



LIFE EXPECTANCY

Progetto di Marta Zecchini, Simone Paloschi, Andrea Rella e Elisa Quadrini

DOMANDA DI RICERCA:

Come può uno stato migliorare l'aspettativa di vita dei cittadini?

Tra tutti i fattori che abbiamo selezionato, si vuole studiare quali di questi si relazionano effettivamente all'aspettativa di vita

PRESENTAZIONE DEL DATASET

Variabile risposta:
Life expectancy

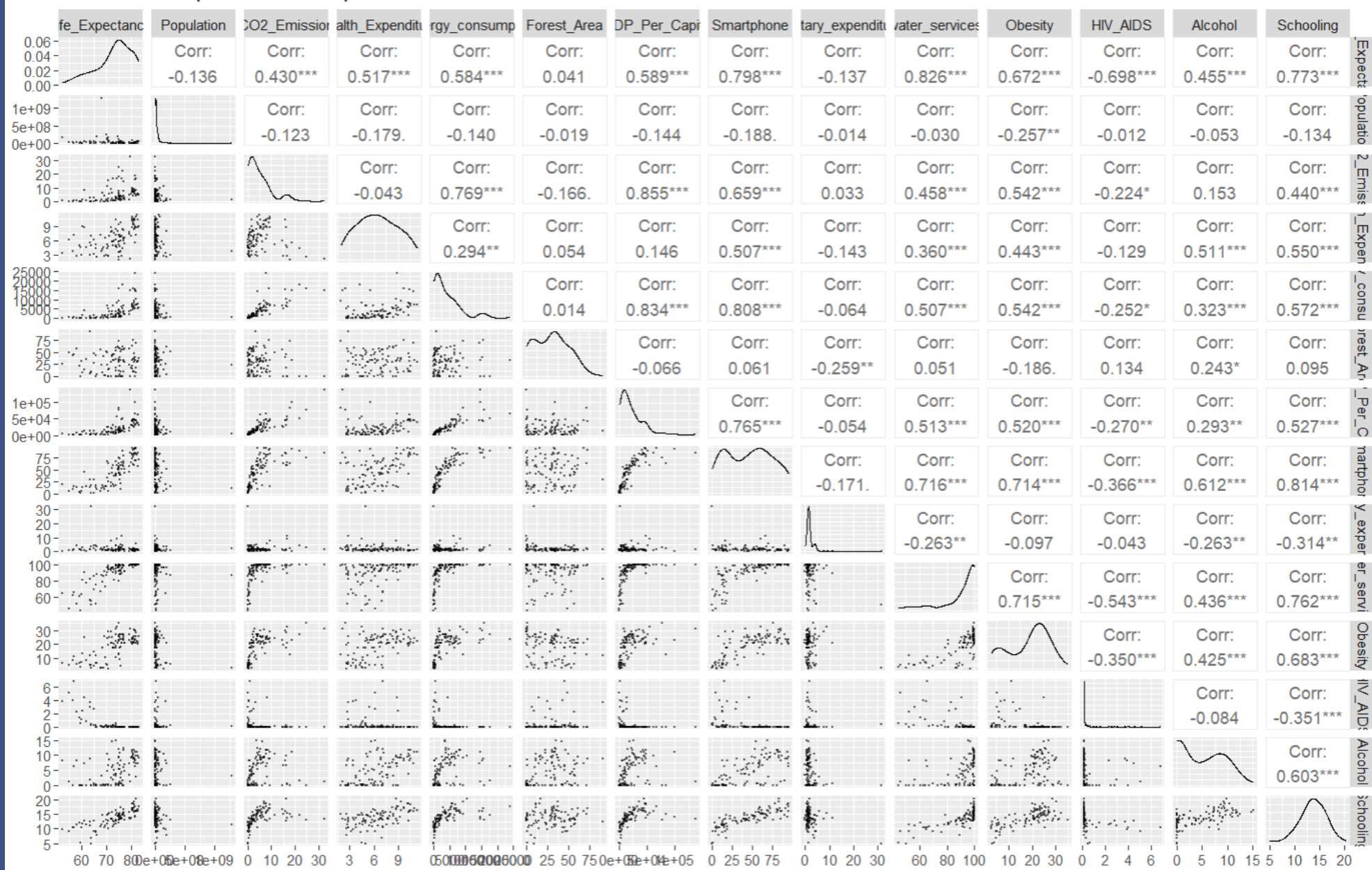
Arco temporale:
15 anni

Modello lineare:
anno 2013

17 covariate
119 osservazioni

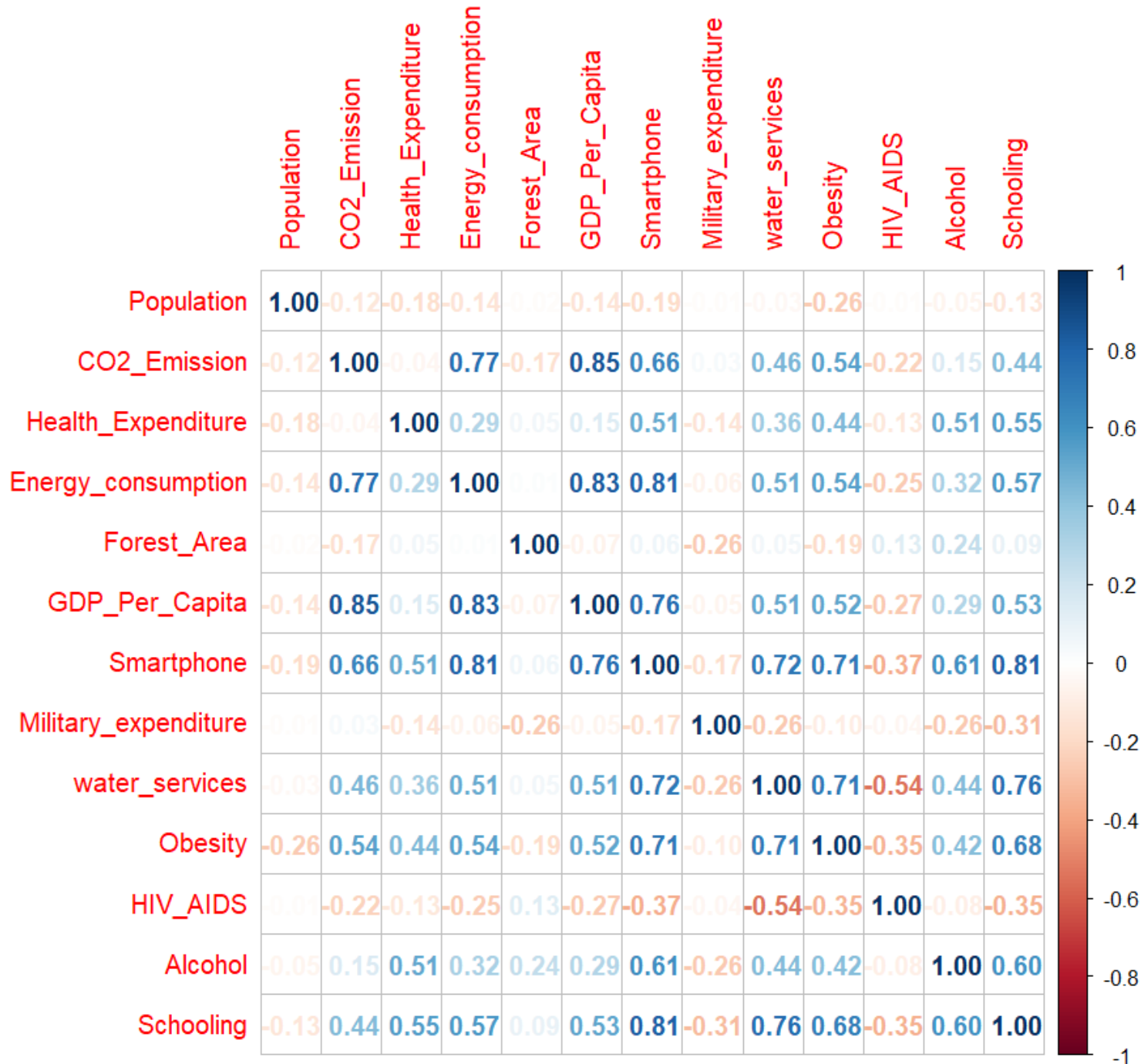
Country	Year	Least developed	Life expectancy	Population	CO ₂ emission
Health expenditure	Electric power consumption	Forest area	GDP per capita	Smartphone access	Military expenditure
People using basic drinking water services	Obesity among adults	HIV/AIDS	Alcohol	Schooling	

Relazione tra i predittori e la risposta



ANALYSIS DEL

CORRELATION PLOT



VIF Smartphone

10,10052

MODELLO COMPLETO

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.038e+01	2.540e+00	19.836	< 2e-16	***
Population	-1.221e-09	1.916e-09	-0.637	0.525565	
CO2_Emission	-2.460e-01	9.675e-02	-2.542	0.012781	*
Health_Expenditure	4.285e-01	1.476e-01	2.904	0.004674	**
