Costruzione di un modello lineare robusto relativo a un e-commerce francese

Martina Chiesa 837484, Carlo Saccardi 839641, Davide Valoti 846737

19/11/2020

Il dataset scelto contiene informazioni relative ad un e-commerce francese di successo, presente in diversi Paesi, basato sul modello economico C2C (customer to customer), in cui ogni utente è sia venditore, sia acquirente. Il dataset è composto da 98913 righe e 24 colonne. Ciascuna riga corrisponde a un utente registrato, quindi, nel file non sono presenti clienti non registrati, infatti la variabile *type* presenta una sola modalità. Sono presenti sia variabili quantitative sia qualitative: *identifierHash* comprende i codici identificativi corrispondenti a ciascun utente; *type* indica la tipologia di cliente; *country* e *countryCode* corrispondono a nome e codice ISO del Paese di appartenenza dell'utente; *language* si riferisce alla lingua selezionata come preferita tra le cinque opzioni proposte; *socialNbFollowers* numero di utenti iscritti all'attività di questo user; *socialNbFollows* numero di utenti seguiti dallo user; *socialProductsLiked* numero di prodotti graditi dall'utente; *productsListed* numero di prodotti attualmente non venduti ma caricati dall'utente; *productsSold* numero di prodotti venduti; *productsPassRate* percentuale di prodotti la cui descrizione è coerente col bene offerto; *productsWished* numero di prodotti aggiunti alla lista dei desideri; *productsBought* numero di prodotti acquistati; *gender* genere dell'utente; *civilityTitle* e *civilityGenderId* indica lo stato civile e la rispettiva codifica in numeri da 1 a 3; *hasAnyApp* indica se l'utente ha mai utilizzato l'app ufficiale dello store, in caso affermativo, se è la versione Android *hasAndroidApp* o Ios *hasIosApp*; *hasProfilePicture* segnala se è presente l'immagine del profilo; *daysSinceLastLogin* è il numero di giorni trascorsi dall'ultimo login; *seniority*, *seniorityAsMonths*, *seniorityAsYears* corrispondono rispettivamente al numero di giorni, mesi e anni decorsi dalla registrazione.

```
##  identifierHash         type              country      language
##  Min.   :-9.223e+18   user:98913   France     :25135   de: 7178
##  1st Qu.:-4.623e+18                Etats-Unis :20602   en:51564
##  Median :-1.338e+15                Royaume-Uni:11310   es: 6033
##  Mean   :-6.692e+15                Italie     : 8015   fr:26372
##  3rd Qu.: 4.616e+18                Allemagne  : 6567   it: 7766
##  Max.   : 9.223e+18                Espagne    : 5706
##                                    (Other)    :21578
##  socialNbFollowers socialNbFollows     socialProductsLiked productsListed
##  Min.   :  3.000   Min.   :    0.000   Min.   :    0.00    Min.   :  0.000
##  1st Qu.:  3.000   1st Qu.:    8.000   1st Qu.:    0.00    1st Qu.:  0.000
##  Median :  3.000   Median :    8.000   Median :    0.00    Median :  0.000
##  Mean   :  3.432   Mean   :    8.426   Mean   :    4.42    Mean   :  0.093
##  3rd Qu.:  3.000   3rd Qu.:    8.000   3rd Qu.:    0.00    3rd Qu.:  0.000
##  Max.   :744.000   Max.   :13764.000   Max.   :51671.00    Max.   :244.000
##                                                            NA's   :4944
##   productsSold      productsPassRate   productsWished     productsBought
##  Min.   :  0.0000   Min.   :  0.0000   Min.   :   0.000   Min.   :  0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:   0.000   1st Qu.:  0.0000
##  Median :  0.0000   Median :  0.0000   Median :   0.000   Median :  0.0000
##  Mean   :  0.1216   Mean   :  0.8123   Mean   :   1.555   Mean   :  0.1719
##  3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:   0.000   3rd Qu.:  0.0000
##  Max.   :174.0000   Max.   :100.0000   Max.   :2635.000   Max.   :405.0000
##                                        NA's   :3208
##  gender    civilityGenderId civilityTitle hasAnyApp     hasAndroidApp
```

```
##  F:76121   Min.   :1.000    miss:  437    False:72739   False:94094
##  M:22792   1st Qu.:2.000    mr  :22792    True :26174   True : 4819
##            Median :2.000    mrs :75684
##            Mean   :1.774
##            3rd Qu.:2.000
##            Max.   :3.000
##
##  hasIosApp     hasProfilePicture daysSinceLastLogin   seniority
##  False:77386   False: 1895       Min.   : 11.0      Min.   :2852
##  True :21527   True :97018       1st Qu.:572.0      1st Qu.:2857
##                                  Median :694.0      Median :3196
##                                  Mean   :581.3      Mean   :3064
##                                  3rd Qu.:702.0      3rd Qu.:3201
##                                  Max.   :709.0      Max.   :3205
##
##  seniorityAsMonths seniorityAsYears  countryCode
##  Min.   : 95.07    Min.   :7.92     fr     :25135
##  1st Qu.: 95.23    1st Qu.:7.94     us     :20602
##  Median :106.53    Median :8.88     gb     :11310
##  Mean   :102.13    Mean   :8.51     it     : 8015
##  3rd Qu.:106.70    3rd Qu.:8.89     de     : 6567
##  Max.   :106.83    Max.   :8.90     es     : 5706
##                                     (Other):21578
```

Tramite la funzione summary si ottengono alcune delle statistiche descrittive univariate relative ad ogni variabile. Si osserva che non sono presenti valori negativi nel dataset, infatti per tutte le variabili il valore minimo è pari o superiore a 0. Il maggior numero di utenti è francese, come ci aspettavamo, dato che l'e-commerce è nato proprio in questo Stato, ma la lingua più ricorrente, tra le 5 presenti, è l'inglese. Notiamo inoltre la presenza di valori mancanti per le variabili *proudctsListed* e *productsWished*. Dando uno sguardo alla variabile dipendente scelta (*productsSold*), si riscontra che il valor medio assegnato a questa è 0.12. Dal terzo quartile si constata che, almeno il 75% degli utenti non ha venduto nessun prodotto.

L'intero file contiene dati relativi agli utenti registrati, i quali possono sia aver venduto, che non. Vogliamo quindi indagare la differenza tra queste due categorie di utenti.

```
## FALSE  TRUE
## 96877  2036
```

Coloro che hanno venduto almeno un prodotto sono 2036 e corrispondono solo al 2% circa di tutti gli utenti presenti nel dataset. Procediamo con la creazione di un nuovo dataset che contiene unicamente questi ultimi, ovvero i venditori.

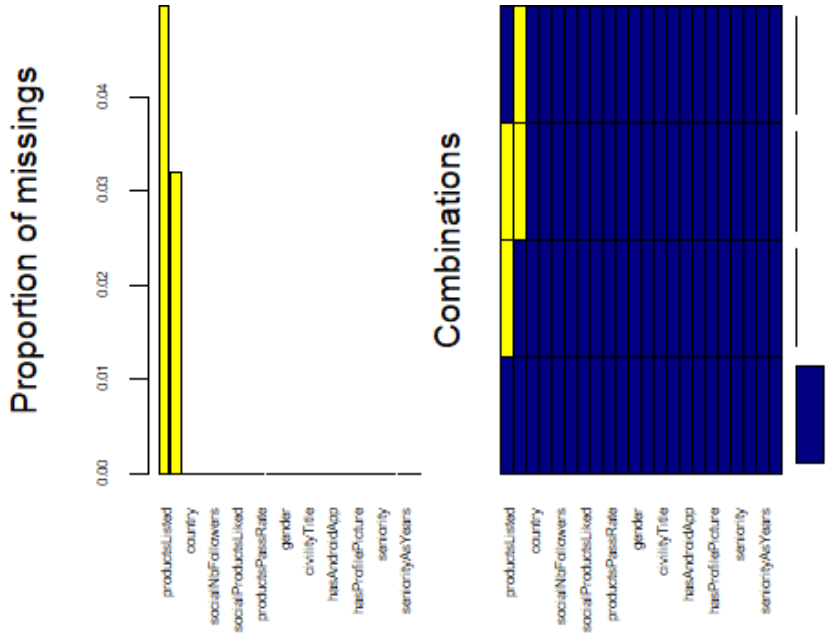
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   5.907   5.000 174.000
```

Il dataset ridotto presenta lo stesso numero di variabili di quello completo. I cambiamenti della variabile dipendente sono evidenti, infatti, ora dalla mediana si osserva che la metà degli utenti ha venduto al massimo 2 prodotti. Un quarto ha venduto un solo prodotto; un altro quarto, invece 3, 4 o 5 prodotti (terzo quartile). La media del numero di prodotti venduti è pari a circa 6 prodotti.

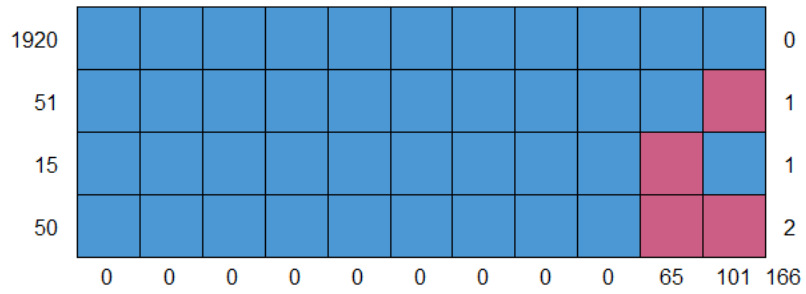
Tramite la funzione df_status si osservano le quantità in termini assoluti e percentuali del numero di 0, di NA e di valori unici presenti nel dataset per ogni variabile. Le due percentuali di valori mancanti (4.96 e 3.19) sono poco elevate, quindi decidiamo di conservare le rispettive variabili nel dataset e procedere successivamente con l'imputazione. La variabile *civilityGenderId* assume valore 1, 2 o 3, tuttavia non si tratta di una variabile quantitativa, poichè queste tre cifre sono codifiche dei tre livelli presenti in

civilityTitle, ovvero 'miss', 'mr', e 'mrs'. Dunque, procediamo con la correzione, trasformandola in fattore a tre livelli. Inoltre, rimuoviamo la variabile identificativa *identifierHash*, e *type*, poichè presenta un solo livello, quindi si tratta di una variabile non discriminante.

NA analysis



In giallo si evidenzia la proporzione di dati mancanti e la combinazione con cui questi si presentano. Sono presenti osservazioni che hanno NA per entrambe le variabili ed altre che sono caratterizzate da un solo valore mancante. Per conferma utilizziamo un'altra tecnica di verifica, che mostra 101 dati mancanti per *productsListed* e 65 per *productsWished*. Dunque, decidiamo di procedere con la mice imputation.

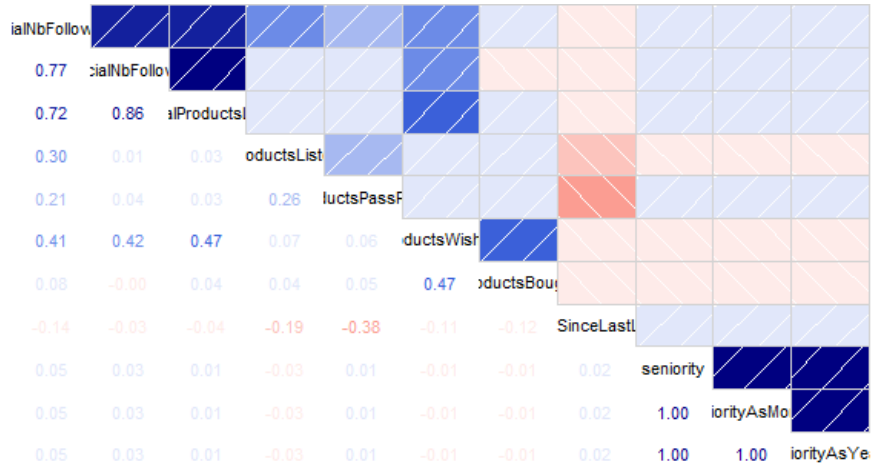


Consideriamo solo le variabili numeriche escludendo quella dipendente e osserviamo che ci sono 1920 righe complete, 15 presentano valori mancanti solo per la variabile *productsListed*, 51 per *productsWished* e 50 per entrambe. Quindi sarà necessario stimare 166 valori. Scegliamo di utilizzare il metodo pmm (predictive mean matching) con 5 ripetizioni. Dopo esserci assicurati che l'imputazione è andata a buon fine, in quanto è stata raggiunta convergenza, creiamo un nuovo dataset che non riporta più valori mancanti, in quanto questi vengono sostituiti con i valori imputati.

