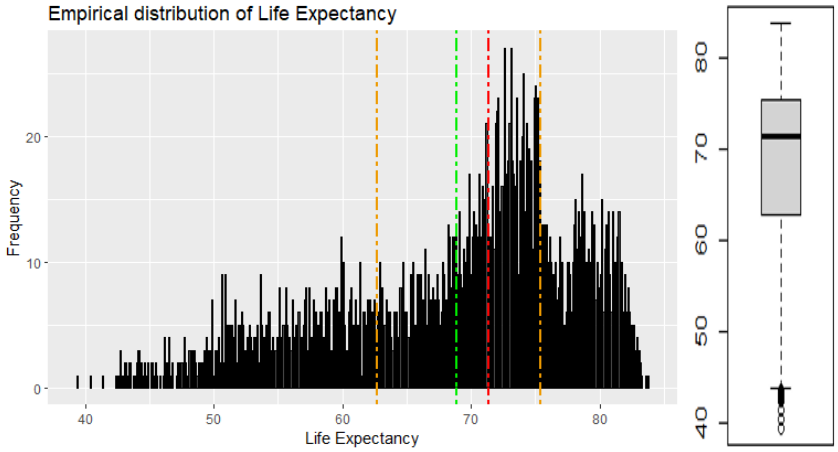
Factors Affecting Global Life Expectancy: A 2000-2015 Analysis

The goal of this investigation is to understand the impact of multiple factors influencing life expectancy across 193 countries from 2000 to 2015. Our analysis takes into account health-related variables, including immunization coverage for some diseases like Hepatitis B and Diphtheria, alongside economic, social, and mortality factors.

Descriptive Statistics for Life Expectancy

We start by looking closely at the dependent variable in question, which is Life Expectancy. We computed its summary statistics, and we plotted its empirical distribution on a histogram. The green line represents the mean, while the red one the median and the orange ones the first and the third quartiles, respectively.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	interq_range	std_dev_LE	var_LE
39.40	62.70	71.40	68.86	75.40	83.80	12.7	9.405608	88.46546



We notice that the mean is smaller than the median: this happens because the mean is not robust to outliers which, as we see from the boxplot, are massively present on the left side of the distribution.

Regression analysis

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	81.4712992	0.6638006	122.735	< 2e-16	***
InfDth	-0.0419476	0.0061711	-6.797	1.29e-11	***
U5dth	-0.0542583	0.0037691	-14.396	< 2e-16	***
AdultmiMort	-0.0474055	0.0006263	-75.693	< 2e-16	***
Alcohol	0.0405160	0.0098829	4.100	4.25e-05	***
HepatitisB	-0.0087120	0.0025318	-3.441	0.000588	***
Measles	-0.0013345	0.0017128	-0.779	0.435958	
BMI	-0.2034056	0.0193302	-10.523	< 2e-16	***
Polio	0.0091951	0.0058231	1.579	0.114431	
Diphtheria	-0.0001260	0.0058504	-0.022	0.982816	
HIV	0.0583790	0.0185886	3.141	0.001703	**
lGDP	0.5164385	0.0384819	13.420	< 2e-16	***
population	-0.0001902	0.0001980	-0.961	0.336715	
Thinness1019	-0.0326969	0.0169806	-1.926	0.054261	.
Thinness59	-0.0046030	0.0166514	-0.276	0.782235	
School	0.0917330	0.0166031	5.525	3.59e-08	***
Developed	0.7264282	0.1033081	7.032	2.54e-12	***

The initial approach we have taken to understand the relationship between the dataset's explanatory factors and the dependent variable is through Ordinary Least Squares. This estimator minimizes the sum of squared differences between the observed and predicted values of the dependent variable (so minimizing the part that the model cannot explain about the relation between the variables) and providing a linear equation that best fits the given data. The coefficients derived by the model represent the estimated impact of each independent variable on the dependent variable, assuming a linear relationship. Based on the R output, it appears that several variables may not have statistical significance, given their notably low t-values and high p-values (Measles, Polio, Diphtheria, population, thinness 1019 and thinness59).

Significant attention should be given to the variable "Developed," which functions as a dummy variable with a value of 1 denoting a developed country and 0 representing a non-developed one. On average, the life expectancy in industrialized states is higher than in underdeveloped states by 0.726 years, *ceteris paribus*. The intercept indicates that, if all the control variables are equal to zero, the average life expectancy in the world is around 81.5 years. Unfortunately, this coefficient does not have an economic significance, since it is meaningless to think about a country whose population is zero.

