Progetto modelli demografici e previdenza

Giorgio Marletta matricola 1000007271

Analisi sull’aspettativa di vita OMS

Premessa: la tematica e l’esame preliminare dei dati

Il presente studio si propone di analizzare l'aspettativa di vita, esaminando le correlazioni tra questo indicatore e diversi fattori socioeconomici e sanitari. Mediante l’utilizzo di dati provenienti dall'Organizzazione Mondiale della Sanità (OMS), l'obiettivo principale è quello di indagare come variabili quali l'accesso ai servizi sanitari, l'istruzione, la spesa per la salute e altre caratteristiche socioeconomiche possano influenzare l'aspettativa di vita nei vari paesi. Ai fini di detto studio, sono stati presi in considerazione i dati relativi agli anni 2000 e 2014 di ciascun paese, consentendo così di osservare l'evoluzione dell'aspettativa di vita e dei fattori correlati nel corso di questo periodo.

L'aspettativa di vita è un indicatore cruciale per valutare lo stato di salute di una popolazione e la qualità della vita all'interno di una nazione. Tuttavia, essa è influenzata da un ampio insieme di fattori, tra cui le condizioni economiche, le politiche sanitarie adottate, i determinanti sociali e le malattie prevalenti. Questo studio mira a esplorare come tali variabili interagiscano tra loro, determinando in che misura ciascuna di esse contribuisca alla durata media della vita in vari contesti globali, sia nel 2000 che nel 2014.

Onde conseguire tale obiettivo, viene utilizzato un modello di regressione lineare multipla, che permette di analizzare le interazioni tra diverse variabili esplicative e la variabile di risposta, ossia l'aspettativa di vita.

Inoltre, il confronto tra questi due anni offre l'opportunità di evidenziare i cambiamenti significativi che si sono verificati nell’ intervallo temporale. I risultati ottenuti non solo contribuiranno a una maggiore comprensione delle dinamiche che influenzano la salute delle popolazioni, ma forniranno anche spunti pratici per sviluppare politiche sanitarie più mirate ed efficaci, in grado di migliorare la qualità della vita nei vari paesi.

In sintesi, l'analisi delle variabili che determinano l'aspettativa di vita offre una visione complessa delle sfide sanitarie ed economiche globali e, al contempo, fornisce orientamenti concreti per una gestione più efficace della salute pubblica a livello internazionale.

Nella seguente tabella vengono riportate le variabili incluse nel modello:

Variabile	Descrizione
Life_EXP (Y)	*Aspettativa di vita in età*: La durata media della vita in una determinata popolazione, utilizzata come variabile dipendente.
thinness_5_9 (X1)	*Prevalenza della magrezza tra i bambini (5-9 anni)*: Percentuale di bambini di età compresa tra 5 e 9 anni con indice di massa corporea (IMC) basso.
Diphtheria (X2)	*Copertura vaccinale DTP3*: Percentuale di bambini di 1 anno vaccinati contro difterite, tetano, tossoide e pertosse.
exp_tot (X3)	*Spesa pubblica per la salute (%)*: Percentuale della spesa pubblica totale destinata alla salute nel paese.
HIV (X4)	*Decessi per HIV/AIDS (0-4 anni)*: Tasso di mortalità infantile dovuto all'HIV/AIDS, misurato per 1.000 nati vivi.
GDP (X5)	*Prodotto interno lordo pro capite (USD)*: Misura del reddito medio per persona, utilizzato come indicatore economico del paese.
Schooling (X6)	*Anni di istruzione*: Numero medio di anni di istruzione formale completata dalla popolazione adulta di un paese.

Le variabili incluse nel modello riflettono una serie di fattori socioeconomici e sanitari che condizionano l'aspettativa di vita. L'aspettativa di vita (Life_EXP) è la variabile dipendente, mentre le variabili indipendenti esplorano determinanti chiave del benessere e della salute.

La prevalenza della magrezza tra i bambini (thinness_5_9) e la copertura vaccinale (Diphtheria) sono indicatori cruciali dello stato di salute infantile e dell'efficacia dei programmi sanitari. La spesa pubblica per la salute (exp_tot), come percentuale della spesa totale, evidenzia l'impegno dei paesi nell'investire nel benessere sanitario.

Il tasso di mortalità infantile per HIV/AIDS (HIV) e il prodotto interno lordo pro capite (GDP) sono utilizzati per analizzare l'impatto delle malattie infettive e dello sviluppo economico sulla salute della popolazione.

Infine, il numero di anni di istruzione (Schooling) è considerato un ulteriore fattore determinante, poiché una maggiore educazione è generalmente associata a migliori condizioni di salute.

Il dataset utilizzato per questa analisi copre il periodo dal 2000 al 2014, ma essa si concentra specificamente su due anni di riferimento: il 2000 e il 2014. Per ciascuno di questi due anni, il dataset include 183 osservazioni, ognuna delle quali rappresenta un'unità statistica. Ogni unità è descritta attraverso sette variabili indipendenti, scelte per il loro potenziale impatto sull'aspettativa di vita (Life_EXP), che costituisce la variabile dipendente del modello.

Oltre a queste variabili, il dataset comprende le colonne "Country" (Paese) e "Year" (Anno), che identificano rispettivamente il paese e l'anno di osservazione. La struttura complessiva dei dati che si sviluppa, quindi, su 183 righe e 7 colonne per ciascun anno analizzato, offre così una base robusta per l'elaborazione del modello di regressione lineare multipla.

	Country	Year	Life_EXP	exp_tot	Diphtheria	thinness_5_9	HIV	GDP	Schooling
1	Afghanistan	2000	54.8	8.20	24	2.5	0.1	114.560000	5.5
2	Albania	2000	72.6	6.26	97	2.2	0.1	1175.788981	10.7
3	Algeria	2000	71.3	3.49	86	6.4	0.1	1757.177970	10.7
4	Angola	2000	45.3	2.79	28	1.9	2.0	555.296942	4.6
5	Antigua and Barbuda	2000	73.6	4.13	95	3.6	0.1	9875.161736	0.0
6	Argentina	2000	74.1	9.21	83	1.1	0.1	7669.273916	15.0
7	Armenia	2000	72.0	6.25	93	2.2	0.1	622.742748	11.2
8	Australia	2000	79.5	8.80	9	0.7	0.1	2169.921000	20.4
9	Austria	2000	78.1	1.60	81	1.9	0.1	24517.267450	15.4
10	Azerbaijan	2000	66.6	4.67	76	3.1	0.1	655.974326	10.1
11	Bahamas	2000	72.6	5.21	99	2.6	0.1	*NA*	12.0
12	Bahrain	2000	74.5	3.51	97	6.0	0.1	13636.346680	13.2
13	Bangladesh	2000	65.3	2.33	82	21.5	0.1	45.633710	7.3
14	Barbados	2000	73.3	5.16	93	4.2	0.9	11568.111100	14.0
15	Belarus	2000	68.0	6.13	99	2.8	0.1	1276.288340	13.1
16	Belgium	2000	77.6	8.12	95	0.8	0.1	2327.459100	18.0
17	Belize	2000	68.3	3.98	91	3.7	0.3	3364.423711	11.7
18	Benin	2000	55.4	4.34	78	9.6	2.0	374.192394	6.4
19	Bhutan	2000	62.0	6.91	92	19.9	0.1	765.863236	7.3
20	Bolivia (Plurinational State of)	2000	62.6	5.67	75	1.4	0.1	*NA*	13.3
21	Bosnia and Herzegovina	2000	74.6	7.90	85	3.2	0.1	1461.755200	0.0

	Country	Year	Life_EXP	exp_tot	Diphtheria	thinness_5_9	HIV	GDP	Schooling
1	Afghanistan	2014	59.9	8.18	62	17.5	0.1	612.69651	10.0
2	Albania	2014	77.5	5.88	98	1.3	0.1	4575.76379	14.2
3	Algeria	2014	75.4	7.21	95	5.8	0.1	547.85170	14.4
4	Angola	2014	51.7	3.31	64	8.3	2.0	479.31224	11.4
5	Antigua and Barbuda	2014	76.2	5.54	99	3.3	0.2	12888.29667	13.9
6	Argentina	2014	76.2	4.79	94	0.9	0.1	12245.25645	17.3
7	Armenia	2014	74.6	4.48	93	2.1	0.1	3994.71236	12.7
8	Australia	2014	82.7	9.42	92	0.6	0.1	62214.69120	20.4
9	Austria	2014	81.4	11.21	98	2.0	0.1	51322.63997	15.9
10	Azerbaijan	2014	72.5	6.40	94	2.9	0.1	7891.29978	12.2
11	Bahamas	2014	75.4	7.74	96	2.5	0.1	*NA*	12.6
12	Bahrain	2014	76.8	4.98	98	6.0	0.1	24983.37920	14.5
13	Bangladesh	2014	71.4	2.82	97	18.6	0.1	184.56543	10.0
14	Barbados	2014	75.4	7.47	94	3.7	0.1	15359.66971	15.3
15	Belarus	2014	72.0	5.69	97	2.0	0.1	8318.42929	15.7
16	Belgium	2014	89.0	1.59	99	1.0	0.1	47439.39684	16.3
17	Belize	2014	70.0	5.79	95	3.4	0.2	4852.22367	12.8
18	Benin	2014	59.7	4.59	78	6.9	1.1	943.68657	10.7
19	Bhutan	2014	69.4	3.57	99	16.2	0.5	2522.79680	12.5
20	Bolivia (Plurinational State of)	2014	74.0	6.33	98	1.1	0.1	*NA*	13.8
21	Bosnia and Herzegovina	2014	77.2	9.57	86	2.4	0.1	5193.94932	14.2

Prima di procedere con la formulazione del modello di regressione lineare multipla, è fondamentale avviare un'analisi esplorativa dei dati, esaminando le proprietà statistiche delle variabili per ciascun anno di riferimento, ossia il 2000 e il 2014. Per i due anni, quindi, è necessario analizzare anzitutto le principali statistiche descrittive, tra cui i valori minimi e massimi, la media, la mediana, e i quartili (primo e terzo quartile). Questo approccio consente di ottenere una visione d'insieme della distribuzione delle variabili in entrambi gli anni, facilitando l'identificazione di eventuali differenze o tendenze particolari.

Le statistiche relative all'anno 2000 e al 2014 sono presentate nelle immagini sottostanti, rispettivamente, in ordine cronologico, per consentire una comparazione diretta tra i due periodi e tale ordine cronologico sarà ripetuto in tutta l'analisi.