```
##   socialNbFollowers     socialNbFollows socialProductsLiked     productsListed
##                   0                   0                   0                  0
##    productsPassRate      productsWished      productsBought  daysSinceLastLogin
##                   0                   0                   0                  0
##           seniority   seniorityAsMonths    seniorityAsYears
##                   0                   0                   0
```

Collinearità

L'analisi della collinearità è uno step necessario per la costruzione di un modello robusto, dato che variabili correlate comportano problemi di efficienza e non permettono di ottenere stime OLS precise. Tramite la funzione corrgram viene proposta una rappresentazione grafica delle correlazioni presenti tra le variabili quantitative considerate.



Notiamo come *seniority*, *senorityAsMonths* e *seniorityAsYears* sono perfettamente correlate, come era lecito aspettarsi. Anche tra le variabili *socialNbFollowers*, *socialNbFollows* e *socialProductsLiked* sono presenti forti correlazioni positive. Questa rappresentazione mediante matrice delle collinearità ha alcuni limiti, risulta infatti, di difficile interpretazione quando il numero di variabili è elevato, inoltre, analizza solamente le correlazioni bivariate. Procediamo allora con l'analisi tramite altre metodologie (TOL e VIF), che permettano di considerare anche la presenza di multicollinearità tra le variabili quantitative.

```
##                              VIF    TOL           Wi           Fi Leamer
## socialNbFollowers         3.3524 0.2983 4.763572e+02 5.295471e+02 0.5462
## socialNbFollows           5.0413 0.1984 8.183555e+02 9.097329e+02 0.4454
## socialProductsLiked       4.1977 0.2382 6.475348e+02 7.198384e+02 0.4881
## productsListed            1.3000 0.7692 6.074810e+01 6.753120e+01 0.8771
## productsPassRate          1.2616 0.7926 5.297530e+01 5.889050e+01 0.8903
## productsWished            1.7694 0.5651 1.558135e+02 1.732116e+02 0.7518
## productsBought            1.3915 0.7187 7.927040e+01 8.812170e+01 0.8477
## daysSinceLastLogin        1.1982 0.8346 4.013640e+01 4.461810e+01 0.9136
## seniority           4623537.0283 0.0000 9.362660e+08 1.040809e+09 0.0005
## seniorityAsMonths   4541805.0301 0.0000 9.197153e+08 1.022411e+09 0.0005
## seniorityAsYears      25611.1679 0.0000 5.186059e+06 5.765133e+06 0.0062
##                             CVIF Klein   IND1   IND2
## socialNbFollowers   2.756480e+01     1 0.0015 1.1722
## socialNbFollows     4.145160e+01     1 0.0010 1.3391
## socialProductsLiked 3.451540e+01     1 0.0012 1.2726
## productsListed      1.068910e+01     0 0.0038 0.3855
## productsPassRate    1.037350e+01     0 0.0039 0.3464
## productsWished      1.454920e+01     0 0.0028 0.7264
## productsBought      1.144120e+01     0 0.0035 0.4700
## daysSinceLastLogin  9.852200e+00     0 0.0041 0.2763
## seniority           3.801684e+07     1 0.0000 1.6705
## seniorityAsMonths   3.734480e+07     1 0.0000 1.6705
## seniorityAsYears    2.105868e+05     1 0.0000 1.6704
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

```
##
## productsWished , productsBought , daysSinceLastLogin , seniority , seniorityAsMonths , senio
rityAsYears , coefficient(s) are non-significant may be due to multicollinearity
##
## R-square of y on all x: 0.6604
##
## * use method argument to check which regressors may be the reason of collinearity
## ==================================
```

La colonna Klein presenta il valore 1 in corrispondenza di sei variabili, tale numero è sinonimo di collinearità. Convenzionalmente, i valori di VIF devono essere inferiori a 5 mentre i valori di TOL superiori a 0.30. Procediamo, quindi, all'eliminazione della variabile che presenta un valore di VIF più elevato, ovvero *seniority* (VIF: 4623537.0283), aggiorniamo il modello e ripetiamo l'operazione fino a quando le soglie risultano rispettate. Progressivamente vengono rimosse *seniorityAsYears* (24988.1983) e *socialNbFollows* (5.0401).

```
##                         VIF    TOL       Wi       Fi Leamer    CVIF Klein   IND1
## socialNbFollowers    2.5895 0.3862 460.4895 537.5027 0.6214 13.8459     0 0.0013
## socialProductsLiked 2.5285 0.3955 442.8159 516.8733 0.6289 13.5197     0 0.0014
## productsListed       1.2292 0.8135  66.4090  77.5153 0.9020  6.5727     0 0.0028
## productsPassRate     1.2542 0.7973  73.6551  85.9733 0.8929  6.7064     0 0.0028
## productsWished       1.7646 0.5667 221.5159 258.5627 0.7528  9.4354     0 0.0020
## productsBought       1.3709 0.7294 107.4658 125.4386 0.8541  7.3304     0 0.0025
## daysSinceLastLogin   1.1976 0.8350  57.2433  66.8168 0.9138  6.4035     0 0.0029
## seniorityAsMonths    1.0095 0.9905   2.7664   3.2291 0.9953  5.3981     0 0.0034
```

Osserviamo che la colonna Klein non individua più collinearità, i valori di VIF sono tutti inferiori a 5 e quelli di TOL superiori a 0.30. Concludiamo che non risultano ulteriori eliminazioni da compiere e rimangono otto covariate quantitative.

Analizziamo ora le associazioni tra le variabili qualitative. Se il chi quadro normalizzato associato a una coppia di variabili presenta un valore superiore a 0.90, allora è presente una forte associazione, quindi escludiamo una tra le due variabili coinvolte.

```
##    X1              Row            Column   df p.value Chi.Square.norm
## 1   1          country          language  164   0.000    8.323279e-01
## 2   2          country            gender   41   0.409    2.084214e-02
## 3   3          country  civilityGenderId   82   0.000    3.989435e-02
## 4   4          country     civilityTitle   82   0.000    3.989435e-02
## 5   5          country         hasAnyApp   41   0.003    3.483680e-02
## 6   6          country     hasAndroidApp   41   0.002    3.515391e-02
## 7   7          country         hasIosApp   41   0.279    2.250983e-02
## 8   8          country hasProfilePicture   41   0.000    4.590621e-02
## 9   9          country       countryCode 1681   0.000    1.000000e+00
## 10 10         language            gender    4   0.010    6.576547e-03
## 11 11         language  civilityGenderId    8   0.000    1.554454e-02
## 12 12         language     civilityTitle    8   0.000    1.554454e-02
## 13 13         language         hasAnyApp    4   0.000    1.431394e-02
## 14 14         language     hasAndroidApp    4   0.000    1.389756e-02
## 15 15         language         hasIosApp    4   0.011    6.408124e-03
## 16 16         language hasProfilePicture    4   0.000    1.371945e-02
## 17 17         language       countryCode  164   0.000    8.323279e-01
## 18 18           gender  civilityGenderId    2   0.000    1.000000e+00
## 19 19           gender     civilityTitle    2   0.000    1.000000e+00
## 20 20           gender         hasAnyApp    1   0.005    3.876045e-03
## 21 21           gender     hasAndroidApp    1   0.241    6.742038e-04
## 22 22           gender         hasIosApp    1   0.048    1.914484e-03
## 23 23           gender hasProfilePicture    1   0.781    3.803161e-05
```

```
## 24 24            gender       countryCode   41   0.409    2.084214e-02
## 25 25  civilityGenderId     civilityTitle    4   0.000    1.000000e+00
## 26 26  civilityGenderId         hasAnyApp    2   0.000    1.353952e-02
## 27 27  civilityGenderId     hasAndroidApp    2   0.448    7.887358e-04
## 28 28  civilityGenderId         hasIosApp    2   0.000    1.110371e-02
## 29 29  civilityGenderId hasProfilePicture    2   0.550    5.864470e-04
## 30 30  civilityGenderId       countryCode   82   0.000    3.989435e-02
## 31 31     civilityTitle         hasAnyApp    2   0.000    1.353952e-02
## 32 32     civilityTitle     hasAndroidApp    2   0.448    7.887358e-04
## 33 33     civilityTitle         hasIosApp    2   0.000    1.110371e-02
## 34 34     civilityTitle hasProfilePicture    2   0.550    5.864470e-04
## 35 35     civilityTitle       countryCode   82   0.000    3.989435e-02
## 36 36         hasAnyApp     hasAndroidApp    1   0.000    6.805782e-02
## 37 37         hasAnyApp         hasIosApp    1   0.000    6.808121e-01
## 38 38         hasAnyApp hasProfilePicture    1   0.000    3.292200e-02
## 39 39         hasAnyApp       countryCode   41   0.003    3.483680e-02
## 40 40     hasAndroidApp         hasIosApp    1   0.000    7.966971e-02
## 41 41     hasAndroidApp hasProfilePicture    1   0.622    1.194873e-04
## 42 42     hasAndroidApp       countryCode   41   0.002    3.515391e-02
## 43 43         hasIosApp hasProfilePicture    1   0.000    3.147240e-02
## 44 44         hasIosApp       countryCode   41   0.279    2.250983e-02
## 45 45 hasProfilePicture       countryCode   41   0.000    4.590621e-02
```