Size of a test

```
Mean Test Size (t-test): 0.0514
Mean Test Size (z-test): 0.0516
```

We then computed the size of a test to check if our chosen significance level is also the one that resulted by doing tests on the coefficients of the model. Knowing that the significance level is the probability of rejecting the null hypothesis when it is true, we took all the coefficients from OLS, and we put the variable we were studying equal to zero. Then we used the parametric bootstrapping method, by which we derived ten thousand of fitted values of the dependent variable with the previous coefficients and we regressed each one of them in respect to all the regressors by OLS method. After that, we calculated a test on the coefficient of variable "Alcohol" for every OLS fitted line and we stored the results. We finally divided the times that we rejected the null hypothesis (in which the coefficient of "Alcohol" was 0) by 10000 (the number of tests we computed), and we found the size of a test about the chosen variable. This statistic can be derived both with a t-statistic and z-statistic, which are respectively distributed as a t-Student and as a standardized normal. While t-distribution is preferred to the z one in small samples because it is an exact distribution of T test, the z is exact only asymptotically, thing that can be seen in huge samples like ours where the two tests converge to the real value alpha (0.05).

Robust analysis

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.8582481  0.8213396  99.664  < 2e-16 ***
InfDth       -0.0353529  0.0054150  -6.529 7.83e-11 ***
U5dth        -0.0614223  0.0032992 -18.617  < 2e-16 ***
AdultmiMort  -0.0479243  0.0007451 -64.321  < 2e-16 ***
Alcohol       0.0340126  0.0116820   2.912 0.003624 **
HepatitisB   -0.0077014  0.0019885  -3.873 0.000110 ***
Measles      -0.0015876  0.0015497  -1.024 0.305725
BMI          -0.1901632  0.0232816  -8.168 4.66e-16 ***
Polio        -0.0012108  0.0097104  -0.125 0.900775
Diphtheria    0.0042552  0.0090229   0.472 0.637249
HIV           0.0902132  0.0290286   3.108 0.001904 **
lGDP          0.5300483  0.0461097  11.495  < 2e-16 ***
population   -0.0002157  0.0001942  -1.111 0.266779
Thinness1019 -0.0309065  0.0113925  -2.713 0.006710 **
Thinness59   -0.0053904  0.0113844  -0.473 0.635903
School        0.0684440  0.0184912   3.701 0.000218 ***
Developed     0.7343360  0.1088528   6.746 1.83e-11 ***
```

The robust MM-estimator was computed through the utilization of the *lmrob* command in R. This regression approach was undertaken due to the non-robust nature of Ordinary Least Squares (OLS) estimation, where the impact of outliers is disproportionately influential on residuals. The MM-estimator was employed to mitigate this issue by using the parameter estimate $\hat{\beta}$ as the solution to a specific ψ-function. In the *lmrob* function within R, the bisquare redescending score function was utilized, ensuring a diminished influence of outliers or extreme observations. The formulation of the function incorporates the standardized residuals ($\tilde{r}=r/\sigma$), underscoring the importance of robustly estimating the scale parameter σ. This robust estimation is in fact achieved through the utilization of a high breakdown point estimator.

An exploratory analysis was conducted, involving the examination of both the standardized residual plot and the QQ plot. These graphical representations offer valuable insights into the presence of outliers. A well-behaved residual probability density function (PDF), approaching normality, provides an indication that the linear model employed for the regression is a suitable choice. We can notice a few differences between this model and the previous one. Thinnes1019 is now statistically significant even at a 5% significance level. The

coefficient of Polio changes its sign but is still not statistically significant. The coefficient of Diphtheria is now larger and positive but still not statistically significant. The impact of the number of years of schooling is smaller but still statistically significant. Overall, however, the coefficients estimated with the MM-estimator are very similar and even the level of significance is the same for most of them.

Lasso

(Intercept)	8.056389e+01
InfDth	-4.727418e-02
U5dth	-5.147483e-02
AdultmiMort	-4.657310e-02
Alcohol	3.732745e-02
HepatitisB	-6.394416e-03
Measles	.
BMI	-1.729444e-01
Polio	6.628377e-03
Diphtheria	.
HIV	2.968111e-02
lGDP	5.136822e-01
population	-3.309377e-05
Thinness1019	-3.014995e-02
Thinness59	.
School	8.559578e-02
Developed	7.746637e-01

The underlying rationale of the lasso estimator is to trade off estimation unbiasedness for a smaller variance, with the aim of optimizing the mean squared error (MSE) or prediction error. The lasso offers a significant advantage by setting a specific set of coefficients (βj) to be precisely zero. Consequently, the lasso estimator serves as both a regularized estimator and a means of model subset selection, as we can see by the coefficients of Measles, Diphtheria and Thinness59 that become 0. Lasso regression is primarily suited for predictive modelling rather than the interpretation of slope coefficients. The regularization employed in lasso regression, characterized by the λ penalty term, introduces a degree of bias in coefficient estimates, leading to shrinkage and variable selection. Furthermore, the emphasis in lasso regression lies on optimizing predictive accuracy and generalization to new data, rather than providing precise interpretations of individual coefficient values.

Post Lasso