2000

```
   Life_EXP      thinness_5_9      Diphtheria       exp_tot           HIV              GDP
Min.   :39.00   Min.   : 0.100   Min.   : 3.00   Min.   : 1.100   Min.   : 0.10   Min.   :    3.69
1st Qu.:58.65   1st Qu.: 1.600   1st Qu.:58.75   1st Qu.: 4.165   1st Qu.: 0.10   1st Qu.:  263.15
Median :71.00   Median : 3.400   Median :86.00   Median : 5.420   Median : 0.10   Median :  828.99
Mean   :66.75   Mean   : 5.245   Mean   :73.63   Mean   : 5.585   Mean   : 2.53   Mean   : 4708.52
3rd Qu.:74.45   3rd Qu.: 7.900   3rd Qu.:96.00   3rd Qu.: 6.940   3rd Qu.: 1.10   3rd Qu.: 3361.50
Max.   :81.10   Max.   :28.600   Max.   :99.00   Max.   :13.700   Max.   :46.40   Max.   :48736.00
                NA's   :2        NA's   :3       NA's   :4                        NA's   :29
  Schooling
Min.   : 0.00
1st Qu.: 8.00
Median :11.40
Mean   :10.51
3rd Qu.:13.20
Max.   :20.40
NA's   :10
```

2014

```
   Life_EXP      thinness_5_9      Diphtheria       exp_tot           HIV              GDP
Min.   :48.10   Min.   : 0.100   Min.   : 2.00   Min.   : 1.210   Min.   :0.100   Min.   :    12.28
1st Qu.:65.60   1st Qu.: 1.500   1st Qu.:83.00   1st Qu.: 4.480   1st Qu.:0.100   1st Qu.:   617.99
Median :73.60   Median : 3.400   Median :94.00   Median : 5.840   Median :0.100   Median :  3154.51
Mean   :71.54   Mean   : 4.676   Mean   :84.08   Mean   : 6.201   Mean   :0.682   Mean   : 10015.57
3rd Qu.:76.85   3rd Qu.: 6.600   3rd Qu.:97.00   3rd Qu.: 7.740   3rd Qu.:0.400   3rd Qu.:  8239.95
Max.   :89.00   Max.   :27.400   Max.   :99.00   Max.   :17.140   Max.   :9.400   Max.   :119172.74
                NA's   :2                        NA's   :2                        NA's   :28
  Schooling
Min.   : 4.90
1st Qu.:10.80
Median :13.00
Mean   :12.89
3rd Qu.:14.90
Max.   :20.40
NA's   :10
```

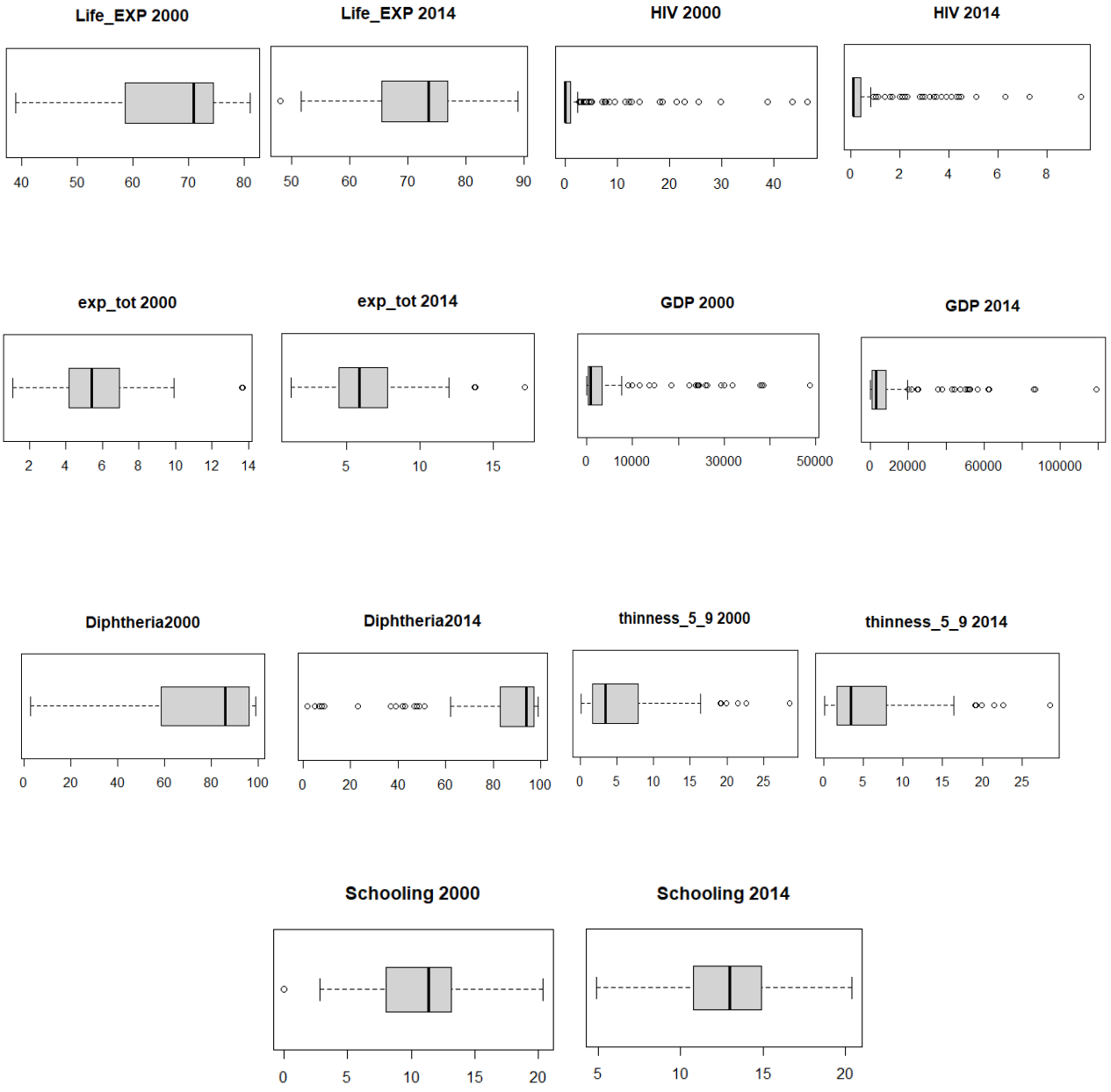
Dalla prima analisi descrittiva di tali variabili, emerge anzitutto una tendenza positiva nell'aspettativa di vita (Life_EXP), come evidenziato dall'incremento dei principali indicatori statistici della distribuzione. Nel 2014, infatti, si registra un aumento dei valori minimo, massimo, della media, del primo quartile e del terzo quartile rispetto al 2000, suggerendo un miglioramento generale delle condizioni di vita.

Un andamento simile si osserva per le variabili GDP (PIL pro capite) e Schooling (anni di istruzione), le quali riflettono un progresso economico e un accesso più ampio all'istruzione, fattori che contribuiscono significativamente al benessere della popolazione.

Di contro, la variabile HIV, rappresentativa del tasso di mortalità infantile associato all'HIV/AIDS, mostra una diminuzione nel periodo considerato. Questo calo potrebbe essere attribuito ai progressi nei programmi di prevenzione e trattamento della malattia, nonché a una maggiore consapevolezza sanitaria.

Per le altre variabili analizzate, tuttavia, non si riscontrano variazioni significative tra i due anni presi in esame, suggerendo una relativa stabilità negli indicatori ad esse associati.

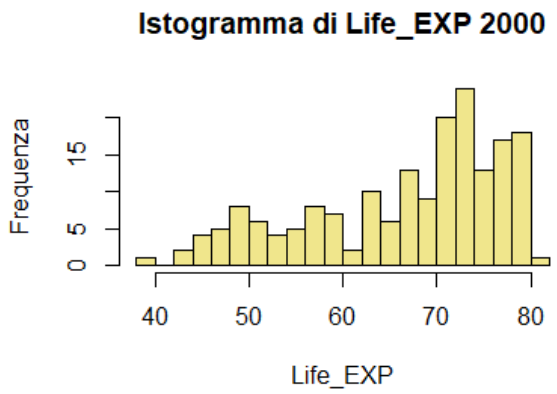
Per rappresentare graficamente le variabili analizzate, è stato utilizzato per prima cosa il diagramma a scatola e baffi (box plot), uno strumento efficace per sintetizzare visivamente le caratteristiche principali di una distribuzione. Il box plot, in sintesi, evidenzia la forma della distribuzione, il grado di dispersione dei dati e la presenza di eventuali valori anomali (outlier). Ogni box rappresenta l'intervallo interquartile (IQR), che include il 50% centrale dei dati, mentre i "baffi" si estendono fino ai valori massimi e minimi che non sono considerati outlier. La linea all'interno del box indica la mediana, un importante indicatore di posizione centrale.

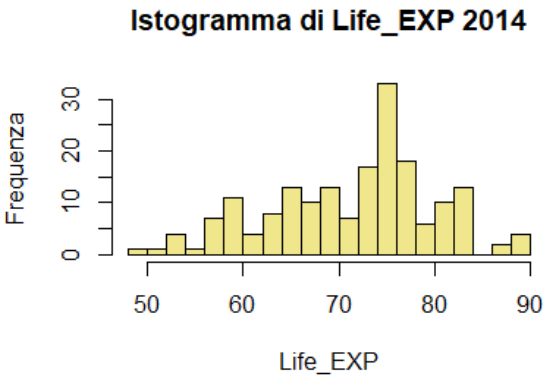


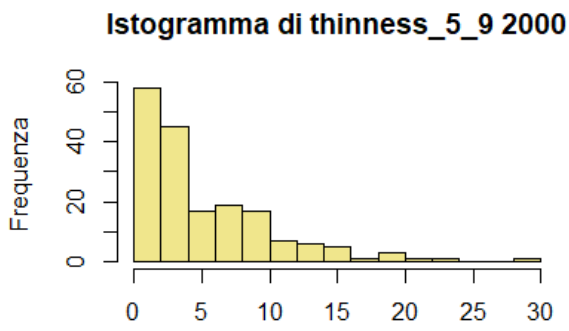
Il confronto dei box plot tra il 2000 e il 2014 per la variabile Life_EXP conferma quanto osservato nelle statistiche descrittive: infatti, l'intera scatola, che rappresenta la distribuzione dei dati, si sposta verso destra nel 2014, indicando un aumento dell'aspettativa di vita. Una tendenza simile è visibile anche per le variabili GDP e Schooling, che mostrano una traslazione verso destra, suggerendo un miglioramento economico e nell'accesso all'istruzione.