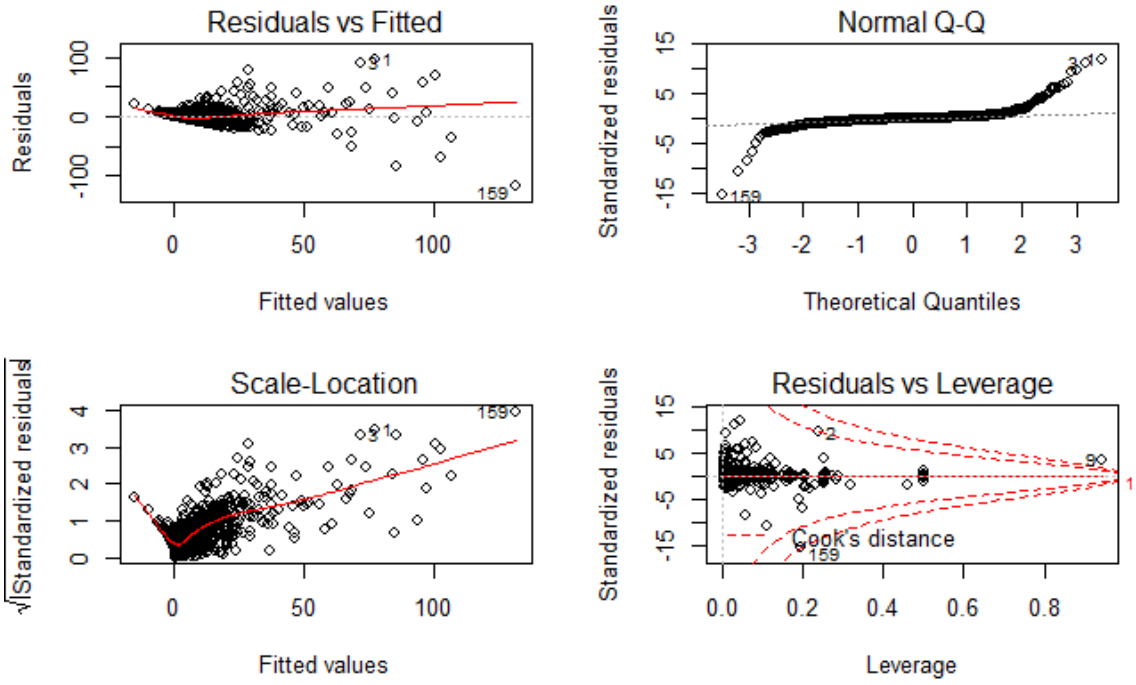
Notiamo dalla colonna Chi.Square.norm un valore pari a 1 per le coppie di variabili *gender* con *civilityTitle*, *gender* con *civilityGenderId* e *civilityTitle* con *civilityGenderId*.

```
##           F    M
##   miss  134    0
##   mr      0  486
##   mrs  1416    0


##        F    M
##   1    0  486
##   2 1416    0
##   3  134    0
```

Dalla tabella di contigenza notiamo infatti massima associazione tra queste variabili. Tale risultato era prevedibile, poichè tutte e tre considerano il genere, di conseguenza eliminiamo *gender* e *civilityGenderID,* poichè entrambe perfettamente associate a *civilityTitle*. Ricalcolando i chi quadrati normalizzati tra le variabili qualitative rimaste, si osservano valori tutti inferiori alla soglia 0.90, pertanto non procediamo a ulteriori eliminazioni. Manteniamo dunque otto variabili qualitative in considerazione.

Starting model

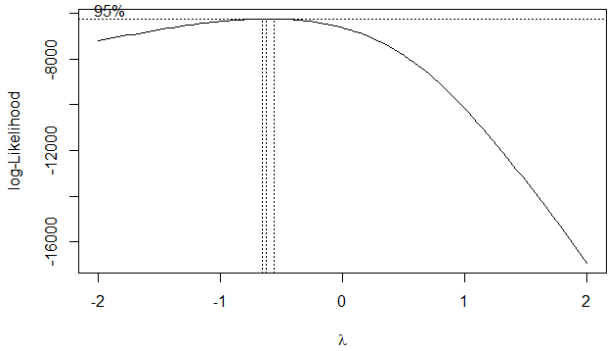
Procediamo alla formulazione dello starting model.



Esprimiamo la variabile risposta in funzione delle variabili quantitative e qualitative rimaste in analisi. Il modello presenta un R^2 aggiustato di circa 0.63 e un elevato numero di coefficienti, molti non significativi. Analizzando i grafici del modello di partenza si rilevano diverse problematiche. Notiamo in particolare valori di leverage molto elevati (quarto grafico) e una situazione di forte eteroschedasticità (terzo grafico).

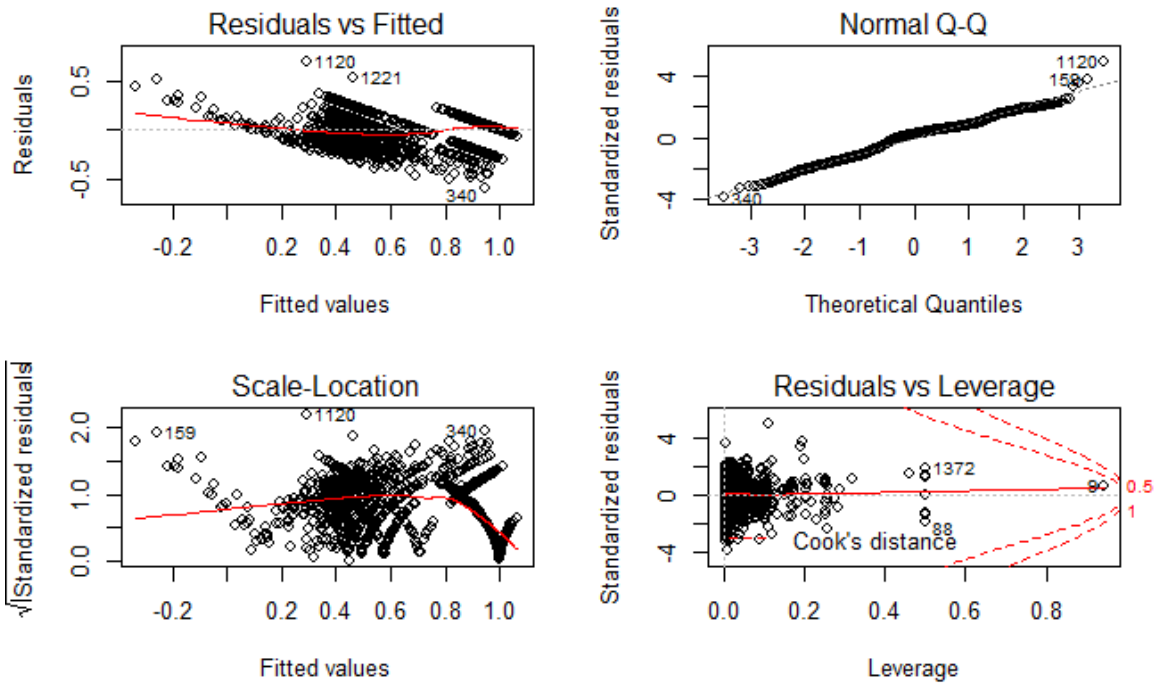
Linearità

Procediamo con l'analisi della linearità per il modello iniziale, dapprima valutando la variabile risposta. L'ipotesi di linearità è un'assunzione molto forte e se non rispettata comporta stime non BLUE, distorte ed inefficienti. Tramite la funzione Box-Cox otteniamo il valore di lambda che minimizza l'SSE (sum of squares errors), a cui dobbiamo elevare la variabile risposta per ottenere la miglior trasformazione di questa.