Energy_consumption	-9.370e-05	1.140e-04	-0.822	0.413521	
Forest_Area	2.085e-02	1.326e-02	1.572	0.119501	
GDP_Per_Capita	8.932e-05	2.486e-05	3.593	0.000541	***
Smartphone	6.857e-02	2.624e-02	2.613	0.010577	*
Military_expenditure	1.001e-01	7.741e-02	1.293	0.199593	
water_services	1.336e-01	3.322e-02	4.021	0.000123	***
Obesity	3.144e-02	5.222e-02	0.602	0.548637	
HIV_AIDS	-2.267e+00	2.491e-01	-9.098	2.83e-14	***
Alcohol	-1.052e-01	8.128e-02	-1.294	0.198963	
Schooling	3.591e-01	1.739e-01	2.065	0.041902	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.322 on 87 degrees of freedom

Multiple R-squared: 0.9112, Adjusted R-squared: 0.8979

F-statistic: 68.66 on 13 and 87 DF, p-value: < 2.2e-16

RICERCA DEL MODELLO OTTIMALE

Applicazione del comando leaps rispetto all' R^2_{adj} → modello più semplice con R^2_{adj} solo poco diminuito

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.929e+01	2.030e+00	24.285	< 2e-16	***
Health_Expenditure	6.874e-01	1.212e-01	5.673	1.51e-07	***
GDP_Per_Capita	6.624e-05	1.318e-05	5.025	2.36e-06	***
water_services	1.438e-01	3.018e-02	4.767	6.73e-06	***
HIV_AIDS	-2.393e+00	2.455e-01	-9.747	5.74e-16	***
Schooling	4.636e-01	1.583e-01	2.928	0.00426	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