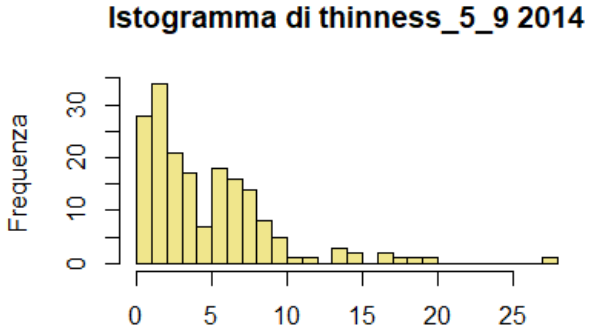
Al contrario, per la variabile HIV, la scatola si sposta verso sinistra, evidenziando una riduzione del tasso di mortalità infantile legato all'HIV/AIDS, in linea con i risultati precedenti dell'analisi descrittiva.

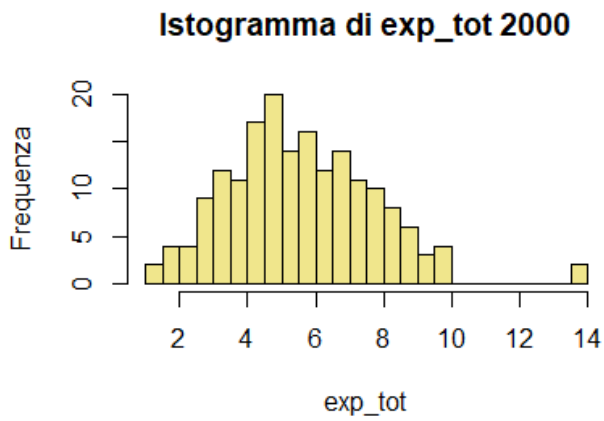
In secondo luogo, un altro step nell'analisi descrittiva delle variabili consiste nella creazione degli istogrammi, i quali forniscono una rappresentazione visiva della forma e dell'andamento della distribuzione.

