Term Project

Francesco Colzi - Daniel Vito Lobasso - Danny Prest - Ciro Spallitta - Roberto Vacante
0001127782 0001125016 0001140688 0001140660 0001131429

Data Presentation[1]

The dataset focuses on life expectancy and health-related variables for 179 countries. The information was gathered from the WHO data repository website, covering the time span from 2000 to 2015. However, the objective of this analysis is to focus on life expectancy, by using cross-sectional data. For this reason, we will only deal with the last year (2015), analysing an observation for each country. In particular, the dataset includes variables such as economic-demographic information (GDP, Population), health-related factors such as life expectancy, immunization coverage (measles, hepatitis B, diphtheria, polio), deaths/illness per 1000 (HIV, infant mortality, adult mortality, etc ...) and other health-status information about the population (BMI, thinness, alcohol usage).

Descriptive Statistics

Table 1 shows a summary statistic of the aforementioned variables. Moreover, it might be more informative to include other statistics from the empirical probability density function of the variable of interest, such as the followings

	Life_expect
Lower Quartile	66.3
Median	73
Upper Quartile	76.85
Mode	74.25

Notably, the median is not that close to the mean (equal to 71.46). As a matter of fact, these statistics would suggest that the distribution is not symmetric (it is indeed left-skewed).

Data Visualization

A first insight into the relationships between the variables in this dataset can be drawn by looking at some scatter plots, which can help us visualize the data (See Figs. 3, 4 and 5). As life expectancy is our variable of interest, we can have a closer look at the scatter plots regarding its relationship with mortality, healthcare, and wealth. For example, plotting against *Infant Deaths* or *Thinness* show a negative relationship, indicating that countries with higher infant mortality or a higher rate of malnourished children have a lower life expectancy. The relationship between life expectancy and immunization rates seems to be positive, as countries with high life expectancy all show high rates of vaccinations.
To further evaluate the relationships between the variables a correlation matrix can be used, as in Figure 2. As shown, some of the variables (e.g. *Thinness5_9* and *Thinness10_19*, *Under5Deaths* and *Infant Deaths*) exhibit a really high correlation as they capture the same kind of observations. Variables capturing immunization rates (*Polio*, *Hepatitis_B*, *Diphteria*) are also highly correlated (intuitively, it is more likely to be vaccinated against multiple diseases). As for *life expectancy*, it shows a high negative correlation with mortality rates. The visualization of the variables in boxplots (See again Figure. 1) is an efficient method of spotting outliers. For example, as some countries (e.g. China or India) have a significantly

[1]The hierarchy of this project groups figures and tables in the last few pages. It is organized in a way that allows quick linking (by clicking on the reference) between the figures and the text.

higher population, we can observe a large quantity of outliers in the corresponding boxplot. Similarly, as some countries have lower healthcare conditions or low immunization rates, large quantities of outliers can be spotted.

Regression Analysis

Simple Regression and Multiple Regression

Based on the data collected from previous analyses, some variables are strongly related to each other: they might represent a situation of collinearity, the phenomenon for which one predictor variable in a multiple regression model can be perfectly predicted from the others. For this reason, we choose a linear regression model that excluded the following variables:

- *Infant_deaths*, *Under_five_deaths* and *Adult_mortality*: these variables have not been considered because they are strongly connected with themselves and also, they give the same information of the dependent variable *Life_expectancy* that includes by its definition all previous data.
- *Thinness_10_19*: this variable has been excluded in favor of the variable *Thinness_5_9* which is redundant for explaining *Life_expectancy*, so maintaining both variables doesn't improve the regression model.
- *Hepatitis_B* and *Diphtheria*: following the same reasoning as before, the high correlation of these two variables with the third predictor *Polio* shows how the countries that resort to vaccinations carry out the latter for all three cases of immunization, so keeping all the variables in the model does not increase its predictive power. This is the reason why we decided to keep only immunization data about *Polio*.
- *Population*: it involves very spread data between themselves for the natural differences among the countries and it doesn't help to make any consideration on Life_expectancy.

Instead, for what concerns the variable *GDP*, the scatterplot shows us the situation according to which for low *GDP* values, small increases in the explanatory variable lead to a sharp increase in life expectancy. This relationship reverses as the same values of *GDP* grow, so that inversely, for high *GDP* values, large increases in the independent variable imply only a small increase in the independent variable. This is why we consider that the logarithm of *GDP* better explains the trend of the variable (See Fig. 6).