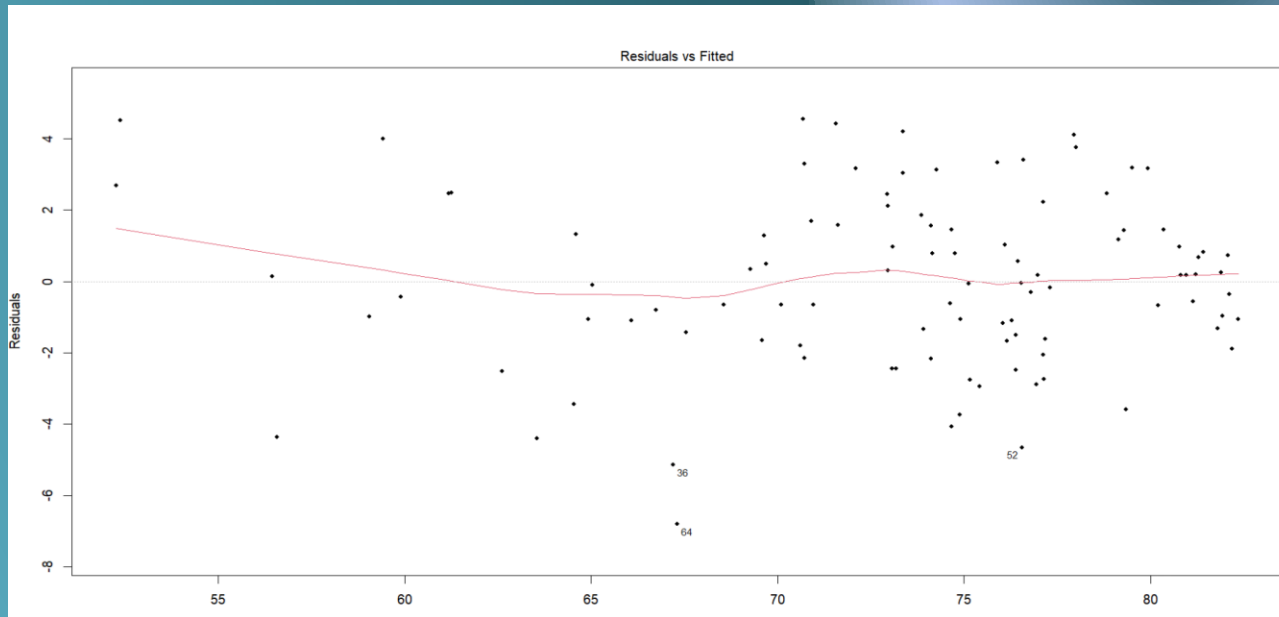
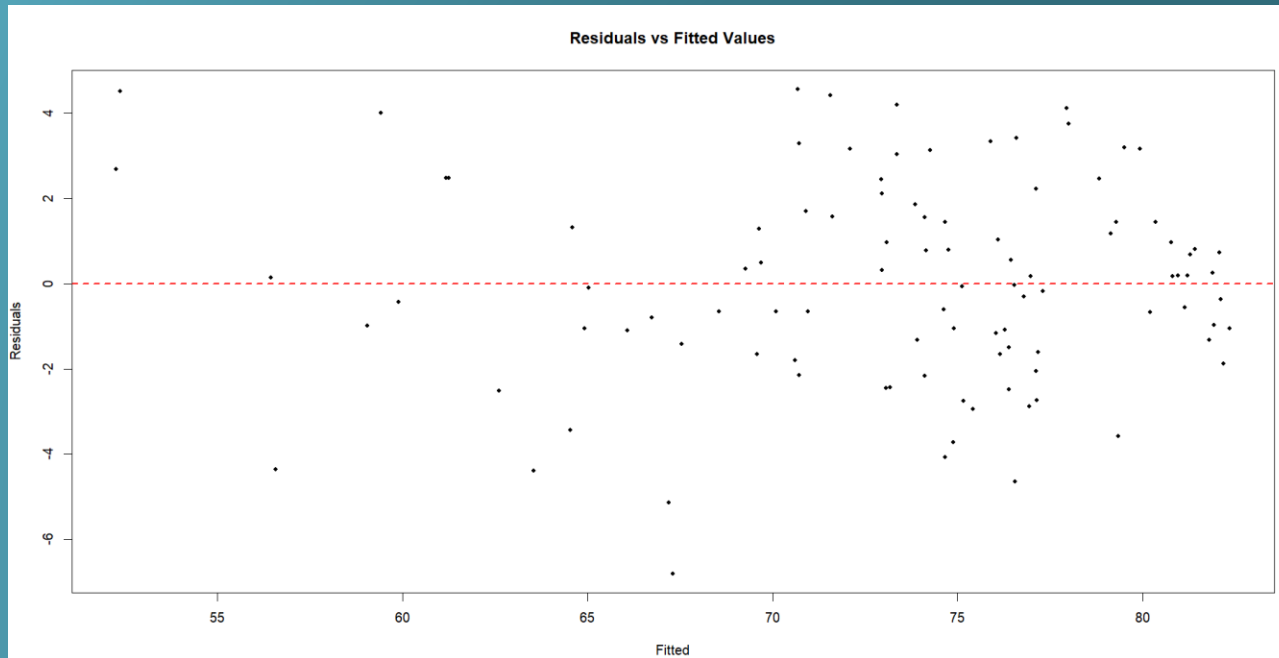
Residual standard error: 2.451 on 95 degrees of freedom

