird	-0.0455770095	Location_G7	0.1749819129
OwnerUnRegistered Car	0.1880101802	Make_G2	0.2448443591
DrivetrainFWD	-0.1144893991	Make_G3	0.5764284663
DrivetrainRWD	-0.1095865791	Make_G4	0.7829750216
Seating.Capacity	-0.0334192092	Make_G5	1.6827318296
		PC1	0.5194173653

I grafici diagnostici sono:



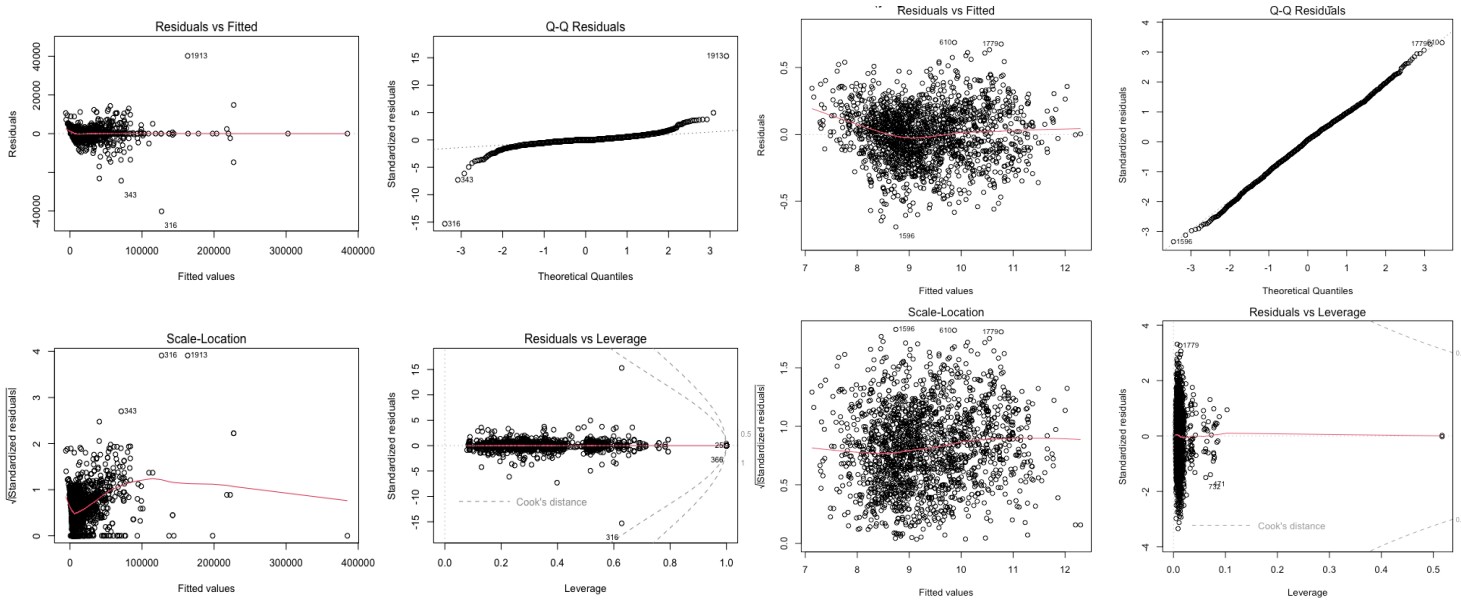
In particolare osserviamo il grafico dei valori osservati contro quelli fittati dal modello finale:



Confronto modello iniziale e modello finale

Si riportano per il confronto i grafici diagnostici relativi a modello iniziale e finale.

iniziale finale



Modello logistico target binario

Si impiega il dataset con la collinearità già risolta e dove non ci sono fattori o variabili continue con *zero-variance* o *near-zero-variance*, quindi l'unico problema da controllare è la *separation* o *quasi-separation*. Dopo aver convertito il target continuo *Prezzo_eu* in una variabile binaria è stata eseguita una regressione logistica:

```
rl_0 ← glm(Prezzo_alto ~ Year + Kilometer + Transmission + Make_G +
          Owner + Drivetrain + Seating.Capacity + Location_G + PC1, data = r_rl1,
family = binomial)
```

che restituisce un *warning* che suggerisce *separation*, vedremo che è su *Make_G*.

Le relative frequenze con cui si distribuiscono le osservazioni a seconda del prezzo (maggiore o minore del prezzo mediano) sono:

0	1	0	1
882	883	0.4997167	0.5002833

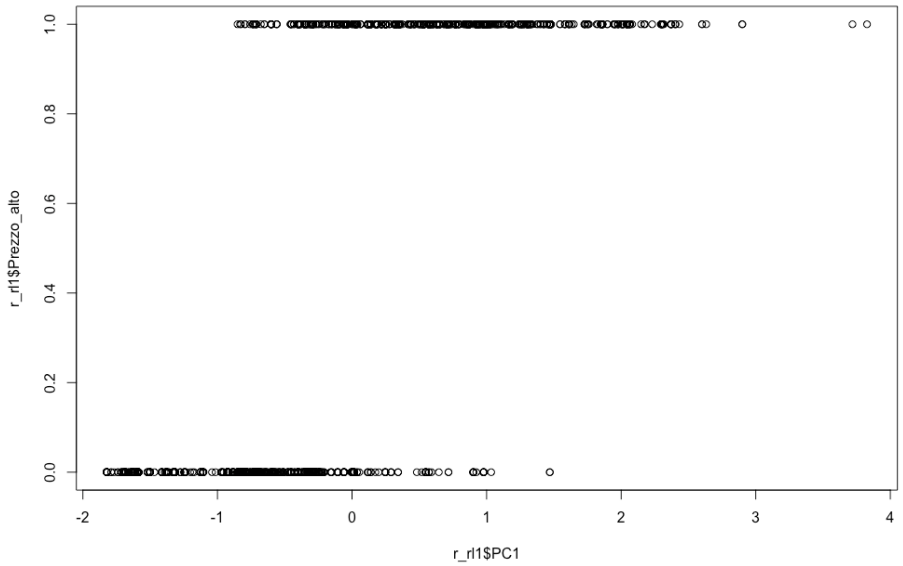
Calcolando poi gli *odds ratio* si nota che quelli relativi a *Make_G* nei livelli 3, 4 e 5 sono estremamente elevati; è inoltre sospetto anche quello di PC1. Si decide allora di valutare un eventuale problema di *separation*.

Separation

Dalla tabella delle frequenze di *Make_G*, riportata di seguito, si nota che soffre di *separation*, causata proprio dai livelli 3, 4 e 5, caratterizzati da OR estremi e da assenza di significatività. Dato che costituisce di fatto una regola classificativa deterministica, la rimuoviamo dal modello.

	1	2	3	4	5
0	822	60	0	0	0
1	289	313	243	36	2

Estendendo il controllo anche alle altre covariate, si osserva che *Owner* ha due livelli senza osservazioni (presumibilmente a causa della precedente pulizia del dataset). Per mantenere la covariata si eliminano i livelli vuoti. Anche PC1 sembra problematica, tuttavia, guardando alla sua distribuzione rispetto a *Prezzo_alto*, si decide di conservarla nel modello, in quanto non si evidenzia grave *separation*.



Aggiorniamo il modello come segue:

rl_1 ← glm(Prezzo_alto ~ Year + Kilometer + Transmission + Owner +
Drivetrain + Seating.Capacity + Location_G + PC1, data = r_rl1, family = binomial)

Owner	First	1438		Reference	
	Second	294		0.79 (0.37, 1.69)	0.543
	Third	19		9.31 (0.78, 123.57)	0.098
	UnRegistered Car	14		73.00 (0.05, 62953.40)	0.610

Dato che *Owner* non è significativo, lo si rimuove dal modello. Fittando un altro modello senza *Owner* si nota invece che *Kilometer* è scarsamente significativo; quindi la si rimuove e si crea un nuovo modello:

rl_3 ← glm(Prezzo_alto ~ Year + Transmission + Drivetrain + Seating.Capacity +
Location_G + PC1, data = r_rl1, family = binomial)

Variable		N	Odds ratio	p
Year		1765	3.33 (2.81, 4.03)	<0.001
Transmission	Automatic	772	Reference	
	Manual	993	0.23 (0.13, 0.40)	<0.001
Drivetrain	AWD	237	Reference	
	FWD	1260	1.15 (0.32, 3.77)	0.829
	RWD	268	5.89 (1.42, 24.38)	0.014
Seating.Capacity		1765	0.28 (0.17, 0.45)	<0.001
Location_G	1	186	Reference	
	2	238	6.51 (2.36, 18.73)	<0.001
	3	293	4.87 (1.74, 14.23)	0.003
	4	286	5.92 (2.10, 17.44)	<0.001
	5	362	19.38 (6.98, 57.35)	<0.001
	6	311	4.54 (1.62, 13.25)	0.005
	7	89	11.50 (2.44, 57.20)	0.002
PC1		1765	1555.51 (586.77, 4674.18)	<0.001

Di seguito si riportano gli *odds ratio* relativi alle covariate.

	OR	2.5 %	97.5 %
(Intercept)	0.00	0.00	0.00
Year	3.33	2.81	4.03
TransmissionManual	0.23	0.13	0.40
DrivetrainFWD	1.15	0.32	3.77
DrivetrainRWD	5.89	1.42	24.38
Seating.Capacity	0.28	0.17	0.45
Location_G2	6.51	2.36	18.73
Location_G3	4.87	1.74	14.23
Location_G4	5.92	2.10	17.44
Location_G5	19.38	6.98	57.35
Location_G6	4.54	1.62	13.25
Location_G7	11.50	2.44	57.20
PC1	1555.51	586.77	4674.18

Si presentano alcuni esempi di interpretazione degli *odds ratio*:

La propensione, cioè l'*odds*, per un'automobile prodotta nell'anno n di presentare un prezzo sopra la mediana è 3.33 volte superiore di quella di un'automobile prodotta nell'anno $n - 1$.

La propensione di un'automobile con cambio automatico di avere un prezzo superiore alla mediana è pari a circa 4 volte (*1/0.23*) quella di un'automobile con cambio manuale.

La propensione, cioè l'*odds,* di avere una macchina a trazione anteriore con un prezzo sopra la mediana è superiore del 15% rispetto all'*odds* analogo di una macchina con la trazione totale.