Regarding the analysis, we ask ourselves what covariates have an influence on life expectancy among the 179 countries. For this reason, one can progressively carry out different regression models (See Table 2) in this sense. As shown, we start with a simple regression model, which shows a positive significant effect of *log(GDP)* on life expectancy, but might suffers from omitted variable bias. Higher *log(GDP)* generally reflects better access to healthcare, nutrition, and overall living conditions. The coefficient decreases as more variables are included, suggesting that some of the impact of GDP on life expectancy might be given by factors similar to schooling and health interventions. By controlling for immunization coverages (such as polio, measles and diphtheria vaccines) we find that only *Polio* (among the three) suggests a significant impact on life expectancy. Still, even testing for joint significance (F-test) on *Measles* and *Diphtheria*, we find a non-significant impact on the dependent variable. Furthermore, we continue to specify the model by adding *Schooling* and *BMI*. Among the added variables, only *Schooling* seems to have a significant coefficient. Finally, the last model includes other factors that do not influence the significance of the previous model's variable, though we can see their magnitude has decreased. Not surprisingly, the variable *HIV* deaths has a significant (at even the 1% level) negative impact on Life Expectancy.

Residual Analysis

Once we estimated the model, the residual values are obtained by subtracting the true value from the predicted one. If the model is correct, we should expect that the extracted residuals follow a distribution close to the normal. Indeed, one can directly check this fact through a graphical representation. It is also possible to use standardised residuals, given by $\frac{r_i}{\hat{\sigma}}$. However, the kernel density estimator (in Fig. 7) allowed us to directly confirm some facts. First of all, the distribution of residuals is not that far from a normal distribution. It is indeed quite symmetric and its expected value is approximately equal to zero. Moreover, by plotting residuals against explanatory variable (See Fig. 8), we see the lack of a relationship, which is in line with the zero conditional mean assumption of Ordinary Leasts Square (*i.e.* $\mathbb{E}(u|x) = 0$).

Inference and Bootstrap Analysis

Given the previous section, we see that model diagnostic performed quite good. This might indicate that OLS assumptions are not violated. Therefore, even if the use of non-parametric bootstrap is less critical in this scenario, there are still some reasonable justification for its construction.
On top of this, the non-parametric bootstrap can be particularly beneficial in cases where we deal with a not that large sample size. Indeed, it would provide more reliable confidence intervals by resampling directly from the observed data. This would help mitigate issues related to finite sample sizes by providing a more accurate estimate of parameter variability.
As a matter of fact, even if this does not validate the analysis, it is still useful to assess the stability of parameter estimates. After constructing CIs from a non-parametric bootstrap, it results to be really close to OLS confidence intervals. The consistency observed between OLS and bootstrap intervals could be partially attributed to a well-fitted model. Since diagnostic checks indicate a reasonable fit and the assumptions underlying OLS are not severely violated, the intervals derived from both methods converge (See Table 4).

Testing with Robustness

We know that the OLS model is sensitive to outliers, which can potentially lead to biased results (As shown in the example in Fig. 11). The robust MM-estimators model is designed to limit the influence of such outliers, providing a more reliable estimate of the coefficients by giving less weight to these extreme observations. In Table 3 and Figs. 9 - 10, when comparing the coefficients between the OLS and robust models, we observe some differences that indicate the influence of outliers on the OLS estimates. For example, the *log(GDP)* coefficient slightly increases from 2.59 in the OLS model to 2.69 in the robust model. The coefficient for *Developed* decreases from 1.61 to 1.24 in the robust model, which could imply that outliers were artificially inflating its effect in the OLS model. However, coefficients for *Schooling*, *Measles*, *Polio* and *HIV* remain approximately of the same effect size and statistical significance. In conclusion, the robust approach seems to provide a slightly better fit for the data, while maintaining the overall direction and magnitude of the relationships seen in the initial OLS model.

Test Size simulation

The test size, denoted as λ_α, is a critical concept in hypothesis testing. It is the probability of incorrectly rejecting the null hypothesis (H_0) when it is, in fact, true, which is known as a Type I error. The test size should ideally be equal to the significance level if the test is properly calibrated. In the context of the robust MM-estimators model, the empirical test size is obtained by repeatedly simulating samples under the null hypothesis and performing robust *t-tests* on each sample. The empirical test size is the proportion of times H_0 is rejected ($p-value < \alpha$) across these simulations. Looking at the empirical test sizes given for the robust MM-estimators model: for *log(GDP)*, *BMI*, *Developed*, *Schooling*, *Measles*, *Polio*, and *Alcohol*, the empirical test sizes are slightly above the 0.05 threshold but still quite close. Regarding *Thinness_5_9*

(0.069) and *HIV* (0.14), we see that the empirical test sizes are higher than the significance level. This could imply that the tests are too liberal, leading to a higher-than-acceptable probability of Type I errors. In summary, while the empirical test sizes for most variables in the robust MM-estimators model are close to the nominal level, suggesting an adequate sample size and well-calibrated tests, the test for *HIV* shows a need for caution and perhaps a re-evaluation of the model or the robustness approach for this variable.

