LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

REGRESSION MODELLING PROJECT

December 2022

1. EXPLORATORY DATA ANALYSIS

1.1 CLEANING AND TIDING THE DATA

The original data (nrows=2938) was reduced by 43.87% after eliminating all the NA values (Fig. 1). The new data comprised 1649 rows. The details of the original number of NA values per column can be observed in the Appendix, Fig. 1.

The highest proportion of NA values is in the Population variable (652), followed by Hepatitis B (553) and GDP (448).

Columns were renamed and relocated to facilitate further work (Fig. 1). The head of the resulting dataset is shown in the Appendix, Tab. 1. It must be pointed out that even when the Country column was kept, it was not included in the analysis. Of the 21 variables to be analysed, only one is categorical: SD.

Life Expectancy	**LE**	Income composition of resources	**ICR**	Diphtheria	**DIPH**
Year	**YR**			Hepatitis B	**HEPB**
Status	**SD**	Schooling	**SCH**	HIV/AIDS	**AIDS**
Population	**POP**	Infant deaths	**U1D**	Alcohol	**ETOH**
GDP	**GDP**	Under-five deaths	**U5D**	Thinness 5-9 years	**THIN5_9**
Percentage Expenditure	**PE**	Adult Mortality	**AM**	Thinness 1-19 years	**THIN1_19**
		Measles	**MSS**	BMI	**BMI**
Total Expenditure	**TE**	Polio	**POL**		

FIG. 1 RENAMING VARIABLES

1.2 DESCRIPTIVE STATISTIC OF NUMERICAL VARIABLES

Table 1. shows the descriptive statistics of all numerical variables.

Given that it is a percentage, the Percentage Expenditure (PE) values cannot exceed 100. In this case, it has a median (145.10), a mean (698.97) and a max (18961.35) over 100.

Something similar happened with the rate of infant deaths per 1000 population (UI1) and the rate of deaths of individuals under five years old per 1000 (UI5) since these are rates of death per 1000, and both have maximums bigger than 1000 (1600 and 2100 respectively). There is a similar issue with Reported cases of measles per 1000 population (MSS), with a mean (2224) and a max (131441) above 1000. In this variable, it is also significant the difference between the mean and the median (15).

TABLE 1. DESCRIPTIVE STATISTIC

col_name	min	q1	median	mean	q3	max	sd
LE	44.00	64.40	71.70	69.30	75.00	89.00	8.80
YR	2000	2005	2008	2008	2011	2015	4
POP	34	191897	1419631	14653626	7658972	1293859294	70460393
GDP	1.68	462.15	1592.57	5566.03	4718.51	119172.74	11475.90
PE	0.00	37.44	145.10	698.97	509.39	18961.35	1759.23
TE	0.74	4.41	5.84	5.96	7.47	14.39	2.30
ICR	0.00	0.51	0.67	0.63	0.75	0.94	0.18
SCH	4.20	10.30	12.30	12.12	14.00	20.70	2.80
U1D	0	1	3	33	22	1600	121
U5D	0	1	4	44	29	2100	163
AM	1	77	148	168	227	723	125
MSS	0	0	15	2224	373	131441	10086
POL	3	81	93	84	97	99	22
DIPH	2	82	92	84	97	99	22
HEPB	2	74	89	79	96	99	26
AIDS	0.10	0.10	0.10	1.98	0.70	50.60	6.03
ETOH	0.01	0.81	3.79	4.53	7.34	17.87	4.03
THIN5_9	0.10	1.70	3.20	4.91	7.10	28.20	4.65
THIN1_19	0.10	1.60	3.00	4.85	7.10	27.20	4.60
BMI	2.00	19.50	43.70	38.13	55.80	77.10	19.75

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

The variable BMI has questionably very high values, considering that more than 30 Is obesity and more than 70 is morbid obesity. In this dataset, this variable has a median=43.70, a mean=38.13 and a max=77. All these values are inconsistent with the range of BMI.

The variable Population has problems, too. The min is 34, an odd figure because no country has so few inhabitants. When checked further, the variable shows enormous fluctuations in the same country between years. In addition, 22.2% of it has NA values. It was considered a low-quality variable for all these issues and was not considered for the following analysis.

1.3 CORRELATION BETWEEN VARIABLES

Figure 2 explores the relationship between all the numerical variables through a correlation matrix. Life expectancy (LE) showed strong positive correlations with Income composition of resources (ICR) (r=0.72) and Schooling (SCH) (r=0.73) and a strong negative correlation with Adult Mortality (AM) (r=-0.70). Also, it has a weak positive correlation with BMI (-0.54) and a weak negative correlation with HIV/AIDS (AIDS) (r=-0.59).

There are several variables in the range between 0.40 and 0.50, showing either a weak negative or positive correlation: GDP(0.44), PE (0.41), ETOH(0.40), THIN5_9 (-0.46) and THIN1_19 (-0.46). The rest of the variables have values lower than 0.40.

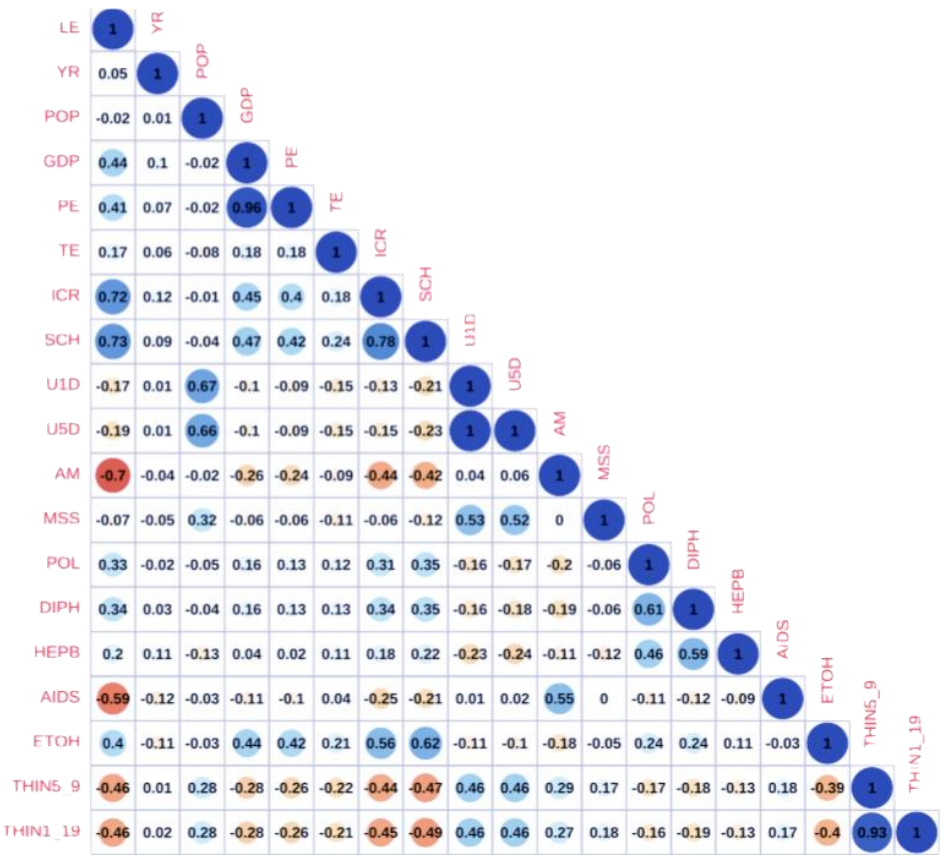


FIG. 2 CORRELATION MATRIX

All these findings seem logical considering the interrelation between these factors in human life.

The correlation matrix (Fig.3) suggested multicollinearity between some of the variables: U1D and U5D (r=1.00), THIN5_9 and THIN1_19 (0.93), and GDP and PE(r=0.96). It is logical because each of these pairs of variables is related.

