UNIVERSITÀ DELLA CALABRIA

BDP I Bimbi Di Pearson

# Progetto Modelli Statistici e Statistical Learning

## INTRODUZIONE

Il repository di dati del Global Health Observatory (GHO) sotto l'Organizzazione mondiale della sanità (OMS) tiene traccia dello stato di salute e di molti altri fattori correlati per tutti i paesi. Il set di dati utilizzato, relativo all'aspettativa di vita e ai fattori di salute per 193 paesi, è stato raccolto dal sito Web del repository di dati dell'OMS, mentre i relativi dati economici sono stati raccolti dal sito Web delle Nazioni Unite (ONU). Tra tutte le categorie di fattori relativi alla salute sono stati scelti solo i fattori critici più rappresentativi.

È stato osservato che negli ultimi 15 anni c'è stato un enorme sviluppo nel settore sanitario con un conseguente miglioramento dei tassi di mortalità soprattutto nelle nazioni in via di sviluppo rispetto agli ultimi 30 anni.

L'obiettivo è identificare le caratteristiche significative il cui trattamento consente di migliorare in modo efficiente l'aspettativa di vita della popolazione di un Paese. Ci si pone quindi la domanda, se un'organizzazione sanitaria volesse migliorare l'aspettativa di vita della popolazione in uno specifico Paese, quali caratteristiche dovrebbe cambiare per raggiungere l'obiettivo?

Essendo la variabile dipendente continua, si procede eseguendo una regressione lineare multipla.

Il dataset è composto da 1649 osservazioni su 20 variabili, descritte di seguito:

| | |
|---|---|
| Status | Variabile binaria con due categorie: 'Developing' e 'Developed'. Fa riferimento allo stato del Paese preso in considerazione. Viene successivamente ricodificata: vale 1 se il Paese è sviluppato, 0 se il Paese è in via di sviluppo. |
| life.expectancy | Aspettativa di vita del soggetto preso in considerazione (in anni). Sarà la nostra variabile dipendente. |
| adult.mortality | Tasso di mortalità per entrambi i sessi, misura la probabilità (per mille) di morire tra i 15 e i 60 anni. |
| infant.deaths | Numero di morti infantili per 1000 abitanti. |
| Alcohol | Consumo di alcol registrato pro-capite tra gli abitanti aventi più di 15 anni (in litri di alcol puro). |
| percentage.expenditure | Percentuale del Prodotto Interno Lordo destinato a spese sanitarie (pro-capite). |
| hepatitisB | Percentuale dei bambini di un anno coperti dal vaccino contro l'epatite B. |
| Measles | Numero di casi di morbillo segnalati per 1000 abitanti. |

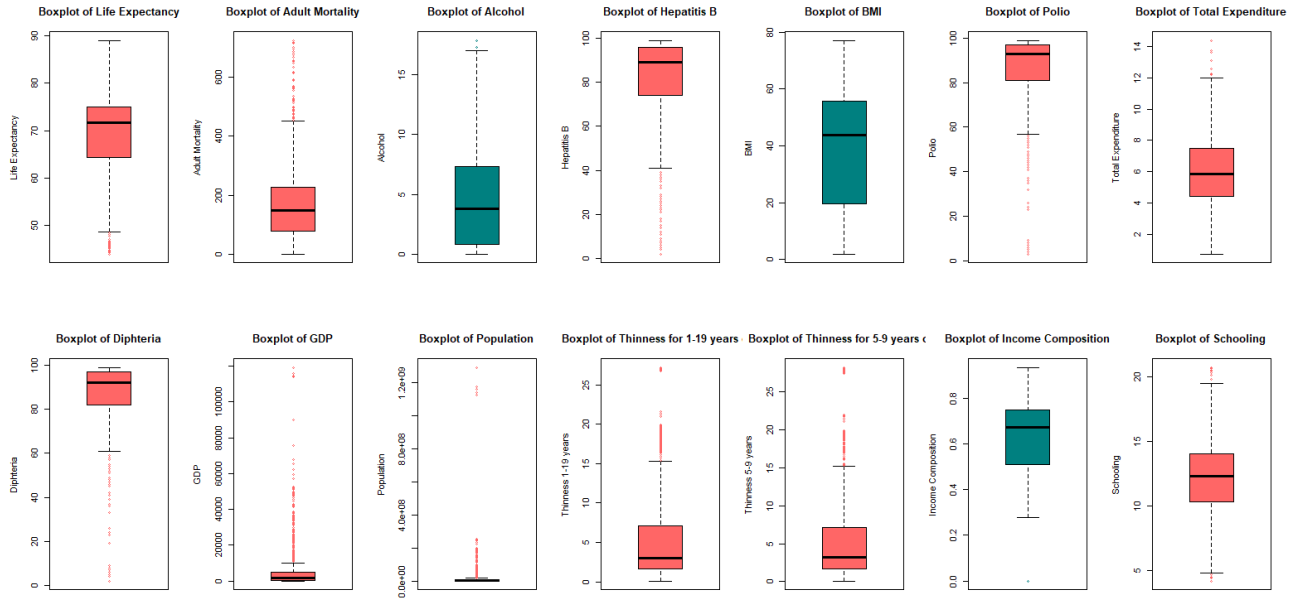
**Componenti**: Bessioud Ghofran, Ferrari Gianluca, Gallo Debora, Iannicelli Stefano, Lavecchia Nicola, Pascuzzi Eugenio, Pascuzzi Michela, Sorrenti Alessandro

*Anno Accademico 2022/2023*

| | |
|---|---|
| BMI | Indice di massa corporea media dell'intera popolazione. |
| under_five.deaths | Numero di decessi di abitanti aventi età inferiore a 5 anni per 1000 abitanti. |
| Polio | Percentuale dei bambini di un anno coperti dal vaccino contro il Polio. |
| total.expenditure | Percentuale della Spesa pubblica totale destinata alla sanità. |
| Diphtheria | Percentuale dei bambini di un anno coperti dal vaccino contro la difterite, il tossoide tetanico e la pertosse. |
| HIV/AIDS | Numero di morti per 1000 bambini tra 0 e 4 anni nati con HIV/AIDS. |
| GDP | Prodotto Interno Lordo pro-capite (USD). |
| Population | Numero di abitanti. |
| thinness1_19 | Percentuale di prevalenza di magrezza tra bambini e adolescenti (età compresa tra i 10 e i 19 anni). |
| thinness5_9 | Percentuale di prevalenza di magrezza tra bambini aventi età tra i 5 e i 9 anni. |
| income.composition.resources | Indice di Sviluppo Umano in termini di composizione reddituale delle risorse. La variabile assume valori compresi tra 0 e 1. |
| Schooling | Numero di anni di scolarizzazione. |

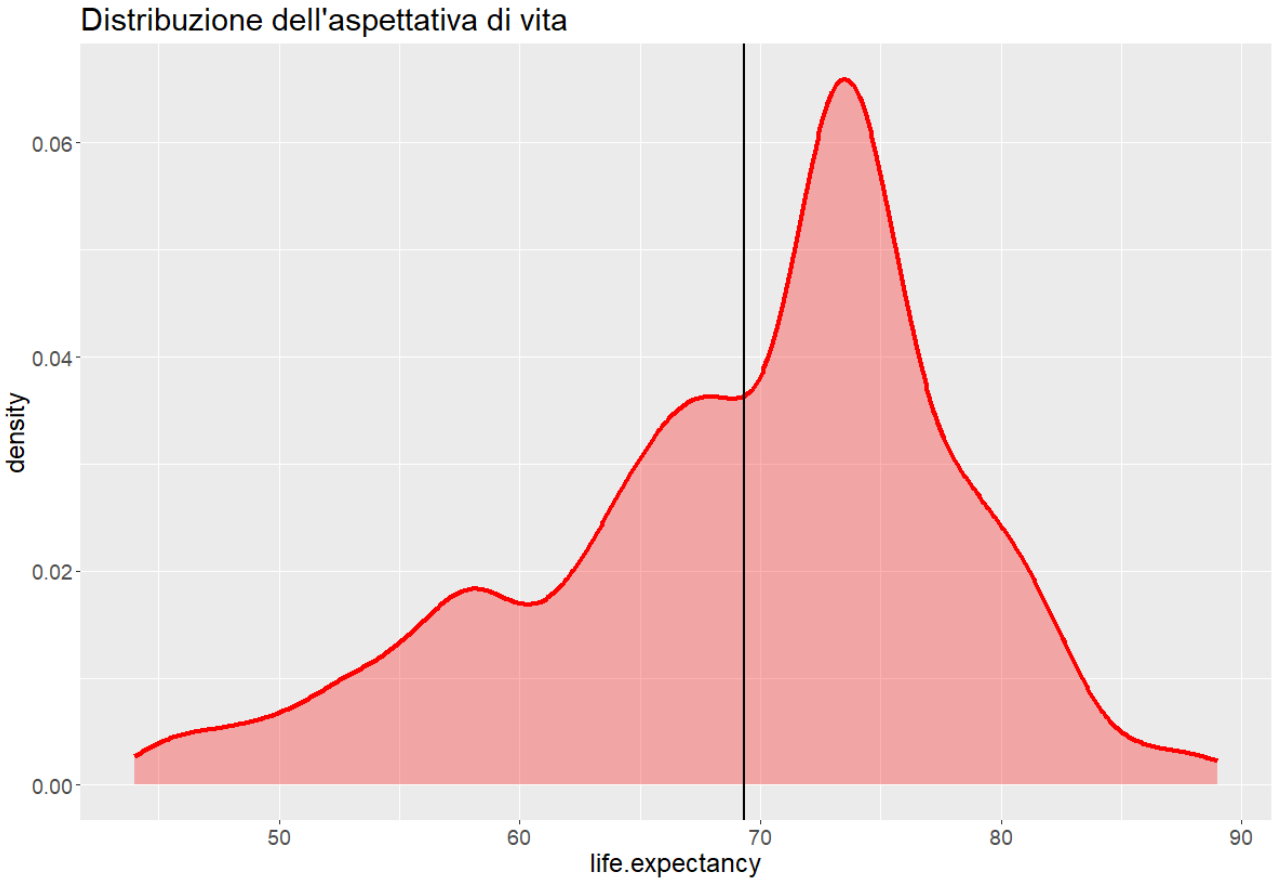
# ANALISI DESCRITTIVE

- Tramite i boxplot applicati a ciascuna variabile, si può procede ad una prima visualizzazione grafica che consente di verificare la presenza o meno di valori anomali.




Si può subito notare l'assenza (o quasi) di valori anomali per le variabili Alcohol, BMI e income.composition.resources, mentre per le altre variabili la presenza di valori anomali sembra essere più rilevante. Indice di sviluppo umano

- Il grafico riportato di seguito consente di visualizzare la distribuzione dell'aspettativa di vita. Viene inoltre aggiunta un'asse sul valore della media della variabile, per valutare la presenza o meno di asimmetria.