Bias-Variance Tradeoff

In this context, the Lasso Estimator (L1 regularization) has been employed to build a tool for feature selection, in order to prevent overfitting. The Lasso Estimator penalizes the less influential variables by pulling the coefficients toward zero. However, it is important to say that this dataset might not need such penalization on the predictors, due to a not large number of the latter.

Lasso Estimation

As said above, throughout this section, Lasso Estimator will be mostly used for its shrinkage effect, which involves some steps.
The process starts by choosing a tuning parameter λ, devoted to control the strength of the penalty. To perform this process, we rely on ten-fold cross-validation, which is the common choice for choosing the tuning parameter for the lasso regression. This method splits the sample between training and testing dataset respectively in the proportions of 90% and 10%.

One can plot the value of the slopes with respect to different values of λ (See Fig. 12). The λ parameter chosen is 0.064 and it will describe the strength of the penalization. As depicted in the graph, this lambda suggests to drop one variable ($Thinness\_5\_9$) by setting its coefficient toward zero. Not surprisingly, in the original model, the t-statistic of the coefficient (estimated with OLS) was already close to zero.

After comparing the two models in Table 5, we see that some variables appear with slightly different slopes. As one could expect, the parameters are likely to contain a source of bias. A direct visual representation is provided in Figure 13. As depicted, the coefficient associated with the variable *Alcohol* has slightly changed and its distribution has shifted away from the original OLS.

Regarding the shrank variable from the Lasso estimator (*Thinness_5_9)*, we see from its Box-Plot that its probability mass is now concentrated toward negative values.

Generally, the shrank model does not seem to exhibit big changes in the variability of the distribution than the unrestricted model.

List of Tables and Figures

Table 1: **Summary Statistic**

	Obs	Mean	SD	Min	Max
Life expectancy	179	71.46	7.83	50.90	83.80
Schooling	179	8.36	3.15	1.40	14.10
Thinnes10-19	179	4.55	4.12	0.10	26.70
Thinness5-9	179	4.59	4.20	0.10	27.30
GDP	179	12617.30	17719.61	306.00	1.1e+05
HIV	179	0.61	1.62	0.01	14.30
Diphtheria	179	87.92	14.69	16.00	99.00
Polio	179	88.26	13.02	37.00	99.00
BMI	179	25.60	2.19	20.50	32.10
Measles	179	80.23	16.19	21.00	99.00
Hepatatitis B	179	87.10	14.17	22.00	99.00
Alcohol	179	4.73	3.74	0.00	16.72
Adult mortality	179	163.67	89.95	49.38	513.48
Infant deaths	179	23.56	21.49	1.80	95.10