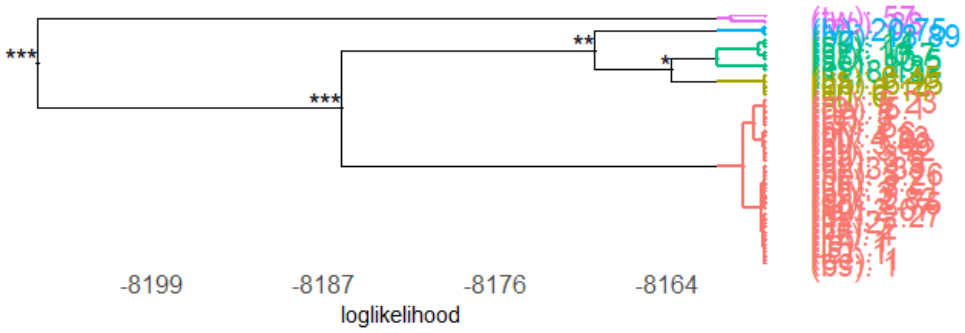

```
## [1] -0.6262626
```

Il valore di lambda risulta essere prossimo a -0.5.



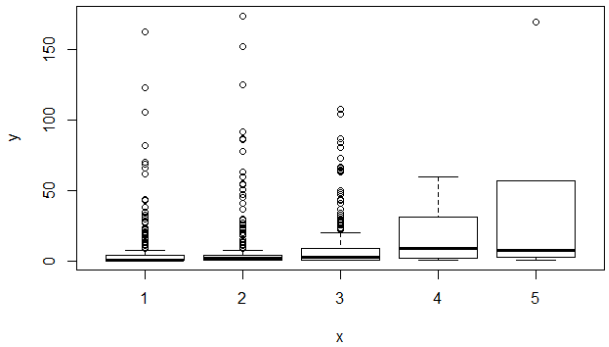
L'R^2 aggiustato di questo nuovo modello con la variabile risposta trasformata aumenta rispetto al precedente. I grafici appaiono molto differenti rispetto a quelli del modello iniziale. In particolare, si evince un netto miglioramento del terzo grafico relativo ai fitted values che ora sembrano essere più distribuiti e più lineari.

A questo punto dell'analisi, ci concentriamo sulla linearità delle singole covariate, nello specifico operiamo con l'optimal grouping per le variabili qualitative. Le variabili *country* e *countryCode* presentano un numero elevato di livelli (42), quindi ci serviamo della procedura di optimal grouping per ridurli.




```
## [1] 2036

##    1    2    3    4    5
## 1056  447  514   13    6
```

Il Merging Path Plot panel mostra la struttura gerarchica delle similarità tra i gruppi. Le stelle indicano quanto sono significative le differenze tra due cluster. I 2036 utenti sono classificati in base alla sigla dello Stato in cinque gruppi di ampiezza differente: il più numeroso comprende 1056 utenti e il più piccolo ne caratterizza solo 6. Otteniamo quindi *optimal_countrycode*, una variabile qualitativa ricodificata in livelli identificati con numeri interi da 1 a 5.



I boxplot confermano la divisione in gruppi effettuata ed è possibile osservare che i valori delle mediane aumentano in corrispondenza del passaggio al livello successivo. Sono presenti anche dei possibili valori anomali, in particolare per i primi tre livelli, ovvero quelli che comprendono più osservazioni.

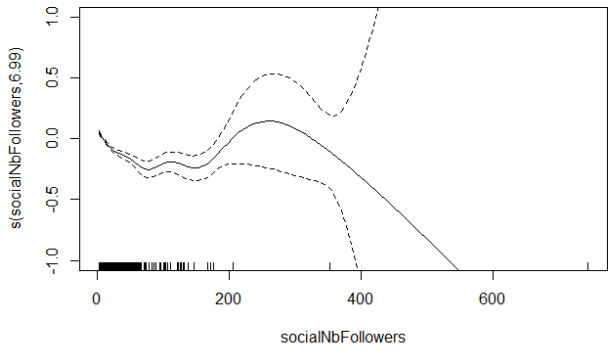
Sempre nel contesto dell'analisi della linearità, formuliamo un modello gam con la stessa struttura del precedente, seguito da uno analogo, con l'aggiunta di "s" che precedono le variabili quantitative, che, a nostro avviso, potrebbero avere una relazione non lineare con la dipendente. Dal summary del modello ricaviamo la significatività approsimativa delle trasformazioni. Valutiamo in particolare *socialNbFollowers*, *socialProductsLiked*, *productsListed* e *productsPassRate* che presentano p-value significativi.

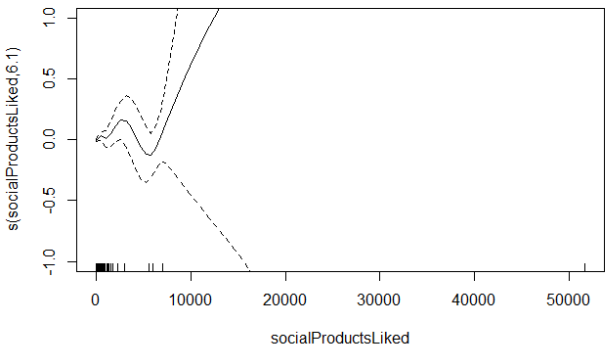
Eseguiamo ora il test del rapporto di verosomiglianza (Likelihood Ratio Test) per effettuare un confronto tra il modello lineare ottenuto in seguito alla trasformazione Box-Cox (1) e il modello gam appena formulato (2).

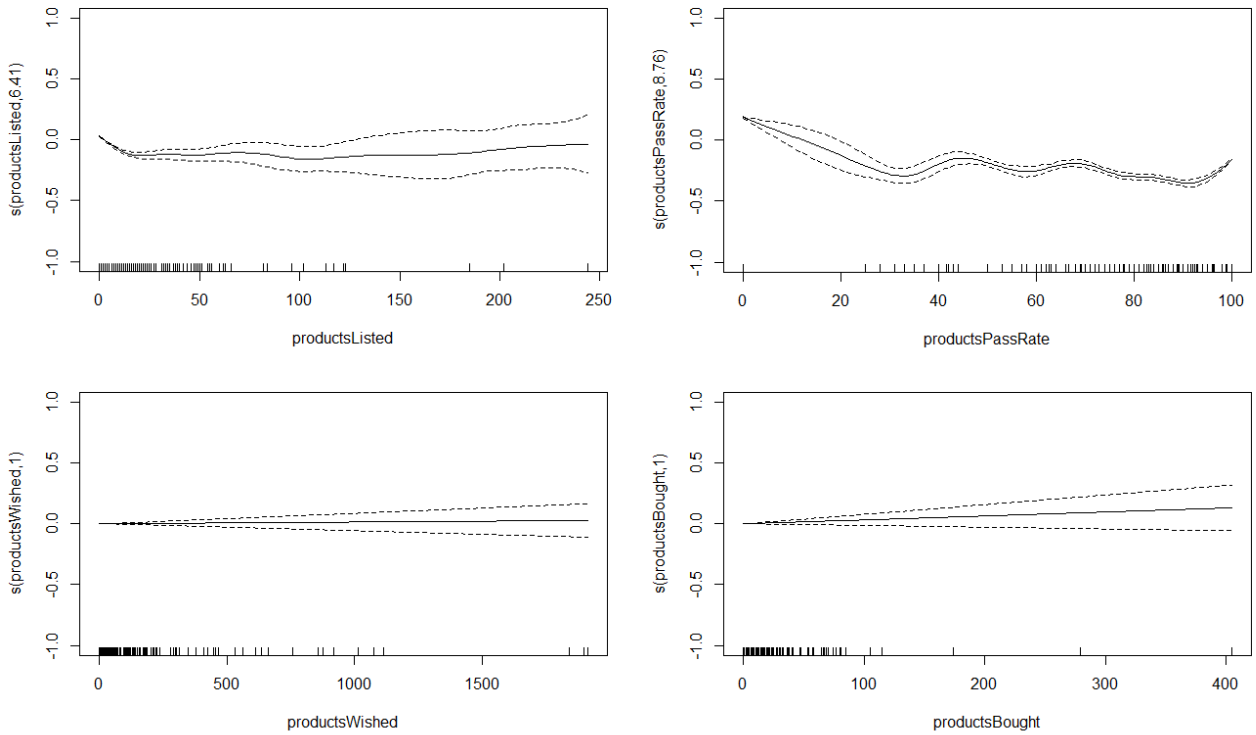
```
## Analysis of Deviance Table
##

##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1      2013     47.696
## 2      1986     28.734 27.018   18.962 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notiamo che il p-value osservato per la statistica Chi-quadrato è prossimo a 0 pertanto possiamo concludere che il modello gam è significativamente migliore in termini di likelihood.

Analizziamo i plot del modello gam per osservare graficamente se l'andamento delle covariate è lineare o segue un'altra distribuzione. I grafici sono ottenuti tramite procedura splines.



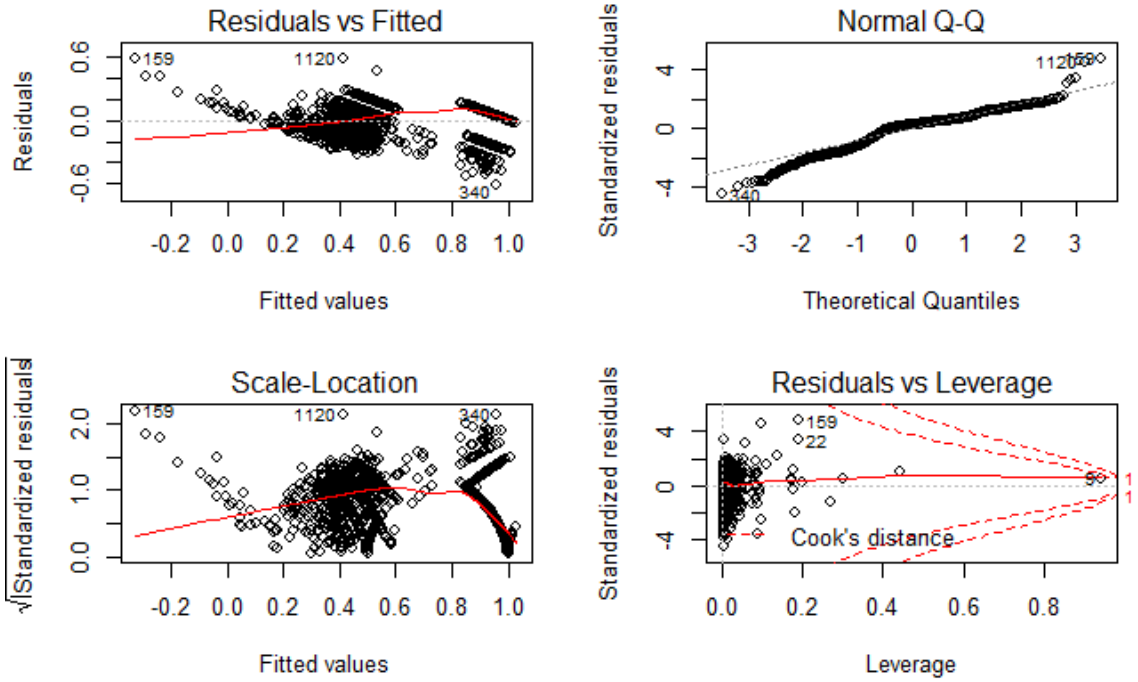




Poniamo particolare attenzione all'andamento delle funzioni in corrispondenza di concentrazione maggiore di trattini, posizionati sull'asse delle ascisse, che indicano le osservazioni del dataset. L'unica variabile che presenta un andamento non lineare è *productsPassRate*, quindi formuliamo un nuovo modello che presenta tale variabile anche al secondo e terzo grado. Manteniamo queste trasformazioni in seguito, poiché il coefficiente di terzo grado risulta significativo.

