Dataset Suicidi

Il dataset master.csv, estratto dal sito kaggle.com, è costituito da 4 diversi dataset combinati sulla base di parametri spazio-temporali. Il dataset è formato da 27820 osservazioni e 12 variabili:

- ***Country***: nazione in cui è avvenuto il suicidio.
- ***Year***: anno in cui si è verificato l'evento
- ***Sex***: sesso dell'individuo
- ***Age***: classe di età di appartenenza dell'individuo.
- ***Suicides_no***: numero di suicidi dello specifico sottogruppo per la riga.
- ***Population***: popolazione dello specifico sottogruppo per la riga.
- ***Sucides/100k pop***: tasso di suicidi dello specifico sottogruppo per la riga.
- ***Country-year***: unione delle variabili country e year.
- ***HDI for year***: indice di sviluppo umano relativo all'anno dell'evento.
- ***Gdp_for_year ($)***: prodotto interno lordo nell'anno del suicidio.
- ***Gdp_per_capita ($)***: prodotto interno lordo pro capite nell'anno del suicidio.
- ***Generation***: generazione di appartenenza dell'individuo in base all'anno di nascita.

Nel dataset sono presenti 101 nazioni. Gli anni considerati vanno dal 1985 al 2016. Nella popolazione in esame le componenti maschili e femminili sono identiche. La variabile *age* comprende 6 classi di età: 5-14, 15-24, 25-34, 35-54, 55-74, 75+. Anche *Generation* è divisa in 6 diverse categorie: G.I.Generation (1901-1927), Silent (1925-1942), Boomers (1946-1964), Generation X (1960-1980), Millennials (1980-primi2000), Generation Z (1990-2000s).

Il nostro intento è studiare le caratteristiche che influisco maggiormente sulla tendenza al suicidio. L'obiettivo è quello di stimare un modello lineare robusto che permetta di fare previsioni. Si decide quindi di utilizzare come variabile target *suicides/100k pop*.

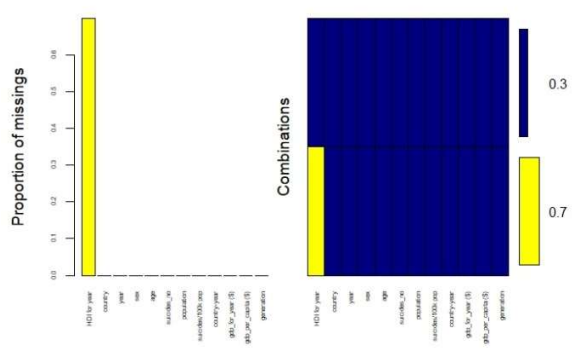
Vengono riportate le tipologie e le statistiche descrittive delle variabili

```
> str(suicidi)
Classes 'tbl_df', 'tbl' and 'data.frame':       16168 obs. of  15 variables:
 $ country           : chr  "Albania" "Albania" "Albania" "Albania" ...
 $ year              : num  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
 $ sex               : chr  "male" "male" "female" "male" ...
 $ age               : chr  "25-34 years" "55-74 years" "75+ years" "75+ years" ...
 $ suicides_no       : num  17 10 2 1 6 5 5 3 4 1 ...
 $ population        : num  232000 177400 37800 24900 263900 ...
 $ suicides/100k pop : num  7.33 5.64 5.29 4.02 2.27 2.08 1.51 1.22 1.07 0.27 ...
 $ country-year      : chr  "Albania2000" "Albania2000" "Albania2000" "Albania2000" ...
 $ HDI for year      : num  0.656 0.656 0.656 0.656 0.656 0.656 0.656 0.656 0.656 0.656 ...
 $ gdp_for_year ($)  : num  3.63e+09 3.63e+09 3.63e+09 3.63e+09 3.63e+09 ...
 $ gdp_per_capita ($): num  1299 1299 1299 1299 1299 ...
 $ generation        : chr  "Generation X" "Silent" "G.I. Generation" "G.I. Generation" ...
 $ more6             : num  1 0 0 0 0 0 0 0 0 0 ...
 $ predicted_p       : num  0.758 0.847 0.392 0.834 0.246 ...
 $ predicted_y       : num  1 1 0 1 0 1 0 0 1 0 ...
```

```
> summary(suicidi)
   country               year          sex                 age              suicides_no
 Length:27820       Min.   :1985   Length:27820       Length:27820       Min.   :    0.0
 Class :character   1st Qu.:1995   Class :character   Class :character   1st Qu.:    3.0
 Mode  :character   Median :2002   Mode  :character   Mode  :character   Median :   25.0
                    Mean   :2001                                         Mean   :  242.6
                    3rd Qu.:2008                                         3rd Qu.:  131.0
                    Max.   :2016                                         Max.   :22338.0

   population       suicides/100k pop country-year        HDI for year   gdp_for_year ($)
 Min.   :     278   Min.   :  0.00    Length:27820       Min.   :0.483   Min.   :4.692e+07
 1st Qu.:   97498   1st Qu.:  0.92    Class :character   1st Qu.:0.713   1st Qu.:8.985e+09
 Median :  430150   Median :  5.99    Mode  :character   Median :0.779   Median :4.811e+10
 Mean   : 1844794   Mean   : 12.82                       Mean   :0.777   Mean   :4.456e+11
 3rd Qu.: 1486143   3rd Qu.: 16.62                       3rd Qu.:0.855   3rd Qu.:2.602e+11
 Max.   :43805214   Max.   :224.97                       Max.   :0.944   Max.   :1.812e+13
                                                         NA's   :19456

 gdp_per_capita ($)  generation
 Min.   :   251     Length:27820
 1st Qu.:  3447     Class :character
 Median :  9372     Mode  :character
 Mean   : 16866
 3rd Qu.: 24874
 Max.   :126352
```

Missing values

Da un'attenta analisi si osserva la presenza di valori mancanti. Graficamente:




```
Variables sorted by number of missings:
            Variable    Count
       HDI for year 0.699353
            country 0.000000
               year 0.000000
                sex 0.000000
                age 0.000000
         suicides_no 0.000000
          population 0.000000
  suicides/100k pop 0.000000
       country-year 0.000000
    gdp_for_year ($) 0.000000
  gdp_per_capita ($) 0.000000
         generation 0.000000
```

Sono presenti valori mancanti solo per la variabile *HDI for year*, pari al 70% dei casi. Da un primo tentativo di eliminare tutte le osservazioni con valori mancanti, si otterrebbe un dataset con solo 8364 obs, perdendone più di 2/3. Di conseguenza, si decide di eliminare la variabile *HDI for year*.

Data preparation

Osservando il dataset notiamo che la variabile *country-year* è formata dall'unione di variabili già presenti nel dataset. Siccome, nel corso delle analisi, causerebbe problemi di collinearità decidiamo di rimuoverla.