Già dalla visualizzazione grafica si nota che la distribuzione risulta essere solo leggermente inclinata verso sinistra. A conferma viene utilizzata la funzione Skewness, che misura, appunto, l'asimmetria di una distribuzione.

```
sprintf("Skewness: [%s]", toString(skewness(life.expectancy, na.rm = TRUE)))
```

```
## [1] "Skewness: [-0.628185990981033]"
```

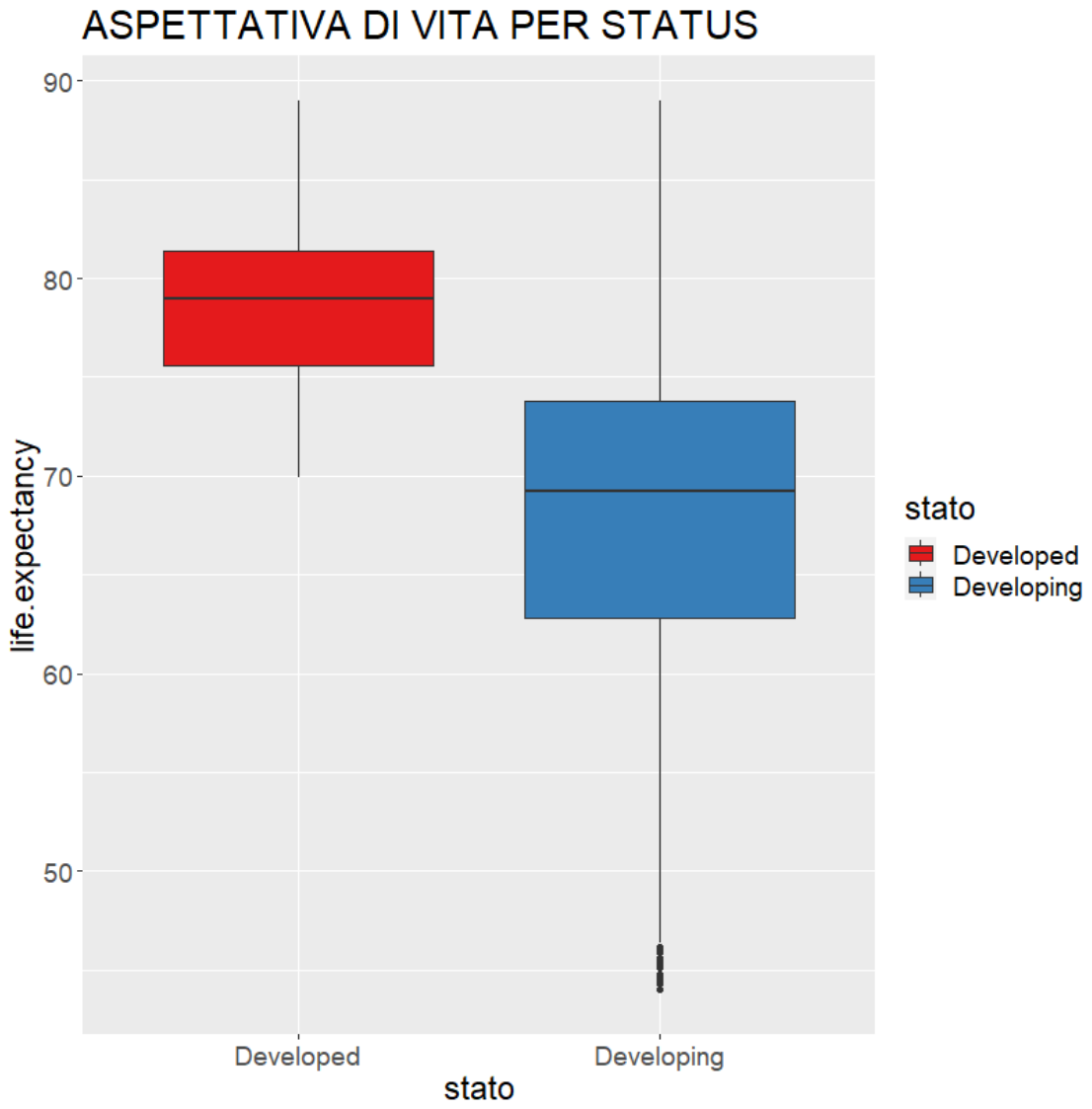
Come previsto, il risultato della funzione conferma la leggera asimmetria della distribuzione.

Inoltre, si utilizza anche l'indice di Kurtosis, il quale presenta un valore maggiore di 0, perciò la distribuzione ha forma leptocurtica:

```
sprintf("Indice Kurtosis: [%s]", toString(kurtosis(life.expectancy, na.rm = TRU
E)))
```

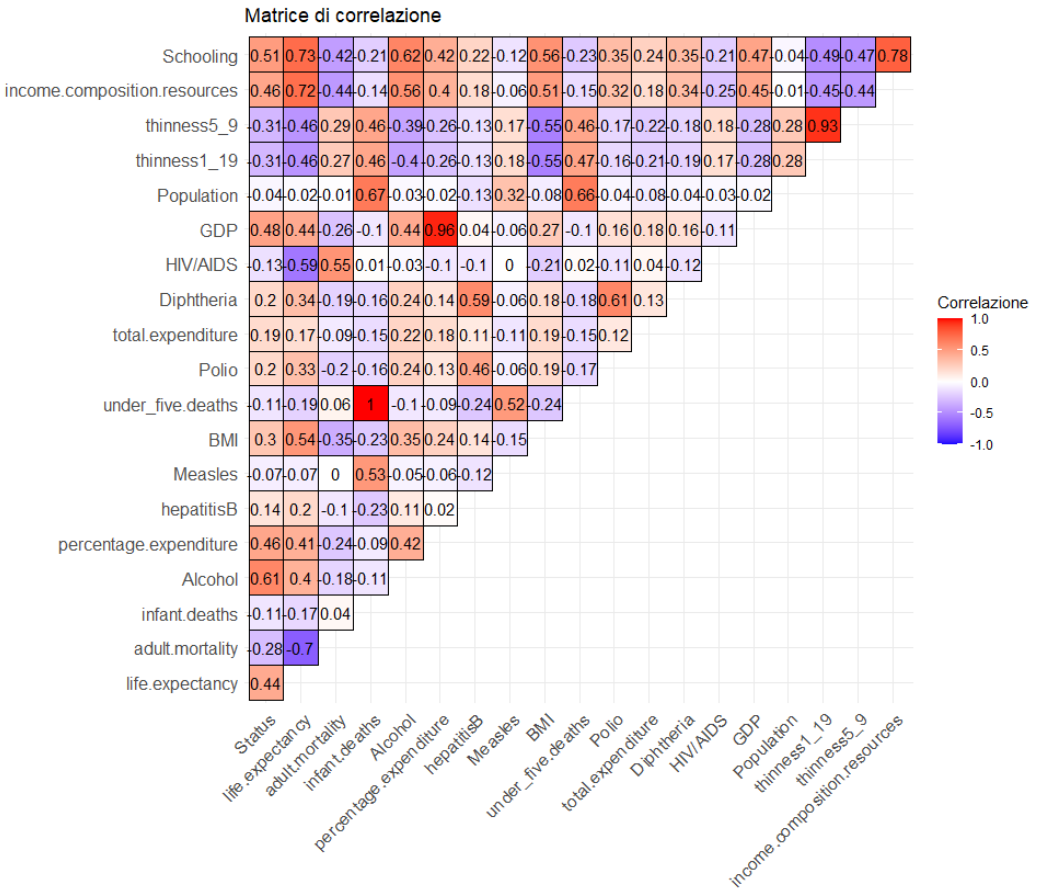
```
## [1] "Indice Kurtosis: [3.03656198417518]"
```

- Successivamente, è stato controllato il boxplot dell'aspettativa di vita (variabile dipendente) nelle due categorie di 'Status', che fanno riferimento allo stato del Paese considerato: 'Sviluppato' e 'In via di sviluppo' ('Developed' e 'Developing'). Questa procedura consente di avere un'idea circa la loro relazione.




Il risultato rappresenta esattamente ciò che ci aspettavamo: l'aspettativa di vita nei Paesi sviluppati è più alta dell'aspettativa di vita nei Paesi in via di sviluppo. Questa potrebbe essere una conferma del fatto che la variabile 'Status' sia un buon predittore per il modello.

- Dalla descrizione delle variabili, si possono notare alcune relazioni ambigue. Ad esempio, prendendo in considerazione lo specifico caso delle variabili 'infant deaths' e 'under-five deaths', esaminando la descrizione di entrambe, risulta evidente che una delle due variabili 'comprende' l'altra.
  Per valutare le relazioni esistenti (o meno) tra i regressori e la variabile dipendente, e nello specifico per avere un'idea circa la presenza o assenza di multicollinearità, si procede al calcolo della matrice di correlazione.




Il grafico mostra i coefficienti di correlazione presenti nella matrice di correlazione. Ovviamente, essendo la matrice in questione simmetrica, con valori pari ad 1 sulla diagonale principale, vengono riportati esclusivamente i valori presenti sotto (o sopra) la diagonale principale.

Da subito si nota la presenza di un indice di correlazione estremamente elevato tra la 'infant.deaths' e 'under-five deaths' (pari esattamente a 0.99690562). Questo risultato era prevedibile fin da subito, considerato che, come detto in precedenza, le due variabili misurano quasi la stessa caratteristica.

Una situazione simile si verifica tra la variabile 'GDP' e 'percentage.expenditure'. In questo caso l'alto indice di correlazione è dovuto al fatto che entrambe le variabili sono indici economici (si ricorda che la prima misura il Prodotto Interno Lordo pro-capite, mentre la seconda la Percentuale del Prodotto Interno Lordo destinato a spese sanitarie).