In the previous analysis, certain variables exhibited either statistical insignificance or manifested multicollinearity. Notably, the scatter plots of "Thinnes1019" and "Thinness59" revealed a high degree of correlation, indicating a potential issue of multicollinearity. Additionally, within the ordinary least squares (OLS) regression framework, both "Population" and "Diphtheria" were identified as statistically insignificant due to elevated p-values.

To solve these issues we used a Lasso regression, which enforces the reduction of some of the insignificant or multicollinear variables to zero. Subsequently, we did an OLS regression after eliminating the coefficients forced to zero by the Lasso regression. While the outcomes of this last OLS regression may lack statistical correctness in a conventional sense, they make the model more comprehensible by selectively holding only pertinent variables, thereby enhancing interpretability and focusing on factors deemed relevant to the overall model.

Interpretation of Results

To analyse the relationship between the explanatory variables and life expectancy we decided to use the MM-estimator since it is robust to outliers, which we know to be present in our case by looking at the boxplots of the variables. We also included the variable GDP in a logarithmic form to mitigate heteroscedasticity, which can be observed by looking at the scatter plot between the variable and Life Expectancy. The variable GDP can only assume positive values and among all the independent variables is the one with highest magnitude so a one-unit change could be meaningless to analyse, while through a logarithmic transformation it becomes meaningful since it measures a percentage change.

Many variables considered in our model are health-related: infant deaths, under-5 deaths, adult mortality, thinness (both between 5 and 9 and between 10 and 19), hepatitis B. The last one represents the immunization coverage among 1-year-olds, and one could expect that an increase in this variable has a positive effect on life

expectancy. Strangely, this seems not to be the case, since its coefficient is not only negative, but even strongly statistically significant. This could have happened since data about immunization are reported by national authorities, which could be politically interested into overestimating it. It could also be the case that different countries adopted different methods to gather immunization data, making them not comparable: this could result in a biased estimator.

Apart from hepatitis B (and Polio, even though it is not significant), all the other health-related variables are correlated with life expectancy in the way common sense would suggest. Since most of them are very significant, it is advisable for governments to try to influence them by spending more on healthcare, to increase life expectancy. Moreover, healthcare spending positively affects GDP, which in turn contributes to a higher life expectancy. Therefore, a country with a relatively low life expectancy (<65) could benefit from higher healthcare expenditure.

Infant and adult mortality rates play significant roles in determining life expectancy within a population. Both coefficients have negative effect on life expectancy (-0.035 and –0.048, respectively). This is attributed to the fact that our variable "adult mortality" accounts for deaths occurring among individuals aged 15 to 60, values that lower the global life expectancy average (69).

Another way in which governments seem to be able to increase the life expectancy of their citizens is by investing in education and by incentivising people to study for more years. Indeed, the impact of schooling on lifespan is positive and significant: each additional year of education contributes to an increase in life expectancy of 0.068 years. There probably exists a relationship between individuals' years of schooling and wealth (a variable that we did not control for), which can guarantee a better access to healthcare; so maybe there is a multicollinearity problem and the direct impact of schooling on life expectancy derived in our model is biasedly stronger than it is in the population model.

Surprisingly, alcohol consumption shows a significant positive coefficient at 5% level in our regression analysis, seemingly indicating a longer life expectancy with a higher alcohol intake. This outcome appears inconsistent with common knowledge, as alcohol consumption is widely recognized for its detrimental health effects and increased risk of various illnesses. However, it is noteworthy that a remarkable correlation exists between alcohol consumption and variables such as schooling, GDP, and the level of development. This correlation may be attributed to the fact that developed and rich countries tend to have greater financial resources, potentially facilitating increased alcohol consumption. In light of this, it can be argued that the positive coefficient associated with alcohol in the regression analysis may be influenced by the presence of multicollinearity with other variables. Therefore, caution is recommended in attributing a direct and causal relationship between higher alcohol consumption and an extended lifespan, as the observed association may be confounded by multiple other factors.

In conclusion, factors like mortality rates, healthcare spending, education and GDP significantly impact life expectancy. Careful consideration is essential when interpreting intricate relationships where numerous influences and confounding factors may come into play.

- Andrea Faramondi (1125550)
- Federico Monaci (1131561)
- Giovanni Gentilcore (1130975)
- Riccardo Pondini(1137045)
- Nicolas Romagnoli (1131264)