Multiple R-squared: 0.8919, Adjusted R-squared: 0.8862

F-statistic: 156.8 on 5 and 95 DF, p-value: < 2.2e-16

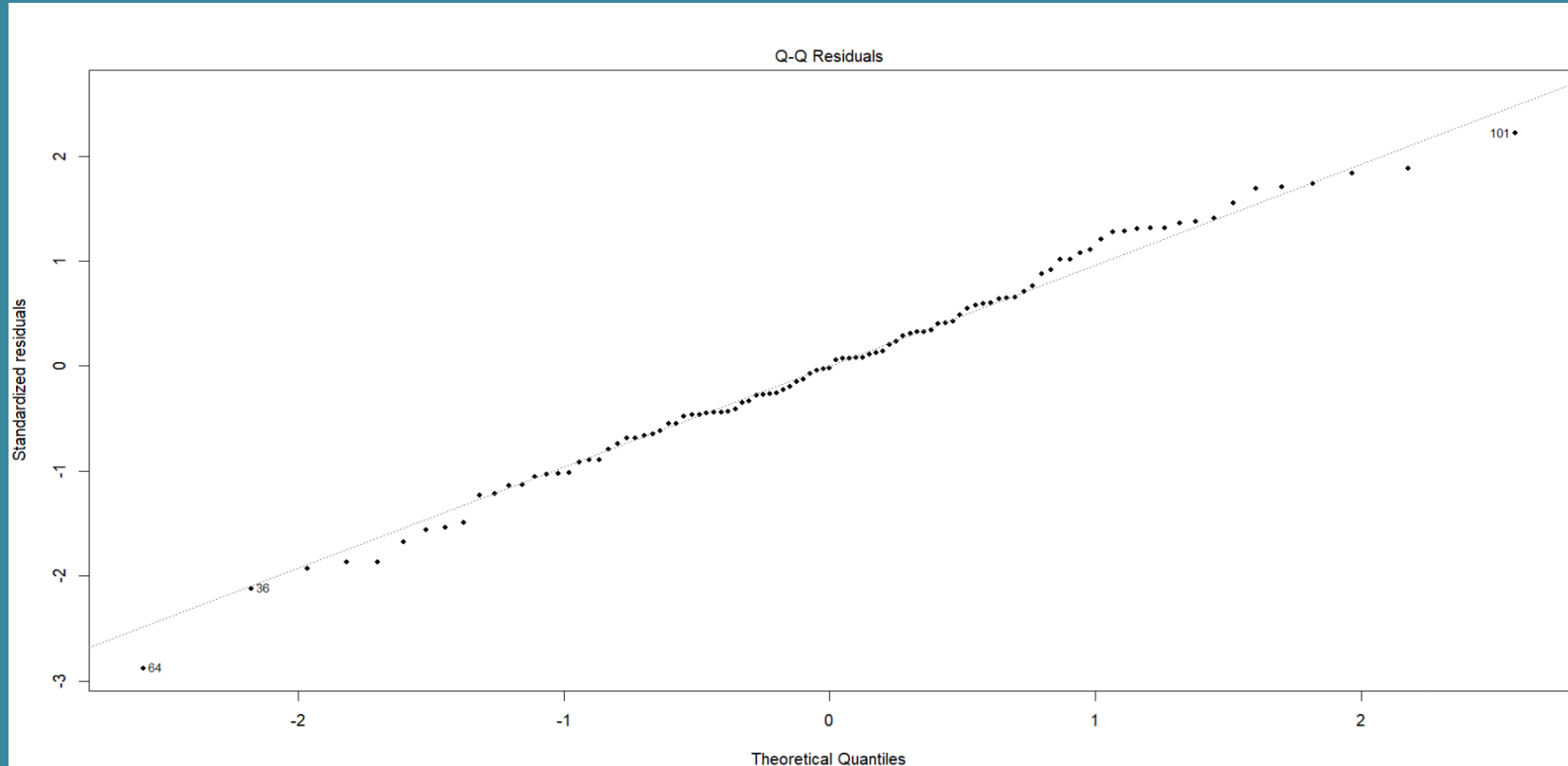
VERIFICA DELLE IPOTESI

CONTROLLO OMOSCHEDASTICITÀ



VERIFICA DELLE IPOTESI

CONTROLLO NORMALITÀ



Shapiro-Wilk normality test

```
data: g$res  
W = 0.98876, p-value = 0.5586
```

Il modello rispetta le ipotesi di omoschedasticità e normalità dei residui

Cross Validation / Previsione

Cross Validation

$mean(abs(\hat{Y} - Y))$	1.490523
$max(abs(\hat{Y} - Y))$	3.420973
$min(abs(\hat{Y} - Y))$	0.0434405

Modello fittato sull'80% delle osservazioni del dataset del 2013 e testato sul restante 20%

Avendo in mano anche i valori effettivi delle life expectancies siamo in grado di valutare in prima approssimazione la capacità del nostro modello di adattarsi ai dati e prevedere

$$\hat{Y} = Z\hat{\beta}$$

Previsione sui dati del 2014

$mean(abs(\hat{Y} - Y))$	1.872448
$max(abs(\hat{Y} - Y))$	6.309184
$min(abs(\hat{Y} - Y))$	0.04131859