# STIMA DEI COEFFICIENTI DI REGRESSIONE CON IL METODO DEI MINIMI QUADRATI ORDINARI E MULTICOLLINEARITA'

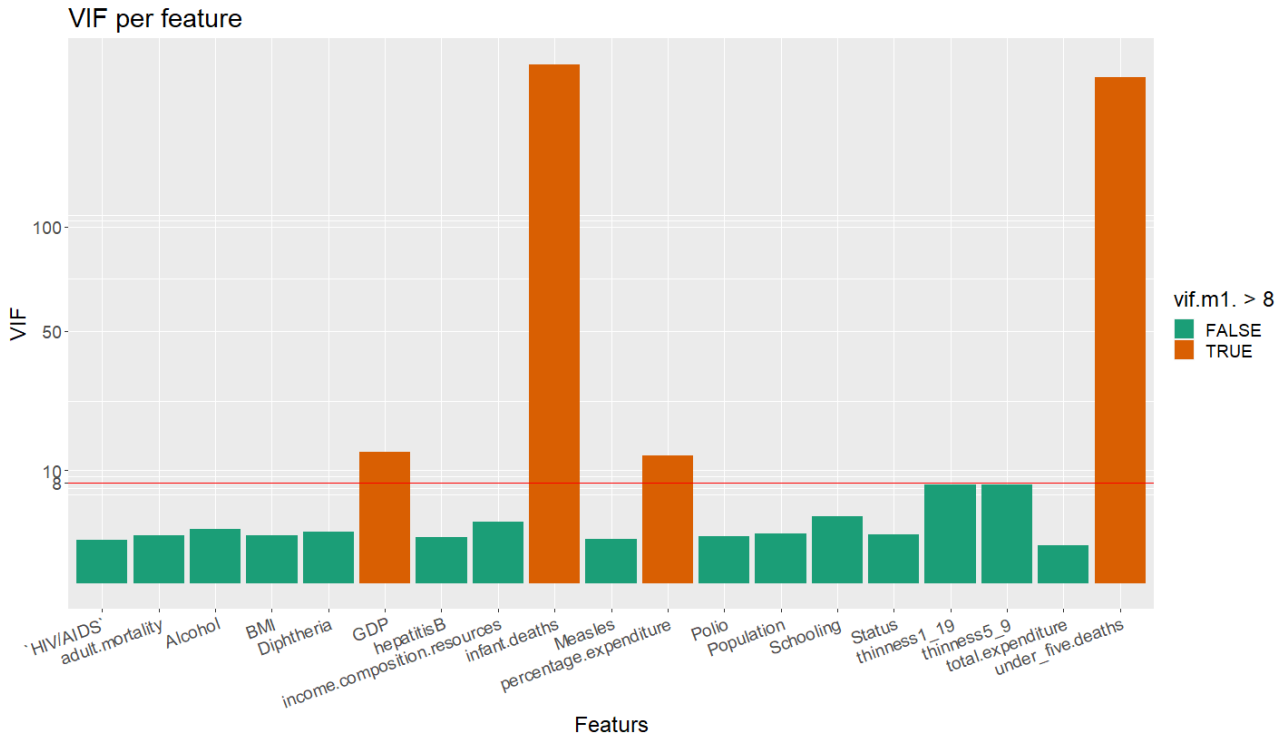
```
m1<-lm(life.expectancy~ ., data=dati)
summary(m1)

##
## Call:
## lm(formula = life.expectancy ~ ., data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9597  -2.0621  -0.0147   2.2751  11.7115
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.348e+01  7.375e-01  72.515  < 2e-16 ***
## Status                         9.684e-01  3.379e-01   2.865  0.00422 **
## adult.mortality               -1.663e-02  9.494e-04 -17.517  < 2e-16 ***
## infant.deaths                  9.350e-02  1.065e-02   8.777  < 2e-16 ***
## Alcohol                       -9.140e-02  3.316e-02  -2.756  0.00592 **
## percentage.expenditure         3.673e-04  1.801e-04   2.040  0.04156 *
## hepatitisB                    -6.525e-03  4.449e-03  -1.467  0.14265
## Measles                       -7.865e-06  1.079e-05  -0.729  0.46597
## BMI                            3.376e-02  5.998e-03   5.628 2.15e-08 ***
## under_five.deaths             -7.035e-02  7.711e-03  -9.123  < 2e-16 ***
## Polio                          7.935e-03  5.152e-03   1.540  0.12370
## total.expenditure              7.586e-02  4.067e-02   1.865  0.06236 .
## Diphtheria                     1.490e-02  5.928e-03   2.513  0.01205 *
## `HIV/AIDS`                    -4.370e-01  1.784e-02 -24.490  < 2e-16 ***
## GDP                            8.738e-06  2.837e-05   0.308  0.75813
## Population                    -6.425e-10  1.749e-09  -0.367  0.71337
## thinness1_19                  -1.238e-02  5.300e-02  -0.234  0.81527
## thinness5_9                   -4.798e-02  5.231e-02  -0.917  0.35917
## income.composition.resources   9.817e+00  8.321e-01  11.797  < 2e-16 ***
## Schooling                      8.665e-01  5.940e-02  14.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.588 on 1629 degrees of freedom
## Multiple R-squared:  0.8356, Adjusted R-squared:  0.8336
## F-statistic: 435.7 on 19 and 1629 DF,  p-value: < 2.2e-16
```

Il valore dell'indice di determinazione è pari a 0.8336, ovvero, la percentuale di variabilità spiegata simultaneamente da tutti i regressori è pari all'83,36%. Tuttavia, la presenza di multicollinearità potrebbe influire sulla precisione delle stime ai minimi quadrati. Dunque, viene calcolato l'indice VIF. Valori di $\max\{VIF_j\} > 10$ segnalano presenza di multicollinearità.

Componenti: Bessioud Ghofran, Ferrari Gianluca, Gallo Debora, Iannicelli Stefano, Lavecchia Nicola, Pascuzzi Eugenio, Pascuzzi Michela, Sorrenti Alessandro

Tuttavia, nel grafico viene inserito un valore limite minore, pari ad 8:



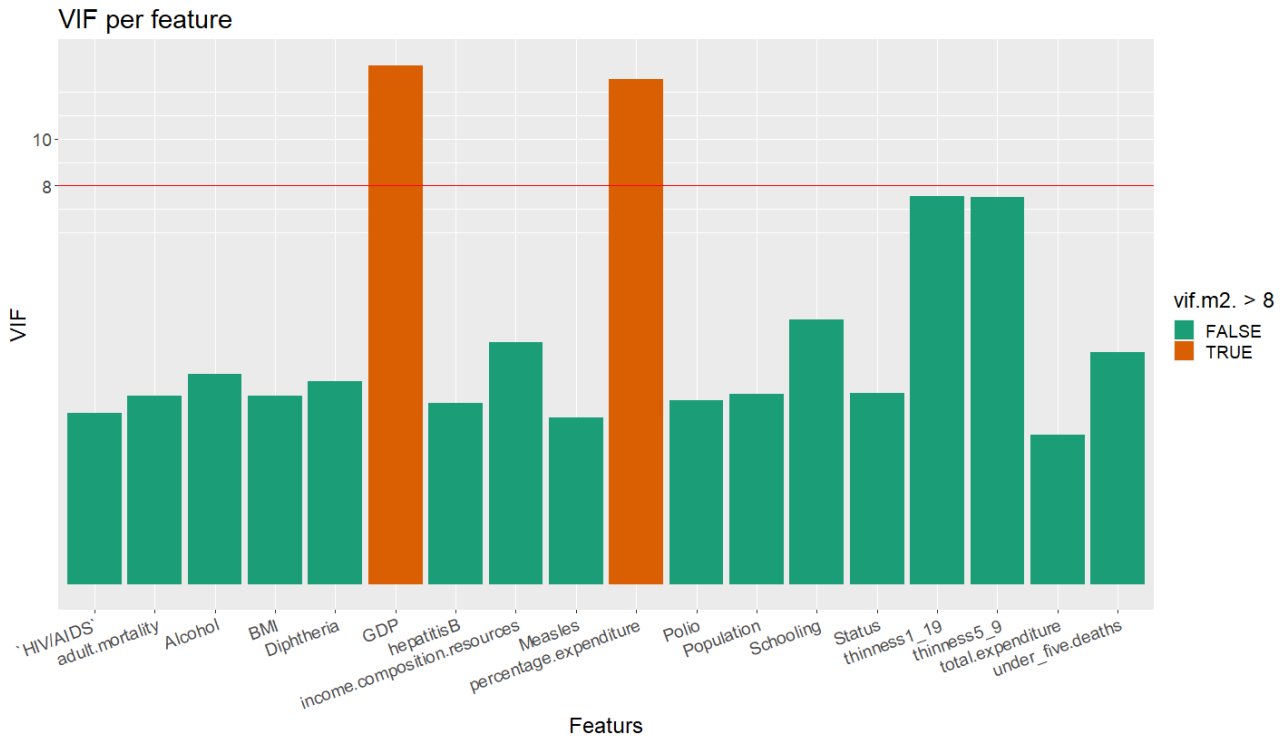

Come si poteva immaginare, il valore dell'indice risulta essere molto elevato per i regressori citati precedentemente. Si tenta di risolvere la situazione andando ad eliminare tra i regressori correlati fra loro, quello con il valore dell'indice VIF maggiore.

Si effettua un primo tentativo eliminando dal modello la variabile 'infant.deaths':

```
m2<-update(m1, .~.-infant.deaths)
summary(m2)
```

```
. . .
```

```
## Residual standard error: 3.671 on 1630 degrees of freedom
## Multiple R-squared:  0.8278, Adjusted R-squared:  0.8259
## F-statistic: 435.3 on 18 and 1630 DF,  p-value: < 2.2e-16
```