Other independent variables have some multicollinearities, such as POP with U5D and AM, SCH with ICR and THIN5_9 or POL with DIPH. This point is checked in the following step.

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

1.4 MULTICOLLINEARITY

The multicollinearity of the data was also checked using the Variance Inflation Factor (VIF). As a result of this analysis, the multicollinearity between GDP and PE, U1D and U5D, and THIN5_9 and THIN1_19 (Tab. 2) was confirmed.

TABLE 2. VIF VALUES

GDP	13.649710
PE	12.904426
U1D	213.609554
U5D	203.591034
THIN5_9	7.584832
THIN1_19	7.606109

Only one of the variables of the pair that shows multicollinearity is kept for the next steps in the following way:

- Pair U1D and U5D: both show similar multicollinearity to the other variables. They have 0% of NA data. Finally, it is estimated that U5D contributes more to our data because it covers a broader life period.
- Pair THIN5_9 and THIN1_19: both show similar multicollinearity to the other variables. They have similar percentages of NA data. Finally, it is estimated that THIN1_19 contributes more to our data because it covers a broader life period.
- Pair GDP and PE: both show similar multicollinearity with the rest of the variables, even when GDP is slightly more correlated. GDP has a high percentage of NA data, with 15.2%, the third variable with more NA values of all variables in the dataset. On the other hand, PE does not have any NA value. Also, PE is a variable linking health and GDP (that is more general) and could contribute more to our data.

1.5 OUTLIERS

The outlier detection is conducted through boxplots (Fig. 3). It is noticed that most numerical variables have outliers. This result is logical because it is a dataset with numerous entries worldwide. The outliers will be kept in our data because they result from the world's variability. However, considering the result of the descriptive statistic, some outliers are the product of problems with our data. In the cases of U1D, U5D and MSS, because they are rated per 1000, they cannot have values above 1000. Something similar happened with PE which should have a maximum value of 100. It is noticed that in the previous step, the variable U1D was dedicated. All these inconsistent values in U5D, MSS and PE were deleted before the analysis continued.

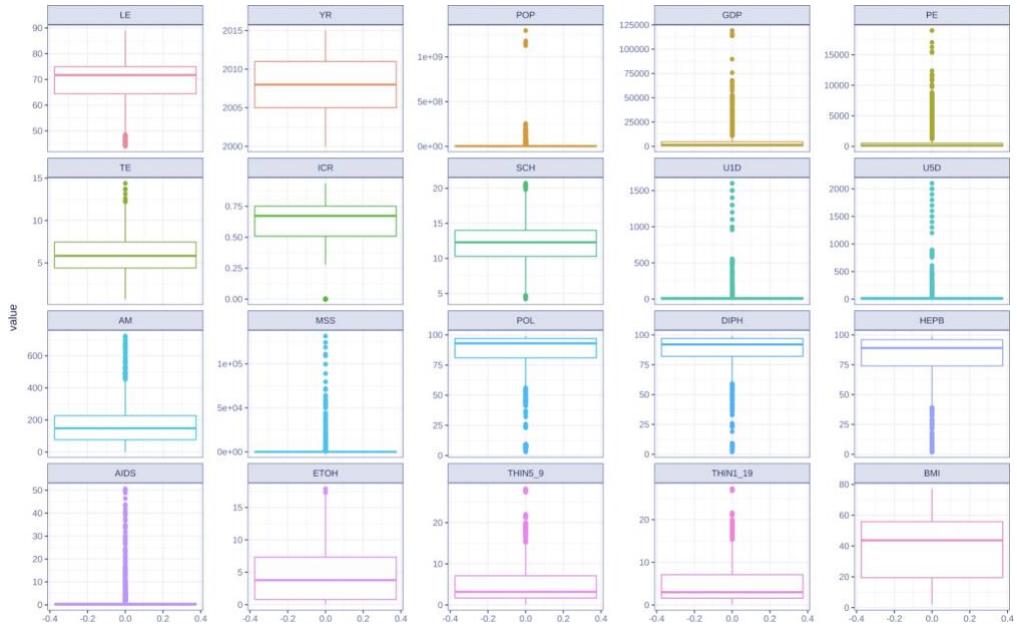


FIG. 3 OUTLIERS IN ALL THE VARIABLES CONSIDERED UNTIL THIS POINT

The data was reduced to 569 entries, mainly for values over 100 in PE. This had an impact on the data, decreasing its size substantially. However, even when PE could have made more contributions to the analysis, this variable in the dataset has so many inconsistent values that it is impossible to work with it. For this reason, in the pair of variables GDP-PE, which shows multicollinearity, it was decided to change the previous decision and choose GDP.

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

1.6 LINEARITY

Figure 4 shows scatterplots of LE versus all the independent variables considered until this point, and after deleting, the outliers were checked. Again, curvatures in several independent variables were detected. They are especially important in AIDS, GDP, and U5D. Therefore, at this stage, the need for transformation is likely.

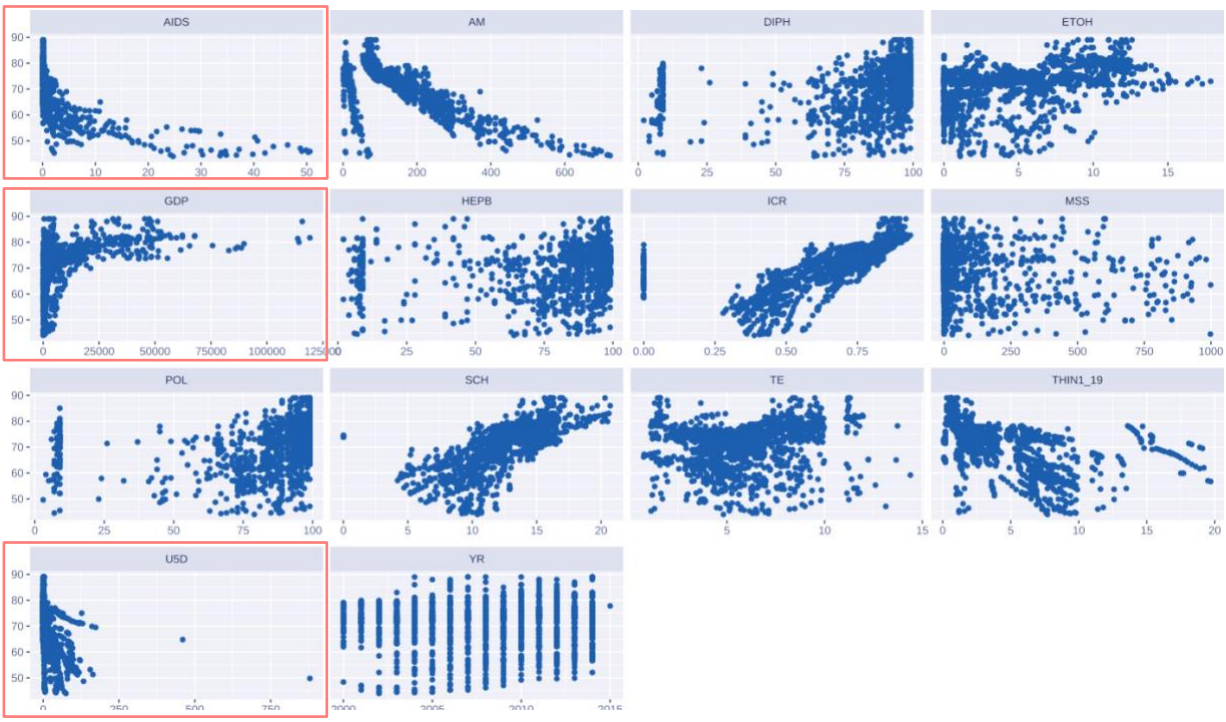


FIG. 4 RELATION BETWEEN LIFE EXPECTANCY AND ALL THE NUMERICAL VARIABLES

1.7 TRANSFORMATIONS

The transformation of the independent variables is carried out through Tukey's Ladder way (Fig. 5). Different transformations (λ=0, 0.5, 2, -0.5) were applied to various variables, and the results are presented in Appendix, Fig. 2, and Table 2. Following the analysis of R2 and adjusted R2 before and after each transformation, it is concluded that transforming GDP, AIDS, and U5D is necessary. It is decided to apply the same type of transformation to these three variables to avoid further complicating the model.