```
## Call:
## lm(formula = (productsSold)^(-0.5) ~ socialNbFollowers + socialProductsLiked +
##     productsListed + productsPassRate + I(productsPassRate^2) +
##     I(productsPassRate^3) + productsWished + productsBought +
##     daysSinceLastLogin + seniorityAsMonths + language + civilityTitle +
##     hasAnyApp + hasAndroidApp + hasIosApp + hasProfilePicture +
##     optimal_countrycode, data = data_completo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60795 -0.07517  0.03589  0.07749  0.58357
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.595e-01  5.819e-02  16.488  < 2e-16 ***
## socialNbFollowers      -2.626e-03  2.088e-04 -12.577  < 2e-16 ***
## socialProductsLiked     3.149e-05  4.132e-06   7.619 3.91e-14 ***
## productsListed         -2.451e-03  2.575e-04  -9.520  < 2e-16 ***
## productsPassRate       -7.172e-03  1.578e-03  -4.546 5.79e-06 ***
## I(productsPassRate^2)  -7.655e-05  4.005e-05  -1.911  0.05610 .
## I(productsPassRate^3)   1.089e-06  2.476e-07   4.397 1.15e-05 ***
## productsWished         -3.335e-06  3.859e-05  -0.086  0.93114
## productsBought          2.795e-05  2.532e-04   0.110  0.91211
## daysSinceLastLogin      9.324e-05  1.490e-05   6.259 4.72e-10 ***
## seniorityAsMonths       7.289e-05  5.403e-04   0.135  0.89270
## languageen              2.213e-02  1.591e-02   1.391  0.16433
## languagees              2.861e-02  2.416e-02   1.184  0.23640
## languagefr              2.462e-03  1.447e-02   0.170  0.86490
```

```
## languageit              2.678e-02  2.192e-02   1.222  0.22182
## civilityTitlemr        -3.365e-02  1.364e-02  -2.468  0.01367 *
## civilityTitlemrs       -2.743e-02  1.264e-02  -2.169  0.03017 *
## hasAnyAppTrue          -3.512e-02  2.844e-02  -1.235  0.21708
## hasAndroidAppTrue       3.000e-02  2.615e-02   1.147  0.25140
## hasIosAppTrue           2.604e-02  2.772e-02   0.939  0.34759
## hasProfilePictureTrue   3.399e-03  7.040e-03   0.483  0.62933
## optimal_countrycode2   -2.757e-02  1.084e-02  -2.543  0.01106 *
## optimal_countrycode3   -4.905e-02  1.614e-02  -3.039  0.00241 **
## optimal_countrycode4   -9.779e-02  3.928e-02  -2.490  0.01286 *
## optimal_countrycode5   -1.636e-02  5.648e-02  -0.290  0.77210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1354 on 2011 degrees of freedom
## Multiple R-squared:  0.7905, Adjusted R-squared:  0.788
## F-statistic: 316.2 on 24 and 2011 DF,  p-value: < 2.2e-16
```



I grafici non mostrano significativi miglioramenti per quanto riguarda eteroschedasticità e residui. Rimaniamo comunque soddisfatti delle trasformazioni effettuate nell'analisi della linearità, in particolare per i miglioramenti apportati dalla trasformazione Box-Cox della variabile risposta.

Model selection

Spesso, alcune variabili inserite nel modello di regressione non sono significativamente associate con la variabile risposta. Se queste vengono incluse si rende il modello più complesso del necessario, complicando l'interpretazione degli output. La model selection mira dunque a rimuovere tali variabili, così da ottenere un modello facilmente interpretabile. Applichiamo la funzione di R stepAIC con direzione "both" (stepwise selection), che ci permette di ottenere un modello parsimonioso, ovvero semplice e allo stesso tempo performante per prevedere la y con accuratezza.

```
## Start:  AIC=-7647.59
## (productsSold)^(-0.5) ~ socialNbFollowers + socialProductsLiked +
##     productsListed + productsPassRate + I(productsPassRate^2) +
##     I(productsPassRate^3) + productsWished + productsBought +
##     daysSinceLastLogin + seniorityAsMonths + language + civilityTitle +
##     hasAnyApp + hasAndroidApp + hasIosApp + hasProfilePicture +
##     optimal_countrycode
##
##                          Df Sum of Sq    RSS     AIC
## - language                4   0.06181 34.907 -7652.2
## - seniorityAsMonths       1   0.00028 34.846 -7649.6
## - productsBought          1   0.00089 34.846 -7649.5
## - hasProfilePicture       1   0.00092 34.846 -7649.5
## - productsWished          1   0.00988 34.855 -7649.0
## - hasIosApp               1   0.02105 34.866 -7648.4
## - hasAndroidApp           1   0.02575 34.871 -7648.2
## - hasAnyApp               1   0.03082 34.876 -7647.9
## <none>                                34.845 -7647.6
## - I(productsPassRate^2)   1   0.06784 34.913 -7645.9
## - civilityTitle           2   0.13977 34.985 -7643.9
## - optimal_countrycode     4   0.28957 35.135 -7639.7
## - productsPassRate        1   0.33041 35.176 -7631.5
## - I(productsPassRate^3)   1   0.33393 35.179 -7631.3
## - daysSinceLastLogin      1   0.69643 35.542 -7611.6
## - socialProductsLiked     1   1.06058 35.906 -7592.0
## - productsListed          1   1.41610 36.261 -7573.1
## - socialNbFollowers       1   2.82542 37.671 -7499.9

...

## Step:  AIC=-7662.94
## (productsSold)^(-0.5) ~ socialNbFollowers + socialProductsLiked +
##     productsListed + productsPassRate + I(productsPassRate^2) +
##     I(productsPassRate^3) + daysSinceLastLogin + civilityTitle +
##     optimal_countrycode
##
##                          Df Sum of Sq    RSS     AIC
## <none>                                34.966 -7662.9
## + hasAnyApp               1   0.01617 34.950 -7661.8
## + productsWished          1   0.01412 34.952 -7661.7
## + hasIosApp               1   0.01231 34.954 -7661.6
## - I(productsPassRate^2)   1   0.06861 35.035 -7661.2
## + hasAndroidApp           1   0.00206 34.964 -7661.1
## + hasProfilePicture       1   0.00136 34.965 -7661.0
## + productsBought          1   0.00113 34.965 -7661.0
## + seniorityAsMonths       1   0.00001 34.966 -7660.9
## + language                4   0.06102 34.905 -7658.3
## - civilityTitle           2   0.16046 35.127 -7658.1
## - optimal_countrycode     4   0.39038 35.357 -7649.6
## - productsPassRate        1   0.33197 35.298 -7646.8
## - I(productsPassRate^3)   1   0.33663 35.303 -7646.5
## - daysSinceLastLogin      1   0.68592 35.652 -7627.6
## - socialProductsLiked     1   1.30037 36.267 -7594.8
## - productsListed          1   1.40448 36.371 -7589.3
## - socialNbFollowers       1   3.12228 38.089 -7500.7
```

In alternativa eseguiamo un altro stepwise basandoci sull'indice SBC, molto simile ad AIC, ma più severo per i modelli con un numero di covariate maggiore. Ci aspettiamo dunque un modello più semplice, con meno covariate.

Il modello ricavato rispetta le aspettative, infatti presenta due covariate in meno (*civilityTitle* e *optimalcountryCode*) rispetto a quello ottenuto con l'AIC.