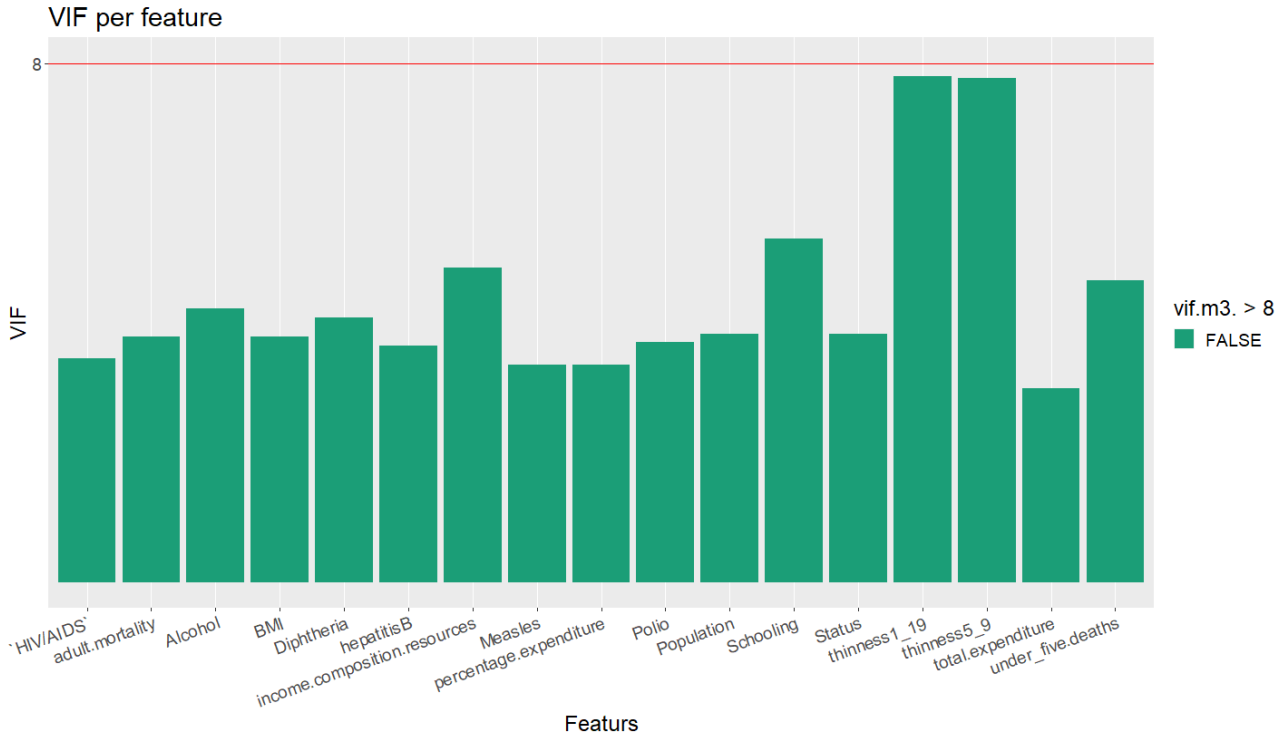



Per lo stesso motivo, viene eliminata dal modello la variabile 'GDP':

```
m3<-update(m2, .~.-GDP)
summary(m3)

##
## Call:
## lm(formula = life.expectancy ~ Status + adult.mortality + Alcohol +
##     percentage.expenditure + hepatitisB + Measles + BMI + under_five.deaths
+
##     Polio + total.expenditure + Diphtheria + `HIV/AIDS` + Population +
##     thinness1_19 + thinness5_9 + income.composition.resources +
##     Schooling, data = dati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.068  -2.133   0.029   2.418  11.551
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.246e+01  7.424e-01  70.664  < 2e-16 ***
## Status                         9.943e-01  3.450e-01   2.883 0.003996 **
## adult.mortality               -1.763e-02  9.640e-04 -18.290  < 2e-16 ***
## Alcohol                       -1.392e-01  3.346e-02  -4.159 3.36e-05 ***
## percentage.expenditure         4.111e-04  6.086e-05   6.755 1.98e-11 ***
## hepatitisB                    -7.269e-03  4.548e-03  -1.598 0.110148
## Measles                        1.804e-05  1.061e-05   1.700 0.089252 .
## BMI                            3.590e-02  6.128e-03   5.858 5.67e-09 ***
## under_five.deaths             -3.120e-03  9.136e-04  -3.414 0.000655 ***
## Polio                          1.023e-02  5.261e-03   1.944 0.052015 .
## total.expenditure              6.898e-02  4.156e-02   1.659 0.097208 .
```

```
## Diphtheria                      2.010e-02  6.032e-03   3.333 0.000879 ***
## `HIV/AIDS`                     -4.389e-01  1.825e-02 -24.054  < 2e-16 ***
## Population                      3.082e-09  1.735e-09   1.777 0.075829 .
## thinness1_19                   -2.296e-02  5.419e-02  -0.424 0.671922
## thinness5_9                    -1.250e-02  5.334e-02  -0.234 0.814710
## income.composition.resources    1.039e+01  8.464e-01  12.270  < 2e-16 ***
## Schooling                       8.903e-01  6.048e-02  14.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 1631 degrees of freedom
## Multiple R-squared:  0.8278, Adjusted R-squared:  0.826
## F-statistic: 461.1 on 17 and 1631 DF,  p-value: < 2.2e-16
```




L'indice VIF risulta essere in un range di valori accettabili per tutti i regressori.

Il modello di regressione costruito è il seguente:

$$
\begin{aligned}
life.expectancy \\
&= \beta_0 + \beta_1\, Status + \beta_2\, adult.mortality + \beta_3\, Alcohol \\
&+ \beta_4\, percentage.expenditure + \beta_5\, hepatitisB + \beta_6\, Measles + \beta_7\, BMI \\
&+ \beta_8\, underfive.deaths + \beta_9\, Polio + \beta_{10}\, total.expenditure + \beta_{11}\, Diphtheria \\
&+ \beta_{12}\, HIV/AIDS + \beta_{13}\, Population + \beta_{14}\, thinness1\_19 + \beta_{15}\, thinness5\_9 \\
&+ \beta_{16}\, income.composition.resources + \beta_{17}\, Schooling
\end{aligned}
$$

Tuttavia, si può notare che vi sono alcuni regressori che risultano non avere un effetto significativo sulla variabile dipendente.

Dopo vari tentativi, sono stati eliminati tutti i regressori non significativi. Il modello risultante è il seguente:

```
m6<-update(m5, .~.-thinness1_19)
summary(m6)

##
## Call:
## lm(formula = life.expectancy ~ Status + adult.mortality + Alcohol +
##     percentage.expenditure + Measles + BMI + under_five.deaths +
##     Polio + total.expenditure + Diphtheria + `HIV/AIDS` + Population +
##     income.composition.resources + Schooling, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0887  -2.1200   0.0584   2.4107  11.8864
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.201e+01  6.880e-01  75.599  < 2e-16 ***
## Status                          9.478e-01  3.440e-01   2.755  0.00593 **
## adult.mortality                -1.773e-02  9.622e-04 -18.430  < 2e-16 ***
## Alcohol                        -1.288e-01  3.294e-02  -3.909 9.64e-05 ***
## percentage.expenditure          4.216e-04  6.060e-05   6.958 4.98e-12 ***
## Measles                         2.005e-05  1.054e-05   1.903  0.05721 .
## BMI                             3.827e-02  5.726e-03   6.684 3.18e-11 ***
## under_five.deaths              -3.392e-03  8.589e-04  -3.950 8.16e-05 ***
## Polio                           8.549e-03  5.184e-03   1.649  0.09932 .
## total.expenditure               6.999e-02  4.143e-02   1.690  0.09131 .
## Diphtheria                      1.584e-02  5.435e-03   2.915  0.00360 **
## `HIV/AIDS`                     -4.392e-01  1.819e-02 -24.149  < 2e-16 ***
## Population                      3.089e-09  1.733e-09   1.782  0.07494 .
## income.composition.resources    1.050e+01  8.442e-01  12.436  < 2e-16 ***
## Schooling                       8.914e-01  6.039e-02  14.760  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.671 on 1634 degrees of freedom
## Multiple R-squared:  0.8273, Adjusted R-squared:  0.8258
## F-statistic: 559.2 on 14 and 1634 DF,  p-value: < 2.2e-16
```

L'indice di determinazione risulta essere quasi il medesimo e lo stesso vale per la statistica F.

*COMMENTI:*

- *Se il valore di tutti i regressori fosse pari a 0, si avrebbe un'aspettativa di vita pari a 52;*
- *Il consumo di alcol, la mortalità degli adulti e dei bambini, le morti dovute a HIV/AIDS hanno un coefficiente angolare negativo, dunque la loro presenza influisce negativamente sull'aspettativa di vita;*
- *L'82,58% della variabilità dell'aspettativa di vita è spiegata simultaneamente dai regressori, un ottimo risultato;*

Il risultato ottenuto, tuttavia, non è soddisfacente. La presenza di multicollinearità ci ha costretto ad eliminare alcuni regressori, dunque, ad eliminare informazioni che avrebbero potuto essere di fondamentale importanza. Oltre alla perdita informativa, sappiamo che la varianza dei regressori aumenta all'aumentare della multicollinearità e conseguentemente siamo portati ad eliminare regressori anche quando questi ultimi potrebbero avere un effetto significativo.

## AUTOCORRELAZIONE

Una delle ipotesi fondamentali del modello lineare classico è che la matrice delle varianze e covarianze degli errori sia diagonale, ovvero:

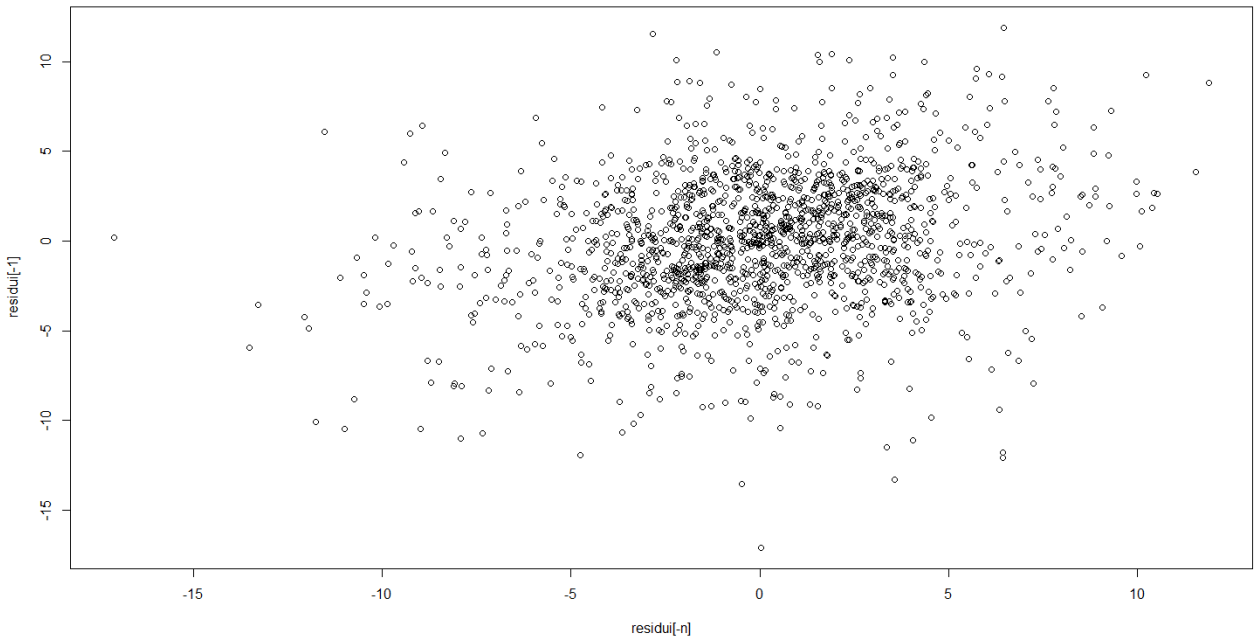
$$E(\varepsilon\varepsilon') = \sigma^2 I_n$$

Nel caso in cui questa ipotesi viene meno, la matrice di varianze e covarianza assume la seguente forma:

$$E(\varepsilon\varepsilon') = \sigma^2 \Omega$$

In tale circostanza, le stime dei coefficienti di regressione tramite il metodo dei minimi quadrati, pur continuando ad essere corrette, non sono le più efficienti. Occorre, dunque, usare i minimi quadrati generalizzati.