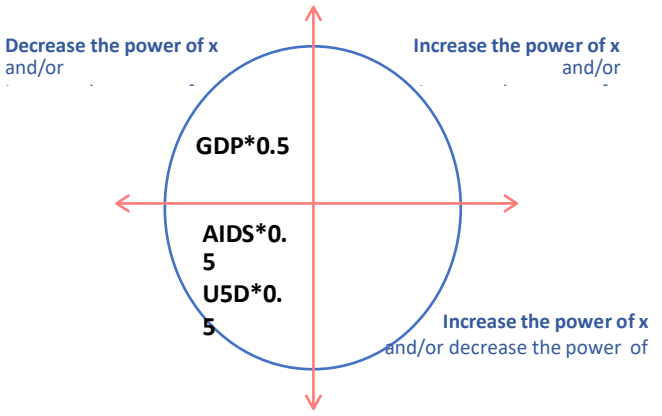


FIG. 5 TUKEY´S LADDER WAY FOR is

In these three variables, the transformation goal is to decrease the power of X. A λ=0, and a λ=0.5 were tested. After transforming GDP, AIDS and U5D, R^2 increases from 0.8193 to 0.8358, with λ=0, but goes up to 0.8514 with λ=0.5. Something similar happened with adjusted R^2, from a value of 0.8176 without transformation, increased to 0.8343 with λ=0 and to 0.8501 with λ=0.5, after the transformation. It is concluded that the square root shows better results even when the log transformation is generally more powerful.

The potency of ICR also increases R2 and adjusted R2. However, this type of transformation leads to a highly complex polynomial model. Therefore, it is decided not to conduct this transformation and evaluate after finding the model. This will be the starting point. Then, after fixing the model, new transformations could be made.

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

1.8 VARIABLE YEAR

The scatterplot of LE against YR (Fig. 6) and the LE median against YR (Fig. 7) show a patron. Therefore, this variable is considered to fix the model.

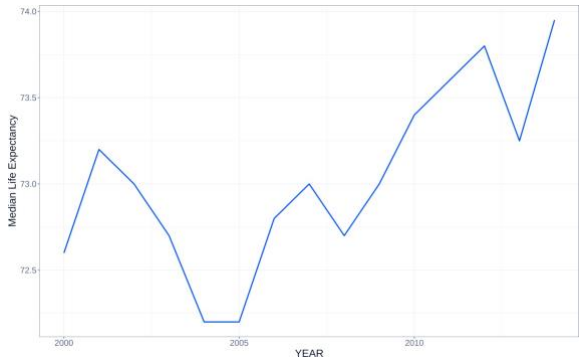


FIG. 6 MEDIAN OF LIFE EXPECTANCY FROM 2000 TO 2014

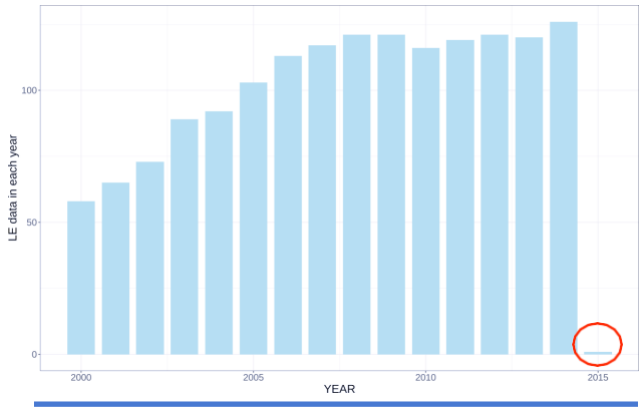


FIG. 7 LACK OF INFORMATION IN YEAR 2015

However, plotting these variables (the last graph in Fig 5 and Fig. 8) shows that 2015 has a vital data problem: it contains only one LE value. Therefore, this year won't be considered in the analysis.

1.9 VARIABLE STATUS DEVELOPING

Status developing is the only categorical variable, with two categories: developed and developing. To conduct the analysis, it is transformed into a factor and renamed as SD_f.

1.10 CONCLUSIONS OF THE EXPLORATORY DATA ANALYSIS

- The data set's BMI and POP variables are not considered in the analysis due to inconsistent values. They are low-quality variables. (POP min=34 / BMI mean, median and max extremely high).
- After detecting multicollinearity between some independent variables (THIN5_9 - THIN1_19/ U1D-U5D / PE- GDP), the variables THIN5_9, U1D and PE were discarded.
- The outliers in all the variables are kept except in U5D and MSS, where the values over 1000 are deleted for inconsistency regarding the definition of the variables (Rate per 1000).
- The categorical variable was transformed into a factor to construct the model. The interaction between SD and the rest of the independent variables will be checked at the end of the process.
- The year 2015 will not be considered. The time series to analyse will be from 2000 to 2014.

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

2. FITTING THE MODEL

2.1 BEST SUBSET SELECTION

This analysis is carried out through the regsubsets( ) function in R Studio.

After plotting the results of regsubset code (FIg. 8), a model that contains around 12 variables could be the best option.

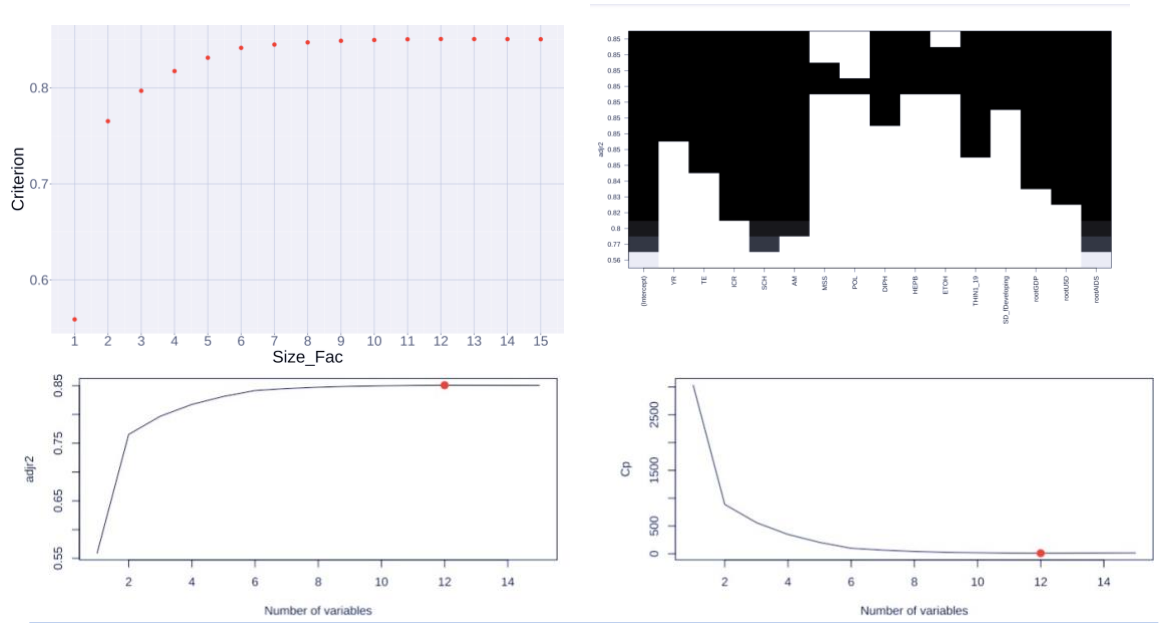


FIG. 8 BEST SUBSET SELECTION

2.2 BEST MODEL SELECTION

To find the best model, six variants were tested: Leaps with 12 variables (11 independent + response); Leaps with 13 variables (12 independent + response); Leaps with 14 variables (13 independent + response); forwards selection, backwards selection and stepwise selection. The details of each test can be seen in the Appendix. The forward, backward, and stepwise selection obtained the same result. This result also matched the one obtained through Leaps with 13 variables.