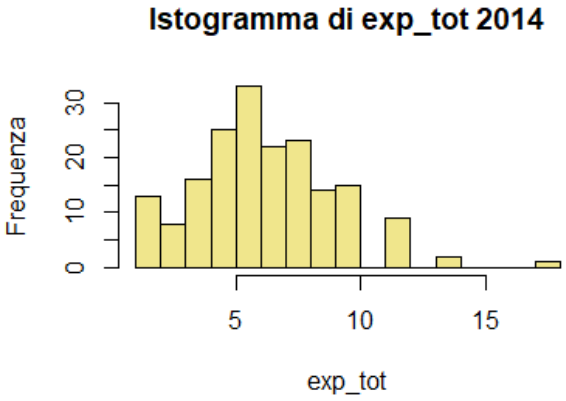


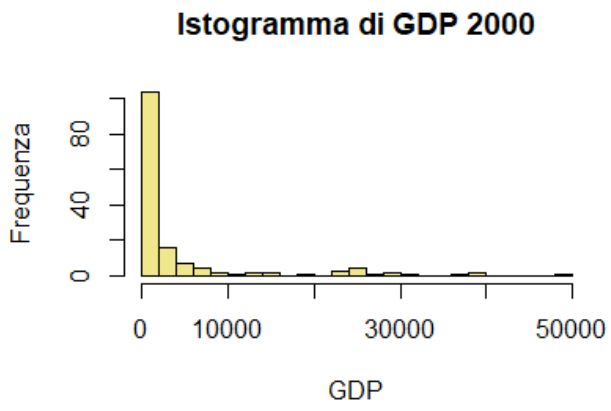


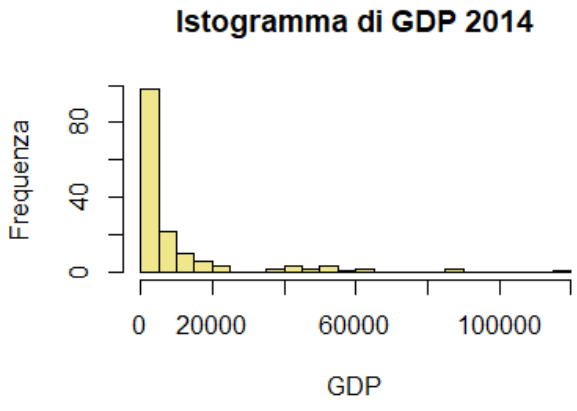


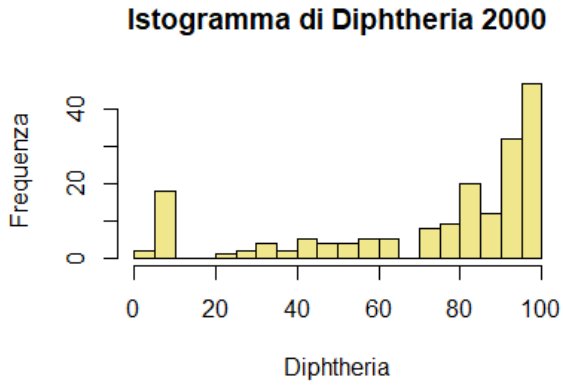


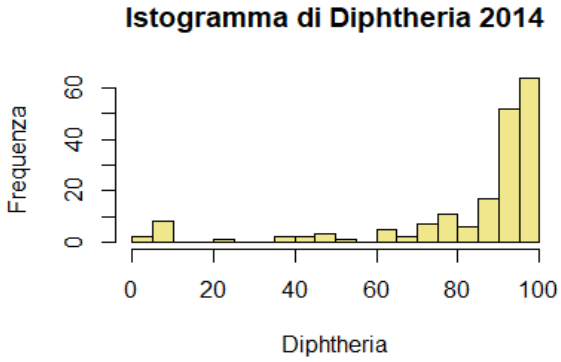


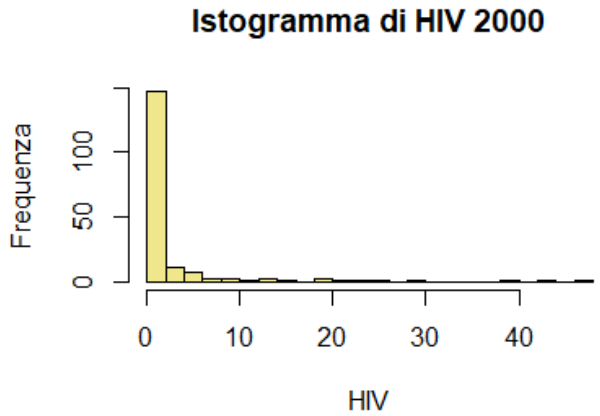


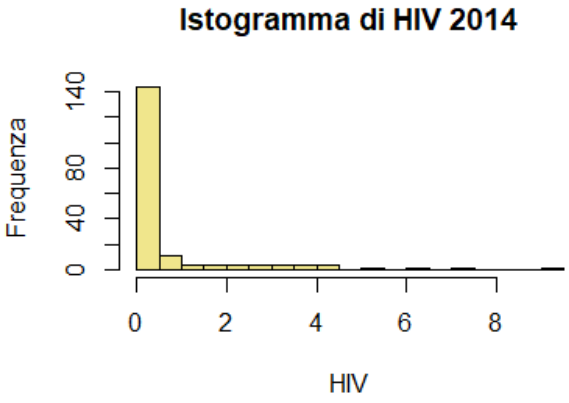

Istogramma di GDP 2000
Frequenza
0
40
80
0
10000
30000
50000
GDP

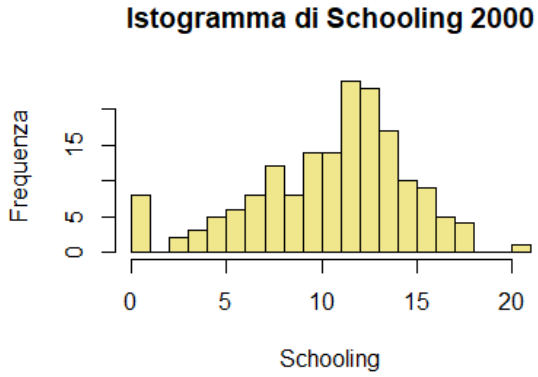

Istogramma di GDP 2014
Frequenza
0
40
80
0
20000
60000
100000
GDP

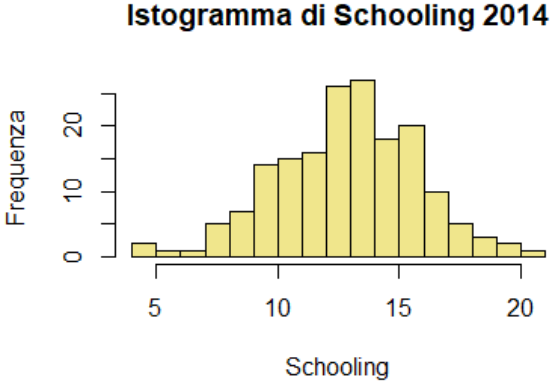

Istogramma di Diphtheria 2000
Frequenza
0
20
40
0
20
40
60
80
100
Diphtheria


Istogramma di Diphtheria 2014
Frequenza
0
20
40
60
0
20
40
60
80
100
Diphtheria


Istogramma di HIV 2000
Frequenza
0
50
100
0
10
20
30
40
HIV


Istogramma di HIV 2014
Frequenza
0
40
80
140
0
2
4
6
8
HIV


Istogramma di Schooling 2000
Frequenza
0
5
15
0
5
10
15
20
Schooling


Istogramma di Schooling 2014
Frequenza
0
10
20
5
10
15
20
Schooling

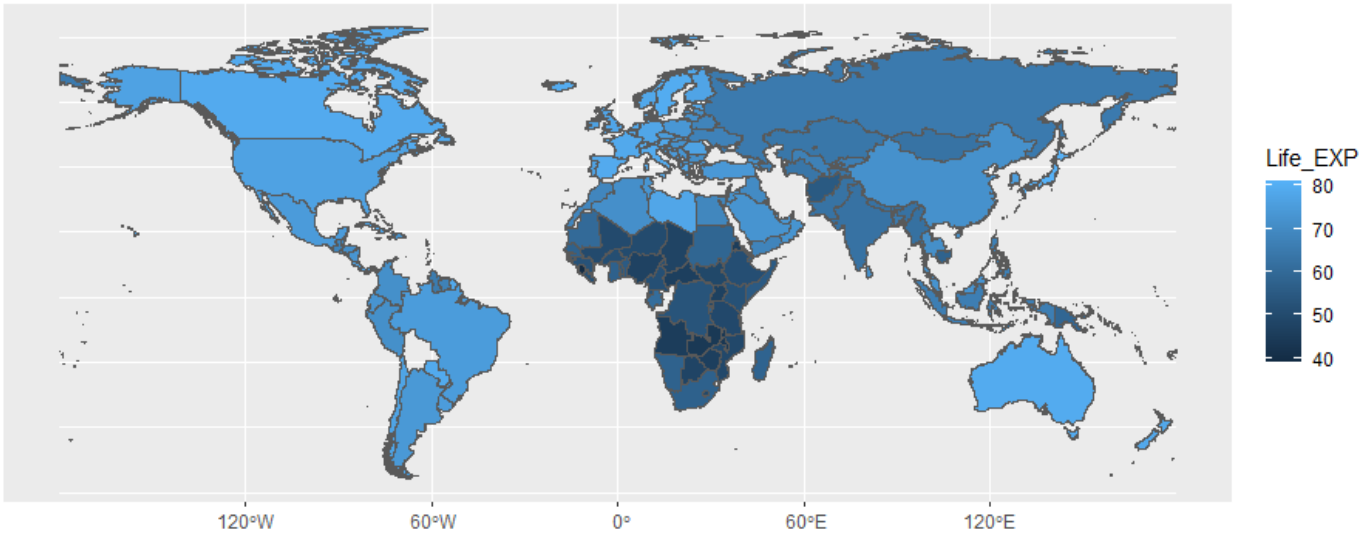
Dall'osservazione degli istogrammi, si rileva che alcune variabili seguono una distribuzione con forma campanulare, come Life_EXP nel 2014 e Schooling, il che indica una distribuzione simmetrica attorno alla media, con la maggior parte dei valori concentrati al centro e una riduzione graduale verso le code.

Al contrario, in generale, quasi tutte le altre variabili presentano un'asimmetria, positiva o negativa. L'asimmetria positiva indica che la distribuzione ha una lunga coda verso destra, con valori più alti meno frequenti ma comunque presenti, mentre l'asimmetria negativa si verifica quando la coda si estende verso sinistra, con valori bassi meno frequenti. Queste osservazioni sulle forme della distribuzione confermano quanto emerso dalle analisi statistiche precedenti, evidenziando ulteriormente le tendenze generali e le peculiarità di ciascuna variabile.

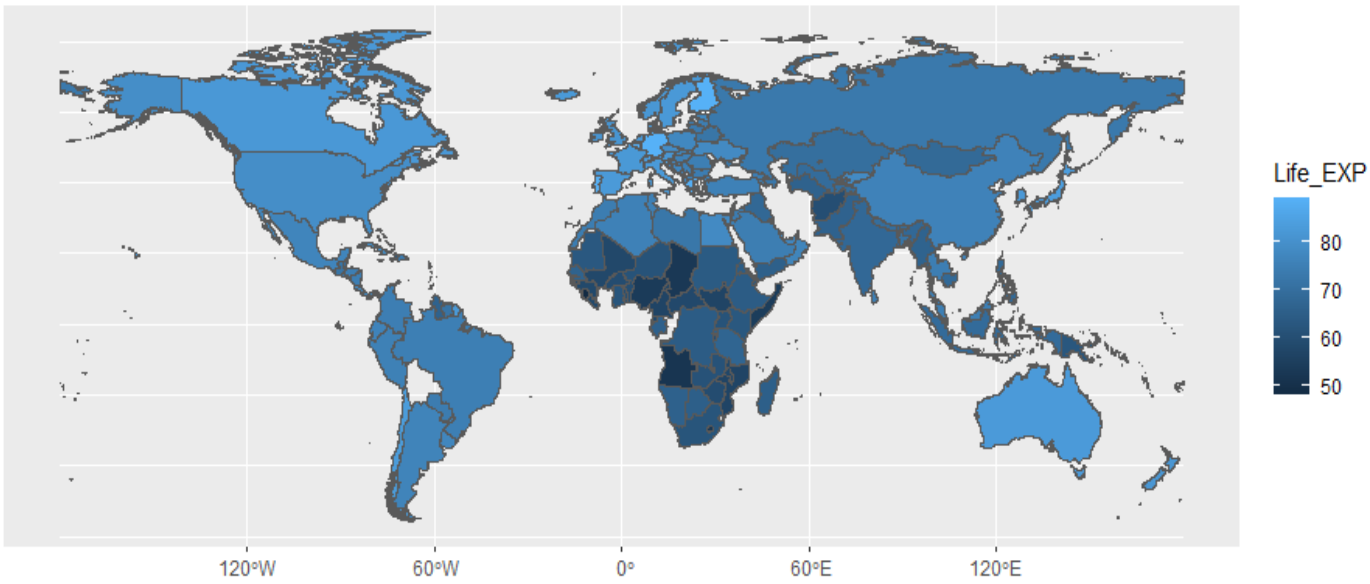
Un’ulteriore estensione interessante di tale analisi è rappresentata dalla rappresentazione di una mappa geografica della distribuzione, che consente di visualizzare come la variabile Life_EXP si distribuisca a livello globale nei due anni di riferimento. Considerando la dimensione cross-country dello studio, questa visualizzazione aggiunge una prospettiva utile per comprendere le disuguaglianze geografiche. La mappa è colorata in sfumature di blu, con un blu chiaro che indica un’aspettativa di vita più alta e un blu scuro che riflette valori più bassi.

È immediatamente intuibile che i paesi con un’aspettativa di vita più bassa si concentrano principalmente nell'Africa subsahariana, mentre i paesi europei e le regioni delle Americhe presentano valori significativamente più elevati. Questa rappresentazione visiva conferma e amplifica le osservazioni fatte in precedenza, fornendo un quadro chiaro delle disparità globali in termini di salute e benessere.

2000



2014



Prima di procedere con la definizione e l'attuazione del modello di regressione, l'ultimo passaggio chiave nell'analisi preliminare dei dati consiste nell'esplorare la struttura di correlazione tra le variabili per identificare potenziali associazioni. Sebbene la matrice varianza-covarianza possa essere uno strumento utile in questo contesto, è stato preferito utilizzare la matrice di correlazione, che fornisce una visione più chiara e intuitiva delle relazioni tra le variabili.

La matrice di correlazione riporta i coefficienti di correlazione, che vanno da -1 a +1, indicando rispettivamente la forza e la direzione di una relazione lineare tra le coppie di variabili. Questo strumento è particolarmente utile perché normalizza le variabili, permettendo di comparare facilmente variabili con scale diverse e facilitando l'individuazione di relazioni lineari.

Per questo studio, abbiamo calcolato e analizzato separatamente la struttura di correlazione per i dati relativi all'anno 2000 e per quelli del 2014. Tale approccio ci ha consentito di confrontare le relazioni tra le variabili nei due periodi, evidenziando eventuali cambiamenti o tendenze nel tempo.

2000

	Life_EXP	thinness_5_9	Diphtheria	exp_tot	HIV	GDP	Schooling
Life_EXP	1.0000000	-0.4219799	0.44360786	0.198390968	-0.553015781	0.45987371	0.6639453
thinness_5_9	-0.4219799	1.0000000	-0.13223665	-0.315684298	0.303122265	-0.30502579	-0.3164418
Diphtheria	0.4436079	-0.1322366	1.00000000	0.147462699	-0.072802967	0.08303392	0.3160808
exp_tot	0.1983910	-0.3156843	0.14746270	1.000000000	0.005979386	0.16939307	0.2600775
HIV	-0.5530158	0.3031223	-0.07280297	0.005979386	1.000000000	-0.14894605	-0.1159210
GDP	0.4598737	-0.3050258	0.08303392	0.169393069	-0.148946047	1.00000000	0.4536317
Schooling	0.6639453	-0.3164418	0.31608082	0.260077481	-0.115920977	0.45363173	1.0000000