In quanto la variabile *gdp_for_year* è definita su una scala diversa rispetto alle altre, si decide di riscriverla con un diverso ordine di grandezza:

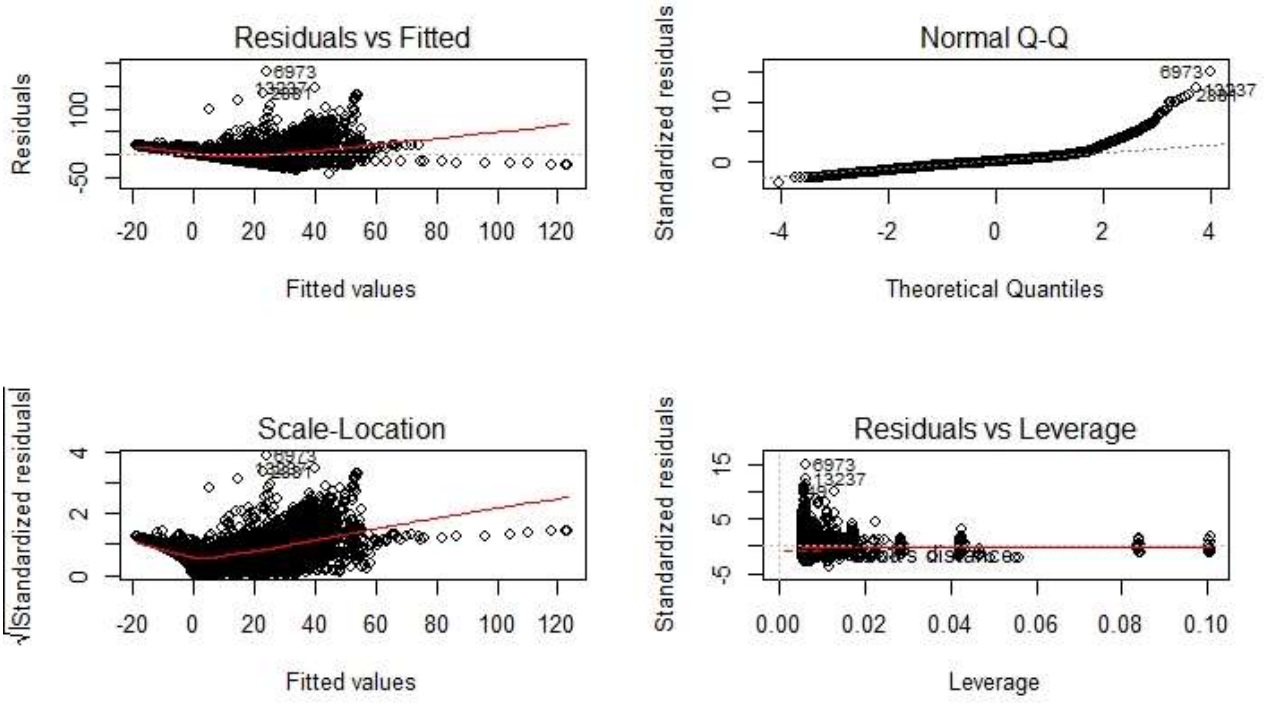
suicidi1$gdp_year = suicidi1$`gdp_for_year ($)`/ 10000000

Siccome il dataset è particolarmente ampio, decidiamo di eseguire l'analisi sulle osservazioni più recenti, eliminando quelle con anno precedente al 2000.

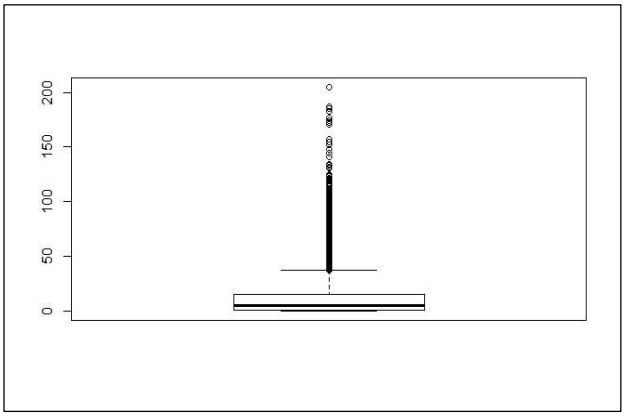
Il nostro dataset di partenza è dunque costituito da 16168 osservazioni e 10 variabili. Da questo fittiamo il nostro modello iniziale:

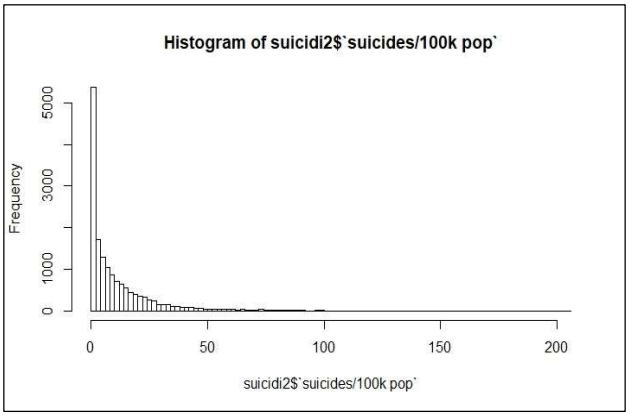
```
Residuals:
    Min      1Q  Median      3Q     Max
-45.179  -6.335  -1.102   4.426 180.880

Residual standard error: 12.02 on 16054 degrees of freedom
Multiple R-squared:  0.5588, Adjusted R-squared:  0.5556
F-statistic: 179.9 on 113 and 16054 DF,  p-value: < 2.2e-16
```

Diagnostiche del modello:



Vediamo anche come si presenta graficamente la variabile target:

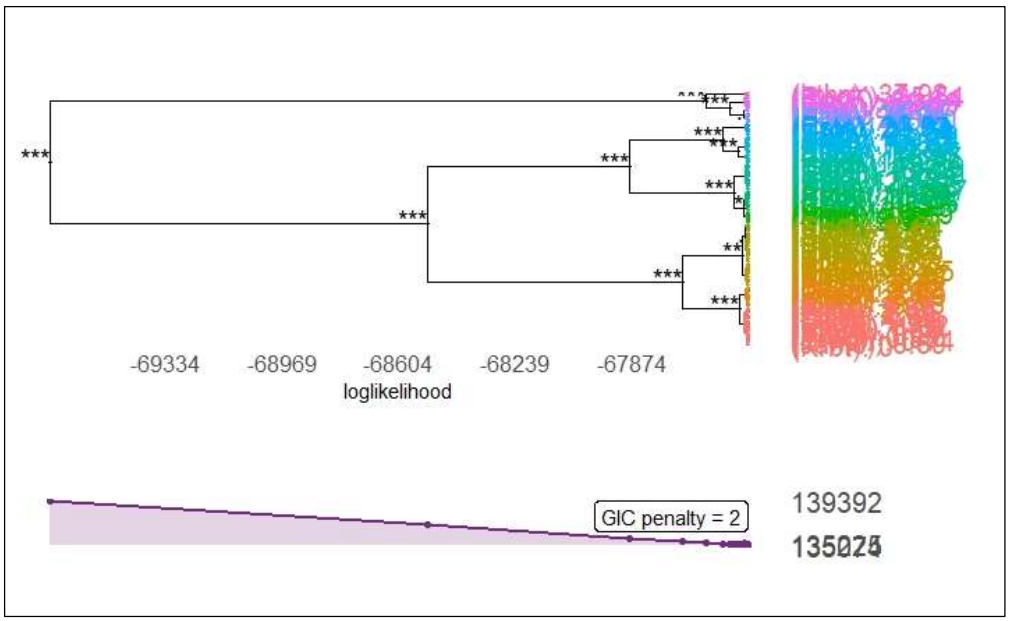




La variabile target presenta una distribuzione asimmetrica positiva con un'elevata presenza di outliers.