Figure 1: Box-Plots

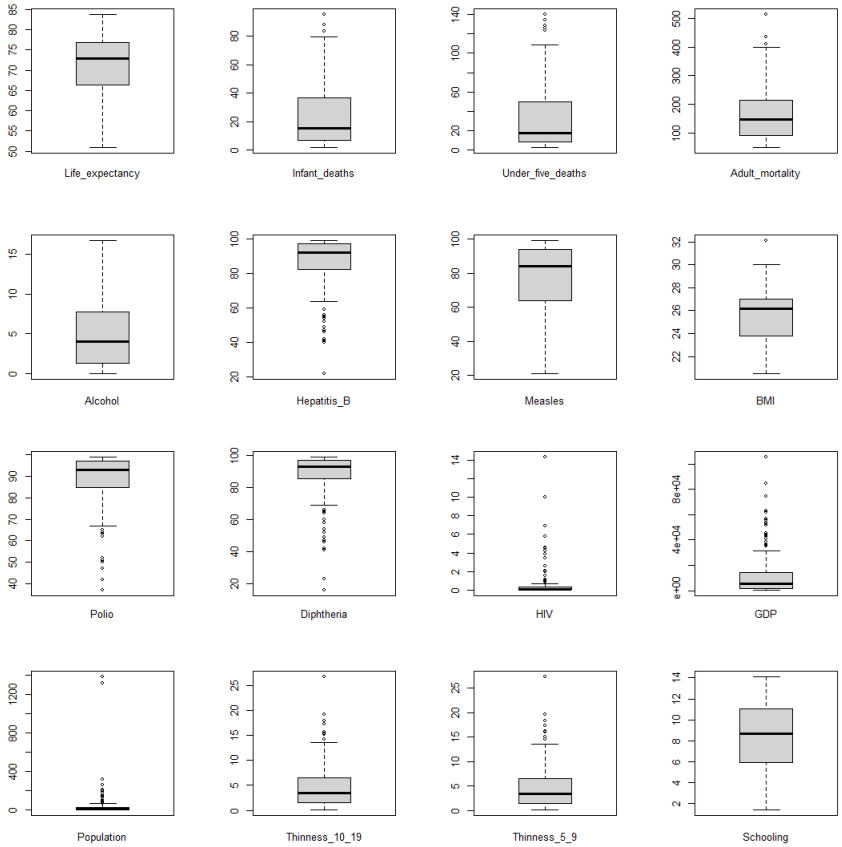


Figure 2: Correlogram

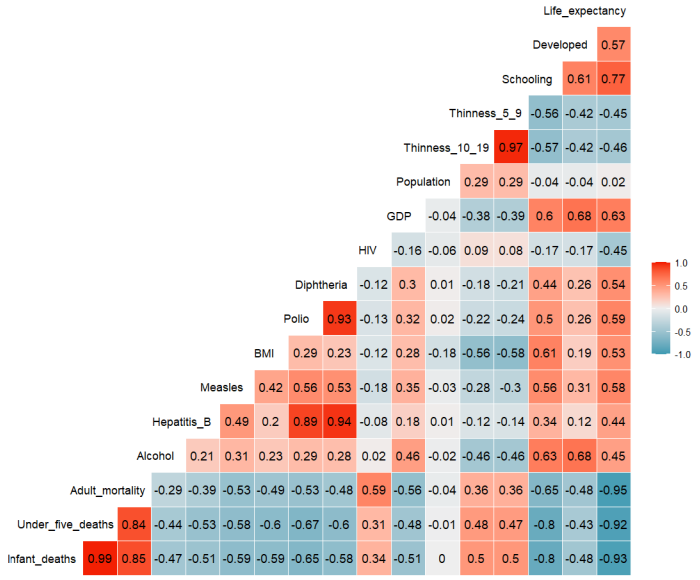


Figure 3: Twoway scatter-plots

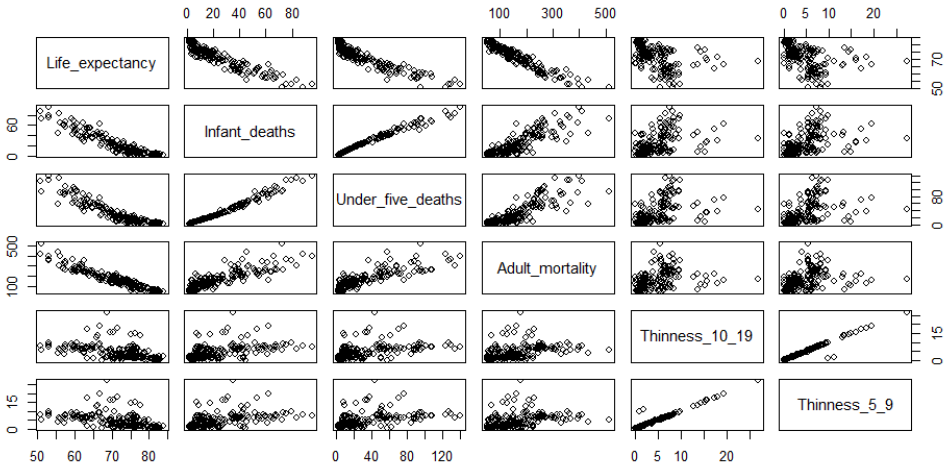


Figure 4: Twoway scatter-plots

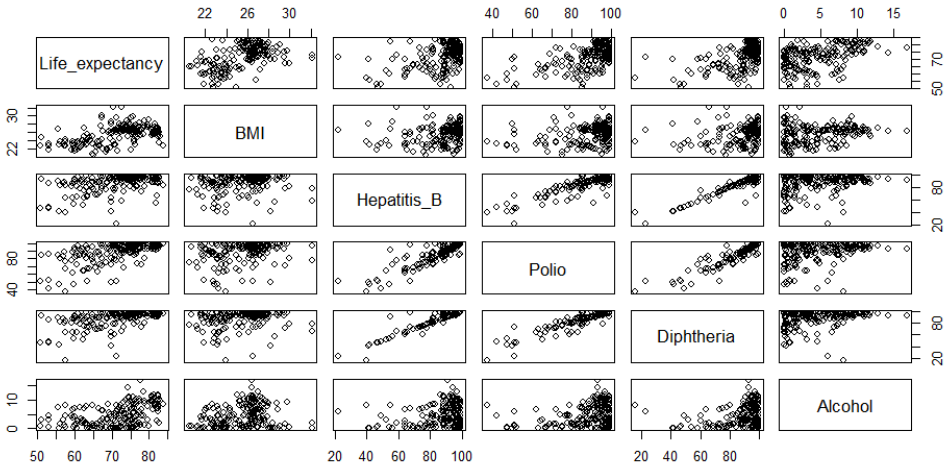


Figure 5: Twoway scatter-plots

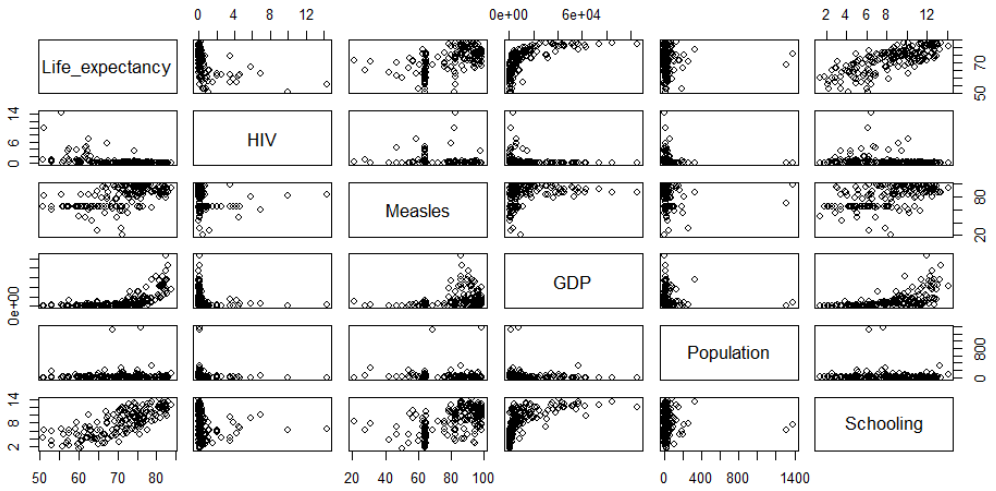


Figure 6: Different functional form

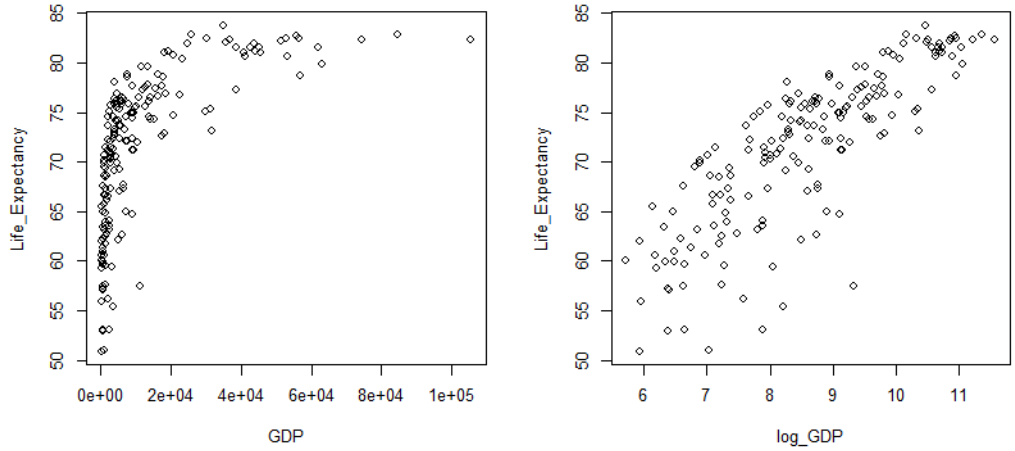


Table 2: **Regression Analysis**

	(1)	(2)	(3)	(4)
	Life Expectancy	Life Expectancy	Life Expectancy	Life Expectancy
log(GDP)	4.624***	3.764***	2.919***	2.596***
	(0.242)	(0.273)	(0.376)	(0.333)
Polio vaccine		0.220***	0.171***	0.130***
		(0.0654)	(0.0656)	(0.0238)
Measles vaccine		0.0363	0.0235	0.00864
		(0.0250)	(0.0249)	(0.0205)
Diphtheria vaccine		-0.0719	-0.0388	
		(0.0565)	(0.0562)	
BMI			0.158	0.216
			(0.175)	(0.172)
Schooling			0.468***	0.462***
			(0.176)	(0.162)
Alcohol				-0.125
				(0.102)
HIV deaths				-1.353***
				(0.160)
Thinness5-9				0.0107
				(0.0805)
Developed				1.612*
				(0.968)
Constant	31.82***	23.22***	24.90***	28.70***
	(2.106)	(2.359)	(4.553)	(4.443)
Observations	179	179	179	179
R-squared	0.673	0.735	0.750	0.834
Adj.R-squared	0.671	0.729	0.741	0.825

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure 7: Non-robust Residuals Analysis