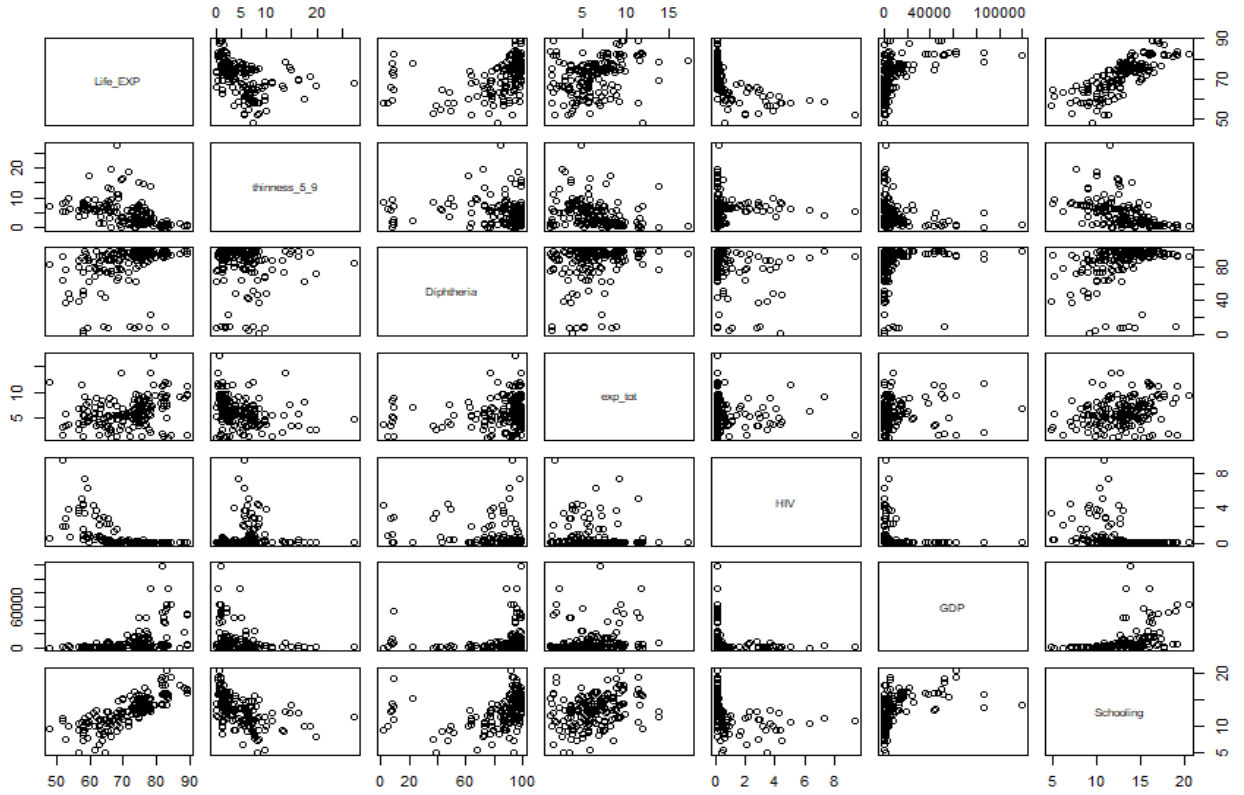
2014

	Life_EXP	thinness_5_9	Diphtheria	exp_tot	HIV	GDP	Schooling
Life_EXP	1.0000000	-0.4885047	0.3759239	0.31293377	-0.6104309	0.46489712	0.8076402
thinness_5_9	-0.4885047	1.0000000	-0.1102223	-0.29080968	0.1743662	-0.27917910	-0.5155893
Diphtheria	0.3759239	-0.1102223	1.0000000	0.20119175	-0.2233369	0.12928682	0.3111725
exp_tot	0.3129338	-0.2908097	0.2011918	1.00000000	-0.1101382	0.08495569	0.2953780
HIV	-0.6104309	0.1743662	-0.2233369	-0.11013818	1.0000000	-0.18871318	-0.3904042
GDP	0.4648971	-0.2791791	0.1292868	0.08495569	-0.1887132	1.00000000	0.4359035
Schooling	0.8076402	-0.5155893	0.3111725	0.29537798	-0.3904042	0.43590355	1.0000000

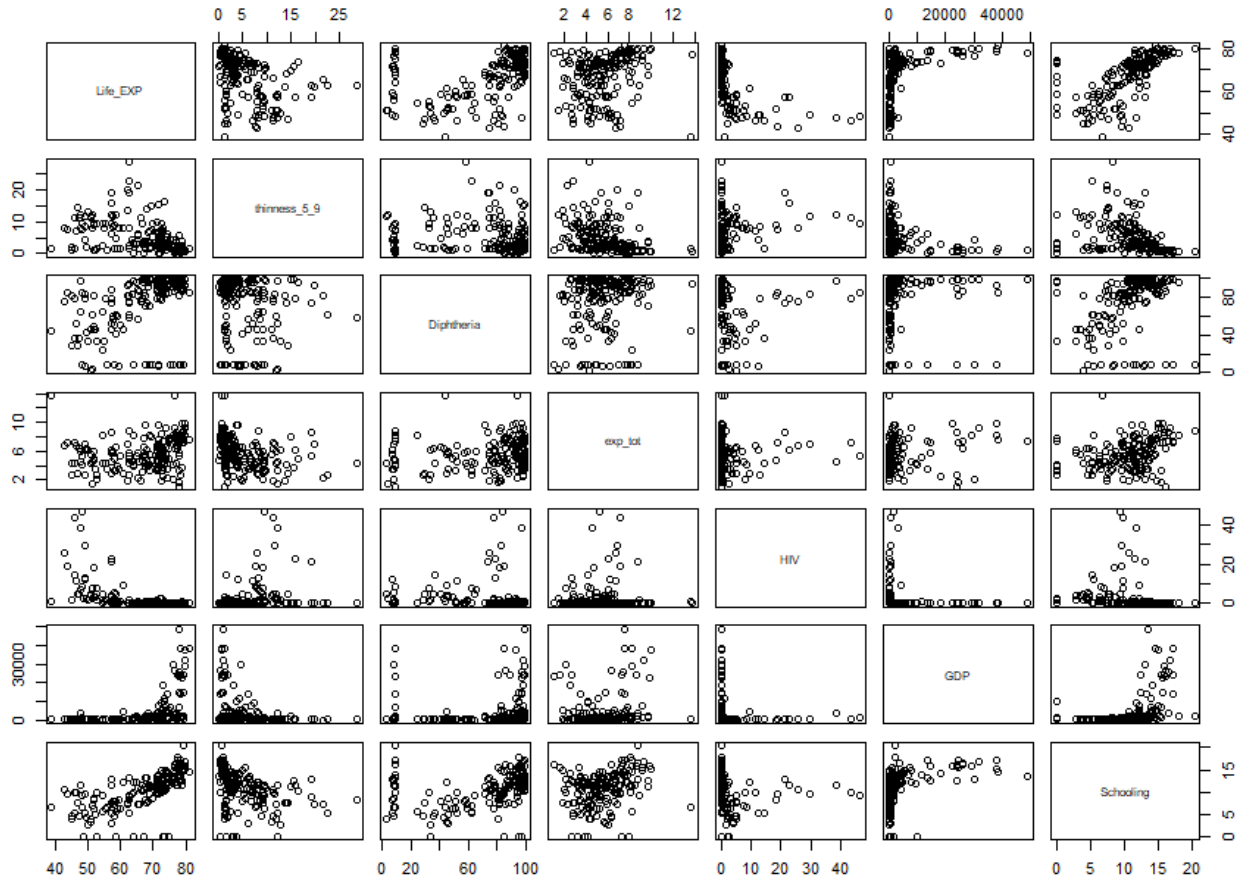
In supporto a questo studio, è stato realizzato il grafico a dispersione, noto come "Scatterplot", per visualizzare graficamente la relazione tra coppie di variabili. Il grafico a dispersione rappresenta ogni punto come una coppia di variabili per una specifica osservazione nel dataset, con una variabile sull'asse delle ascisse (x) e l'altra sull'asse delle ordinate (y).

L'analisi di più coppie di variabili, in sintesi, consente soprattutto di valutare se la relazione tra esse sia lineare, di tipo curvilineo o più complessa, facilitando la comprensione delle dinamiche tra le variabili.

2000

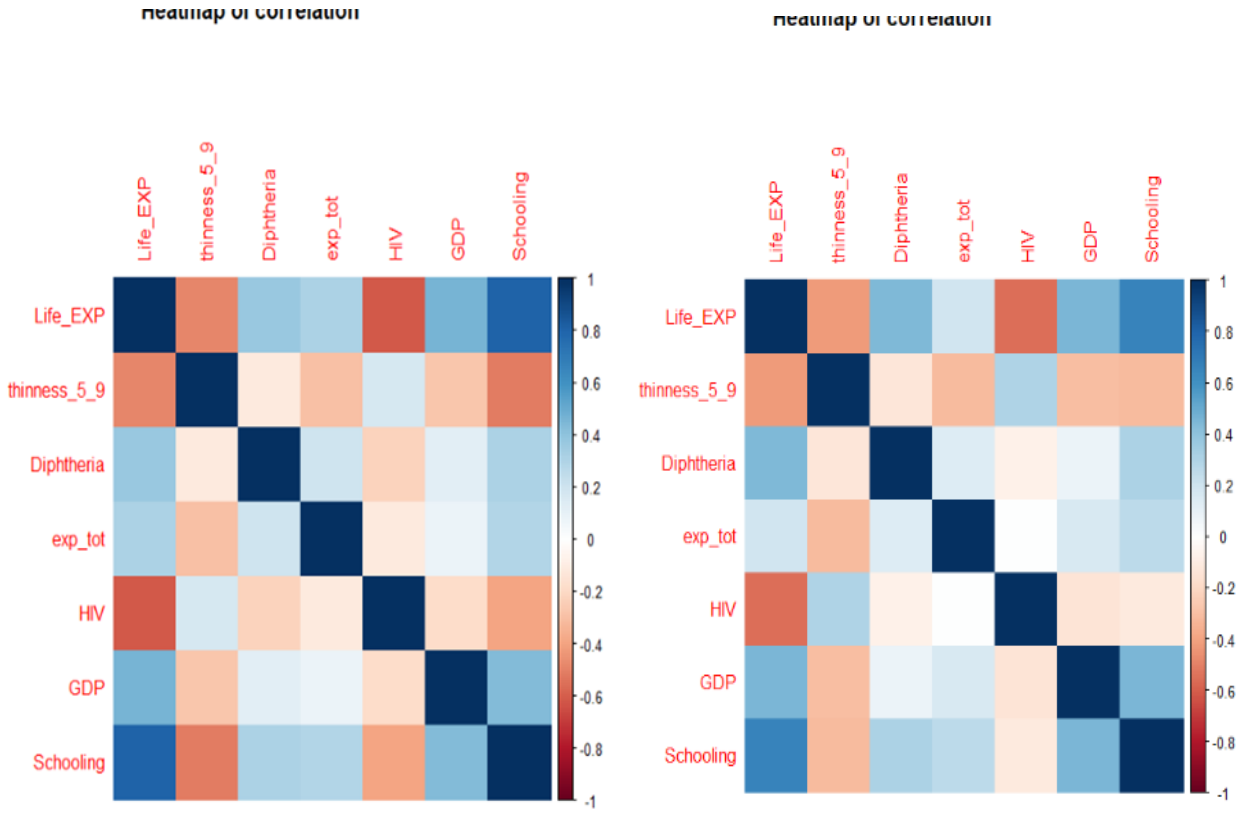

Life_EXP
thinness_5_9
Diphtheria
exp_tot
HIV
GDP
Schooling

2014


Life_EXP
thinness_5_9
Diphtheria
exp_tot
HIV
GDP
Schooling

Infine, in aggiunta alla classica rappresentazione grafica per mezzo dei grafici a dispersione, un mezzo alternativo e di più immediata lettura è la c.d. “Heatmap of correlation”. Esso utilizza una rappresentazione grafica basata sui colori per mostrare i diversi livelli di correlazione tra coppie di variabili nel nostro dataset.

Ogni cella della heatmap è colorata in base al valore della correlazione tra la variabile sull'asse x e quella sull'asse y, rendendo facile e immediata l'interpretazione delle relazioni tra le variabili, rispetto alla tabella di correlazione tradizionale. I colori intensi indicano correlazioni più forti, mentre colori più chiari o neutri indicano correlazioni più deboli o prossime allo zero. Specificando in maniera ancora più particolare, i colori tendenti al blu indicano una correlazione positiva tra variabili, mentre colori tendenti al rosso indicano una correlazione negativa tra quest’ultime.