L'abbondanza dei parametri del modello è dovuta alla variabile *country* poiché è costituita da 101 livelli. Conviene quindi ridurli con l'optimal grouping.

Optimal grouping

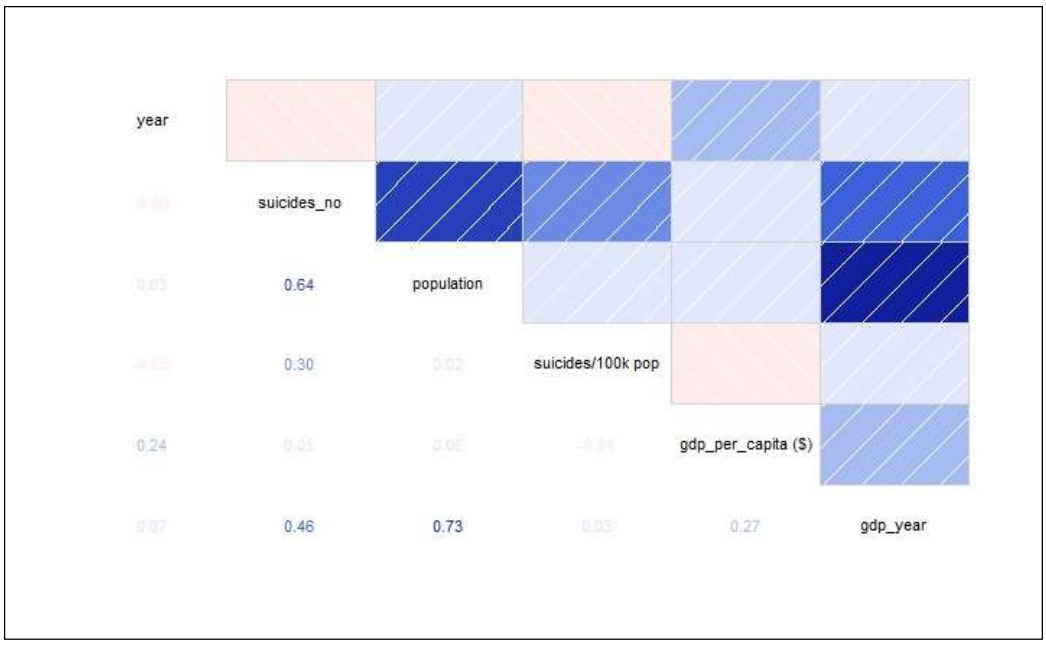


```
(Krbt)(Jamc)(Oman)(Brbd)(AnaB)(SthA)(Kuwt)(Bhms)(UnAE)(Mldv)(Grnd)(Qatr)(Bhrn)(Azrb)(Trky)
                                                                                      1940
                                                                  (Phlp)(Armn)(Gtml)(Albn)(SnMr)(Cypr)(Grec)(Gerg)
                                                                                      1160
                                                        (BsaH)(Prgy)(Fiji)(Mexc)(Panm)(Brzl)(SVaG)(Malt)(Clmb)
                                                                                      1488
                                    (Ncrg)(Ecdr)(UntK)(Trkm)(SntL)(Uzbk)(Thln)(CstR)(Itly)(Isrl)(Belz)(Sych)
                                                                                      2086
                                                                                (Arub)(PrtR)(Span)
                                                                                       528
                                                                           (Mrts)(ElSl)(Nthr)(Mntn)(Argn)
                                                                                       896
                                                                  (Irln)(Cand)(Slvk)(CbVr)(Nrwy)(Kyrg)(Prtg)(Astrl)
                                                                                      1236
                                                  (Romn)(Sngp)(Dnmr)(Icln)(Chil)(UntS)(Swdn)(NwZl)(TraT)(Grmn)(Lxmb)
                                                                                      2070
                                                                                (Blgr)(Plnd)(CzcR)
                                                                                       574
                                                                     (Cuba)(Swtz)(Frnc)(Mngl)(Fnln)(Blgm)(Austr)
                                                                                      1160
                                                                           (Japn)(Serb)(Urgy)(Crot)(Estn)
                                                                                       958
                                                                        (Ukrn)(Srnm)(Slvn)(Latv)
                                                                                       744
                                                                                      (Hngr)
                                                                                       202
                                                                           (Kzkh)(Blrs)(SrLn)(Guyn)(RssF)
                                                                                       732
                                                                                      (RpoK)
                                                                                       192
                                                                                      (Lthn)
                                                                                       202
```

Da 101 livelli si è passati a 16 livelli. Provando a sostituire la variabile *country* con la nuova, *optimal_group*, notiamo che il modello non perde la sua capacità esplicativa e che tutti i nuovi parametri risultano essere molto significativi.

Multicollinearità

VARIABILI QUANTITATIVE



Le due variabili più correlate sono *gdp_year* e *population* con un valore pari a 0.73. Una buona collinearità vi è anche tra *suicides_no* e *population* con 0.64. Tutte le altre variabili sono poco correlate con coefficienti inferiori a 0.5. Ci sono anche casi di incorrelazione (es. *year* e *suicides_no*). I valori calcolati di TOL e VIF non superano le soglie critiche che implicherebbero la rimozione delle variabili.

VARIABILI QUALITATIVE

Esaminando il grado di connessione tra tutte le possibili coppie di variabili, *age* e *generation* risultano essere le più connesse con un chi-quadro normalizzato pari a 0.45. Si decide comunque di tenerle entrambe in quanto la soglia critica (0.8) non viene superata.