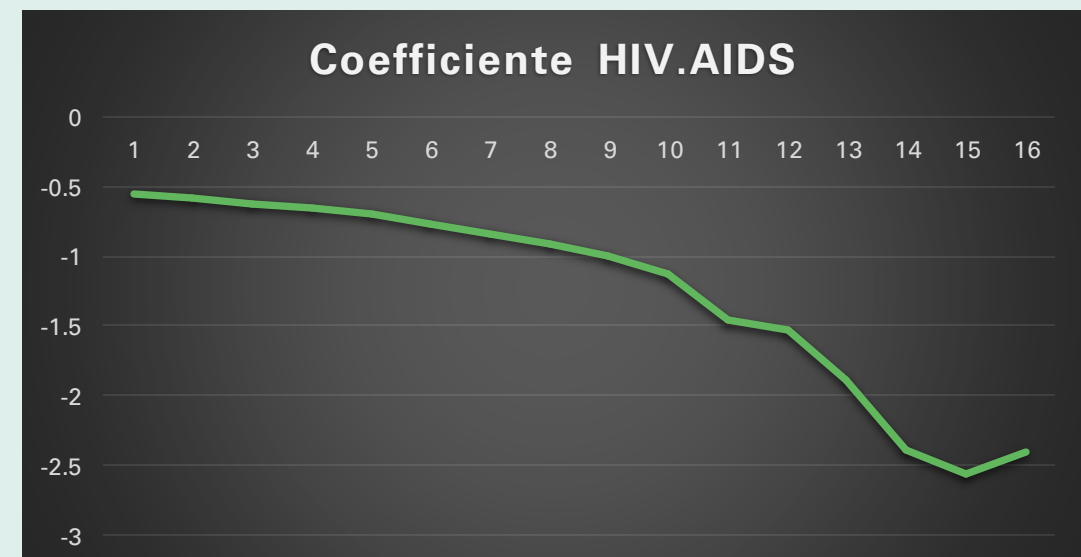
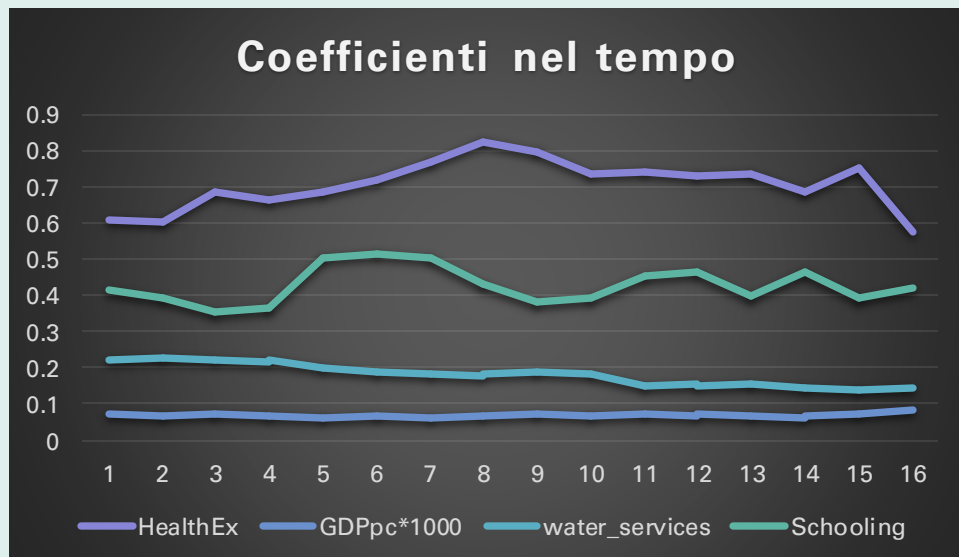
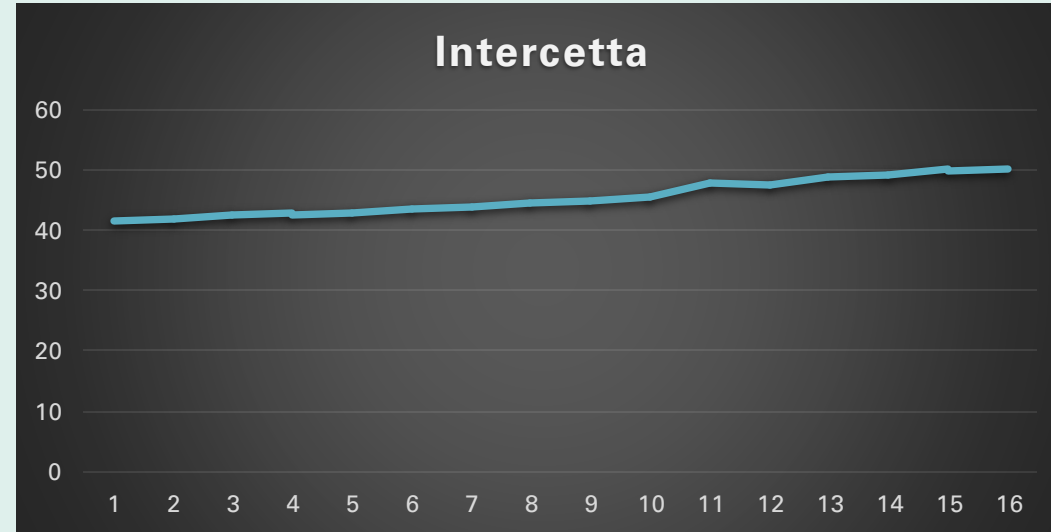
Modello fittato sulla totalità delle osservazioni risalenti al 2013 e testato sui dati del 2014