TAB. 3 SELECTION OF THE FINAL MODEL

Model	p	adjr2	Cp	PRESS
Leaps - Size12	11	0.8506	12.50	16,688.11
Leaps-Size13 & Stepwise	12	0.8509	10.55	16,667.25
Leaps - Size14	13	0.8508	12.16	16,693.37

The final selection was made based on the adjusted R2, Cp and PRESS (Tab. 3). There are slight differences concerning the adjusted r-squared. In the three options, Cp < p+1. Then, only by analysing this aspect can these models be considered good models. However, the smallest values of Cp and PRESS correspond to Leaps Size 13 & Step model. This last model was elected, and it is shown in Table 4.

All the variables are correlated with LE, even when HEPB is too close to 0.05. The R^2 and R^2 adjusted were improved slightly compared to the original model before the selection.

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

2.3 CHECKING ASSUMPTIONS

2.3.1 LINEARITY

The linearity was checked before (see point 1.6), and even when many transformations were probed and root square was applied to AIDS, GDP and U5D, it did not achieve a significant improvement. It was checked for a second time, but the result was the same: the linearity (Appendix, Fig. 2) of several independent variables in the model has problems (THIN5_9, ETOH, MSS, POL, DIPH, TE). This assumption still needs to be fulfilled.

2.3.2 NORMALITY

The Q-Q plot of the residuals (Fig. 9) shows that most of the points lie on the line but are heavy-tailed.

2.3.3 CONSTANT VARIANCE

The plots of the residuals against the fitted values (Fig. 9) show clouds of points with more density in the extreme of the X-axis with the higher values.

A Box-Cox to determine possible transformations to the response will be applied.

In the independent variables, they could point to both extremes. The most notorious, HEPB and DIPH, certainly lie down to the higher extreme of the X-axis, and GDP, U5D and AIDS, to the lower values, even after the root transformation.

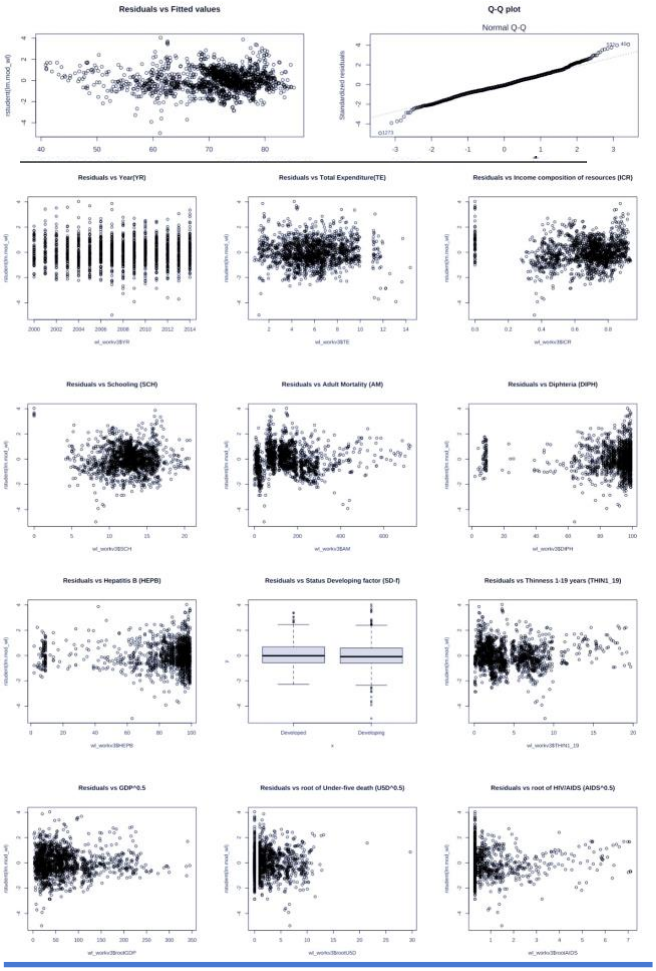


FIG. 9 CHEKING ASSUMPTIONS

2.3.4 INDEPENDENCE

The constant mean is close to 0 (Fig. 9). However, it is known that the data has a temporal (time series data) and spatial structure (data per country). Hence, the data can have problems to fulfil with this assumption.

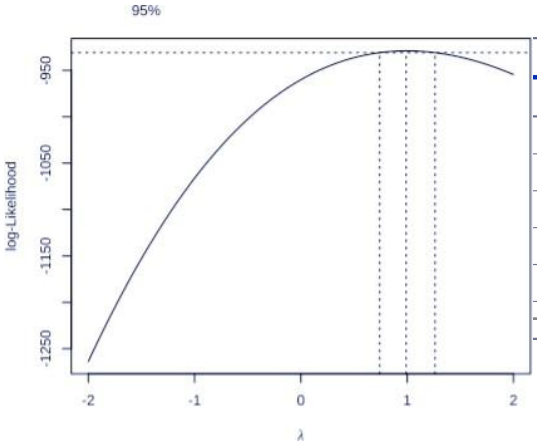
All these graphs can be seen in a bigger size in the Appendix, Fig. 10.

2.4 TESTING TRANSFORM THE RESPONSE

As mentioned before, the response transformation could be an option to improve the model and the assumption fulfilment. The BOX-COX function in R Studio is used to decide about it.

Figure 10 shows that the best lambda obtained was 0.91, close to 1. This result is also evident in the graph.

Hence, it is decided not to transform the response variable (LE).



BEST LAMBDA - BOX-COX

lambda	likelihood
0.9090909	-929.19
0.9494949	-929.04
0.8686869	-929.44
0.9898990	-928.97
0.8282828	-929.79
1.0303030	-929.00

FIG. 10 TESTING BOX-COX TRANSFORMATION OF RESPONSE

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

2.5 TESTING TRANSFORM INDEPENDENT VARIABLES

Different transformations ((λ=0, 0.5, 2, -0.5) were done to each of the independent variables in the model. The model did not improve with any of the four transformations applied in most variables. In a few cases, a slight increase in R2 and adjusted R2 values is observed compared with the basal model. But in any case, there was a substantial increase. Consequently, it was decided not to apply any transformation.

2.6 MULTICOLLINEARITY

The multicollinearity was rechecked after fixing the model. Through de determination of VIF, it was proven that there is no multicollinearity. The maximum VIF obtained, corresponding to ICR and SCH, were 2.41 and 2.76, respectively. The results can be seen in Appendix I, Fig. 11.

2.7 OUTLIERS

This point was checked before to fix the model. As mentioned, they reflected the variability of a dataset containing data from many locations (Fig. 4).

2.8 CONCLUSION ABOUT THE ASSUMPTIONS OF THE MODEL

Even after applying different transformations to the variables, it is concluded that the assumptions are not fulfilled. Try another kind of transformation or build a more complex model (a polynomial one). Another path could be applying a non-linear regression model. The analysis is continuous only for this report.

2.9 INTERACTIONS WITH THE CATEGORICAL VARIABLE

When LE is plotted against each independent variable, grouped by SD, there is a patron: the highest LE figures corresponding to the developed countries.

The equation of the model is (Tab. 4):

LE = -106.69 + 0.08 YR + 420.11 SD-Developing + 0.16 TE + 45.85 ICR -0.56 SCH - AM + 0.02 DIPH - 0.01 HEPB - 1.71 THIN1_19 + ROOT_GDP + 0.34 ROOT_U5D - 3.35 AIDS - 0.20D x YR -38.75DxICR + 1.12D x SCG-0.01D X AM+1.61DXTHIN1_!9+0.02DXrootGDP -0.67DxrootU5D + ε

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

2.10 ANALYSIS OF THE RESULT

It is not sensed by interpreting the intercept in this case (the average LE for someone with the rest of the independent variables = 0 is -106.69) because it is probably irrelevant information.