Linearità

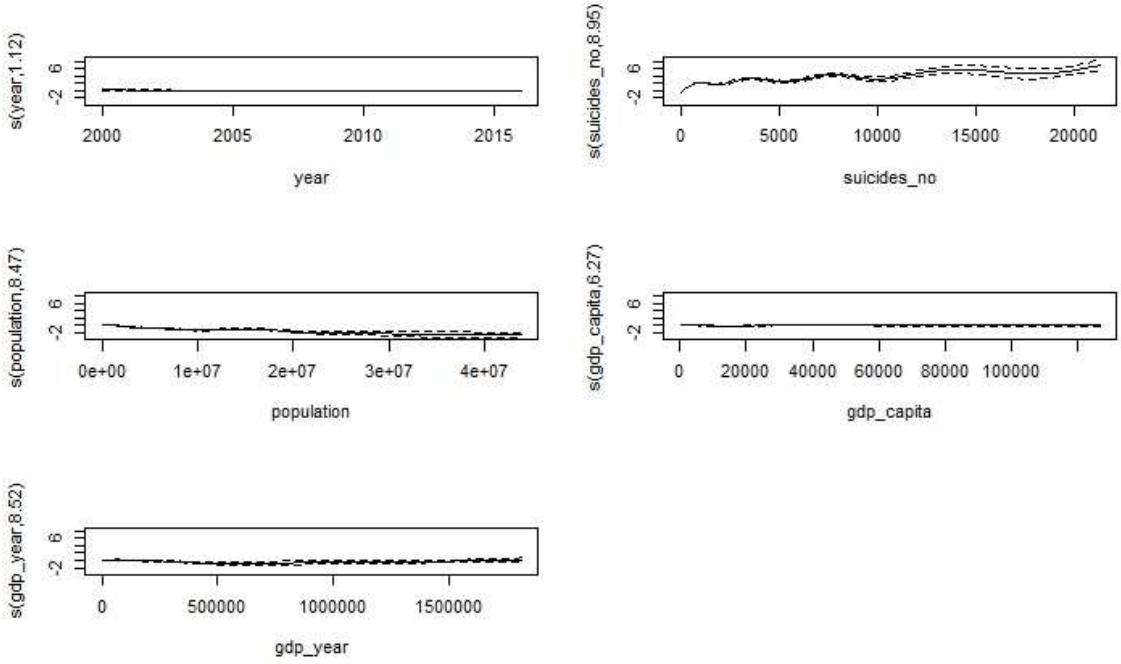
VARIABILE TARGET

Usiamo Box-Cox per vedere se modificando la variabile target il modello potrebbe migliorare. Il valore di Lambda restituito è pari a 0.2626, che approssimiamo a 0.5. Questo ci porta a modificare la variabile target facendone la radice quadrata. Così facendo, l'R quadro del nostro modello aumenta del 13%. In questo modo il range dei residui viene ridotto notevolmente. Rinominiamo la nuova variabile target col nome *tasso*.

COVARIATE

Usando la funzione *gam* vediamo che modificando le covariate la capacità esplicativa del modello migliorerebbe, il che viene confermato dal test *anova*.

Dall'analisi grafica studiamo come trasformare al meglio le covariate:



Le variabili *year, gdp_capita* e *gdp_year* hanno un andamento lineare costante, quindi non necessitano di alcuna modifica. *Population* ha un andamento lineare decrescente, quindi potrebbe essere approssimata con una proporzionalità inversa. La variabile riguardante il numero di suicidi è quella caratterizzata da un andamento più incerto; la maggior parte delle osservazioni però (fino al terzo quartile) sono concentrate nella parte iniziale il cui andamento potrebbe essere approssimato a quello di una radice quadrata. Dopo alcune prove il modello migliore risulta essere

```
tra1 <- lm(tasso ~ year+sex+age+I(sqrt(suicides_no))+population+gdp_capita+generation+gdp_year
+optimal_group, data=suicidi2)
```

L'R quadro così migliora di circa il 3%. Diminuiscono notevolmente i valori di leverage.

Non tutti i parametri però sono significativi. Procediamo quindi con la model selection.

Model selection

AIC

```
tasso ~ year + sex + age + I(sqrt(suicides_no)) +
population +
    generation + gdp_year + optimal_group

                        Df Sum of Sq   RSS     AIC
<none>                                21902  4969.5
+ gdp_capita             1       1.7 21900  4970.2
- generation             5      20.0 21922  4974.2
- gdp_year               1      16.6 21918  4979.7
- year                   1      47.9 21950  5002.8
- population             1     968.3 22870  5667.0
- age                    5    2039.8 23942  6399.3
- I(sqrt(suicides_no))   1    3743.4 25645  7518.6
- sex                    1    7234.5 29136  9582.1
- optimal_group         15   11202.5 33104 11618.4
```

Vediamo che AIC ci suggerisce di rimuovere la variabile *gdp_capita*. E' evidente il basso livello di significatività di *generation*. La variabile più significativa, caratterizzata dall'F-value più elevato, è il sesso.

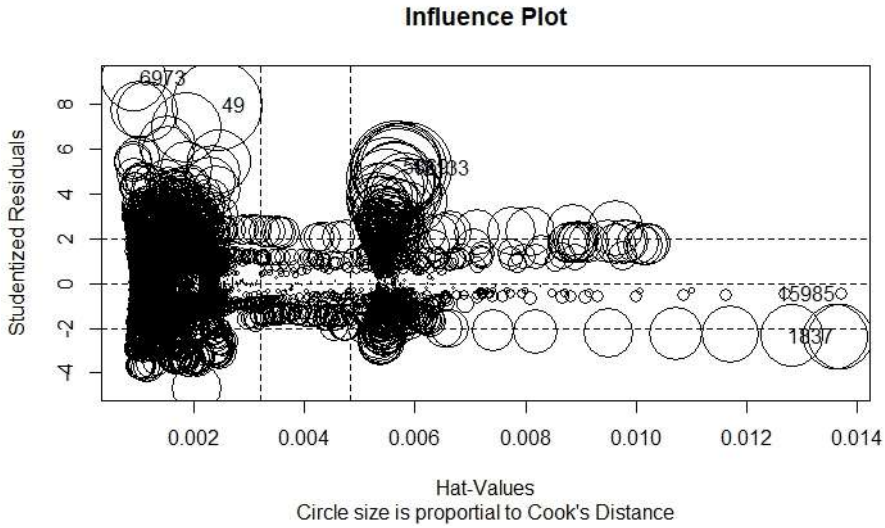
SBC

In questo caso viene proposta l'eliminazione anche della variabile *generation* (infatti SBC è più severo di AIC). Tutte le altre covariate risultano essere pienamente significative.

```
tasso ~ year + sex + age + I(sqrt(suicides_no)) +
population +
    gdp_year + optimal_group

                        Df Sum of Sq   RSS     AIC
<none>                                21922  5174.2
- gdp_year               1      16.2 21938  5176.4
+ gdp_capita             1       1.8 21920  5182.6
+ generation             5      20.0 21902  5207.9
- year                   1      73.6 21995  5218.7
- population             1     972.5 22894  5866.3
- I(sqrt(suicides_no))   1    3746.5 25668  7715.4
- sex                    1    7233.6 29155  9774.9
- age                    5    8558.4 30480 10454.7
- optimal_group         15   11200.7 33122 11701.9
```

Decidiamo di usare come selettore il metodo SBC, ristimando il modello senza le variabili non significative.