Le heatmap of correlation (rispettivamente per il 2000 e 2014) mostrano chiaramente che la nostra variabile di interesse, Life_EXP, è positivamente correlata con Diphtheria, GDP, Schooling ed Exp_tot, mentre è negativamente correlata con Thinness (5-9) e HIV. Questo andamento si riscontra sia nei dati del 2000 che in quelli del 2014, il che rafforza la solidità e la coerenza delle variabili alla base del modello.

Inoltre, dalle heatmap emerge che nel dataset non vi sono evidenti problematiche di perfetta multicollinearità, che potrebbero compromettere l'affidabilità delle stime nei modelli di regressione successivi. La multicollinearità è opportuno ricordare che si verifica quando due o più variabili indipendenti in un modello di regressione sono fortemente correlate tra loro. L’assenza di tale aspetto è fondamentale per garantire che le relazioni tra le variabili siano stabili e le stime siano maggiormente attendibili.

Conclusa l'analisi esplorativa del dataset, si avvia la costruzione dei modelli di regressione, partendo dai modelli lineari semplici. Questi rappresentano il punto di partenza per un'analisi più approfondita e strutturata, finalizzata a comprendere le relazioni tra le variabili.

La regressione lineare semplice analizza la relazione tra la variabile dipendente Life_EXP e un singolo regressore, Schooling.

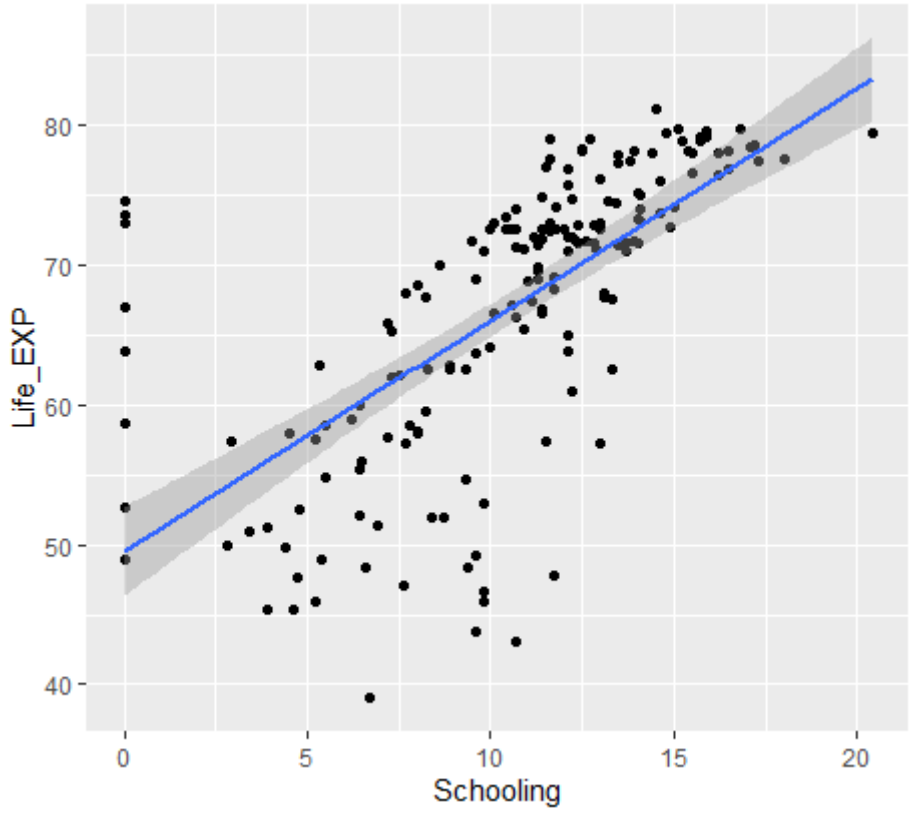
2000

```
Call:
lm(formula = Life_EXP ~ Schooling, data = OMS_2000)

Residuals:
    Min      1Q  Median      3Q     Max
-24.112  -3.481   1.040   4.231  25.081

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.5187     1.6061   30.83   <2e-16 ***
Schooling     1.6536     0.1424   11.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.645 on 171 degrees of freedom
  (10 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.4408,    Adjusted R-squared:  0.4376
F-statistic: 134.8 on 1 and 171 DF,  p-value: < 2.2e-16
```



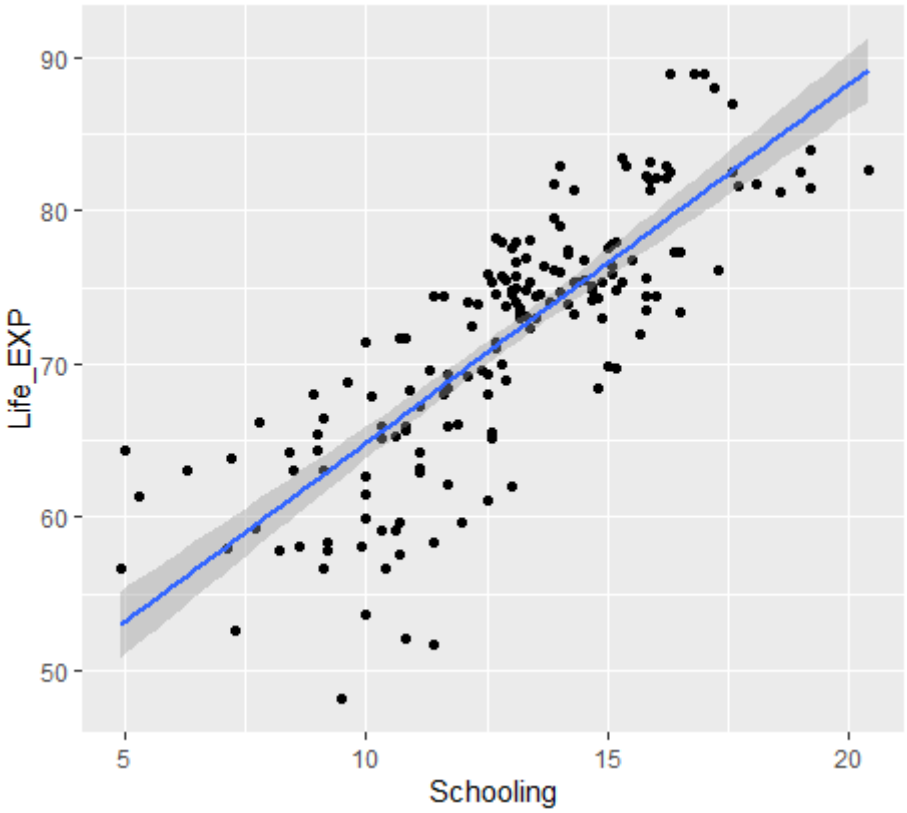
2014

```
Call:
lm(formula = Life_EXP ~ Schooling, data = OMS_2014)

Residuals:
     Min       1Q   Median       3Q      Max
-16.4558  -2.8769   0.3196   3.6196  11.2020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.5123     1.7238   24.08   <2e-16 ***
Schooling     2.3371     0.1305   17.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.983 on 171 degrees of freedom
  (10 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.6523,    Adjusted R-squared:  0.6502
F-statistic: 320.8 on 1 and 171 DF,  p-value: < 2.2e-16
```



Nei modelli analizzati, relativi agli anni 2000 e 2014, Schooling mostra una correlazione positiva e statisticamente significativa con Life_EXP, come evidenziato dai coefficienti stimati (1,6536 per il 2000 e 2,3371 per il 2014) e dai valori di p-value inferiori a 0,001.

Il p-value, o valore di probabilità, ricordiamo, è una misura che quantifica la probabilità di osservare un effetto almeno altrettanto estremo di quello rilevato, assumendo che l’ipotesi nulla sia vera. Un p-value inferiore a 0,001 indica una forte evidenza contro l'ipotesi nulla, confermando la significatività statistica dei risultati.

Entrambi i modelli presentano anche un’intercetta positiva, anch'essa significativa dal punto di vista statistico. Un ulteriore parametro di interesse è l’R^2 (coefficiente di determinazione), che rappresenta la proporzione di

variabilità di Life_EXP spiegata dal modello. I valori di R^2 (0,44 per il 2000 e 0,65 per il 2014) indicano un buon livello di adattamento del modello, sebbene potrebbero aumentare aggiungendo ulteriori regressori, data la natura non decrescente di questo indicatore all'aumentare delle variabili indipendenti che si aggiungono nel modello. La bontà complessiva del modello è confermata dall'F-test, che valuta la significatività del modello rispetto a un modello nullo, mostrando valori sufficientemente elevati in entrambi i casi.

Infine, i modelli sono rappresentati graficamente tramite diagrammi di dispersione della coppia Life_EXP-Schooling, con la retta di regressione calcolata tramite il metodo dei minimi quadrati. L'area grigia attorno alla retta rappresenta l'errore standard, fornendo una misura della precisione delle stime. Questi grafici visualizzano chiaramente la relazione positiva tra le due variabili e rafforzano l'interpretazione dei risultati.

A partire dal modello bivariato semplice, si procede alla costruzione di modelli multivariati che considerano simultaneamente più regressori. Per ottenere questo risultato, si utilizza il metodo della "Forward Selection", una tecnica iterativa che aggiunge progressivamente variabili al modello. Ogni variabile viene inclusa sulla base della sua significatività statistica e della sua capacità di migliorare l'adattamento del modello ai dati, misurato tramite indicatori come l'R^2 e il p-value. Questo approccio consente di identificare il set ottimale di regressori, massimizzando la capacità predittiva del modello e offrendo una visione più completa delle relazioni tra le variabili in esame.