There is a significant relationship between the status development of the country and the LE. The relationship between life expectancy (LE) and year (YR), Human Development Index (ICR), the average number of years of schooling (SCH), adult mortality (AM) prevalence of thinness among 10-19 years (THIN5_19), GDP and rate of deaths of individuals under five years old per 1000 (U5D), is significantly different between developed countries and developing countries.

On the other hand, the relationship between life expectancy and total expenditure on health, diphtheria, and hepatitis B, are not significantly different between developed and developing countries.

The LE (in years) increases on average by 0.08 for each year, 0.16% for a 1% increase in TE, 45.85% for each unit of increase in ICR, 0.02% with 1% of 1-year-olds that received DTP3 inmunization (Diphtheria vaccine). In contrast, it decreased by -1.71 for 1% of the prevalence of thinness between 1 to 19 years.

There are other odd relations: the pessimistic estimates in schooling (-3.35). Also, there could be some multicollinearity (with ICR).

TABLE 4. MODEL: ESTIMATES AND THEIR CONFIDENT INTERVALS (CI)

	LE		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-106.69	-311.53 – 98.15	0.307
YR	0.08	-0.03 – 0.18	0.141
SD f [Developing]	420.11	198.01 – 642.22	**<0.001**
TE	0.16	0.09 – 0.24	**<0.001**
ICR	45.85	31.57 – 60.13	**<0.001**
SCH	-0.56	-0.84 – -0.28	**<0.001**
AM	-0.00	-0.01 – 0.01	0.608
DIPH	0.02	0.01 – 0.03	**0.001**
HEPB	-0.01	-0.02 – 0.00	0.062
THIN1 19	-1.71	-2.33 – -1.09	**<0.001**
rootGDP	0.00	-0.00 – 0.01	0.564
rootU5D	0.34	-0.23 – 0.92	0.237
rootAIDS	-3.35	-3.55 – -3.15	**<0.001**
YR × SD f [Developing]	-0.20	-0.32 – -0.09	**<0.001**
SD f [Developing] × ICR	-38.75	-53.08 – -24.41	**<0.001**
SD f [Developing] × SCH	1.12	0.82 – 1.42	**<0.001**
SD f [Developing] × AM	-0.01	-0.02 – -0.00	**0.041**
SD f [Developing] × THIN1 19	1.61	0.99 – 2.23	**<0.001**
SD f [Developing] × rootGDP	0.02	0.01 – 0.03	**<0.001**
SD f [Developing] × rootU5D	-0.67	-1.25 – -0.10	**0.022**
Observations	1554		
R^2 / R^2 adjusted	0.863 / 0.861		

2.11 CONCLUSION

This model has many shortcomings (the assumptions need to be met, and the dataset has problems with several variables). A new approach must be applied to have a better result.

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

APPENDIX

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

APPENDIX

Variables	1
Year	0
Status	0
Life expectancy	10
Adult Mortality	10
infant deaths	0
Alcohol	194
percentage expenditure	0
Hepatitis B	553
Measles	0
BMI	34
under-five deaths	0
Polio	19
Total expenditure	226
Diphtheria	19
HIV/AIDS	0
GDP	448
Population	652
thinness 1-19 years	34
thinness 5-9 years	34
Income composition of resources	167
Schooling	163

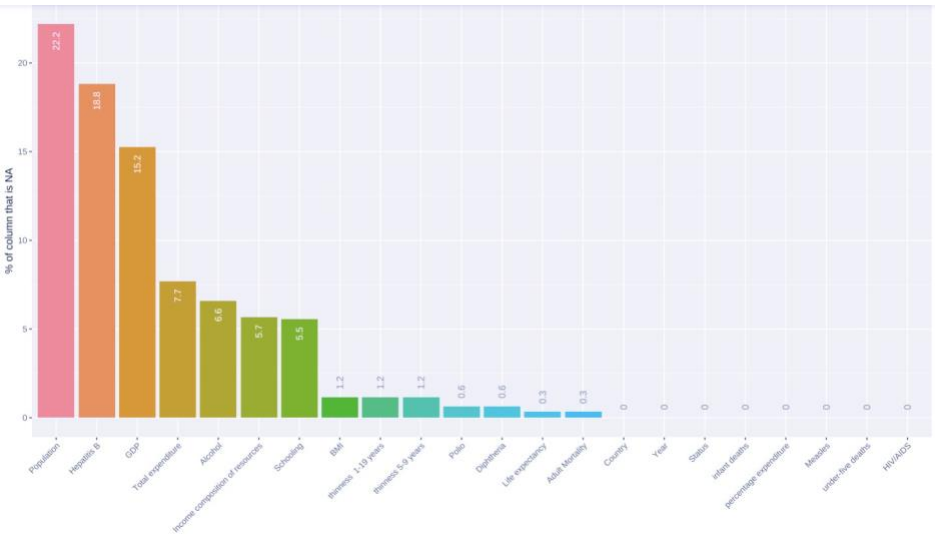


FIG. 1 DISTRIBUTION OF MISSING VALUES IN THE DATASET

TABLE 1 RENAMING VARIABLES

Variable	Renamed as	Description
Country	-	Observation country
Year	**YR**	Observation year
Status	**SD**	Developed or developing status
Life Expectancy	**LE**	Life expectancy in years
Adult Mortality	**AM**	Rate of death for individuals aged between 15 and 60 per 1000 population
Infant Death	**U1D**	Rate of infant deaths per 1000 population
Alcohol	**ETOH**	Litres of (pure) alcohol consumed per member of the population aged over 15
Percentage Expenditure	**PE**	Health expenditure as a percentage of Gross Domestic Product (GDP) per capita
Hepatitis B	**HEPB**	% of 1 year olds who have received immunization against Hepatitis B
Measles	**MSS**	Reported cases on measles per 1000 population
BMI	**BMI**	Average Body Mass Index for the entire population
Under-five deaths	**U5D**	Rate of deaths of individuals under 5 years old per 1000 population
Polio	**POL**	% of 1 year olds who have received immunization against Polio
Total Expenditure	**TE**	Government expenditure on health as a percentage of total government expenditure
Diptheria	**DIPH**	% of 1 year olds who have received DTP3 immunization
HIV/AIDS	**AIDS**	Deaths related to HIV/AIDS per 1000 live births
GDP	**GDP**	Gross Domestic Product per Capita (USD)
Population	**POP**	Population
Thinness 1-19 years	**THIN1_19**	Prevalence (%) of thinness (BMI < 2 SD below the median) among children aged 10 to 19
Thinness 5-9 years	**THIN5_9**	Prevalence (%) of thinness (BMI < 2 SD below the median) among children aged 5-9
Income composition of	**ICR**	Human Development Index in terms of income and composition of resources (index from 0 to 1)
Schooling	**SCH**	Average number of years of schooling

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

TABLE 2. DATASET AFTER RENAMING VARIABLES AND TYDING THE DATA

Country	LE	YR	SD	POP	GDP	PE	TE	ICR	SCH	U1D	U5D	AM	MSS	POL	DIPH	HEPB	AIDS	ETOH	THIN5_9	THIN1_19	BMI
Afghanistan	65.0	2,015	Developing	33,736,494	584.25921	71.279624	8.16	0.479	10.1	62	83	263	1,154	6	65	65	0.1	0.01	17.3	17.2	19.1
Afghanistan	59.9	2,014	Developing	327,582	612.69651	73.523582	8.18	0.476	10.0	64	86	271	492	58	62	62	0.1	0.01	17.5	17.5	18.6
Afghanistan	59.9	2,013	Developing	31,731,688	631.74498	73.219243	8.13	0.470	9.9	66	89	268	430	62	64	64	0.1	0.01	17.7	17.7	18.1
Afghanistan	59.5	2,012	Developing	3,696,958	669.95900	78.184215	8.52	0.463	9.8	69	93	272	2,787	67	67	67	0.1	0.01	18.0	17.9	17.6
Afghanistan	59.2	2,011	Developing	2,978,599	63.53723	7.097109	7.87	0.454	9.5	71	97	275	3,013	68	68	68	0.1	0.01	18.2	18.2	17.2
Afghanistan	58.8	2,010	Developing	2,883,167	553.32894	79.679367	9.20	0.448	9.2	74	102	279	1,989	66	66	66	0.1	0.01	18.4	18.4	16.7