Punti influenti

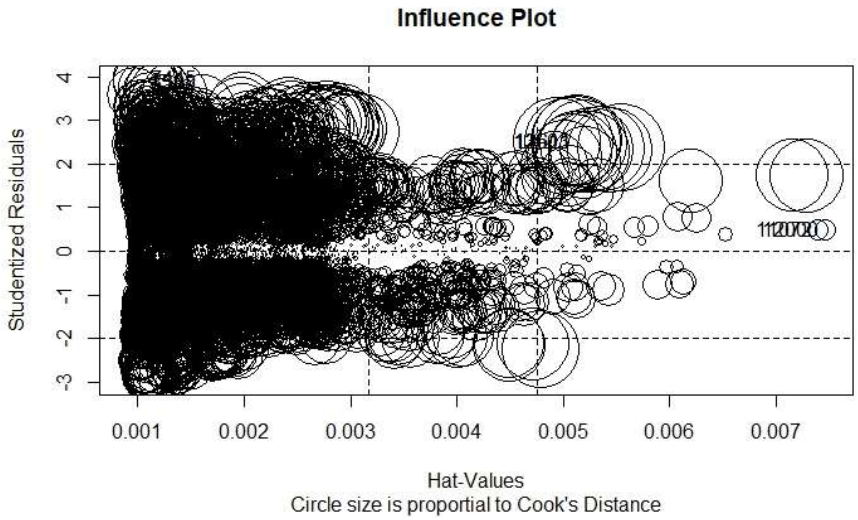


Vediamo una consistente presenza di punti influenti: principalmente si tratta di maschi sopra i 75 anni, non concentrati in una particolare area geografica.

Utilizzando un cutoff pari a 0.000248, si rilevano 1063 osservazioni con distanza di Cook superiore e che quindi risultano essere punti influenti. Trattandosi di circa il 5% delle osservazioni totali si procede alla loro eliminazione.

Rifittando il modello, utilizzando il dataset senza valori influenti, il modello raggiunge un R quadro pari a 0.7866 e il range dei residui studentizzati si restringe.

Per migliorare ulteriormente il dataset procediamo con l'eliminazione dei punti di leva basandoci sulla soglia di accettabilità, in questo caso pari a 0.00344. Ciò porta all'eliminazione di 597 osservazioni.



Dopo le correzioni, il modello di riferimento è:

```
> summary(noninflu2)

Call:
lm(formula = tasso ~ year + sex + age + I(sqrt(suicides_no)) +
    population + gdp_year + optimal_group, data = NOinflu2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6182 -0.5337 -0.0559  0.4902  3.4552

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.489e+01  3.216e+00   4.631 3.67e-06 ***
year                 -7.414e-03  1.602e-03  -4.627 3.74e-06 ***
sexmale               1.173e+00  1.633e-02  71.830  < 2e-16 ***
age25-34 years        2.083e-01  2.469e-02   8.435  < 2e-16 ***
age35-54 years        3.035e-01  2.542e-02  11.937  < 2e-16 ***
age5-14 years        -1.260e+00  2.611e-02 -48.244  < 2e-16 ***
age55-74 years        3.874e-01  2.508e-02  15.445  < 2e-16 ***
age75+ years          7.250e-01  2.594e-02  27.944  < 2e-16 ***
I(sqrt(suicides_no))  1.107e-01  1.603e-03  69.057  < 2e-16 ***
population           -1.958e-07  5.230e-09 -37.432  < 2e-16 ***
gdp_year             -2.094e-06  1.361e-07 -15.391  < 2e-16 ***
optimal_group10       2.001e+00  3.557e-02  56.262  < 2e-16 ***
optimal_group11       2.257e+00  3.824e-02  59.026  < 2e-16 ***
optimal_group12       2.574e+00  4.163e-02  61.822  < 2e-16 ***
optimal_group14       2.549e+00  4.564e-02  55.862  < 2e-16 ***
optimal_group2        6.722e-01  3.280e-02  20.493  < 2e-16 ***
optimal_group3        1.047e+00  3.171e-02  33.031  < 2e-16 ***
optimal_group4        1.124e+00  2.845e-02  39.498  < 2e-16 ***
optimal_group5        1.185e+00  4.550e-02  26.034  < 2e-16 ***
optimal_group6        1.470e+00  3.701e-02  39.723  < 2e-16 ***
optimal_group7        1.657e+00  3.334e-02  49.694  < 2e-16 ***
optimal_group8        1.852e+00  2.974e-02  62.278  < 2e-16 ***
optimal_group9        1.551e+00  4.464e-02  34.749  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8721 on 14485 degrees of freedom
Multiple R-squared:  0.7807, Adjusted R-squared:  0.7804
F-statistic:  2344 on 22 and 14485 DF,  p-value: < 2.2e-16
```

Eteroschedasticità

```
> bptest(noninflu2)

        studentized Breusch-Pagan test

data:  noninflu2
BP = 2017.7, df = 22, p-value < 2.2e-16
```

```
> ncvTest(noninflu2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 8.831021, Df = 1, p = 0.0029615
```

Entrambi i test ci suggeriscono di rifiutare l’ipotesi nulla di omoschedasticità.

Calcoliamo allora gli standard error corretti di White:

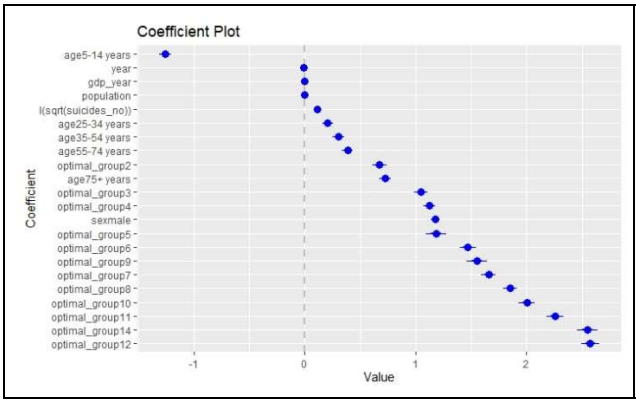
White (1980) Heteroscedasticity-corrected SEs and Tests