2000

```
Call:
lm(formula = Life_EXP ~ GDP + Schooling, data = OMS_2000)

Residuals:
     Min       1Q   Median       3Q      Max
-22.9026  -3.0827   0.6219   4.3821  24.2521

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.003e+01  1.751e+00  28.573  < 2e-16 ***
GDP         2.206e-04  7.463e-05   2.956  0.00362 **
Schooling   1.490e+00  1.665e-01   8.951  1.2e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.553 on 151 degrees of freedom
  (29 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.478
F-statistic: 71.06 on 2 and 151 DF,  p-value: < 2.2e-16
```

2014

```
Call:
lm(formula = Life_EXP ~ GDP + Schooling, data = OMS_2014)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6766  -3.1067   0.5743   3.3876   9.5939

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.153e+01  1.919e+00  21.645  < 2e-16 ***
GDP         6.300e-05  2.406e-05   2.618  0.00974 **
Schooling   2.264e+00  1.515e-01  14.949  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.963 on 151 degrees of freedom
  (29 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.6838,    Adjusted R-squared:  0.6796
F-statistic: 163.2 on 2 and 151 DF,  p-value: < 2.2e-16
```

Ad esempio, al modello bivariato descritto precedentemente viene aggiunta la variabile GDP, che risulta positivamente correlata e statisticamente significativa per entrambi gli anni. L'inclusione di questa variabile porta a un miglioramento del R^2, confermando così quanto affermato in precedenza riguardo alla capacità delle nuove variabili di ottimizzare l'adattamento del modello ai dati.

Questo miglioramento nell'R^2 indica che l'GDP contribuisce significativamente alla spiegazione della variabilità dell'aspettativa di vita, rafforzando ulteriormente la validità e la robustezza del modello multivariato.

Procedendo su questa logica e aggiungendo progressivamente altre variabili indipendenti, si giunge ad un modello più completo con sei regressori, riportato di seguito.

2000

```
Call:
lm(formula = Life_EXP ~ thinness_5_9 + Diphtheria + exp_tot +
    HIV + GDP + Schooling, data = OMS_2000)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0581  -2.5930   0.3009   2.8544  20.1922

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.981e+01  1.986e+00  25.084  < 2e-16 ***
thinness_5_9 -1.288e-01  9.441e-02  -1.364  0.17459
Diphtheria    7.879e-02  1.511e-02   5.213 6.33e-07 ***
exp_tot      -2.337e-01  2.185e-01  -1.069  0.28665
HIV          -6.086e-01  5.889e-02 -10.335  < 2e-16 ***
GDP           1.530e-04  5.172e-05   2.958  0.00362 **
Schooling     1.322e+00  1.316e-01  10.050  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.103 on 144 degrees of freedom
  (32 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.7734,    Adjusted R-squared:  0.764
F-statistic: 81.91 on 6 and 144 DF,  p-value: < 2.2e-16
```

2014

```
Call:
lm(formula = Life_EXP ~ thinness_5_9 + Diphtheria + exp_tot +
    HIV + GDP + Schooling, data = OMS_2014)

Residuals:
     Min       1Q   Median       3Q      Max
-17.9671  -2.2690   0.2413   2.3703   8.5489

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.833e+01  2.456e+00  19.679  < 2e-16 ***
thinness_5_9 -1.872e-01  9.020e-02  -2.075  0.03973 *
Diphtheria    2.846e-02  1.608e-02   1.770  0.07883 .
exp_tot       1.986e-01  1.317e-01   1.509  0.13360
HIV          -2.033e+00  2.455e-01  -8.283 7.29e-14 ***
GDP           5.790e-05  1.972e-05   2.935  0.00388 **
Schooling     1.644e+00  1.613e-01  10.187  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 145 degrees of freedom
  (31 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.7939,    Adjusted R-squared:  0.7854
F-statistic: 93.09 on 6 and 145 DF,  p-value: < 2.2e-16
```

In questo modello, l' R^2 è significativamente aumentato rispetto al modello bivariato di partenza, indicando un miglioramento nella spiegazione della variabile dipendente Y. Analizzando i coefficienti delle singole variabili indipendenti, emergono le seguenti osservazioni:

- Thinness_5_9: La variabile è negativamente correlata con l'aspettativa di vita, come ci si aspetterebbe, dato che l'indice di magrezza tra i bambini riflette uno stato di salute fisica compromesso. Tuttavia, questa variabile risulta statisticamente significativa solo nel modello relativo al 2014, suggerendo un impatto più rilevante in quel periodo.

- Diphtheria: È positivamente correlata e statisticamente significativa in entrambi gli anni. Ciò è coerente con l'idea che una copertura vaccinale più ampia prevenga la diffusione di malattie, contribuendo ad aumentare l'aspettativa di vita.

- Exp_tot: Questa variabile non risulta statisticamente significativa, il che potrebbe essere spiegato dal fatto che, tra i vari Paesi, i sistemi sanitari sono strutturati in modo diverso, e quindi la spesa pubblica per la salute potrebbe non avere un impatto diretto uniforme sull'aspettativa di vita.

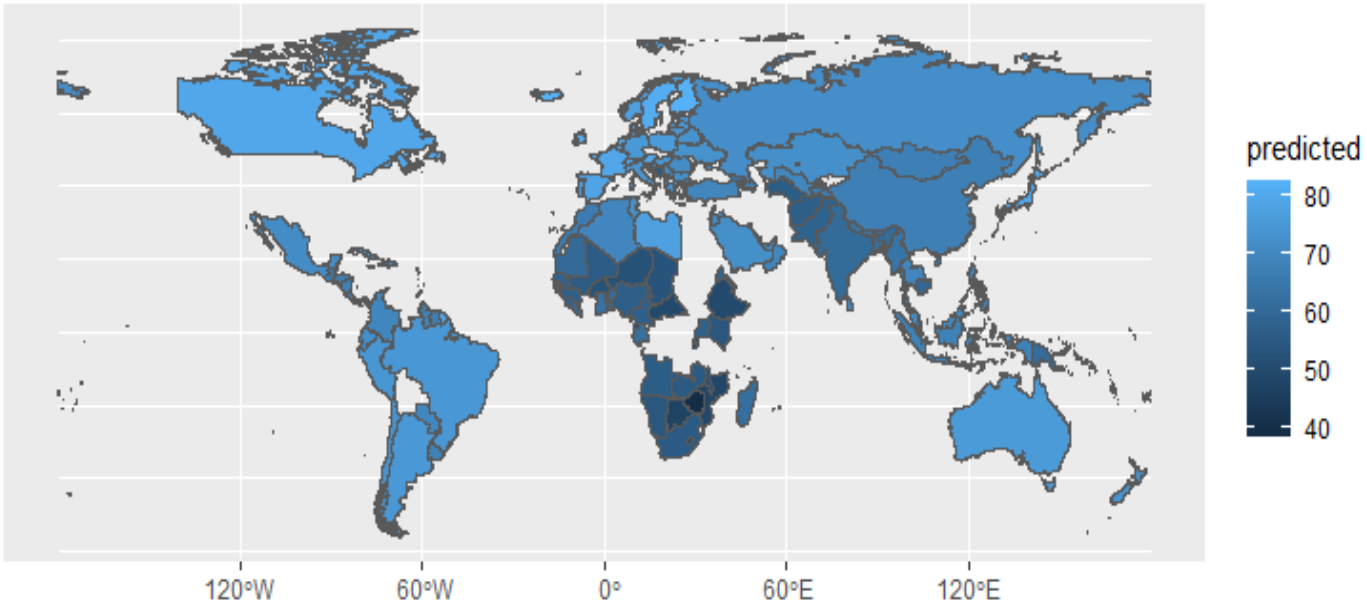
- HIV: La variabile è negativamente correlata e statisticamente significativa. Questo risultato è comprensibile, poiché l'aumento della diffusione del virus comporta una riduzione dell'aspettativa di vita. Tuttavia, dal 2000 al 2014, si osserva una riduzione del valore del coefficiente di regressione, che potrebbe riflettere i progressi nella prevenzione e nel trattamento dell'HIV, nonché i miglioramenti nelle politiche sanitarie globali.

- GDP: La variabile risulta positivamente e statisticamente significativa, indicando che un aumento del PIL pro capite si traduce migliori condizione di vita e di salute e, conseguentemente, in un'aspettativa di vita più alta. Il reddito degli abitanti di un Paese favorisce maggiori e migliori consumi e l'accesso a cure migliori, infrastrutture sanitarie più avanzate e condizioni di vita più favorevoli.

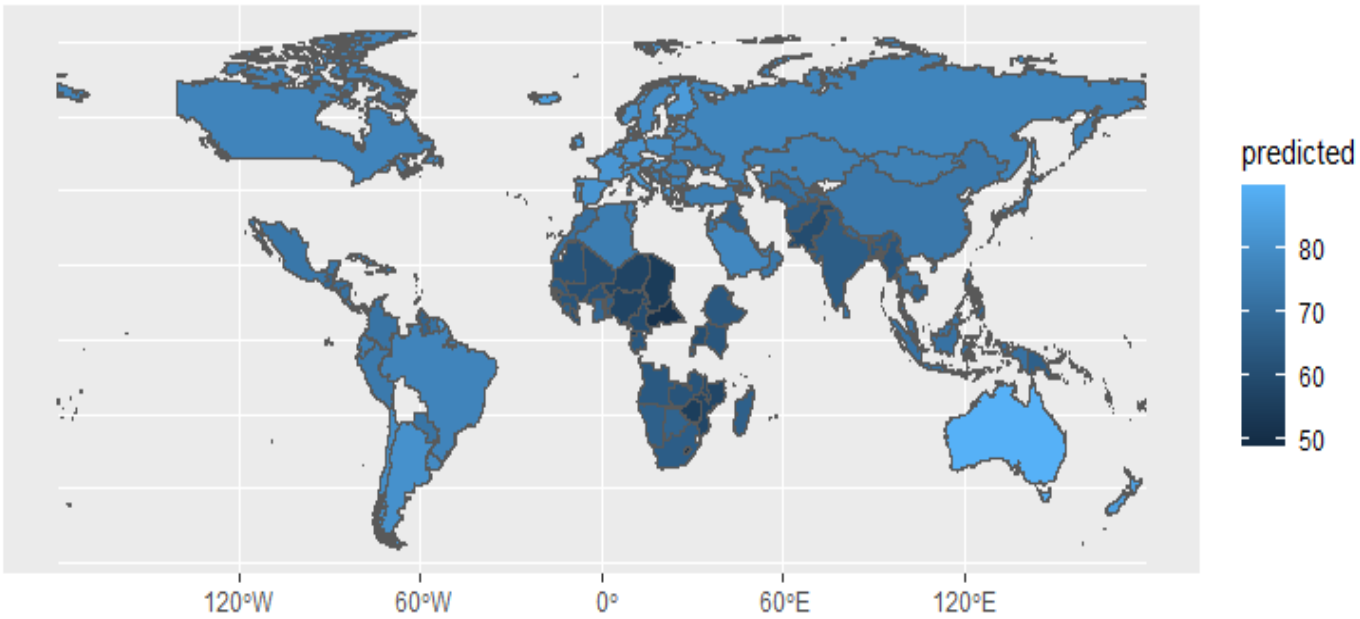
- Schooling: Anche questa variabile è positivamente e statisticamente significativa. Come il GDP, l'istruzione risulta determinante per l'aspettativa di vita, in quanto un livello educativo più elevato è generalmente associato a un migliore stile di vita, a un reddito più alto e, quindi, a una maggiore attenzione alla salute e al benessere, che favoriscono una vita più lunga e prospettive di vita migliori.