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

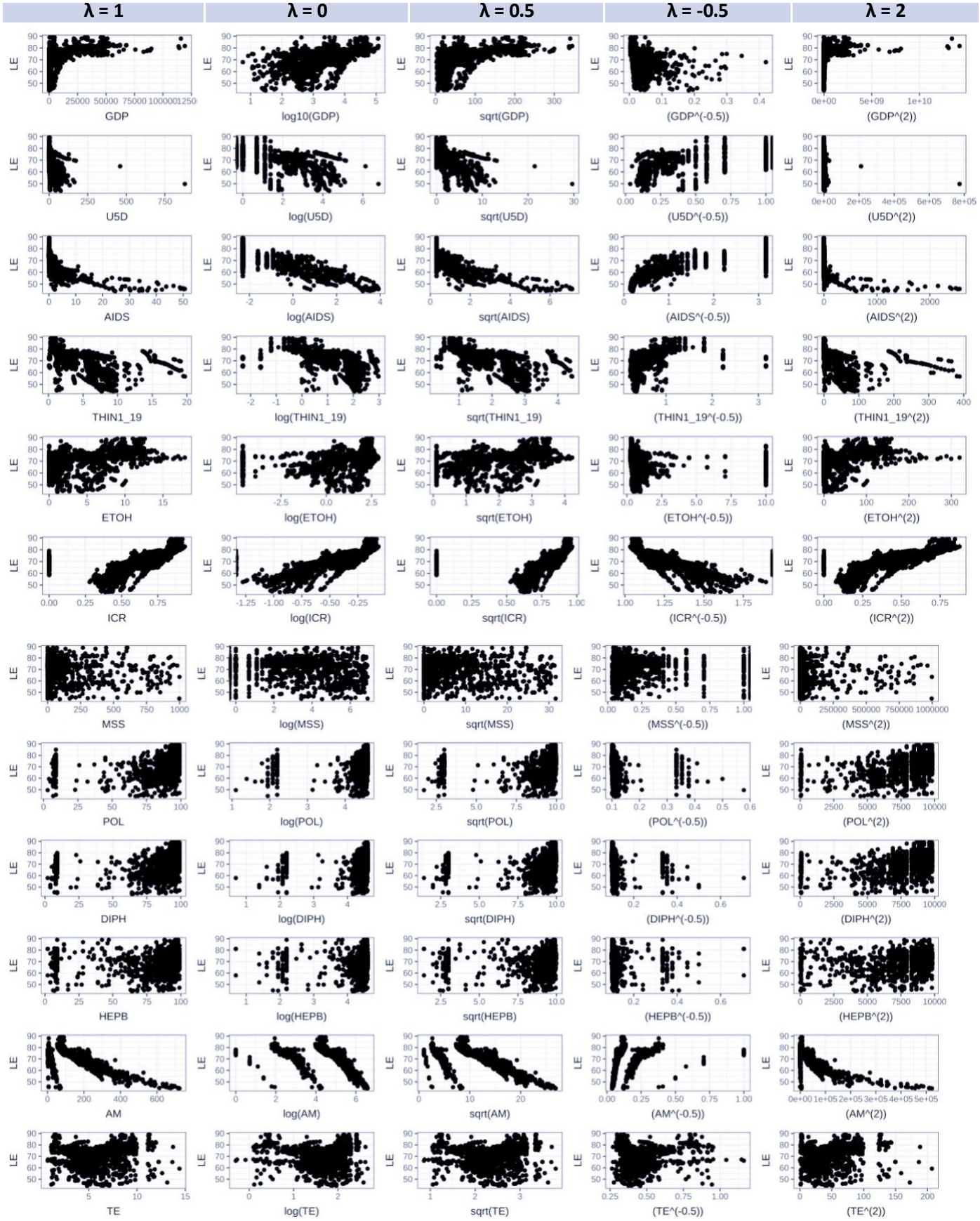


FIG. 2 EVALUATION OF TRANSFORMATION OF THE INDEPENDENT VARIABLES

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

```
Call:
lm(formula = LE ~ ., data = wl_work)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7596  -2.1584  -0.0687   2.1310  16.5877

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.226e+02  4.726e+01   6.826 1.26e-11 ***
YR             -1.313e-01  2.360e-02  -5.564 3.10e-08 ***
GDP             6.630e-05  7.794e-06   8.506  < 2e-16 ***
TE              1.140e-01  4.157e-02   2.741  0.00619 **
ICR             7.805e+00  7.176e-01  10.876  < 2e-16 ***
SCH             7.286e-01  5.432e-02  13.414  < 2e-16 ***
U5D            -2.355e-02  3.003e-03  -7.842 8.22e-15 ***
AM             -1.726e-02  1.029e-03 -16.765  < 2e-16 ***
MSS            -7.809e-04  5.297e-04  -1.474  0.14059
POL             3.113e-03  5.630e-03   0.553  0.58034
DIPH            1.950e-02  6.305e-03   3.093  0.00202 **
HEPB           -4.853e-03  4.870e-03  -0.997  0.31915
AIDS           -4.780e-01  1.909e-02 -25.039  < 2e-16 ***
ETOH            5.508e-03  3.230e-02   0.171  0.86463
THIN1_19       -1.537e-01  3.226e-02  -4.764 2.08e-06 ***
SD_fDeveloping -7.634e-01  3.302e-01  -2.312  0.02093 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.594 on 1538 degrees of freedom
Multiple R-squared:  0.8199,    Adjusted R-squared:  0.8181
F-statistic: 466.8 on 15 and 1538 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = LE ~ ., data = wl_workv2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0612  -1.9982  -0.2394   2.0070  12.9000

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.152e+02  4.281e+01   5.028 5.55e-07 ***
YR             -7.593e-02  2.138e-02  -3.552 0.000395 ***
TE              1.701e-01  3.781e-02   4.500 7.32e-06 ***
ICR             7.586e+00  6.525e-01  11.627  < 2e-16 ***
SCH             4.972e-01  5.044e-02   9.858  < 2e-16 ***
AM             -1.220e-02  9.739e-04 -12.522  < 2e-16 ***
MSS            -1.896e-04  4.914e-04  -0.386 0.699740
POL            -6.853e-04  5.105e-03  -0.134 0.893225
DIPH            1.941e-02  5.700e-03   3.404 0.000681 ***
HEPB           -8.471e-03  4.413e-03  -1.919 0.055125 .
ETOH            1.738e-02  2.927e-02   0.594 0.552669
THIN1_19       -1.169e-01  2.925e-02  -3.995 6.78e-05 ***
SD_fDeveloping -7.625e-01  2.971e-01  -2.567 0.010355 *
rootGDP         1.768e-02  1.827e-03   9.675  < 2e-16 ***
rootU5D        -3.336e-01  3.815e-02  -8.745  < 2e-16 ***
rootAIDS       -3.313e+00  1.064e-01 -31.139  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.257 on 1538 degrees of freedom
Multiple R-squared:  0.8521,    Adjusted R-squared:  0.8506
F-statistic: 590.6 on 15 and 1538 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = LE ~ ., data = wl_work)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5496  -1.8965  -0.1154   1.8834  12.2343

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)    56.7191082 44.7716381   1.267 0.205399
YR              0.0001932  0.0223561   0.009 0.993105
TE              0.1478891  0.0396298   3.732 0.000197 ***
ICR             7.5069119  0.6842129  10.972  < 2e-16 ***
SCH             0.3208281  0.0538062   5.963 3.07e-09 ***
AM             -0.0146263  0.0009991 -14.640  < 2e-16 ***
MSS            -0.0012118  0.0005016  -2.416 0.015823 *
POL             0.0008221  0.0053396   0.154 0.877656
DIPH            0.0215005  0.0059817   3.594 0.000335 ***
HEPB           -0.0135814  0.0046352  -2.930 0.003439 **
ETOH            0.0388497  0.0308924   1.258 0.208733
THIN1_19       -0.1172508  0.0307394  -3.814 0.000142 ***
SD_fDeveloping -1.1639446  0.3101226  -3.753 0.000181 ***
logGDP          0.9741103  0.1419340   6.863 9.74e-12 ***
logU5D         -0.0765724  0.0452148  -1.694 0.090558 .
logAIDS        -5.7031910  0.1961424 -29.077  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.416 on 1538 degrees of freedom
Multiple R-squared:  0.8373,    Adjusted R-squared:  0.8357
F-statistic: 527.5 on 15 and 1538 DF,  p-value: < 2.2e-16
```