Un caso in cui è utile l'applicazione dei minimi quadrati generalizzati è quello in cui vi è la presenza di errori autocorrelati. La forma più comune di correlazione tra gli errori è la correlazione seriale. Per avere un'idea circa la presenza o meno di autocorrelazione seriale, si può inizialmente effettuare un'analisi grafica utilizzando i residui del modello lineare stimato con i minimi quadrati ordinari. Viene di seguito mostrato il grafico dei residui vs i residui 'ritardati'.



Il grafico sembra essere abbastanza informativo, infatti, si può notare una relazione di tipo lineare tra gli errori e nello specifico, i punti sembrano essere concordi. L'analisi grafica, tuttavia, non è sufficiente per trarre delle conclusioni. Si verifica dunque la presenza di autocorrelazione degli errori con il test Durbin-Watson:

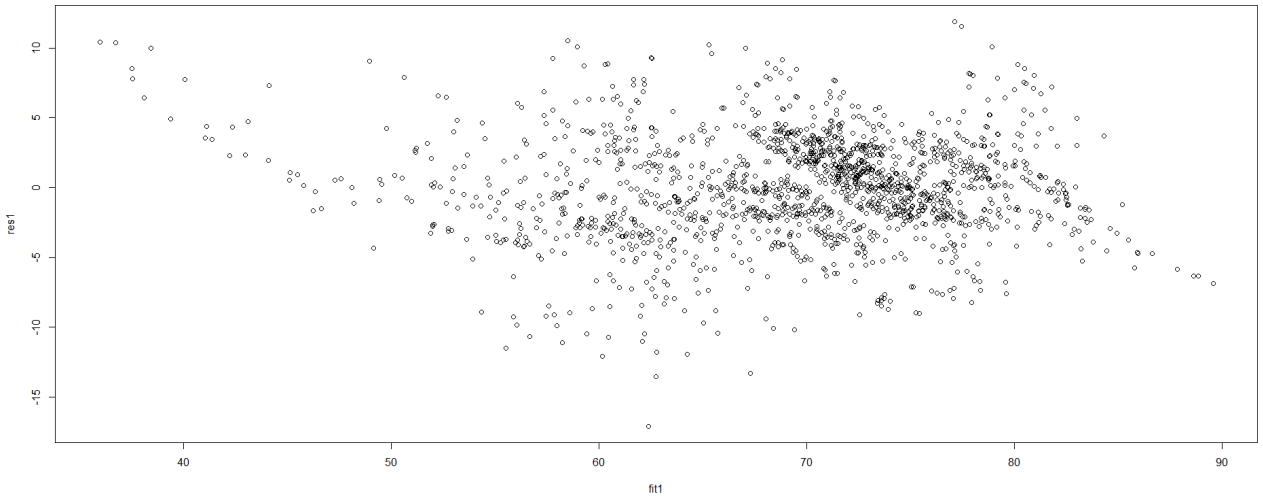
```
lmtest :: dwtest(m6)

##
##  Durbin-Watson test
##
## data:  m1
## DW = 1.5293, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Il p-value è molto basso, dunque il test risulta essere molto significativo: siamo in presenza di correlazione seriale. Per conoscere il verso della correlazione, basta osservare il valore della statistica test, che conferma quanto detto precedentemente tramite la sola visualizzazione grafica dei residui. La statistica, infatti, assume un valore minore di 2, ciò sta ad indicare che vi è la presenza di correlazione positiva.

## ETEROSCHEDASTICITÀ

Un altro tipo di allontanamento dalle ipotesi fondamentali del modello di regressione classico è la presenza di eteroschedasticità. Anche questo è un caso in cui risulta utile l'applicazione dei minimi quadrati generalizzati. Si procede ad una visualizzazione grafica dei residui rispetto alle ordinate stimate:




I punti non sembrano avere un andamento casuale, bensì, sembra che all'aumentare delle ordinate stimate, la variabilità degli errori diminuisca. Come per la verifica della presenza di autocorrelazione, anche in questo caso si procede facendo riferimento ad un test statistico per avere la conferma della presenza di eteroschedasticità.

```
#TEST BREUSCH-PAGAN
res12<-res1^2 # residui al quadrato
modBP<-lm(res12~Status + adult.mortality + Alcohol +
            percentage.expenditure + Measles + BMI + under_five.deaths +
            Polio + total.expenditure + Diphtheria + `HIV/AIDS` + Population +
            income.composition.resources + Schooling)
summary(modBP)

...

## Residual standard error: 21.4 on 1634 degrees of freedom
## Multiple R-squared:  0.1269, Adjusted R-squared:  0.1195
## F-statistic: 16.97 on 14 and 1634 DF,  p-value: < 2.2e-16

#TEST DI WHITE
res12<-res1^2 # residui al quadrato
fit12<-fit1^2

modresW<-lm(res12~fit1+fit12)
summary(modresW)

...

## Residual standard error: 22.46 on 1646 degrees of freedom
## Multiple R-squared:  0.03143,    Adjusted R-squared:  0.03026
## F-statistic: 26.71 on 2 and 1646 DF,  p-value: 3.84e-12
```

Entrambi i test presentano un p-value molto basso, dunque, ci conducono entrambi al rifiuto dell'ipotesi nulla di omoschedasticità.

Sono stati diversi i tentativi di trasformazione effettuati per eliminare l'eteroschedasticità, tra cui: la suddivisione della variabile dipendente e dei regressori per le ordinate stimate; la trasformazione logaritmica della variabile dipendente; l'applicazione di pesi sui regressori. Tuttavia, entrambi i test continuano ad essere significativi.

## CROSS VALIDATION SUL MODELLO LINEARE ORDINARIO

Consideriamo il modello lineare ottenuto con il metodo dei minimi quadrati per effettuare la cross validation. I risultati ottenuti sono i seguenti:

```
ddLOOCV

## [1] 13.07523 13.07512

ddKfold

## [1] 13.06790 13.04903
```

Il risultato è poco informativo poiché bisogna confrontarlo con l'MSE ottenuto con altre tecniche.

# TECNICHE DI REGOLARIZZAZIONE

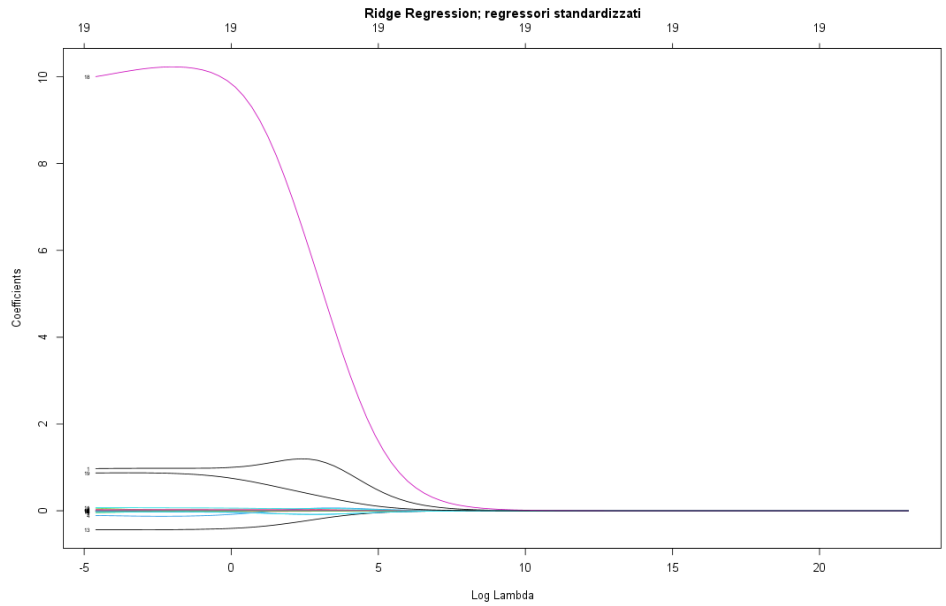
Dalle analisi fatte in precedenza, sappiamo che alcuni regressori sono altamente correlati fra loro. la presenza di multicollinearità può portare a una riduzione della precisione delle previsioni, poiché i modelli basati su dati altamente correlati possono soffrire di un'elevata varianza, il che significa che le previsioni possono essere molto sensibili alle piccole variazioni nei dati di input. Per affrontare il problema della multicollinearità, è possibile utilizzare diverse tecniche, come l'eliminazione di uno dei regressori correlati, tuttavia, come detto in precedenza, tale procedura potrebbe essere pericolosa: il rischio è quello di eliminare informazioni utili. Si procede dunque utilizzando tutte le tecniche di regolarizzazione conosciute al fine di identificare quella con la quale si ottiene un mse minore.

## *RIDGE REGRESSION*

Si crea il modello con la tecnica di Ridge Regression attraverso un modello lineare generalizzato con alpha pari a 0.

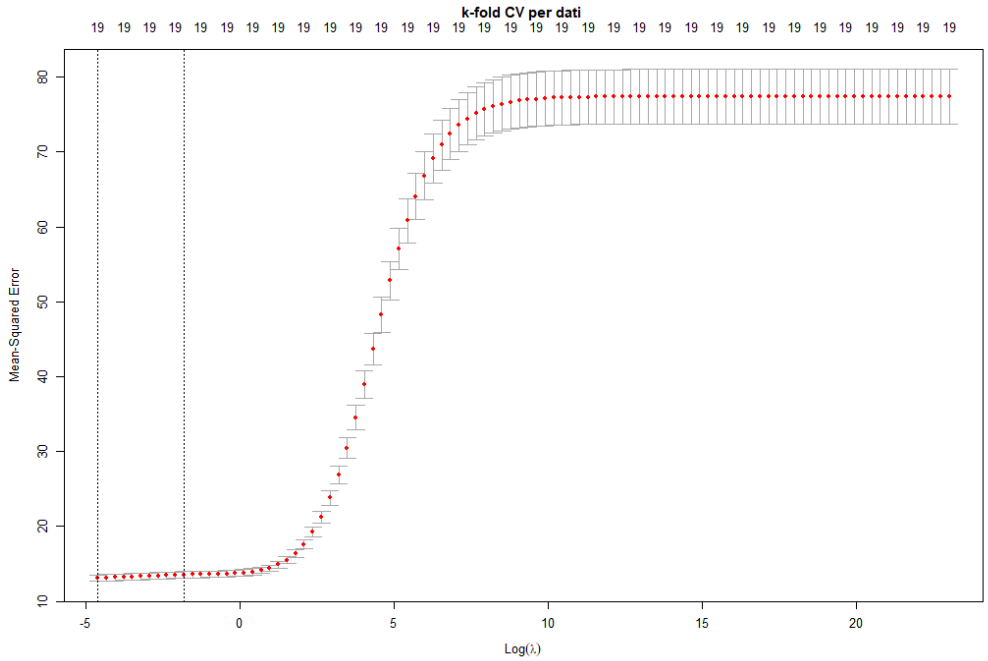
```
ridge.mods.all=glmnet (x,y,alpha=0, lambda=griglia)
```

Otteniamo il seguente grafico:




All'aumentare di $\lambda$, le stime dei parametri tendono tutte a zero, e per valori tendenti a zero, le stime Ridge coincidono con quelle ai minimi quadrati.

Valutiamo ora il Mean Square Error (MSE), al variare del parametro di penalità $\lambda$, del modello ridge.mods.all con la tecnica K-fold Cross Validation

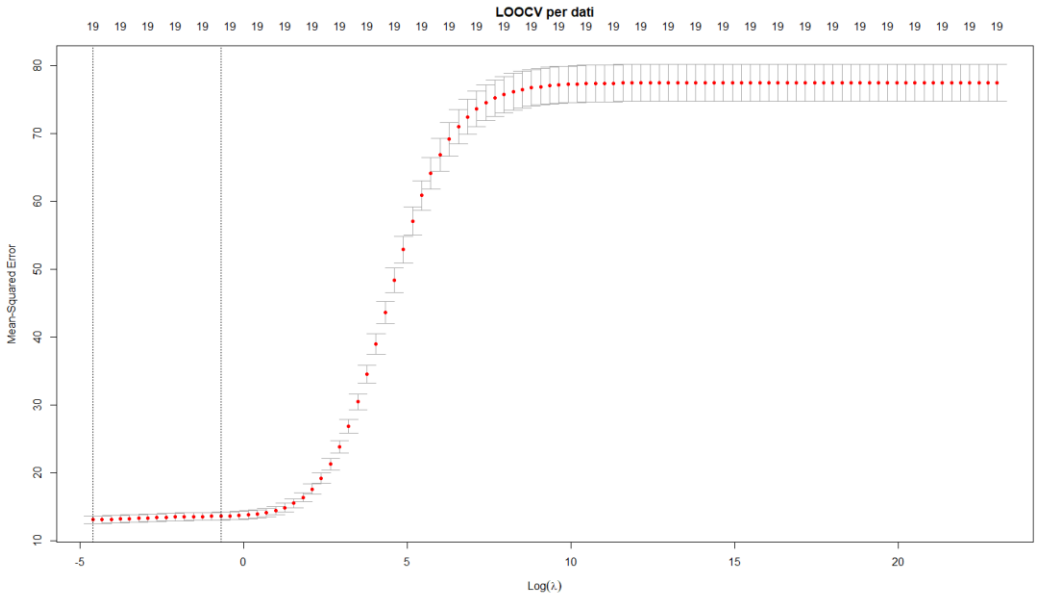



Dal grafico si può notare che il valore di $\lambda$ a cui corrisponde il minimo errore predittivo è 0.01.