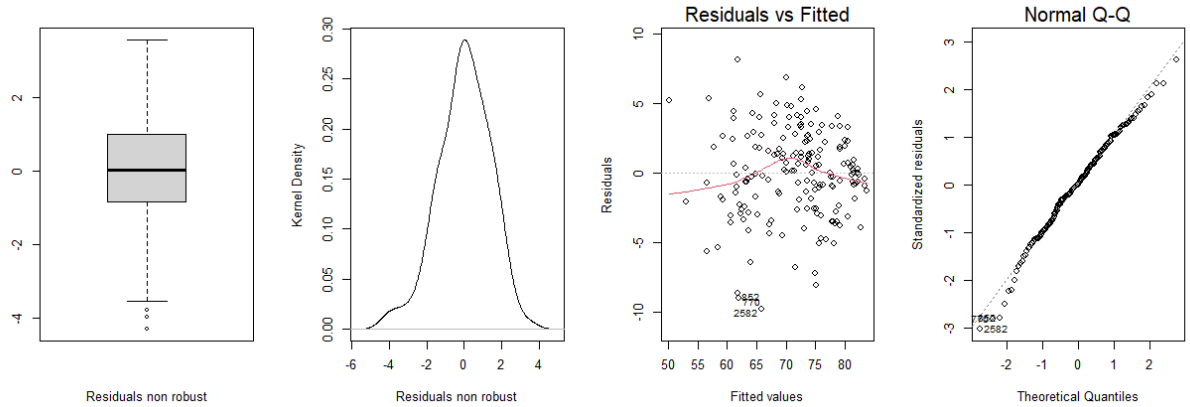


Figure 8: Non-robust Standardised Residuals

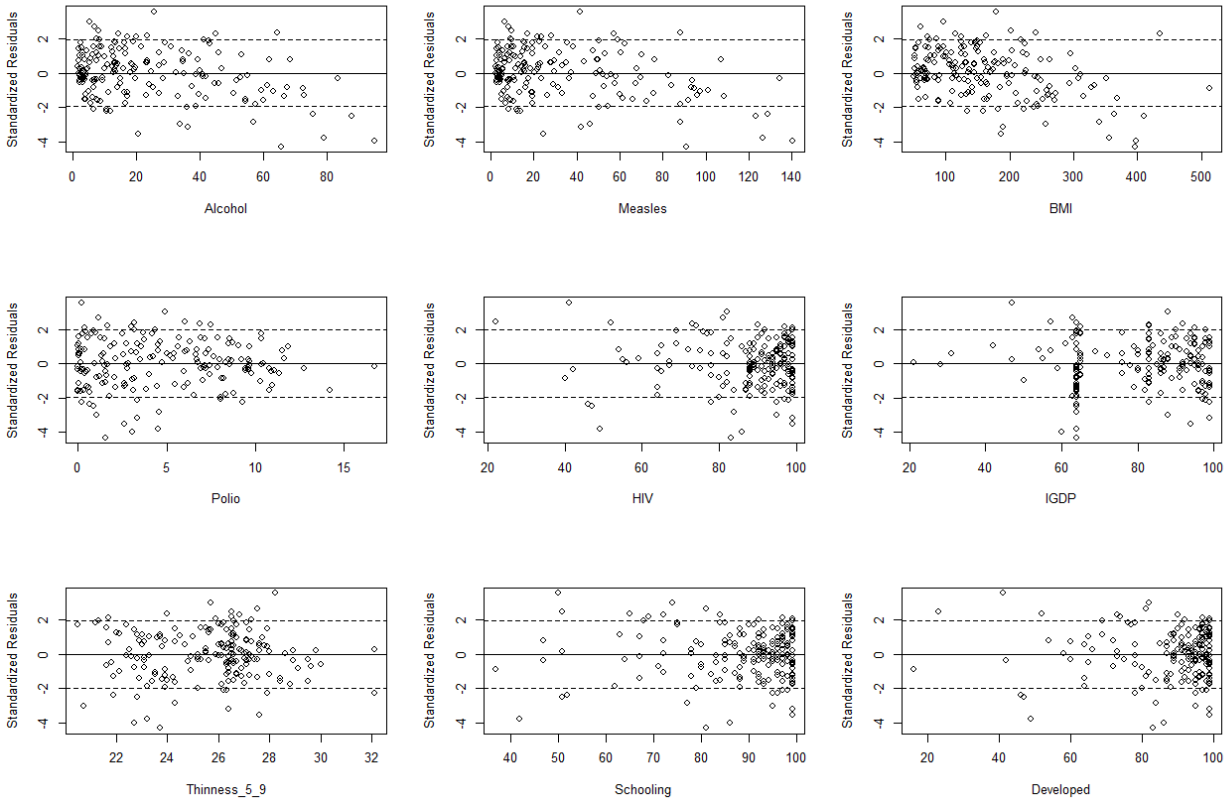


Figure 9: Outliers-Robust Residual Analysis

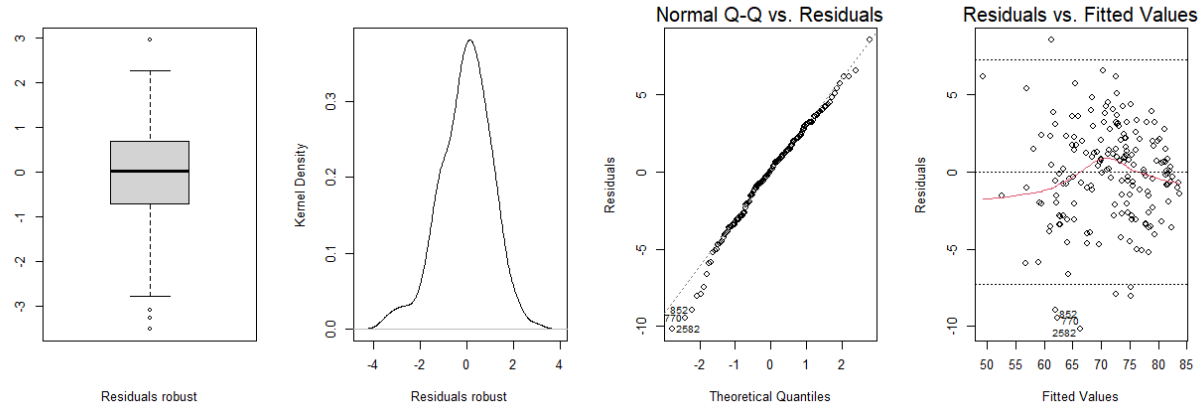


Figure 10: Outliers-Robust Standardised Residuals

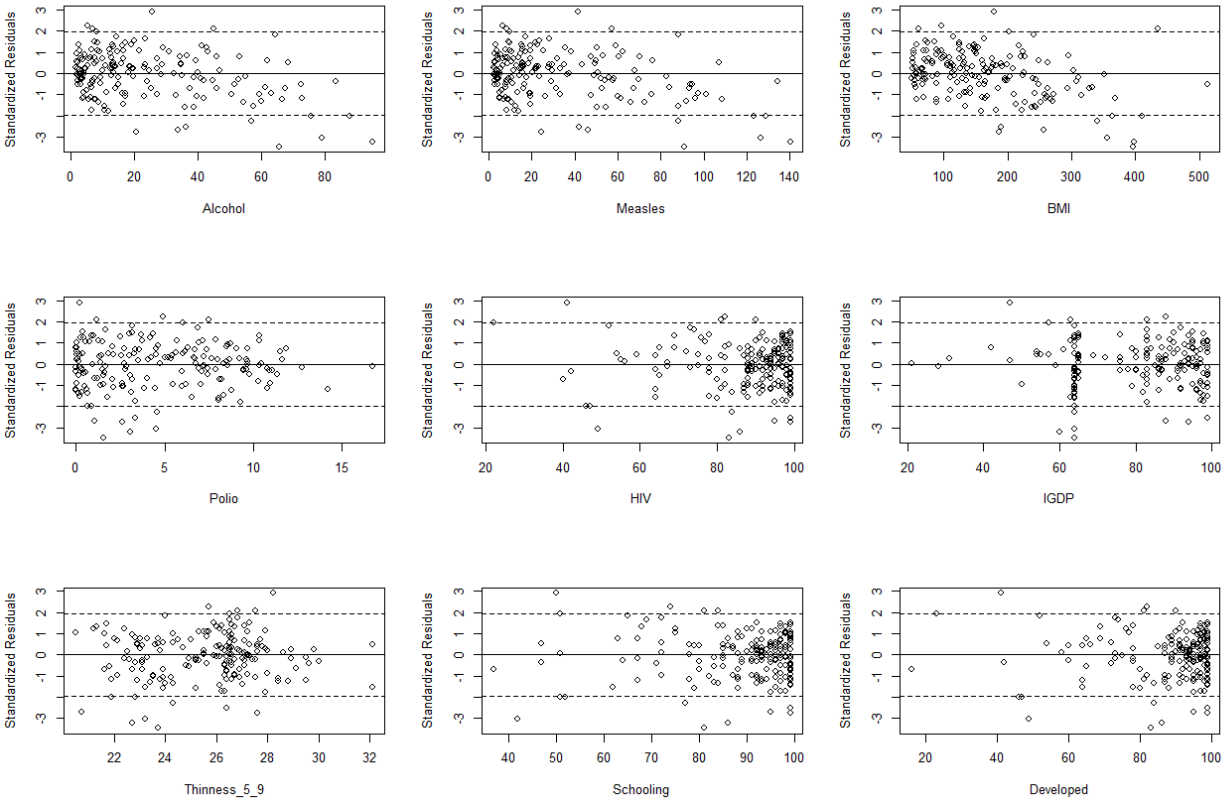


Table 3: Output of Robust Regression and OLS

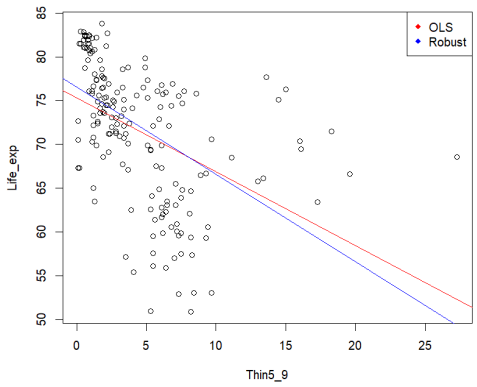
	(1) Life Expectancy	(2) Life Expectancy
Alcohol	-0.123002	-0.125055
	(0.084042)[0.0569]	(0.102000)
Measles	0.011382	0.129899
	(0.016115)[0.0636]	(0.020500)
BMI	0.076630	0.008636
	(0.176178)[0.0621]	(0.172000)
Polio	0.132319**	0.215529***
	(0.039875)[0.0648]	(0.023800)
HIV	-1.425195***	-1.353044***
	(0.289149)[0.1422]	(0.160000)
log(GDP)	2.688931***	2.596416***
	(0.316384)[0.0568]	(0.333000)
Thinness_5_9	0.005994	0.010725