FIG. 3 LINEAR REGRESSION TO TEST TRANSFORMATION OF THREE INDEPENDENT VARIABLES

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

TAB. 3 EVALUATION OF THE BEST SUBSET

Adjusted R^2

Size_Fac	Criterion	X.Intercept.	YR	TE	ICR	SCH	AM	MSS	POL	DIPH	HEPB	ETOH	THIN1_19	SD_fDeveloping	rootGDP	rootU5D	rootAIDS
1	0.55899	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2	0.76535	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
3	0.79696	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
4	0.81748	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
5	0.83136	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
6	0.84166	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
7	0.84507	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
8	0.84736	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
9	0.84899	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
10	0.84982	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
11	0.85058	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
12	0.85087	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
13	0.85081	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
14	0.85073	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
15	0.85063	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

```
Call:
lm(formula = LE ~ YR + TE + ICR + SCH + AM + DIPH + THIN1_19 +
    SD_f + rootGDP + rootU5D + rootAIDS, data = wl_workv2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8842  -2.0025  -0.2442   2.0342  13.3243

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.237e+02  4.189e+01   5.341 1.06e-07 ***
YR             -8.030e-02  2.093e-02  -3.837  0.00013 ***
TE              1.740e-01  3.748e-02   4.642 3.75e-06 ***
ICR             7.617e+00  6.521e-01  11.680  < 2e-16 ***
SCH             5.051e-01  4.888e-02  10.334  < 2e-16 ***
AM             -1.208e-02  9.664e-04 -12.503  < 2e-16 ***
DIPH            1.329e-02  4.259e-03   3.121  0.00183 **
THIN1_19       -1.266e-01  2.829e-02  -4.475 8.20e-06 ***
SD_fDeveloping -8.059e-01  2.703e-01  -2.981  0.00292 **
rootGDP         1.782e-02  1.817e-03   9.806  < 2e-16 ***
rootU5D        -3.360e-01  3.582e-02  -9.379  < 2e-16 ***
rootAIDS       -3.294e+00  1.059e-01 -31.106  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.258 on 1542 degrees of freedom
Multiple R-squared:  0.8516,    Adjusted R-squared:  0.8506
F-statistic: 804.7 on 11 and 1542 DF,  p-value: < 2.2e-16
```

FIG. 4 RESULT CHOOSING RESPONSE + 11 INDEPENDENT

```
Call:
lm(formula = LE ~ YR + TE + ICR + SCH + AM + DIPH + HEPB + THIN
    SD_f + rootGDP + rootU5D + rootAIDS, data = wl_workv2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9443  -1.9929  -0.2351   2.0156  13.0458

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.185e+02  4.193e+01   5.212 2.12e-07 ***
YR             -7.760e-02  2.095e-02  -3.704 0.000220 ***
TE              1.734e-01  3.744e-02   4.632 3.92e-06 ***
ICR             7.589e+00  6.517e-01  11.646  < 2e-16 ***
SCH             5.041e-01  4.884e-02  10.322  < 2e-16 ***
AM             -1.213e-02  9.657e-04 -12.557  < 2e-16 ***
DIPH            1.937e-02  5.239e-03   3.698 0.000225 ***
HEPB           -8.620e-03  4.334e-03  -1.989 0.046882 *
THIN1_19       -1.222e-01  2.835e-02  -4.309 1.74e-05 ***
SD_fDeveloping -8.217e-01  2.702e-01  -3.041 0.002396 **
rootGDP         1.768e-02  1.817e-03   9.732  < 2e-16 ***
rootU5D        -3.375e-01  3.580e-02  -9.430  < 2e-16 ***
rootAIDS       -3.307e+00  1.060e-01 -31.199  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.254 on 1541 degrees of freedom
Multiple R-squared:  0.852,    Adjusted R-squared:  0.8509
F-statistic: 739.4 on 12 and 1541 DF,  p-value: < 2.2e-16
```

FIG. 5 RESULT CHOOSING RESPONSE + 12 INDEPENDENT

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

```
Call:
lm(formula = LE ~ YR + TE + ICR + SCH + AM + DIPH + HEPB + THIN1_19 +
    SD_f + ETOH + rootGDP + rootU5D + rootAIDS, data = wl_workv2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0176  -1.9913  -0.2382   2.0147  12.8856

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.139e+02  4.258e+01   5.025 5.64e-07 ***
YR             -7.531e-02  2.127e-02  -3.540 0.000412 ***
TE              1.703e-01  3.778e-02   4.508 7.03e-06 ***
ICR             7.582e+00  6.519e-01  11.631  < 2e-16 ***
SCH             4.965e-01  5.034e-02   9.863  < 2e-16 ***
AM             -1.220e-02  9.729e-04 -12.539  < 2e-16 ***
DIPH            1.907e-02  5.261e-03   3.625 0.000298 ***
HEPB           -8.453e-03  4.343e-03  -1.946 0.051796 .
THIN1_19       -1.185e-01  2.895e-02  -4.094 4.47e-05 ***
SD_fDeveloping -7.484e-01  2.947e-01  -2.539 0.011208 *
ETOH            1.820e-02  2.917e-02   0.624 0.532837
rootGDP         1.762e-02  1.820e-03   9.679  < 2e-16 ***
rootU5D        -3.381e-01  3.581e-02  -9.440  < 2e-16 ***
rootAIDS       -3.312e+00  1.063e-01 -31.161  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.255 on 1540 degrees of freedom
Multiple R-squared:  0.8521,    Adjusted R-squared:  0.8508
F-statistic: 682.3 on 13 and 1540 DF,  p-value: < 2.2e-16
```

FIG. 6 RESULT CHOOSING RESPONSE + 13 INDEPENDENT VARIABLES

```
          Step Df    Deviance Resid. Df Resid. Dev      AIC
1               NA          NA      1553  110294.72 6625.651
2  + rootAIDS -1 61684.82769      1552   48609.89 5354.414
3       + SCH -1 22763.05654      1551   25846.83 4374.847
4        + AM -1  3495.95071      1550   22350.88 4151.016
5       + ICR -1  2271.23637      1549   20079.64 3986.491
6   + rootU5D -1  1539.67524      1548   18539.97 3864.516
7   + rootGDP -1  1143.52528      1547   17396.44 3767.584
8        + TE -1   385.20974      1546   17011.23 3734.787
9  + THIN1_19 -1   262.47100      1545   16748.76 3712.623
10       + YR -1   189.52158      1544   16559.24 3696.938
11     + DIPH -1   101.83367      1543   16457.41 3689.352
12     + SD_f -1    94.31590      1542   16363.09 3682.421
13     + HEPB -1    41.89903      1541   16321.19 3680.436
> stepFW_wlv2

Call:
lm(formula = LE ~ rootAIDS + SCH + AM + ICR + rootU5D + rootGDP +
    TE + THIN1_19 + YR + DIPH + SD_f + HEPB, data = wl_workv2)

Coefficients:
   (Intercept)        rootAIDS             SCH              AM             ICR
     218.54453         -3.30699         0.50406         -0.01213         7.58901
       rootU5D          rootGDP              TE        THIN1_19              YR
      -0.33754          0.01768         0.17345         -0.12216         -0.07760
          DIPH   SD_fDeveloping            HEPB
       0.01937         -0.82171         -0.00862
```