```
## Call:
## lm(formula = (productsSold)^(-0.5) ~ socialNbFollowers + socialProductsLiked +
##     productsListed + productsPassRate + I(productsPassRate^2) +
##     I(productsPassRate^3) + daysSinceLastLogin, data = d0_nona)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60516 -0.07886  0.03890  0.07577  0.63330
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.360e-01  6.044e-03 154.865  < 2e-16 ***
## socialNbFollowers      -2.769e-03  2.048e-04 -13.519  < 2e-16 ***
## socialProductsLiked     3.405e-05  3.931e-06   8.662  < 2e-16 ***
## productsListed         -2.336e-03  2.629e-04  -8.887  < 2e-16 ***
## productsPassRate       -6.608e-03  1.637e-03  -4.037 5.63e-05 ***
## I(productsPassRate^2) -9.104e-05  4.155e-05  -2.191   0.0286 *
## I(productsPassRate^3)  1.170e-06  2.568e-07   4.558 5.50e-06 ***
## daysSinceLastLogin     9.269e-05  1.514e-05   6.123 1.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1363 on 1912 degrees of freedom
## Multiple R-squared:  0.7864, Adjusted R-squared:  0.7856
## F-statistic:  1006 on 7 and 1912 DF,  p-value: < 2.2e-16
```

La bontà dei due modelli (model_aic e model_sbc) non differisce in modo significativo, scegliamo di mantenere il modello più semplice per le successive analisi, ovvero quello ottenuto con la procedura SBC.

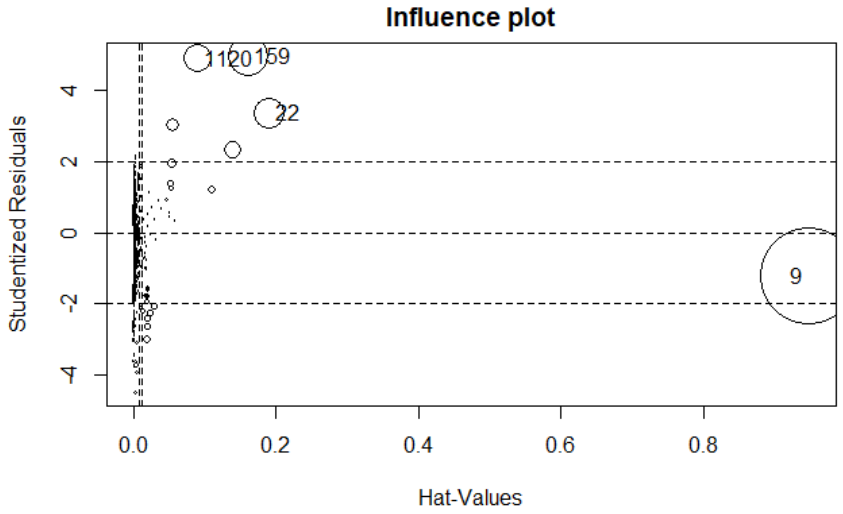
Rifittiamo ora il modello scelto sul dataset di partenza. Questo passaggio è importante poichè, in certi casi, il modello più parsimonioso viene stimato su un numero di osservazioni maggiore rispetto a quello su cui è stato stimato il modello di partenza.

```
## [1] 1920

## [1] 1935
```

Dato che abbiamo rimosso la variabile *productsBought*, il modello viene stimato su un numero di osservazioni maggiore (15 in più).

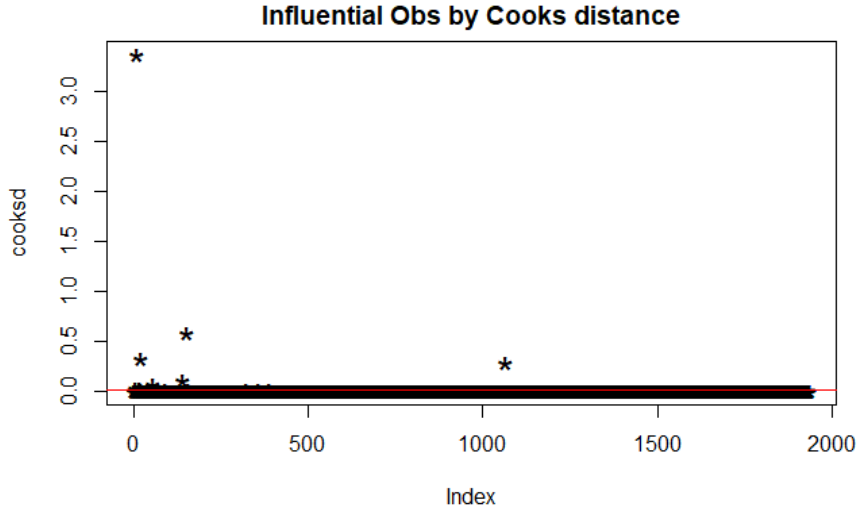
Valori influenti

Osservazioni inusuali all'interno del dataset possono risultare problematiche quando si vuole stimare un modello di regressione lineare tramite stimatori a minimi quadrati. Per identificare queste particolari osservazioni, dette valori influenti, ci serviamo di un Influence Plot che dispone le osservazioni su un grafico con in ascissa i valori degli hat-values e in ordinata i valori dei residui studentizzati.



Dalla dimensione delle bolle nel grafico identifichiamo alcuni dei valori influenti nel nostro dataset. Il venditore 9 risulta avere un alto hat-value ma un basso residuo studentizzato: questo venditore ha un elevato numero di followers, è un venditore che ha messo un gran numero di 'mi piace' a prodotti presenti sul social network, ma il numero di prodotti che vende non è significativamente maggiore rispetto ad altri venditori che sono molto meno attivi sui social. Invece i venditori 159 e 22 hanno un valore non particolarmente elevato di hat-value, ma il loro residuo studentizzato è alto: questi riescono a vendere un numero di prodotti molto più alto rispetto ad altri venditori che hanno simili caratteristiche, tra cui il numero di followers, di 'mi piace' e di prodotti in lista.

Grazie alle distanze di Cook verifichiamo l'eventuale influenza di questi venditori sui parametri del modello. Se ciò si verifica, allora dovremo escluderli dal dataset.



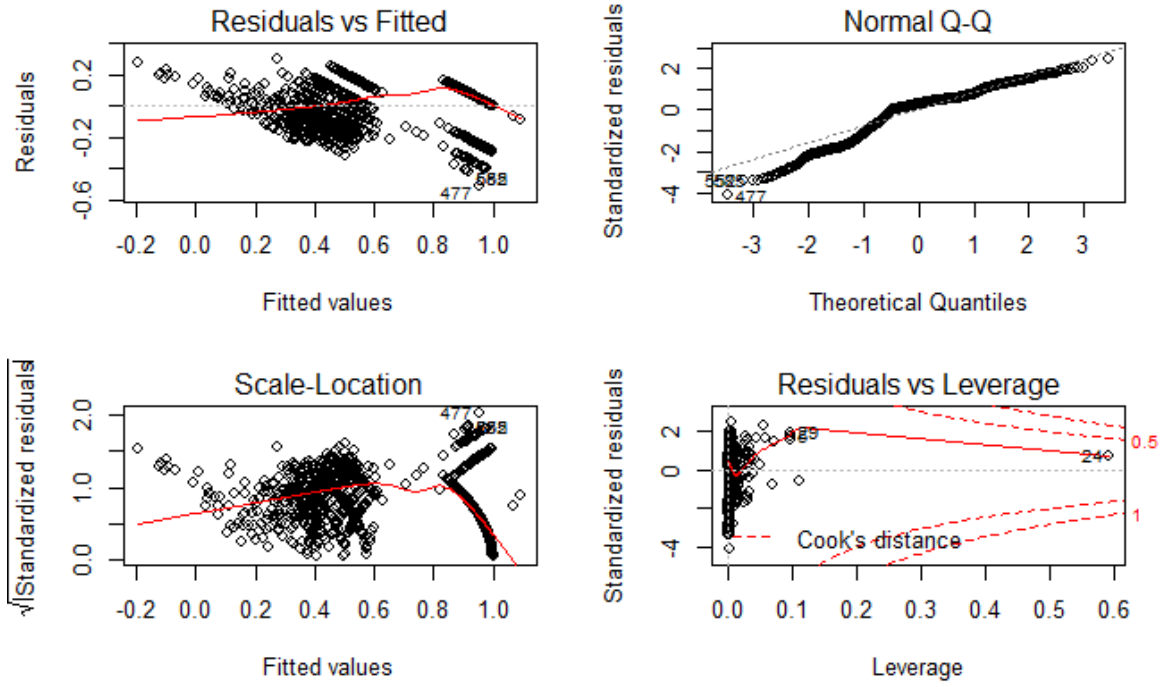
Il grafico mostra con gli asterischi le osservazioni che presentano distanze di Cook che superano la soglia 4/n-p-2 (molto prossima a 0) e procediamo con la loro rimozione dal dataset, per la precisione eliminiamo 67 osservazioni (tra cui il venditore 9 e 159).

Ora confrontiamo questo modello imputato sul dataset senza valori influenti con il modello imputato precedentemente sul dataset completo.

```
## Call:
## lm(formula = (productsSold)^(-0.5) ~ socialNbFollowers + socialProductsLiked +
##     productsListed + productsPassRate + I(productsPassRate^2) +
##     I(productsPassRate^3) + daysSinceLastLogin, data = data_finale)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50889 -0.06199  0.03453  0.07062  0.30168
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.473e-01  5.597e-03 169.240  < 2e-16 ***
## socialNbFollowers      -3.368e-03  2.574e-04 -13.088  < 2e-16 ***
## socialProductsLiked     8.152e-05  1.454e-05   5.606 2.39e-08 ***
## productsListed         -4.708e-03  4.150e-04 -11.345  < 2e-16 ***
## productsPassRate        4.834e-04  1.899e-03   0.255    0.799
## I(productsPassRate^2) -2.650e-04  4.739e-05  -5.591 2.59e-08 ***
## I(productsPassRate^3)  2.216e-06  2.878e-07   7.698 2.23e-14 ***
## daysSinceLastLogin     8.924e-05  1.388e-05   6.429 1.63e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1231 on 1860 degrees of freedom
##   (101 observations deleted due to missingness)
## Multiple R-squared:  0.8212, Adjusted R-squared:  0.8205
## F-statistic:  1221 on 7 and 1860 DF,  p-value: < 2.2e-16
```

È da osservare che i parametri del modello imputato sul detaset senza valori influenti cambiano (anche se non di molto) rispetto ai parametri dell'altro modello. L'adattamento migliora: l'R^2 aggiustato aumenta da 0.78 a 0.82.



Un notevole cambiamento si verifica nel grafico in basso a destra, in quanto non sono più presenti valori che oltrepassano la linea rossa della distanza di Cook. La presenza di outliers e valori influenti è spesso anche fonte di eteroschedasticità, dunque, nel momento in cui questi vengono rimossi dal dataset, questo problema si attenua. Tuttavia, dai grafici tale miglioramento non appare evidente, dunque ci serviamo di test statistici più precisi.

Eteroschedasticità

La presenza di eteroschedasticità implica la violazione sull'assunzione di variabilità costante degli errori del modello stimato. Per capire se il nostro modello finale soffre di questo problema eseguiamo i test di White e Breush-Pagan su più modelli, rispettivamente starting_model, modello stimato sul dataset completo a seguito delle trasformazioni lineari e modello finale.