	(0.067891)[0.0694]	(0.080500)
Schooling	0.444708*	0.461932***
	(0.176113)[0.0579]	(0.162000)
Developed	1.252780	1.611860*
	(0.949055)[0.0523]	(0.968000)
Constant	31.506319***	28.700816***
	(5.438882)	(4.443000)
Observations	179	179
Model	MM-estimates	OLS

Standard errors in round brackets

Empirical test size in square brackets

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure 11: Outliers-robust regression compared to non-robust OLS



	OLS	Robust
(Intercept)	75.34	76.53
Thin_5_9	-0.84	-1.00

Table 4: **Compared Confidence Intervals and Coverage**

	CI (Linear Model)	CI (Bootstrap)
log(GDP)	[1.9398, 3.2529]	[1.9949, 3.2541]
Developed	[-0.2995, 3.5232]	[-0.2170, 3.2735]
Schooling	[0.1422, 0.7816]	[0.1286, 0.7597]
Thinness5-9	[-0.1481, 0.1695]	[-0.1731, 0.1469]
BMI	[-0.1231, 0.5541]	[-0.1324, 0.5867]
Measles	[-0.0318, 0.0490]	[-0.0277, 0.0451]
Polio	[0.0828, 0.1769]	[0.0645, 0.1921]
HIV	[-1.6679, -1.0381]	[-1.9264, -1.0829]
Alcohol	[-0.3267, 0.0766]	[-0.2929, 0.0526]

Figure 12: Feature Selection: the choice of the Tuning parameter λ as a function of MSE (to the left) and the slope of the coefficients (to the right).

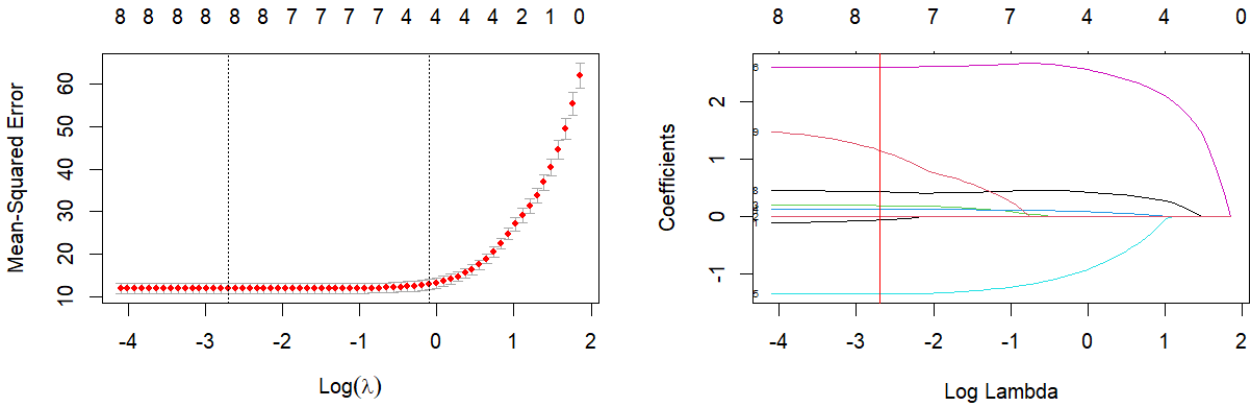


Table 5: **Regression output of Lasso and OLS**

	(1)	(2)
	Life Expectancy	Life Expectancy
Alcohol	-0.058991	-0.125055
Measles	0.007667	0.129899
BMI	0.193969	0.008636
Polio	0.127047	0.215529
HIV	-1.347666	-1.353044
log(GDP)	2.606449	2.596416
Thinness_5_9	.	0.010725
Schooling	0.432297	0.461932
Developed	1.164780	1.611860
Constant	29.569903	28.700816
Observations	179	179
Model	Lasso	OLS

Figure 13: Box-Plots compared: respectively Lasso coefficient, OLS and Post-Lasso OLS