I coefficienti delle covariate e il loro peso sulla risposta

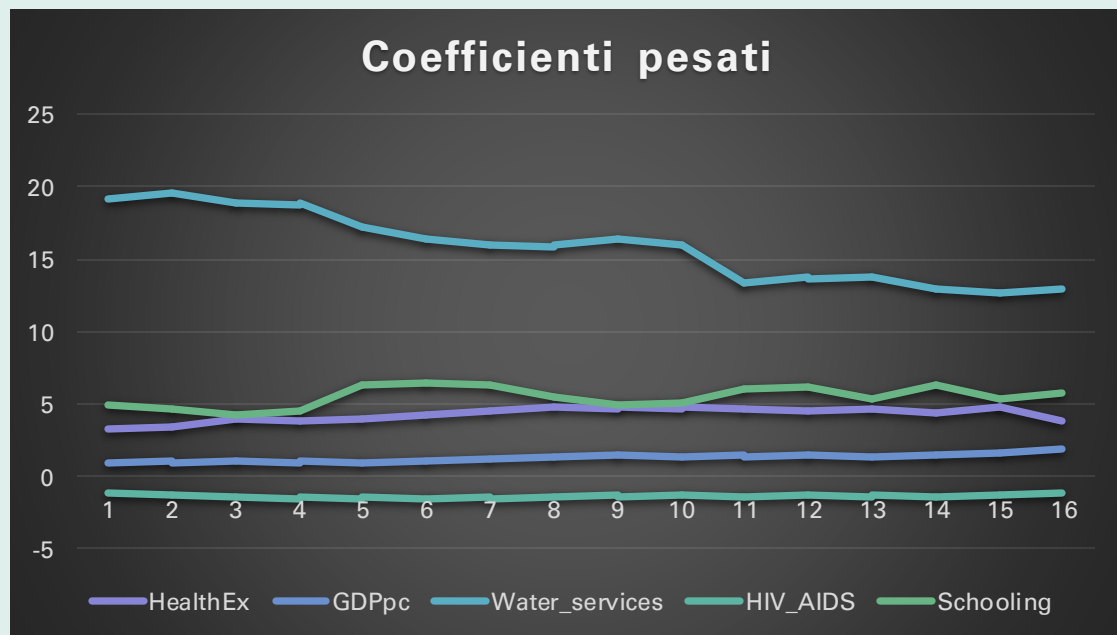
Intercetta	Aspettativa di vita
49,28721	73,15915

	Coefficienti	Medie	Coefficienti pesati
HealthEx	0,6873942	6,443577	4,429277457
GDPpc	6,62366E-05	22951,08	1,520202194
Water	0,1438436	89,85045	12,92441219
HIV.AIDS	-2,393166	0,5425743	-1,298470367
Schooling	0,4635524	13,58317	6,296511053

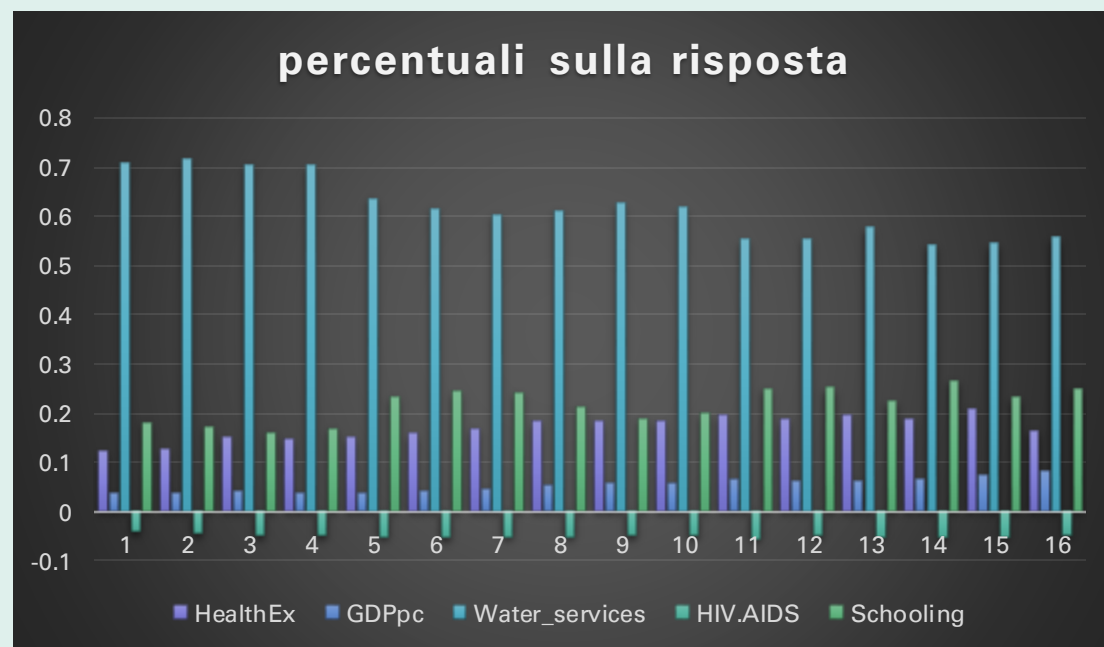
Analisi dell'andamento dei beta al variare del tempo dal 2000 al 2015