```
##  studentized Breusch-Pagan test
## data:  starting_model
## BP = 987.65, df = 59, p-value < 2.2e-16

##  studentized Breusch-Pagan test
## data:  model_5
## BP = 175, df = 24, p-value < 2.2e-16

##  studentized Breusch-Pagan test
## data:  model_final
## BP = 80.401, df = 7, p-value = 1.141e-14



## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 20499.32, Df = 1, p = < 2.22e-16

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 22.11043, Df = 1, p = 2.5741e-06

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 8.40526, Df = 1, p = 0.0037414
```

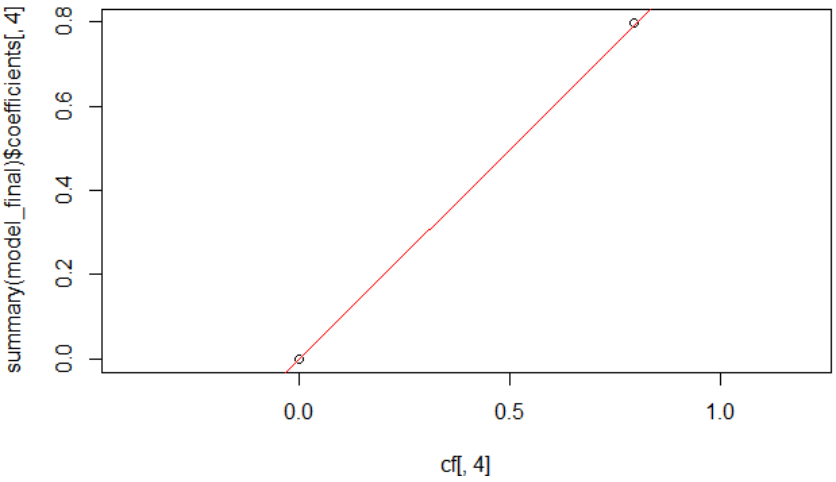
Il modello finale soffre di eteroschedasticità, poichè rifiutiamo l'ipotesi nulla H0 di varianza costante dei residui elaborata da White. Tuttavia, è da sottolineare il miglioramento della statistica test chi-quadrato, da 22.11 a 8.40, ottenuto eliminando i valori influenti.

Procediamo alla stima degli standard error robusti di White, così da poter svolgere un'inferenza corretta sui nostri parametri.

```
## Uncorrected Tests of Coefficients
##
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.47e-01   5.60e-03 169.240 0.00e+00
## socialNbFollowers      -3.37e-03   2.57e-04 -13.088 1.67e-37
## socialProductsLiked     8.15e-05   1.45e-05   5.606 2.39e-08
## productsListed         -4.71e-03   4.15e-04 -11.345 6.79e-29
## productsPassRate        4.83e-04   1.90e-03   0.255 7.99e-01
## I(productsPassRate^2) -2.65e-04   4.74e-05  -5.591 2.59e-08
## I(productsPassRate^3)  2.22e-06   2.88e-07   7.698 2.23e-14
## daysSinceLastLogin      8.92e-05   1.39e-05   6.429 1.63e-10
##
## White (1980) Heteroscedasticity-corrected SEs and Tests
##
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.47e-01   5.85e-03  161.92 0.00e+00
## socialNbFollowers      -3.37e-03   3.33e-04  -10.12 1.79e-23
## socialProductsLiked     8.15e-05   1.57e-05    5.20 2.21e-07
## productsListed         -4.71e-03   4.82e-04   -9.78 4.81e-22
```

```
## productsPassRate       4.83e-04   1.86e-03    0.26 7.95e-01
## I(productsPassRate^2) -2.65e-04   4.61e-05   -5.75 1.05e-08
## I(productsPassRate^3)  2.22e-06   2.79e-07    7.93 3.78e-15
## daysSinceLastLogin     8.92e-05   1.22e-05    7.33 3.33e-13
```

Notiamo che gli standard error robusti di White non sono particolarmente diversi dagli standard error non robusti stimati dal modello, infatti non riscontriamo una pesante eteroschedasticità. Valutiamo graficamente quanto appena osservato:



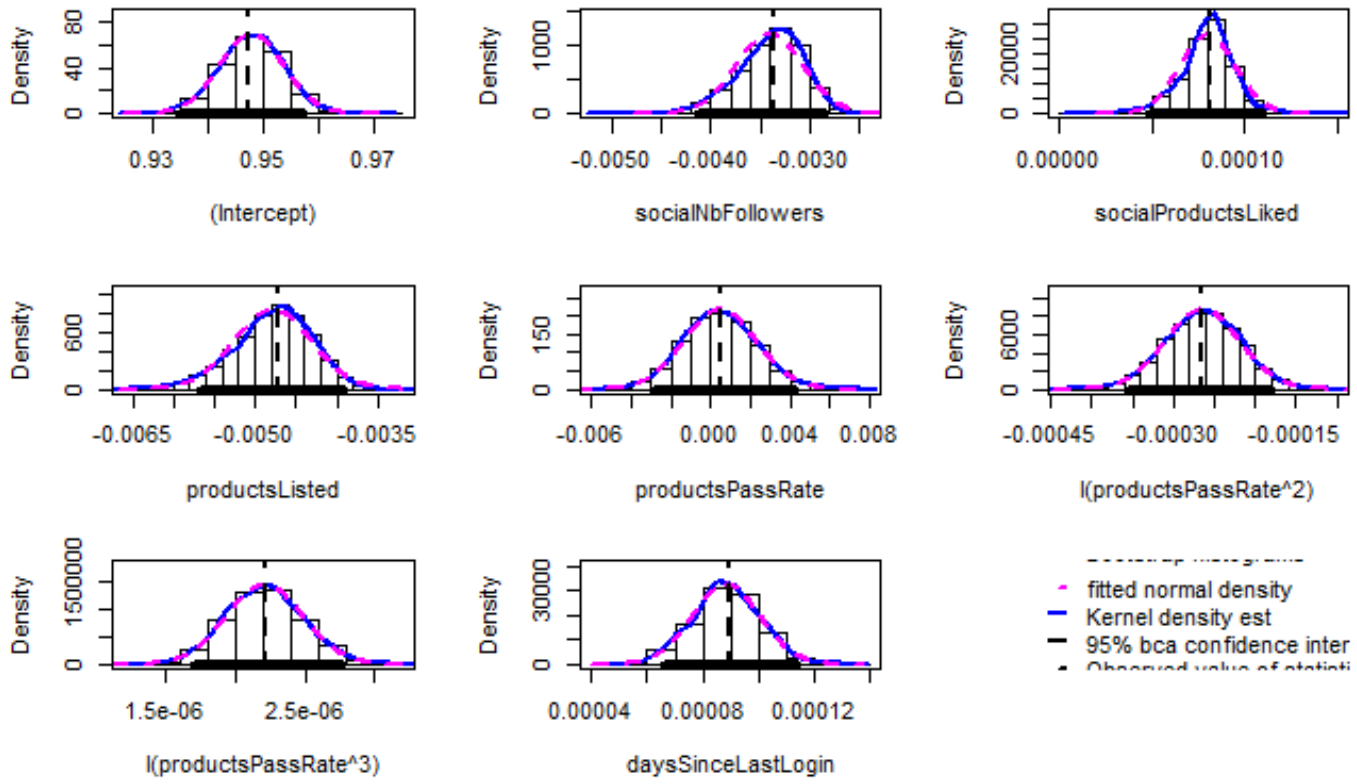
In ascissa sono riportati i valori dei p-value corretti e in ordinata i valori dei p-value calcolati sui parametri del modello. Rispetto alla bisettrice del grafico, la maggior parte di questi è posizionata lungo la linea (sovrapposti l'un con l'altro), eccetto uno che si discosta leggermente. Tale discostamento è sinonimo di eteroschedasticità, a conferma di quanto già considerato.

Bootstrap

Valutiamo la robustezza del modello finale tramite la strategia Bootstrap sui parametri.

```
##
## Number of bootstrap replications R = 1999
##                          original    bootBias     bootSE     bootMed  bootSkew
## (Intercept)            9.4728e-01  6.4534e-04 5.8038e-03  9.4811e-01 -0.046921
## socialNbFollowers     -3.3682e-03 -3.8397e-05 3.4078e-04 -3.3666e-03 -0.687077
## socialProductsLiked    8.1524e-05 -1.2625e-06 1.5093e-05  8.1314e-05 -0.181329
## productsListed        -4.7082e-03 -7.3296e-05 4.7167e-04 -4.7373e-03 -0.396680
## productsPassRate       4.8341e-04 -6.8294e-06 1.8361e-03  4.1424e-04  0.158580
## I(productsPassRate^2) -2.6499e-04  5.7815e-07 4.5628e-05 -2.6373e-04 -0.118308
## I(productsPassRate^3)  2.2156e-06 -4.9617e-09 2.7706e-07  2.2096e-06  0.090253
## daysSinceLastLogin     8.9236e-05 -7.7782e-07 1.2268e-05  8.8165e-05  0.051402
##                       bootKurtosis
## (Intercept)               0.171406
## socialNbFollowers         0.842920
## socialProductsLiked       1.894855
## productsListed            0.124101
## productsPassRate          0.103763
## I(productsPassRate^2)     0.045551
## I(productsPassRate^3)     0.014008
## daysSinceLastLogin        0.095877