A completamento dell'analisi di regressione, sono presentate nuovamente le mappe geografiche che illustrano i valori dell'aspettativa di vita previsti dai modelli di regressione descritti in precedenza. Le mappe confermano quanto osservato in precedenza: i valori più bassi dell'aspettativa di vita si concentrano prevalentemente nei Paesi in via di sviluppo dell'Africa e dell'Asia, a differenza delle nazioni del continente europeo e americano, dove l'aspettativa di vita risulta significativamente più elevata.

2000



2014



In conclusione, i modelli di regressione sviluppati, sia bivariati che multivariati, hanno fornito approfondimenti significativi riguardo ai fattori che influenzano l'aspettativa di vita nei vari Paesi analizzati. Il modello bivariato ha evidenziato una relazione positiva e significativa tra l'istruzione e l'aspettativa di vita, suggerendo che una maggiore educazione è associata a una migliore salute e benessere. L'introduzione di variabili aggiuntive nel modello multivariato, come il GDP, la copertura vaccinale (Diphtheria), HIV, lo stato di salute fisica (Thinness_5_9) e la spesa sanitaria hanno ulteriormente migliorato l'adattamento del modello, confermando l'importanza di fattori socioeconomici e sanitari nel determinare l'aspettativa di vita. L'aumento dell'R^2 nel

modello multivariato ha evidenziato come l'inclusione di più regressori migliori la spiegazione della variabilità dell'aspettativa di vita.

I risultati hanno mostrato che variabili come HIV e Thinness_5_9 esercitano un impatto negativo, con l'aspettativa di vita che diminuisce all'aumentare della diffusione di malattie o della magrezza infantile. Al contrario, GDP e Schooling si sono confermati fattori positivi e significativi, con un impatto diretto sulla qualità della vita e sulla longevità.

Nel complesso, i modelli utilizzati confermano l'importanza di una visione multidimensionale, che considera variabili economiche, sanitarie ed educative, per comprendere le determinanti dell'aspettativa di vita a livello globale. Questi risultati suggeriscono che politiche preordinate a migliorare l'istruzione, la salute e lo sviluppo economico possano avere un impatto positivo e duraturo sul benessere delle popolazioni.

Monte Carlo Simulation

Un'ulteriore analisi condotta successivamente ai modelli di regressione riguarda l'impiego della simulazione Monte Carlo. Questa tecnica è utilizzata, in questo contesto, per prevedere la variabile di interesse (Life_EXP) in un paese in via di sviluppo, il Ghana, che, come emerge dall'analisi del dataset iniziale, presenta una bassa aspettativa di vita nel periodo 2000-2014 (con media pari a 61 anni).

```
  Country              Year         Status            Life_EXP     Adult_Mortality infant_deaths
Length:16          Min.   :2000   Length:16          Min.   :57.20   Min.   : 28.0   Min.   :37.00
Class :character   1st Qu.:2004   Class :character   1st Qu.:58.20   1st Qu.: 38.0   1st Qu.:39.75
Mode  :character   Median :2008   Mode  :character   Median :60.55   Median :253.5   Median :41.00
                   Mean   :2008                      Mean   :60.86   Mean   :180.1   Mean   :40.31
                   3rd Qu.:2011                      3rd Qu.:62.17   3rd Qu.:268.0   3rd Qu.:41.00
                   Max.   :2015                      Max.   :69.00   Max.   :296.0   Max.   :43.00

    Alcohol          exp_proc        HepatitisB       Measles             BMI           deaths5
Min.   :0.010   Min.   :  0.00   Min.   : 8.00   Min.   :    6.00   Min.   : 2.10   Min.   :52.00
1st Qu.:1.355   1st Qu.: 24.33   1st Qu.:27.75   1st Qu.:   96.25   1st Qu.:21.65   1st Qu.:57.50
Median :1.530   Median : 42.17   Median :89.50   Median :  369.50   Median :23.90   Median :61.00
Mean   :1.269   Mean   : 77.25   Mean   :67.50   Mean   : 3419.75   Mean   :21.73   Mean   :59.81
3rd Qu.:1.665   3rd Qu.:134.74   3rd Qu.:93.75   3rd Qu.: 1694.50   3rd Qu.:26.25   3rd Qu.:62.25
Max.   :1.780   Max.   :225.22   Max.   :98.00   Max.   :23068.00   Max.   :28.60   Max.   :65.00
NA's   :1                        NA's   :2
     Polio          exp_tot        Diphtheria         HIV              GDP              Population
Min.   : 8.00   Min.   :3.000   Min.   : 8.00   Min.   :0.700   Min.   :  19.69   Min.   :  215429
1st Qu.:83.25   1st Qu.:3.630   1st Qu.:78.75   1st Qu.:1.200   1st Qu.: 271.02   1st Qu.: 2377839
Median :89.50   Median :4.630   Median :88.00   Median :2.400   Median : 710.55   Median : 2779942
Mean   :78.94   Mean   :4.329   Mean   :73.88   Mean   :2.288   Mean   : 834.38   Mean   :11534783
3rd Qu.:92.25   3rd Qu.:4.830   3rd Qu.:93.25   3rd Qu.:3.250   3rd Qu.:1378.89   3rd Qu.:22865518
Max.   :94.00   Max.   :5.330   Max.   :98.00   Max.   :3.600   Max.   :1814.49   Max.   :27582821
                NA's   :1
thinness_1_19   thinness_5_9        HDI          Schooling
Min.   :6.20   Min.   :6.100   Min.   :0.4800   Min.   : 7.600
1st Qu.:6.85   1st Qu.:6.750   1st Qu.:0.4905   1st Qu.: 7.975
Median :7.60   Median :7.500   Median :0.5245   Median : 9.450
Mean   :7.65   Mean   :7.575   Mean   :0.5259   Mean   : 9.512
3rd Qu.:8.45   3rd Qu.:8.350   3rd Qu.:0.5563   3rd Qu.:10.975
Max.   :9.20   Max.   :9.200   Max.   :0.5760   Max.   :11.700
```

L'analisi si basa su uno scenario ottimistico per il futuro: vengono simulati i valori dei sei regressori utilizzando distribuzioni con media e deviazione standard derivanti da un paese che, pur essendo in via di sviluppo, mostra una migliore aspettativa di vita, ovvero la Tunisia (con una media pari a 74 anni).

Per ciascuna variabile indipendente vengono generati mille valori simulati. L'obiettivo finale della simulazione è prevedere l'aspettativa di vita del Ghana, applicando i coefficienti di regressione ottenuti dal modello multivariato relativo all'anno 2014 ai nuovi dati simulati.

	GDP	HIV	Diphtheria	Schooling	exp_tot	thinness_5_9	predictions
1	5182.36766	0.1	96.07409	13.43557	5.365656	6.444632	73.10082
2	2151.04416	0.1	96.30146	14.48987	6.607987	6.444932	74.91134
3	2695.43337	0.1	99.19985	14.42637	6.239144	6.319268	74.87124
4	2894.01793	0.1	97.71841	14.90104	7.198291	6.291772	75.81640
5	2888.61033	0.1	99.73690	13.15821	4.931815	6.370699	72.54407
6	3380.43703	0.1	96.48957	14.57483	5.970080	6.416756	75.00608
7	3955.68462	0.1	96.28514	14.72600	6.621426	6.342386	75.42532
8	3336.47945	0.1	99.02340	14.40509	6.099095	6.344122	74.83589
9	3469.39743	0.1	97.08505	15.11672	6.496679	6.372476	76.03172
10	2351.87123	0.1	97.83573	14.98444	6.191803	6.370878	75.71071
11	3480.16783	0.1	98.15477	14.47627	6.942610	6.300941	75.11210
12	4906.05353	0.1	96.98221	14.33947	6.082830	6.308196	74.76431
13	3280.27898	0.1	97.87270	13.93284	5.377145	6.444017	73.86161
14	6108.41303	0.1	97.05666	14.03679	6.368662	6.346888	74.38808
15	1597.44001	0.1	97.31074	14.17403	6.086317	6.379033	74.29762
16	2034.86389	0.1	98.66887	14.67794	6.269388	6.272768	75.24607
17	3616.08249	0.1	98.95424	13.81235	7.204044	6.434655	74.07841
18	3571.26175	0.1	98.79108	15.35682	5.587753	6.380970	76.29867
19	3768.48680	0.1	99.65886	13.65441	5.601805	6.354785	73.54441
20	3603.90036	0.1	98.03724	14.54992	5.637008	6.338761	74.97058
21	4194.50152	0.1	100.16579	14.55028	4.868118	6.240173	74.93168

Esempio pratico di simulazione con modello di regressione:

GDP	HIV	Diphtheria	Schooling	exp_tot	thinness_5_9	predictions
5182.36766	0.1	96.07409	13.43557	5.365656	6.444632	73.10082

$$Life\_EXP = 48.33 + (-0.1872) \cdot thinness\_5\_9 + 0.2846 \cdot Diphtheria + 0.1301 \cdot exp\_tot + (-2.033) \cdot HIV + 5.790 \cdot GDP + 1.644 \cdot Schooling + \varepsilon$$

Il risultato finale di questo lavoro mostra che, in uno scenario positivo per il Ghana, paragonabile ai dati della Tunisia nel 2014, l'aspettativa di vita potrebbe aumentare significativamente, passando da una media di 61 anni a 73 anni. Questo miglioramento si verifica grazie all'aumento di variabili chiave come il PIL pro capite e il tasso di istruzione, accompagnato da una riduzione dei decessi per HIV.

Altri fattori, come i valori medi di variabili meno influenti, restano pressoché invariati rispetto al periodo 2000-2014 del Ghana. Questi risultati sottolineano l'importanza di implementare politiche pubbliche mirate che possano sostenere la crescita economica, migliorare l'accesso all'istruzione e ridurre l'impatto di fattori negativi sulla salute pubblica, al fine di favorire un aumento dell'aspettativa di vita.