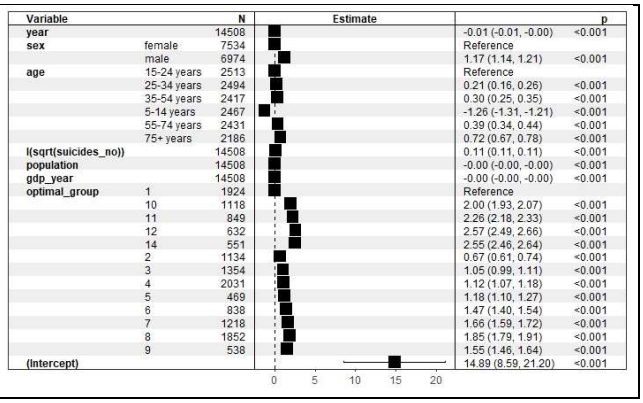
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.49e+01	3.19e+00	4.67	3.09e-06
year	-7.41e-03	1.59e-03	-4.66	3.16e-06
sexmale	1.17e+00	1.82e-02	64.34	0.00e+00
age25-34 years	2.08e-01	2.28e-02	9.14	6.96e-20
age35-54 years	3.04e-01	2.29e-02	13.28	5.26e-40
age5-14 years	-1.26e+00	2.62e-02	-48.13	0.00e+00
age55-74 years	3.87e-01	2.35e-02	16.46	2.37e-60
age75+ years	7.25e-01	2.96e-02	24.50	6.49e-130
I(sqrt(suicides_no))	1.11e-01	1.90e-03	58.15	0.00e+00
population	-1.96e-07	6.37e-09	-30.72	7.15e-201
gdp_year	-2.09e-06	1.19e-07	-17.58	1.72e-68
optimal_group10	2.00e+00	3.55e-02	56.37	0.00e+00
optimal_group11	2.26e+00	4.26e-02	52.95	0.00e+00
optimal_group12	2.57e+00	5.38e-02	47.87	0.00e+00
optimal_group14	2.55e+00	5.20e-02	49.07	0.00e+00
optimal_group2	6.72e-01	3.03e-02	22.19	2.46e-107
optimal_group3	1.05e+00	3.51e-02	29.80	1.87e-189
optimal_group4	1.12e+00	3.17e-02	35.47	2.22e-264
optimal_group5	1.18e+00	4.35e-02	27.26	1.46e-159
optimal_group6	1.47e+00	3.69e-02	39.89	0.00e+00
optimal_group7	1.66e+00	3.22e-02	51.47	0.00e+00
optimal_group8	1.85e+00	3.32e-02	55.80	0.00e+00
optimal_group9	1.55e+00	3.61e-02	42.93	0.00e+00

Anche se lievi, si osservano comunque dei cambiamenti negli standard error (il che ci fa ipotizzare una bassa eteroschedasticità).

Nell'eventualità in cui dovremmo fare delle previsioni, sarà utile fare affidamento a questi standard error, per svolgere un'inferenza corretta.

Interpretazione coefficienti del modello





L'unico parametro che ha coefficiente negativo è *age5-14 years* (ha senso in quanto è difficile che un ragazzo in questa fascia di età si suicidi).

Year, *gdp_year* e *population*, hanno coefficienti prossimi allo zero, quindi esercitano poca influenza sul tasso.

I parametri dell'optimal grouping, *sexmale* e *age75+ years* sono caratterizzate dai coefficienti con valore più elevato.

Tutte le variabili considerate nel modello finale risultano essere molto significative.