## Bootstrap percent confidence intervals
##
##                            Estimate         2.5 %        97.5 %
## (Intercept)            9.472791e-01  9.360829e-01  9.591914e-01
```

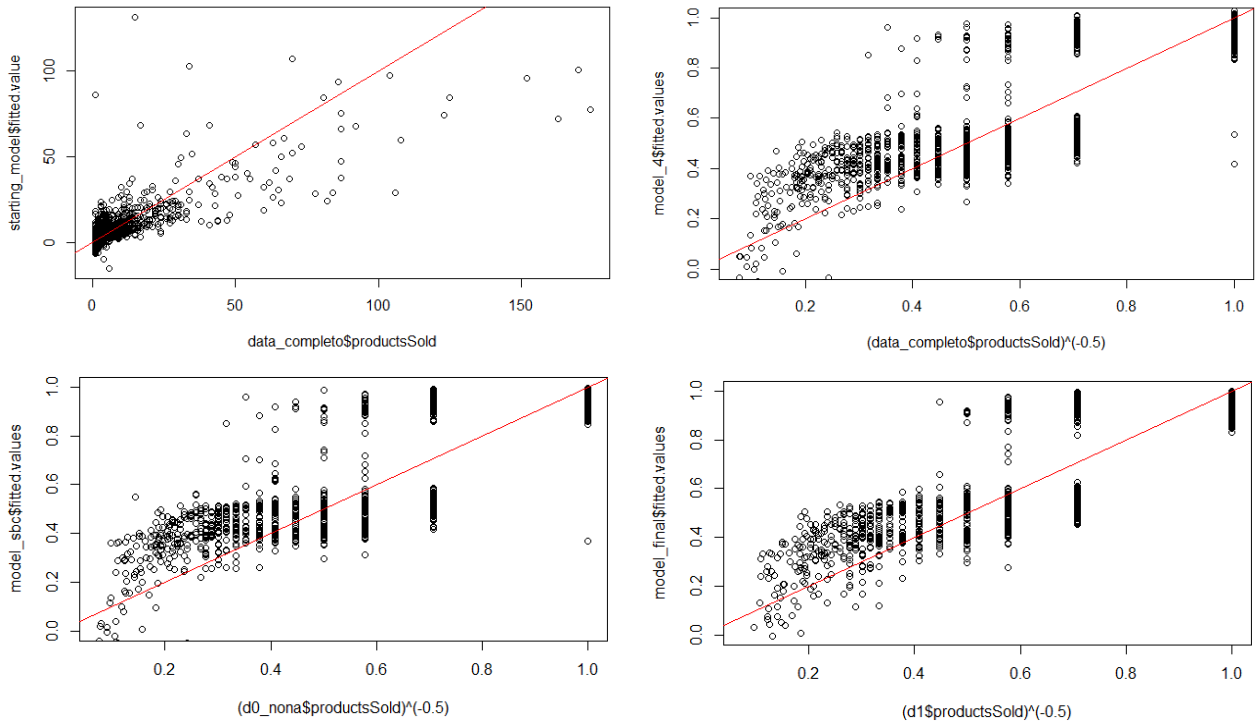
```
## socialNbFollowers     -3.368229e-03 -4.154025e-03 -2.851753e-03
## socialProductsLiked    8.152435e-05  4.775289e-05  1.092004e-04
## productsListed        -4.708216e-03 -5.798410e-03 -3.950712e-03
## productsPassRate       4.834064e-04 -2.990354e-03  4.189039e-03
## I(productsPassRate^2) -2.649876e-04 -3.550622e-04 -1.790858e-04
## I(productsPassRate^3)  2.215640e-06  1.684549e-06  2.756431e-06
## daysSinceLastLogin     8.923643e-05  6.456934e-05  1.120461e-04
```



Da questi grafici possiamo osservare che i parametri del nostro modello finale sono robusti, infatti, per ogni variabile l'intervallo di confidenza Boot (empirico) è centrato sulla stima del rispettivo parametro. Intuiamo, quindi, che queste stime non sovrastimano o sottostimano i vari effetti delle covariate sulla variabile dipendente (*productsSold*). L'intervallo di confidenza Boot della variabile *productsPassRate* comprende il valore 0, quindi possiamo concludere che questa variabile non è significativa all'interno del nostro modello, a conferma di quanto osservato dal summary. Tuttavia, non procediamo alla sua eliminazione, poiché risultano significativi i rispettivi effetti quadratici e cubici.

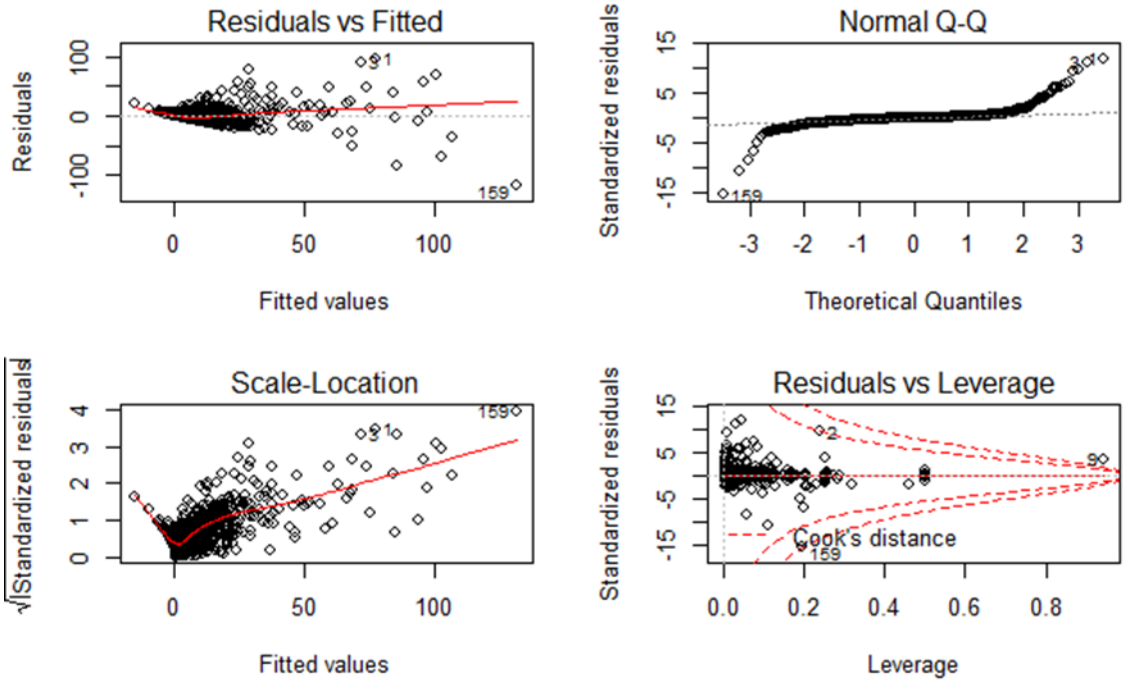
Evoluzione del modello

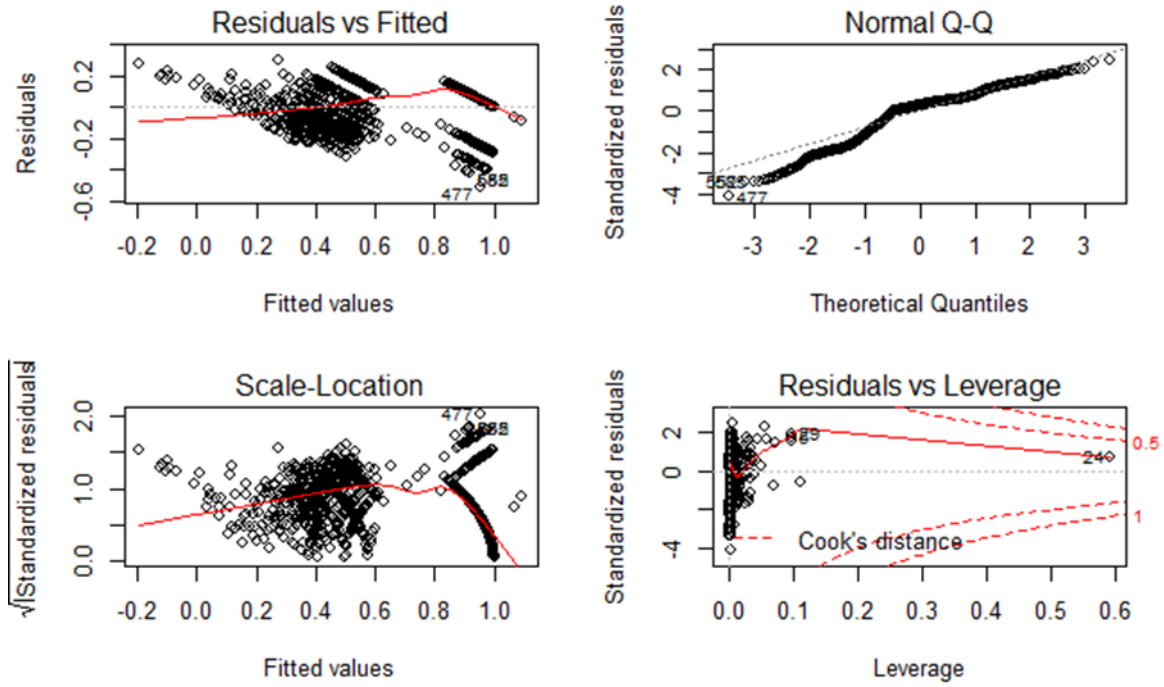
Il modello finale stimato, qualsiasi siano gli step e trasformazioni adottate sulle variabili, resta sempre un modello lineare. Il nostro obiettivo è quello di massimizzare la correlazione tra i valori osservati della variabile dipendente (*productSold*) e i valori previsti del modello, per questo motivo ricaviamo i grafici 'y vs y-fittati' sui diversi modelli.



Il primo grafico presenta una grande concentrazione di punti nella porzione a sinistra del plot. La nuvola dei punti non sembra seguire un preciso andamento lineare e sono presenti diverse osservazioni anomale. Invece, nell'ultimo grafico relativo al modello finale osserviamo che la correlazione tra y e y fittato sembra essersi rafforzata, infatti i punti sono meno dispersi attorno alla retta.

Confrontiamo nuovamente i grafici relativi allo starting model e al modello finale.





I cambiamenti sono evidenti in tutti i grafici, in particolare notiamo che i valori di y-fittata si distribuiscono in modo più uniforme su tutto il range. Osserviamo miglioramenti in termini di eteroschedasticità, infatti per il modello finale, all'aumentare dei fitted values i punti risultano meno dispersi attorno alla loro media. Inoltre, il modello finale non offre più le osservazioni influenti in accordo con la distanza di Cook.