```
coef(ridge.mod.kCV)[,1]
```

```
##                   (Intercept)                        Status
##                  5.312943e+01                  9.737747e-01
##                adult.mortality                infant.deaths
##                  -1.698681e-02                  5.996755e-02
##                       Alcohol       percentage.expenditure
##                  -1.082450e-01                  3.599664e-04
##                     hepatitisB                       Measles
##                  -6.673483e-03                  5.606817e-07
##                            BMI             under_five.deaths
##                   3.461647e-02                 -4.609421e-02
##                          Polio             total.expenditure
##                   8.799760e-03                  7.345771e-02
##                     Diphtheria                       HIV/AIDS
##                   1.678957e-02                 -4.371570e-01
##                            GDP                    Population
##                   9.508941e-06                  5.531480e-10
##                   thinness1_19                   thinness5_9
##                  -1.732183e-02                 -3.573348e-02
## income.composition.resources                       Schooling
##                   1.000297e+01                  8.731635e-01
```

Si può notare che con la tecnica LOOCV il valore di $\lambda$ a cui corrisponde il minimo errore predittivo è 0.01, di conseguenza, i valori dei coefficienti del modello che sono stati ottenuti sono identici a quelli già riportati precedentemente.




```
cv.outLOOCV$lambda.1se
```

```
## [1] 0.4977024
```

```
ridge.mod.kCV=glmnet(x,y,alpha=0,lambda=bestLambdaLOOCV)
coef(ridge.mod.kCV)[,1]
```

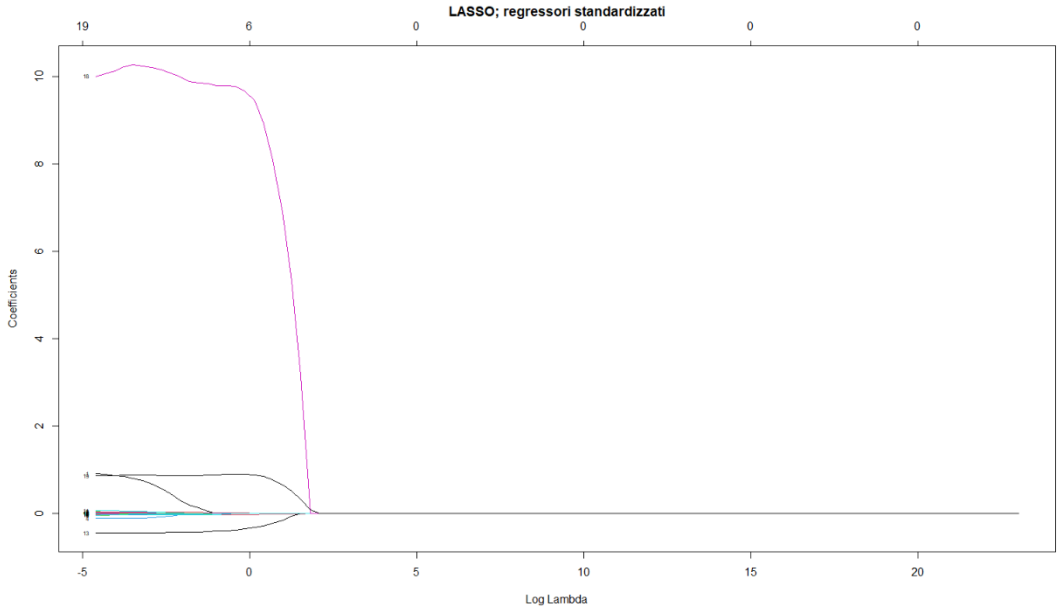
```
##                   (Intercept)                        Status
##                  5.312943e+01                  9.737747e-01
##                adult.mortality                infant.deaths
##                  -1.698681e-02                  5.996755e-02
##                       Alcohol       percentage.expenditure
##                  -1.082450e-01                  3.599664e-04
##                     hepatitisB                       Measles
##                  -6.673483e-03                  5.606817e-07
##                             BMI             under_five.deaths
##                   3.461647e-02                 -4.609421e-02
##                           Polio             total.expenditure
##                   8.799760e-03                  7.345771e-02
##                     Diphtheria                       HIV/AIDS
##                   1.678957e-02                 -4.371570e-01
##                             GDP                    Population
##                   9.508941e-06                  5.531480e-10
##                   thinness1_19                   thinness5_9
##                  -1.732183e-02                 -3.573348e-02
## income.composition.resources                      Schooling
##                   1.000297e+01                  8.731635e-01
```

*LASSO*

Si crea il modello con la tecnica di regolarizzazione LASSO, attraverso un modello lineare generalizzato con alpha pari a 1.

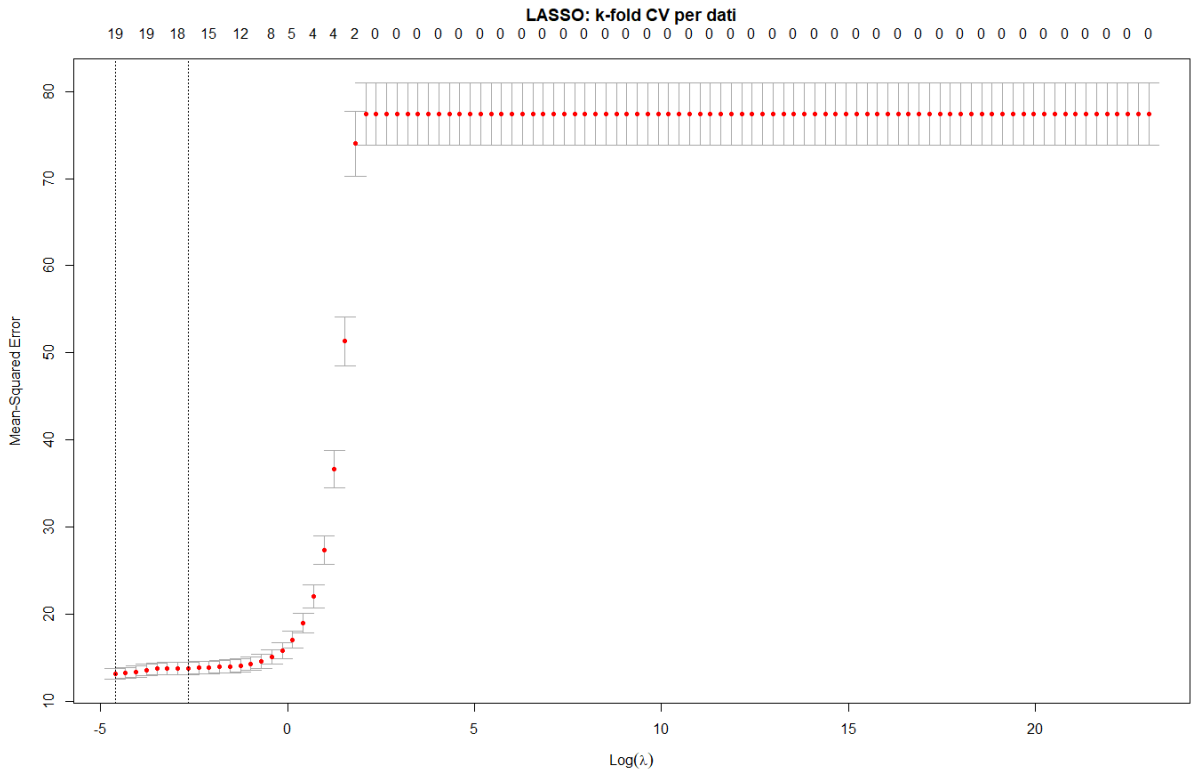
lasso.mods.all=glmnet (x,y, alpha=1, lambda=griglia)

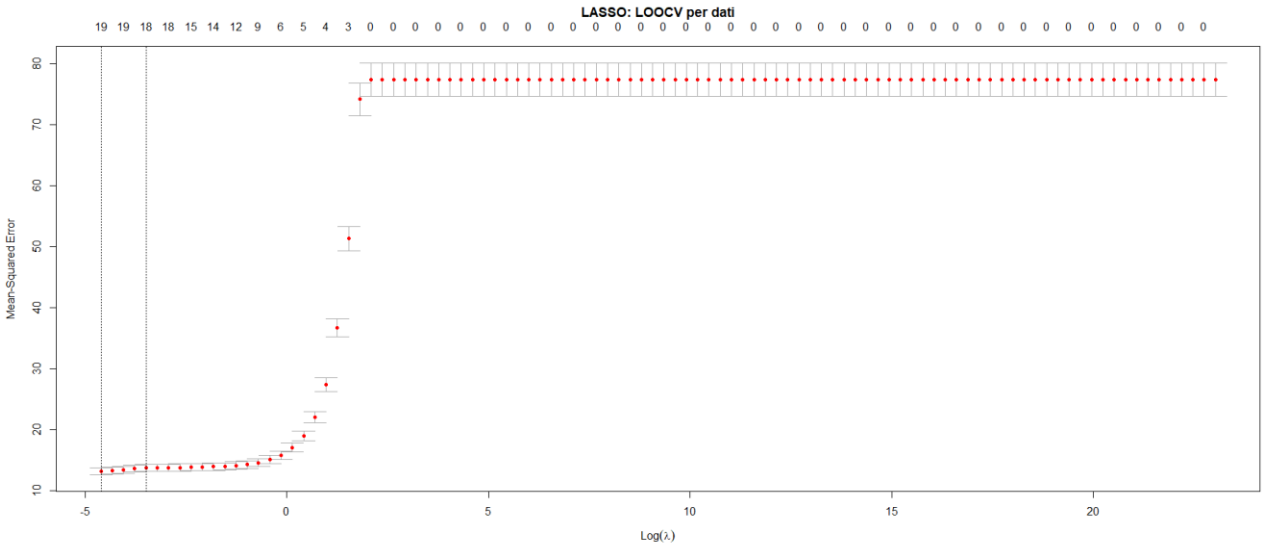
Otteniamo il seguente grafico:




All'aumentare di $\lambda$, le stime dei parametri tendono tutte a zero.

Valutiamo ora il Mean Square Error (MSE), al variare del parametro di penalità $\lambda$, del modello lasso.mods.all con le due tecniche di Cross Validation K-fold e Leave One Out.






Sia con la k-fold Cross Validation che con LOOCV, si ha un valore di $\lambda$ a cui corrisponde il minimo errore predittivo di 0.01. I valori dei coefficienti stimati dalla tecnica LASSO sono i seguenti:

```
coef(LASSO.mod.kCV)[,1]
```

```
##                   (Intercept)                        Status
##                  5.313024e+01                  9.164052e-01
##                adult.mortality                infant.deaths
##                  -1.707813e-02                  5.595023e-02
##                       Alcohol       percentage.expenditure
##                  -1.005951e-01                  3.618195e-04
##                     hepatitisB                       Measles
##                  -5.645655e-03                  2.323537e-07
##                             BMI             under_five.deaths
##                   3.443431e-02                 -4.311834e-02
##                           Polio             total.expenditure
##                   8.431118e-03                  6.884292e-02
##                     Diphtheria                       HIV/AIDS
##                   1.621020e-02                 -4.367531e-01
##                             GDP                    Population
##                   8.598939e-06                  4.816492e-10
##                   thinness1_19                   thinness5_9
##                  -1.547855e-02                 -3.423070e-02
## income.composition.resources                       Schooling
##                   1.000925e+01                  8.735357e-01
```

Si può notare come nessun coefficiente è stato azzerato, una conseguenza del fatto che il valore di $\lambda$ è molto piccolo.

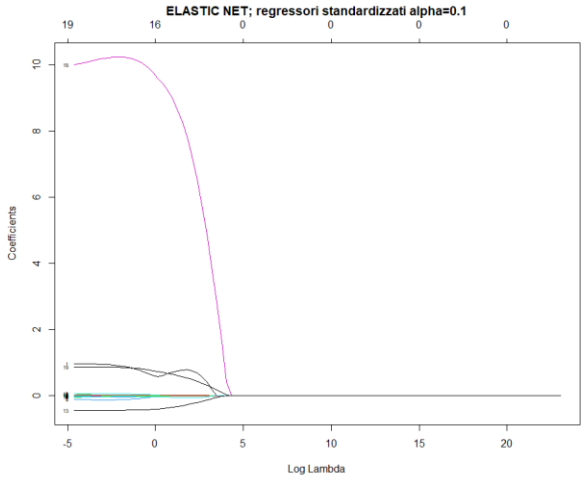
Confronto tra i coefficienti stimati dalla tecnica Ridge Regression e dalla LASSO:

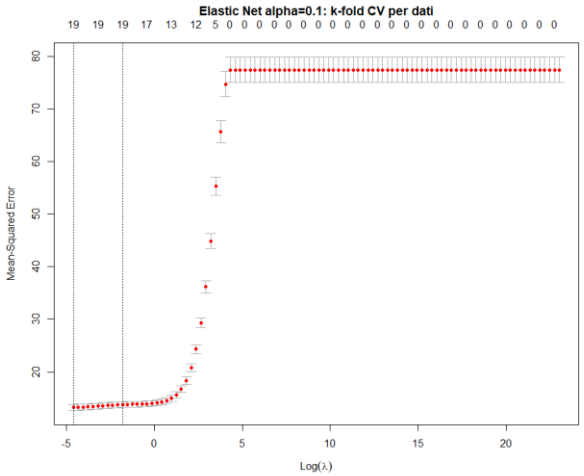
```
cbind(coef(LASSO.mod.kCV)[,1], coef(ridge.mod.kCV)[,1])
```