FIG. 7 RESULT FORWARDS SELECTION METHOD

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

```
    Step Df  Deviance Resid. Df Resid. Dev      AIC
1        NA        NA      1538   16315.32 3685.878
2  - POL  1 0.1911807      1539   16315.52 3683.896
3  - MSS  1 1.5543630      1540   16317.07 3682.044
4 - ETOH  1 4.1232876      1541   16321.19 3680.436
> stepBW_wlv2

Call:
lm(formula = LE ~ YR + TE + ICR + SCH + AM + DIPH + HEPB + THIN1_19 +
    SD_f + rootGDP + rootU5D + rootAIDS, data = wl_workv2)

Coefficients:
   (Intercept)             YR             TE            ICR            SCH
     218.54453       -0.07760        0.17345        7.58901        0.50406
            AM           DIPH           HEPB       THIN1_19  SD_fDeveloping
      -0.01213        0.01937       -0.00862       -0.12216       -0.82171
        rootGDP        rootU5D       rootAIDS
        0.01768       -0.33754       -3.30699
```

FIG. 8 RESULT BACKWARDS SELECTION METHOD

```
          Step Df    Deviance Resid. Df Resid. Dev      AIC
1           NA          NA      1553  110294.72 6625.651
2  + rootAIDS -1 61684.82769      1552   48609.89 5354.414
3       + SCH -1 22763.05654      1551   25846.83 4374.847
4        + AM -1  3495.95071      1550   22350.88 4151.016
5       + ICR -1  2271.23637      1549   20079.64 3986.491
6   + rootU5D -1  1539.67524      1548   18539.97 3864.516
7   + rootGDP -1  1143.52528      1547   17396.44 3767.584
8        + TE -1   385.20974      1546   17011.23 3734.787
9  + THIN1_19 -1   262.47100      1545   16748.76 3712.623
10       + YR -1   189.52158      1544   16559.24 3696.938
11     + DIPH -1   101.83367      1543   16457.41 3689.352
12     + SD_f -1    94.31590      1542   16363.09 3682.421
13     + HEPB -1    41.89903      1541   16321.19 3680.436
> stepBOTH_wlv2

Call:
lm(formula = LE ~ rootAIDS + SCH + AM + ICR + rootU5D + rootGDP +
    TE + THIN1_19 + YR + DIPH + SD_f + HEPB, data = wl_workv2)

Coefficients:
   (Intercept)        rootAIDS             SCH              AM             ICR
     218.54453       -3.30699         0.50406        -0.01213         7.58901
        rootU5D         rootGDP              TE       THIN1_19              YR
      -0.33754         0.01768         0.17345        -0.12216        -0.07760
           DIPH  SD_fDeveloping            HEPB
         0.01937        -0.82171        -0.00862
```

FIG. 9 RESULT STEPWISE SELECTION METHOD

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

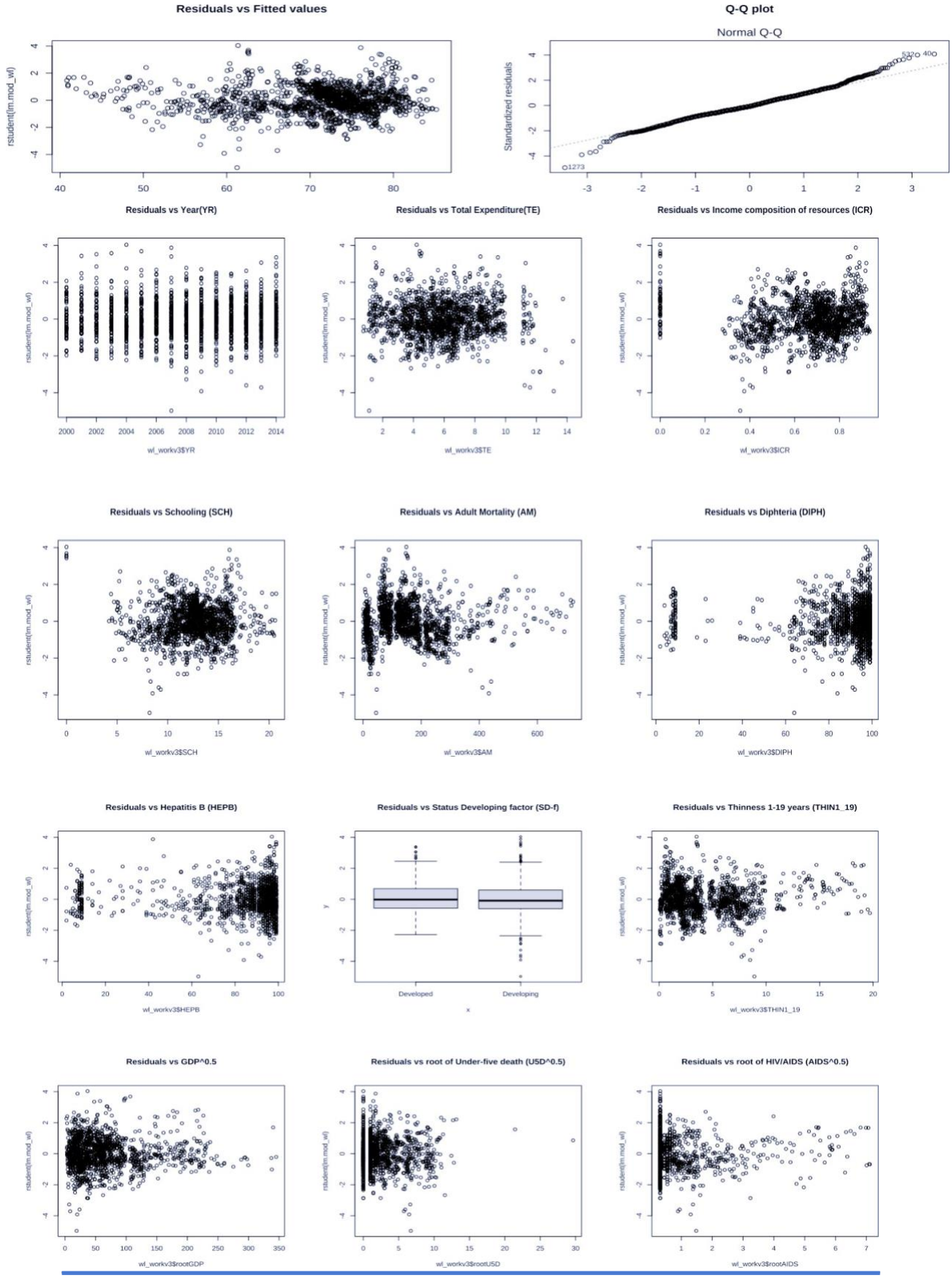


FIG. 10 CHECKING MODEL ASSUMPTIONS

LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

	Variables	Tolerance	VIF
1	YR	0.9352034	1.069286
2	TE	0.8842489	1.130903
3	ICR	0.4149805	2.409752
4	SCH	0.3627258	2.756903
5	AM	0.5101424	1.960237
6	DIPH	0.5869801	1.703635
7	HEPB	0.6246202	1.600973
8	THIN1_19	0.6739175	1.483861
9	SD_fDeveloping	0.6871420	1.455303
10	rootGDP	0.6430855	1.555003
11	rootU5D	0.6714952	1.489214
12	rootAIDS	0.5410876	1.848130

FIG. 11 VALUES OF VARIANCE INFLANTION FACTOR