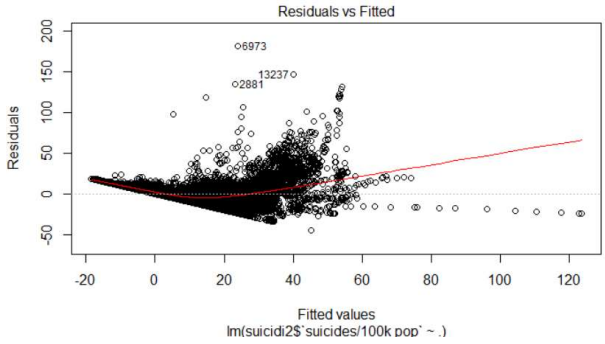
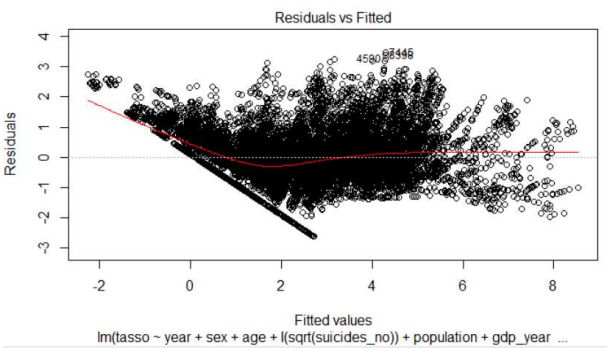
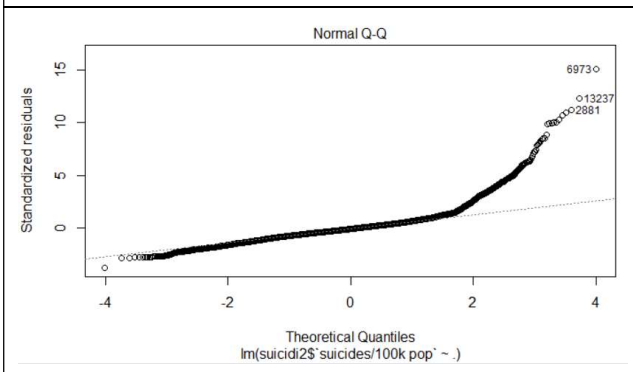
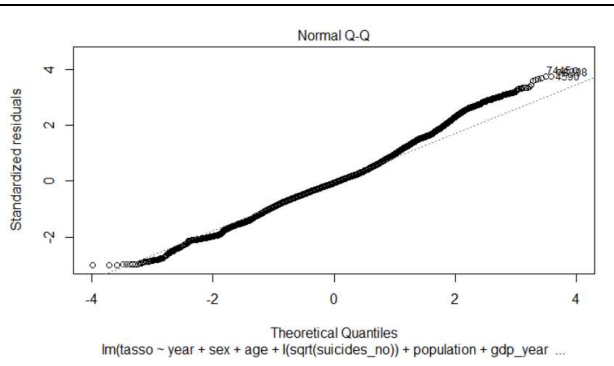
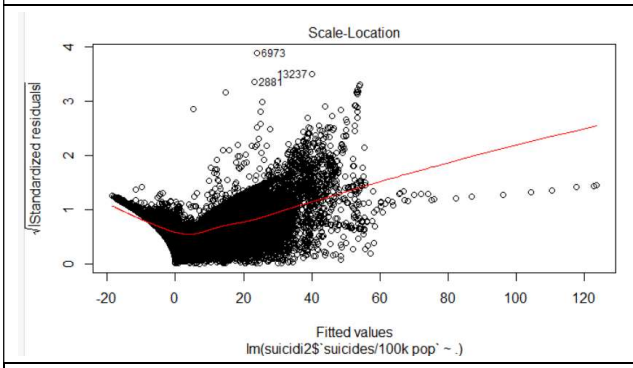
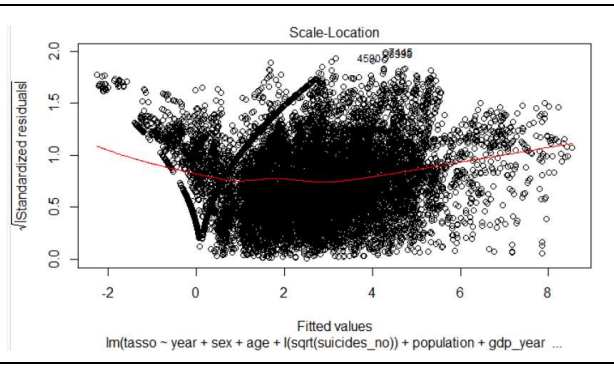
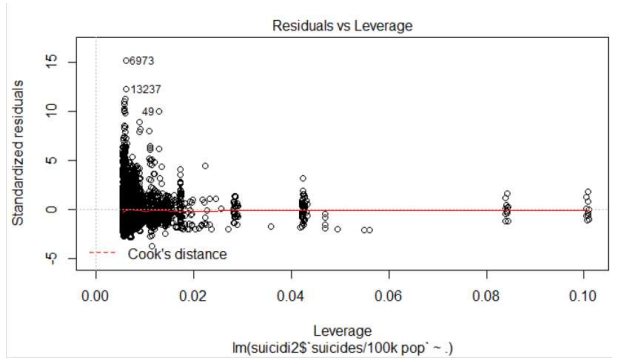
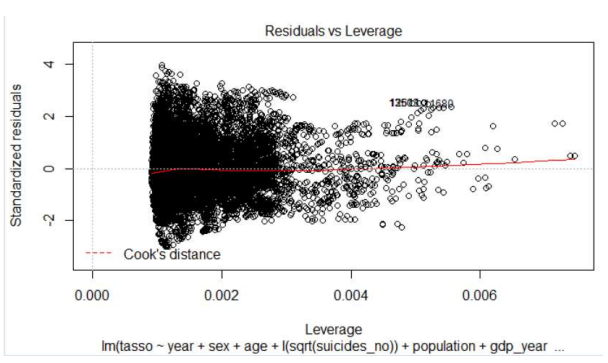
```
> drop1(noninflu2, .~., test="F")
Single term deletions

Model:
tasso ~ year + sex + age + I(sqrt(suicides_no)) + population +
    gdp_year + optimal_group
                     Df Sum of Sq   RSS     AIC  F value    Pr(>F)
<none>                             11017 -3946.9
year                  1      16.3 11034 -3927.5   21.412 3.737e-06 ***
sex                   1    3924.4 14942   471.5 5159.602 < 2.2e-16 ***
age                   5    4418.3 15436   935.3 1161.789 < 2.2e-16 ***
I(sqrt(suicides_no))  1    3627.3 14645   180.1 4768.880 < 2.2e-16 ***
population            1    1065.7 12083 -2609.3 1401.167 < 2.2e-16 ***
gdp_year              1     180.2 11198 -3713.6  236.870 < 2.2e-16 ***
optimal_group        12    5897.4 16915  2248.9  646.128 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confronto tra modello iniziale e modello finale

MODELLO	R QUADRO
full=lm(suicidi2$`suicides/100k pop` ~ .,data=suicidi2)	0.5588
noninflu2 <- lm(tasso ~ year+sex+age+I(sqrt(suicides_no))+population +gdp_year+optimal_group, data=NOinflu2)	0.7807

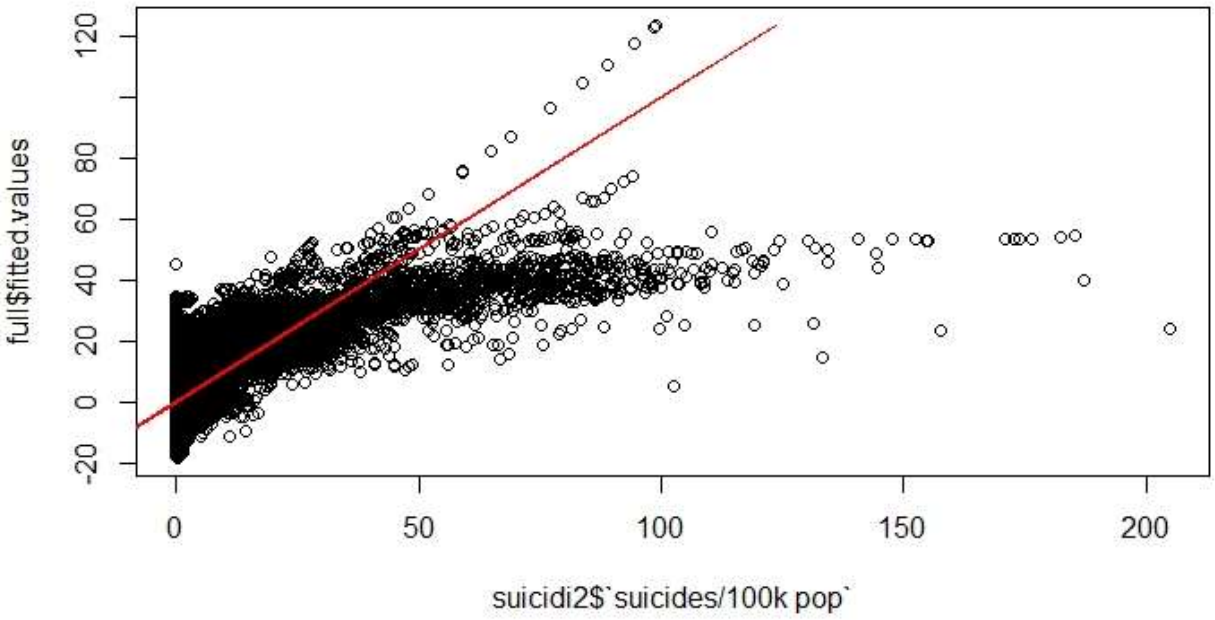
Diminuisce notevolmente il numero dei parametri e risultano tutti significativi. Cambia anche il range dei residui che diminuisce sostanzialmente (da [-45.179;180.880] a [-2.618;3.4552]) assumendo una distribuzione più simmetrica.

Le diagnostiche dei due modelli a confronto:

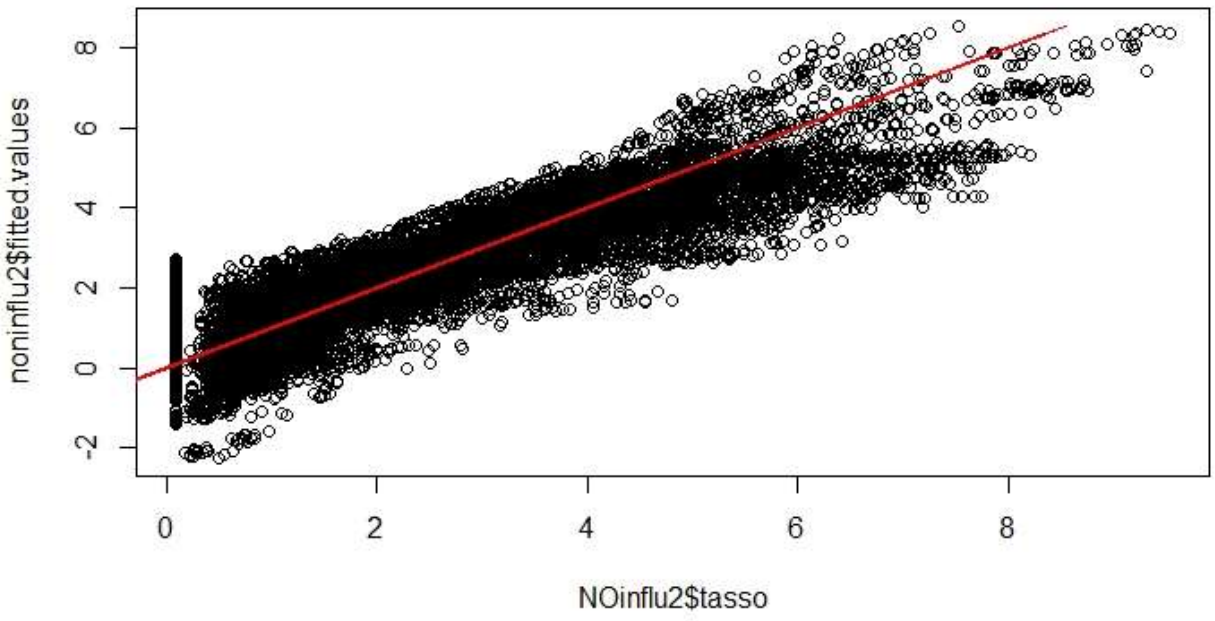
MODELLO INIZIALE *full*	MODELLO FINALE *noninflu2*
	
	
	
	

Grafici valori osservati vs valori fittati a confronto:

MODELLO INIZIALE *full*



MODELLO FINALE *noninflu2*



REGRESSIONE LOGISTICA

Anche in questo caso consideriamo il dataset con le sole osservazioni dall'anno 2000 al 2016.

Il tasso di suicidi in Italia nel 2019 è di 6/100k. Decidiamo quindi di prendere questa soglia per stabilire quali osservazioni hanno un tasso maggiore o minore di quello attuale italiano. Il valore scelto si trova leggermente sopra il valore mediano della variabile target.

In base a questa soglia, la nostra popolazione è così distribuita:

```
> table(suicidi$more6)

   0    1
8379 7789
> prop.table(table(suicidi$more6))

        0         1
0.5182459 0.4817541
```

Si hanno 8379 obs, pari al 51,82% del totale, sotto la soglia stabilita.

Fittiamo il modello solo con alcune variabili di interesse, quelle direttamente riferite alle caratteristiche del singolo individuo: *year, sex, age*.

```
> summary(fitmore6)

Call:
glm(formula = more6 ~ year + sex + age, family = binomial, data = suicidi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9637  -0.8253  -0.0515   0.7409   3.6469

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    14.339812   8.612223   1.665 0.095902 .
year           -0.007702   0.004290  -1.795 0.072630 .
sexmale         2.054528   0.040523  50.701  < 2e-16 ***
age25-34 years  0.208049   0.061836   3.365 0.000767 ***
age35-54 years  0.772187   0.062823  12.291  < 2e-16 ***
age5-14 years  -5.469446   0.255153 -21.436  < 2e-16 ***
age55-74 years  0.780144   0.062847  12.413  < 2e-16 ***
age75+ years    0.683207   0.062579  10.917  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22392  on 16167  degrees of freedom
Residual deviance: 15403  on 16160  degrees of freedom
AIC: 15419
```

Dalle due devianze in output, si calcola l'R quadro del modello, che è pari a 0.312.

Guardando il seguente forestmodel, si nota che tutti i coefficienti sono positivi, tranne *age5-14 years; year* ha un valore prossimo allo 0 e *sexmale* è il parametro con coefficiente più elevato.

Variable		N	Odds ratio	p
year		16168	0.99 (0.98, 1.00)	0.07
sex	female	8084	Reference	
	male	8084	7.80 (7.21, 8.45)	<0.001
age	15-24 years	2700	Reference	
	25-34 years	2700	1.23 (1.09, 1.39)	<0.001
	35-54 years	2700	2.16 (1.91, 2.45)	<0.001
	5-14 years	2668	0.00 (0.00, 0.01)	<0.001
	55-74 years	2700	2.18 (1.93, 2.47)	<0.001
	75+ years	2700	1.98 (1.75, 2.24)	<0.001
(Intercept)			1689277.62 (0.08, 36329976812637.18)	

0.01 1 100 1e+06 1e+10

Vediamo gli Odds Ratio dei parametri ed i loro intervallo di confidenza:

Tutti gli Odds Ratio ricadono negli intervalli di confidenza trovati.
L'intercetta e *year* risultano essere i parametri meno significativi.
I parametri di riferimento sono *sexfemale* e *age15-24 years*.

```
                          OR 2.5 %       97.5 %
(Intercept)     1689277.62  0.08 3.632998e+13
year                  0.99  0.98 1.000000e+00
sexmale               7.80  7.21 8.450000e+00
age25-34 years        1.23  1.09 1.390000e+00
age35-54 years        2.16  1.91 2.450000e+00
age5-14 years         0.00  0.00 1.000000e-02
age55-74 years        2.18  1.93 2.470000e+00
age75+ years          1.98  1.75 2.240000e+00
```

Da qui capiamo che i maschi hanno un'attitudine al suicidio quasi 8 volte superiore a rispetto alle donne; oppure che gli anziani hanno un rischio di sui cidarsi quasi due volte rispetto a quello dei giovani.

Eseguiamo ora il test LRT:

```
> drop1(fitmore6, test="LRT")
Single term deletions

Model:
more6 ~ year + sex + age
       Df Deviance   AIC    LRT Pr(>Chi)
<none>      15403 15419
year    1    15406 15420    3.2  0.07259 .
sex     1    18386 18400 2982.5  < 2e-16 ***
age     5    20030 20036 4626.3  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La variabile *year* è la meno significativa, infatti è caratterizzata da un LRT particolarmente basso. Decidiamo quindi di eliminarla dal modello. La sua rimozione non altera l'R quadro. L'unico coefficiente che muta è l'intercetta che raggiunge un valore pari a 0.3.

Riportiamo infine le statistiche del nostro modello:

```
> round(table(observed=suicidi$more6,predicted=suicidi$predicted_y)/nrow(suicidi),2)
        predicted
observed    0    1
       0 0.44 0.08
       1 0.15 0.33
```

Il nostro modello ha un'accuratezza del 77%, ovvero assegna un valore veritiero nel 77% dei casi. Nel 15% dei casi attribuisce un tasso di suicidi inferiore alla soglia quando in realtà non lo è. Nel restante 8% dei casi, il modello assegna un valore superiore, quando in realtà non lo è.