Evoluzione dei coefficienti dei beta per la media nel corso del tempo



Coefficienti moltiplicati per la media



Coefficienti moltiplicati per la media normalizzati

Analisi dei punti influenti

Studentizzati

Ghana	Namibia	Zimbabwe
-2.118797	-2.877445	2.217303

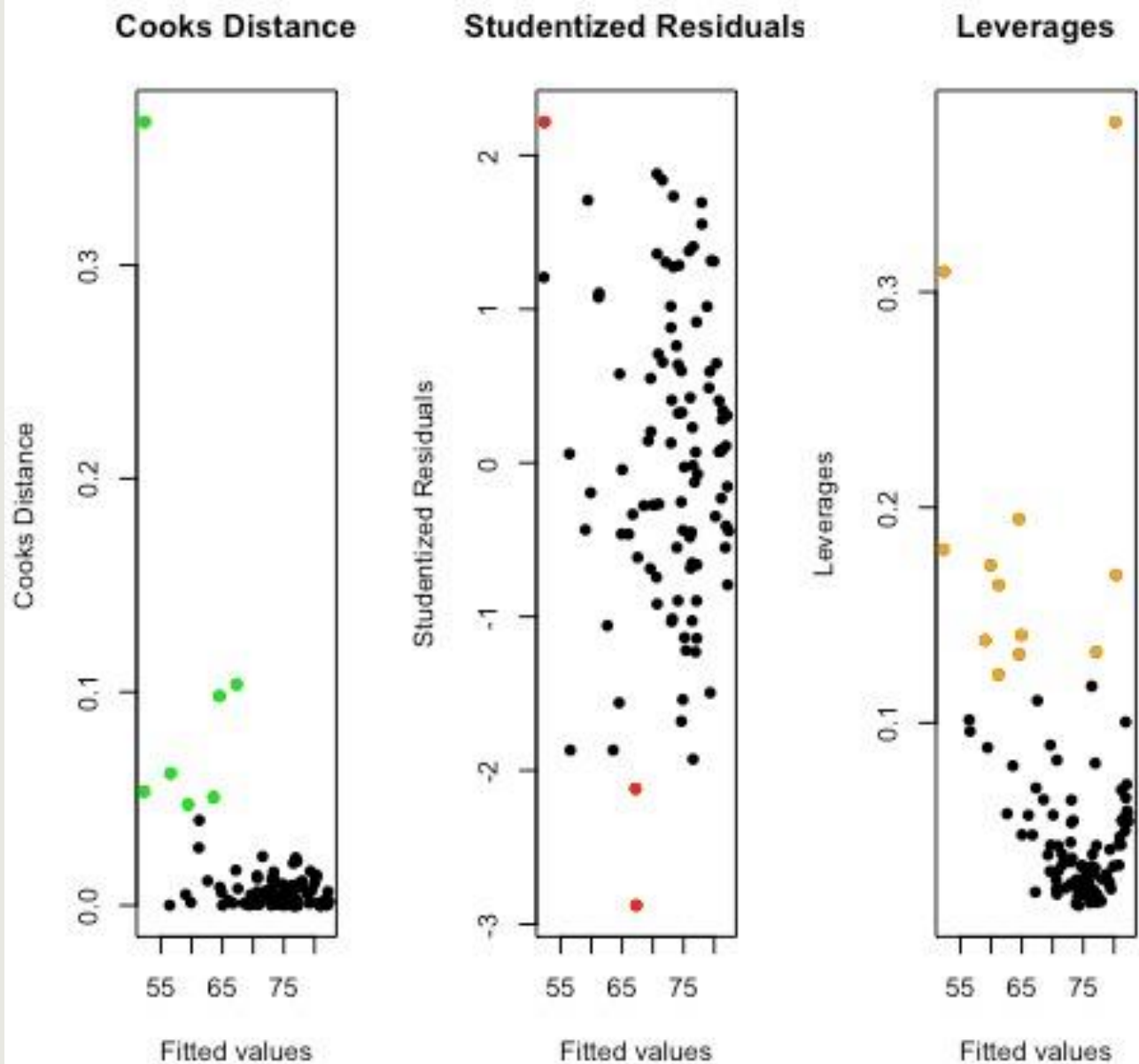
	Criterio di flag
Student	$\text{abs}(\text{stud}) > 2$
Cook	$\text{Cdist} > 0.04210526$
Leva	$\text{Lev} > 0.1188119$

Distanza di Cook

Kenya	Mozambique	Namibia	Nigeria	South Africa	Togo	Zimbabwe
0.04723034	0.05342161	0.10357089	0.06186693	0.09809504	0.05064688	0.36714855

Punti leva

[,1]	[,2]	[,3]	[,4]	[,5]
"Angola"	"Brunei Darussalam"	"Eritrea"	"Gabon"	"Luxembourg"
"0.138321857577194"	"0.132920654718085"	"0.164052473574697"	"0.12247187213929"	"0.168729451716427"
[,6]	[,7]	[,8]	[,9]	[,10]
"Mozambique"	"Niger"	"Qatar"	"South Africa"	"Sudan"
"0.180519350713991"	"0.173139762975819"	"0.378900291661617"	"0.194692638990759"	"0.140700811927311"
[,11]	[,12]			
"Yemen"	"Zimbabwe"			
"0.131885466642916"	"0.309424229695229"			



Analisi dei punti influenti

Strumenti utilizzati:

- 1) Distanza di cook
- 2) Residui studentizzati
- 3) Punti leva

Analisi dei punti influenti

Studentizzati 66.7%

Ghana	<u>Namibia</u>	<u>Zimbabwe</u>
-2.118797	-2.877445	2.217303

Percentuale dei paesi
classificati come
'Least Developed' = 22.8%

Distanza di Cook 85.7%

<u>Kenya</u>	<u>Mozambique</u>	<u>Namibia</u>	<u>Nigeria</u>	South Africa	<u>Togo</u>	<u>Zimbabwe</u>
0.04723034	0.05342161	0.10357089	0.06186693	0.09809504	0.05064688	0.36714855