```
##                                    [,1]          [,2]
## (Intercept)                5.313024e+01  5.312943e+01
## Status                     9.164052e-01  9.737747e-01
## adult.mortality           -1.707813e-02 -1.698681e-02
## infant.deaths              5.595023e-02  5.996755e-02
## Alcohol                   -1.005951e-01 -1.082450e-01
## percentage.expenditure     3.618195e-04  3.599664e-04
## hepatitisB                -5.645655e-03 -6.673483e-03
## Measles                    2.323537e-07  5.606817e-07
## BMI                        3.443431e-02  3.461647e-02
## under_five.deaths         -4.311834e-02 -4.609421e-02
## Polio                      8.431118e-03  8.799760e-03
## total.expenditure          6.884292e-02  7.345771e-02
## Diphtheria                 1.621020e-02  1.678957e-02
## HIV/AIDS                  -4.367531e-01 -4.371570e-01
## GDP                        8.598939e-06  9.508941e-06
## Population                 4.816492e-10  5.531480e-10
## thinness1_19              -1.547855e-02 -1.732183e-02
## thinness5_9               -3.423070e-02 -3.573348e-02
## income.composition.resources  1.000925e+01  1.000297e+01
## Schooling                  8.735357e-01  8.731635e-01
```

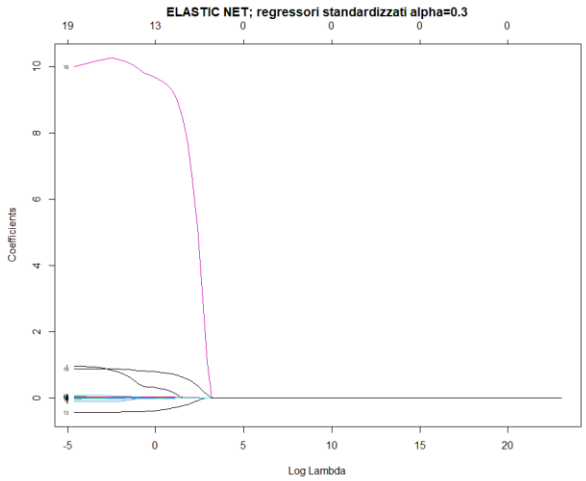
## *ELASTIC NET*

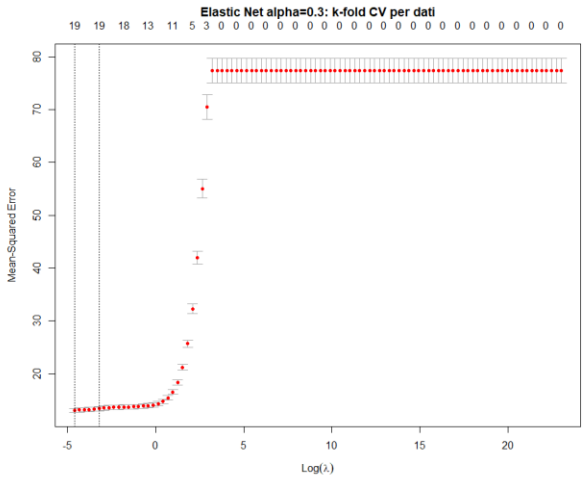
Elastic Net con parametro α=0.1



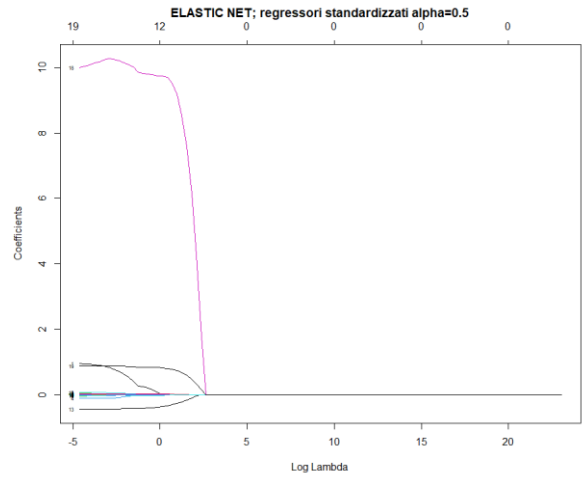



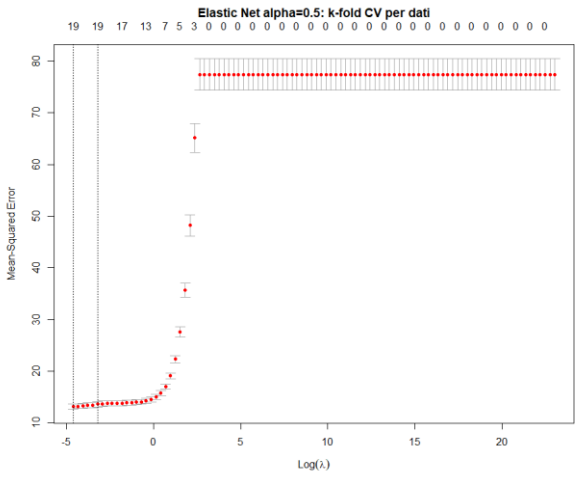

Elastic Net con parametro α=0.3



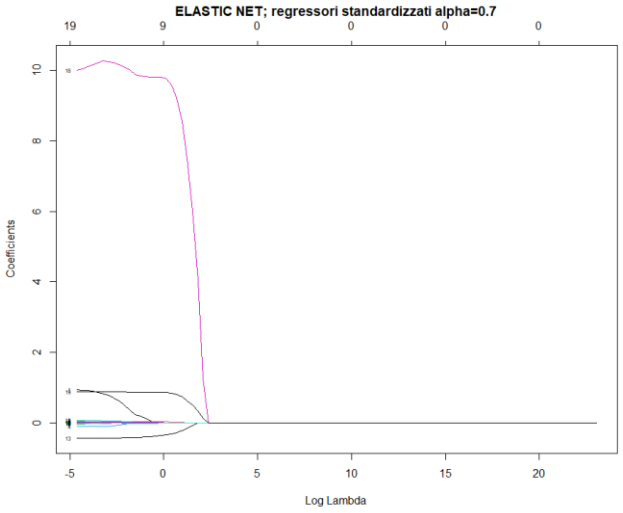

Elastic Net con parametro α=0.5

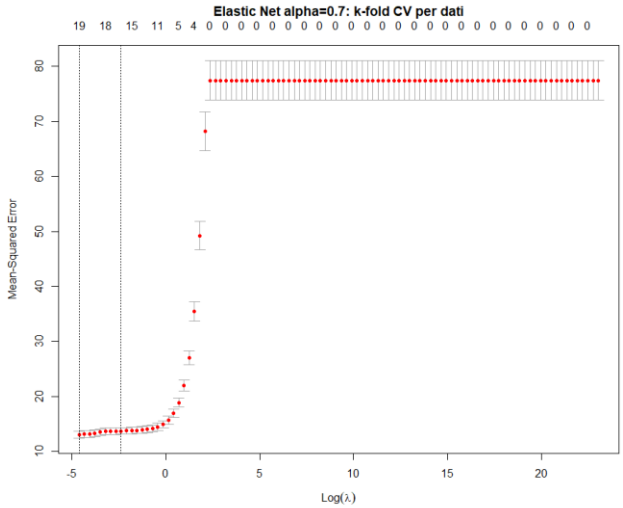


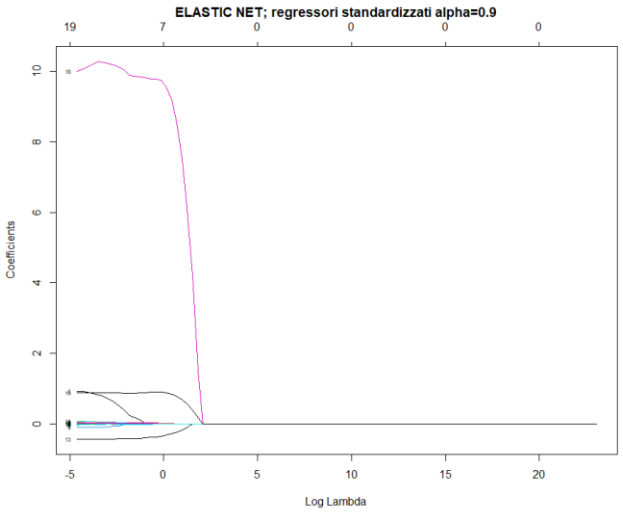

**Componenti**: Bessioud Ghofran, Ferrari Gianluca, Gallo Debora, Iannicelli Stefano, Lavecchia Nicola, Pascuzzi Eugenio, Pascuzzi Michela, Sorrenti Alessandro

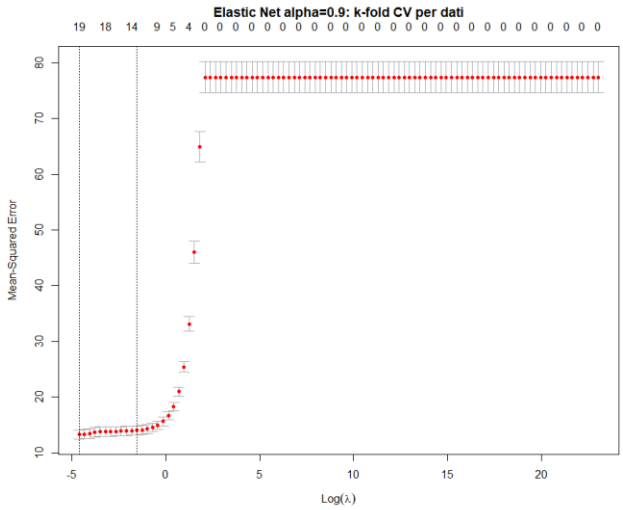
Elastic Net con parametro α=0.7




Elastic Net con parametro α=0.9




Anche con Elastic Net il best $\lambda$ è pari a 0.01 sia con la K-fold cross validation, sia con la Leave-One-Out, dunque la stima dei parametri non cambia nei due casi.

Dopo vari tentativi, è stato scelto di non modificare la griglia per la generazione dei valori di $\lambda$, poiché si è ritenuto che il valore 0.01, ottenuto con tutte le tecniche di regolarizzazione, fosse abbastanza piccolo. Modificando la scala, infatti, si sarebbero ottenuti valori ancora più piccoli del parametro di penalizzazione e conseguentemente avremmo ottenuto coefficienti di regressione eccessivamente simili a quelli ottenuti con il modello lineare ai minimi quadrati.

NOTA: i valori dei coefficienti stimati dalla tecnica Elastic Net sono riportati nel confronto tra le tecniche di regolarizzazione.

| | LASSO | RR | EN α=0.1 | EN α=0.3 | EN α=0.5 | EN α=0.7 | EN α=0.9 |
|---|---|---|---|---|---|---|---|
| (Intercept) | 5.313024e+01 | 5.312943e+01 | 5.313095e+01 | 5.313304e+01 | 5.313412e+01 | 5.313380e+01 | 5.313177e+01 |
| Status | 9.164052e-01 | 9.737747e-01 | 9.680354e-01 | 9.565523e-01 | 9.450705e-01 | 9.335940e-01 | 9.221298e-01 |
| adult.mortality | -1.707813e-02 | -1.698681e-02 | -1.699453e-02 | -1.701092e-02 | -1.702829e-02 | -1.704704e-02 | -1.706744e-02 |
| infant.deaths | 5.595023e-02 | 5.996755e-02 | 5.969661e-02 | 5.906723e-02 | 5.834715e-02 | 5.749919e-02 | 5.649759e-02 |
| Alcohol | -1.005951e-01 | -1.082450e-01 | -1.074139e-01 | -1.057942e-01 | -1.042209e-01 | -1.027127e-01 | -1.012841e-01 |
| percentage.expenditure | 3.618195e-04 | 3.599664e-04 | 3.601517e-04 | 3.605336e-04 | 3.609105e-04 | 3.612845e-04 | 3.616489e-04 |
| hepatitisB | -5.645655e-03 | -6.673483e-03 | -6.570243e-03 | -6.363874e-03 | -6.157888e-03 | -5.952407e-03 | -5.747659e-03 |
| Measles | 2.323537e-07 | 5.606817e-07 | 4.949722e-07 | 3.850700e-07 | 2.982109e-07 | 2.436042e-07 | 2.279707e-07 |
| BMI | 3.443431e-02 | 3.461647e-02 | 3.459518e-02 | 3.455496e-02 | 3.451668e-02 | 3.448135e-02 | 3.444947e-02 |
| under_five.deaths | -4.311834e-02 | -4.609421e-02 | -4.589125e-02 | -4.542211e-02 | -4.488729e-02 | -4.425994e-02 | -4.352139e-02 |
| Polio | 8.431118e-03 | 8.799760e-03 | 8.759671e-03 | 8.681576e-03 | 8.605752e-03 | 8.533099e-03 | 8.464295e-03 |
| total.expenditure | 6.884292e-02 | 7.345771e-02 | 7.300615e-02 | 7.209579e-02 | 7.117877e-02 | 7.025224e-02 | 6.931463e-02 |
| Diphtheria | 1.621020e-02 | 1.678957e-02 | 1.672445e-02 | 1.659886e-02 | 1.647830e-02 | 1.636481e-02 | 1.625992e-02 |
| HIV/AIDS | -4.367531e-01 | -4.371570e-01 | -4.371145e-01 | -4.370308e-01 | -4.369486e-01 | -4.368686e-01 | -4.367911e-01 |
| GDP | 8.598939e-06 | 9.508941e-06 | 9.419640e-06 | 9.238265e-06 | 9.056409e-06 | 8.873317e-06 | 8.689652e-06 |
| Population | 4.816492e-10 | 5.531480e-10 | 5.414023e-10 | 5.211269e-10 | 5.039429e-10 | 4.912315e-10 | 4.838486e-10 |
| thinness1_19 | -1.547855e-02 | -1.732183e-02 | -1.712069e-02 | -1.672902e-02 | -1.634902e-02 | -1.598574e-02 | -1.564315e-02 |
| thinness5_9 | -3.423070e-02 | -3.573348e-02 | -3.562983e-02 | -3.539027e-02 | -3.511935e-02 | -3.480289e-02 | -3.443181e-02 |
| income.composition.resources | 1.000925e+01 | 1.000297e+01 | 1.000286e+01 | 1.000313e+01 | 1.000391e+01 | 1.000542e+01 | 1.000780e+01 |
| Schooling | 8.735357e-01 | 8.731635e-01 | 8.731666e-01 | 8.731936e-01 | 8.732455e-01 | 8.733312e-01 | 8.734581e-01 |

```
cbind(mse.minLASSO, mse.minRR, mse.minEN01, mse.minEN03, mse.minEN05, mse.minEN
07, mse.minEN09)

##      mse.minLASSO mse.minRR mse.minEN01 mse.minEN03 mse.minEN05 mse.minEN07
## [1,]     13.23263   13.1775    13.15378    13.11949     13.1647    13.14357
##      mse.minEN09
## [1,]    13.14628

min(mse.minLASSO, mse.minRR, mse.minEN01, mse.minEN03, mse.minEN05, mse.minEN07
, mse.minEN09)

## [1] 13.11949
```

## RISULTATI

Tra i modelli ottenuti con le tecniche di regolarizzazione, scegliamo la **Elastic Net** con parametro **α** pari a 0.3 poiché presenta il MSE di test più piccolo.

Tuttavia, si può notare come nessun coefficiente è stato spinto verso lo 0. Questo può indicare che tutte le variabili indipendenti nel modello sono effettivamente utili per spiegare la variabile dipendente o che la regolarizzazione non è stata abbastanza forte da eliminare le variabili poco significative o ridondanti (che sappiamo essere presenti per le analisi effettuate precedentemente). Per risolvere il problema, potrebbero essere usate delle tecniche di selezione automatica dei regressori per ridurre la complessità del modello e migliorare la sua capacità di generalizzazione.

Confrontando l'MSE minimo ottenuto con le tecniche di regolarizzazione e quello ottenuto con il modello ai minimi quadrati, si sceglie come tecnica di previsione quest'ultimo, con un MSE pari a 13.06790.