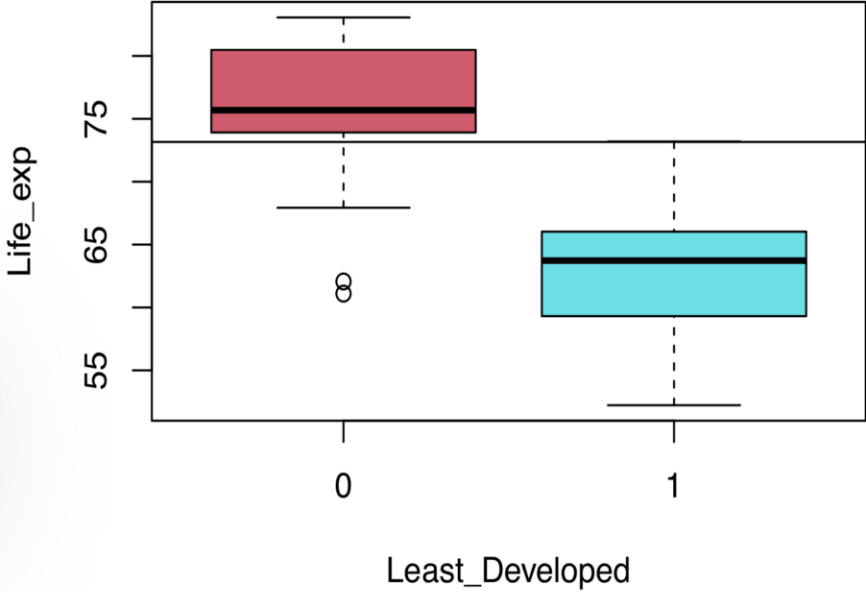
Punti leva 66.7%

[,1] " <u>Angola</u> " "0.138321857577194"	[,2] "Brunei Darussalam" "0.132920654718085"	[,3] " <u>Eritrea</u> " "0.164052473574697"	[,4] " <u>Gabon</u> " "0.12247187213929"	[,5] "Luxembourg" "0.168729451716427"
[,6] " <u>Mozambique</u> " "0.180519350713991"	[,7] " <u>Niger</u> " "0.173139762975819"	[,8] "Qatar" "0.378900291661617"	[,9] "South Africa" "0.194692638990759"	[,10] " <u>Sudan</u> " "0.140700811927311"
[,11] " <u>Yemen</u> " "0.131885466642916"	[,12] " <u>Zimbabwe</u> " "0.309424229695229"			

Anova rispetto allo sviluppo dei paesi

Numerosità dei gruppi	
Least developed	Developed
23	78

Life Expectancy according to Developing status



Anova rispetto allo sviluppo dei paesi

Shapiro test	
Least developed	Developed
0.712593740	0.01383993

Levene test	
F value	Pr(>F)
1.3815	0.2427

Analysis of Variance Table						
Response: Life_exp						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Least_Developed	1	2866.6	2866.60	117.59	< 2.2e-16 ***	
Residuals	99	2413.4	24.38			

Regressione con covariata categorica

```
Call:
lm(formula = Life_exp ~ HealthEx + GDPpp + water_services + HIV.AIDS +
    Schooling + D + HealthEx * D + GDPpp * D + water_services *
    D + HIV.AIDS * D + Schooling * D, data = dataset)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.1075	-1.4267	0.0905	1.5099	4.3111

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.969e+01	5.627e+00	8.830	8.44e-14	***
HealthEx	7.201e-01	1.468e-01	4.907	4.15e-06	***
GDPpp	6.300e-05	1.457e-05	4.325	3.97e-05	***
water_services	1.510e-01	6.595e-02	2.289	0.0245	*
HIV.AIDS	-3.394e+00	5.738e-01	-5.915	6.06e-08	***
Schooling	3.973e-01	2.027e-01	1.960	0.0531	.
D1	8.794e-01	6.503e+00	0.135	0.8927	
HealthEx:D1	-7.124e-02	3.161e-01	-0.225	0.8222	
GDPpp:D1	-8.856e-05	1.816e-04	-0.488	0.6270	
water_services:D1	1.207e-02	8.037e-02	0.150	0.8810	
HIV.AIDS:D1	1.482e+00	6.572e-01	2.255	0.0266	*
Schooling:D1	-2.361e-01	3.949e-01	-0.598	0.5515	

```
Call:
lm(formula = Life_exp ~ HealthEx + GDPpp + water_services + HIV.AIDS +
    Schooling + D * HIV.AIDS, data = dataset)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.8263	-1.4442	-0.0269	1.5082	4.2746

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.137e+01	2.792e+00	18.395	< 2e-16	***
HealthEx	7.385e-01	1.209e-01	6.109	2.31e-08	***
GDPpp	6.536e-05	1.294e-05	5.050	2.20e-06	***
water_services	1.419e-01	3.260e-02	4.354	3.43e-05	***
HIV.AIDS	-3.453e+00	5.531e-01	-6.243	1.27e-08	***
Schooling	3.298e-01	1.681e-01	1.962	0.0528	.
D1	-1.693e+00	1.065e+00	-1.590	0.1153	
HIV.AIDS:D1	1.417e+00	6.268e-01	2.260	0.0261	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.404 on 93 degrees of freedom
Multiple R-squared: 0.8982, Adjusted R-squared: 0.8905
F-statistic: 117.2 on 7 and 93 DF, p-value: < 2.2e-16

Conclusioni e interpretazioni

- Di tutte le covariate quelle che più spiegano l'aspettativa di vita sono la **spesa sanitaria** e l'accesso **l'acqua potabile**. Inoltre si possono interpretare come le covariate più “causali”
- Per quanto riguarda lo **schooling** e il **GDP** la correlazione è molto alta
- I casi di **HIV** nel nostro modello potrebbero essere visti come indice delle malattie più in generale, infatti come abbiamo visto la diminuzione dei casi non ha comportato un effetto “pesato” inferiore sull'intercetta