NANYANG TECHNOLOGICAL UNIVERSITY

Nanyang Business School

Assignment Cover Page

Student Name/ID	Rachyl Tan Syn Yee G2104229F
	Gan Chu G2100385G
	Xu Chaolan G2100651J
	Lin Li Wei G2100374A
Course Code	FE8512
Course Title	Linear Financial Model
Lecturer	Wu Yuan
Date Submitted	31 October 2021

We declare that this assessment item is our own work, except where acknowledged, and has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

Reproduce this assessment item and provide a copy to another member of the University; and/or,

Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

We certify that we have read and understood the University Academic Standing and Grading System in respect of Student Honour Code.

Signed: RT ....date: 31 10 21

Signed: GAN CHU ....date: 31 10 21

Signed: Xu Chaolan ....date: 31 10 21

Signed: Lin Liwei ....date: 31 10 21

Regression Model on Life Expectancy

Abstract

Life expectancy is a key metric for assessing population health. Although much research has been conducted in the past on factors impacting life expectancy, factors such as immunization rates and aspects of human development were not taken into consideration. Hence, this study will explore the relationship between life expectancy and predictive factors across 4 key aspects – immunization-related factors, mortality factors, economical factors and social factors.

Our group selected 19 explanatory variables to build the complete model. Through automatic selection and all possible regression methods, we selected three models with 11 variables as potentially globally optimized models based on Mallow's C_p, $PRESS_p$, $R_{adjusted}^2$ and AIC_p. For these three selected models, we then conducted residual analysis and found that there is no violation of assumptions. As for outlier tests, the results imply that we can fairly say there is no influential outlier within our models.

Finally, we conducted model validation for three potentially globally optimized models. Based on the model validation table, we can conclude that coefficients of all variables are consistent except Thinness factors which are insignificant. We compare the descriptive and predictive powers of each model and choose Model 2 as our global optimal model.

The main conclusion of our projects are: life expectancy can be explained by 11 variables including the status of the country, adult mortality, etc. Based on standardized beta, HIV/AIDS, schooling and adult mortality are the most important factors.

Content Page

Section 1. Introduction

Life expectancy is a key metric for assessing population health. By definition, life expectancy is based on an estimate of the average age that members of a particular population group will be when they die. Globally, life expectancy has increased from 66.8 years in 2000 to 73.4 years in 2019.

This exploratory observational study will focus on immunization-related factors, mortality factors, economical factors and social factors as explanatory variables for life expectancy. This will help in suggesting which area should be given importance in order to efficiently improve the life expectancy of a population.

The outline of the remainder of the report is as follows. In Section 2, we explore the characteristics of the data and prepare the dataset for analysis. Then, we conduct model selection and model validation in Sections 3 to 5 and Section 6 respectively. Lastly, we summarise and conclude our findings in Section 7. More details are provided in the Appendix.

Section 2. Data Characteristics and Preparation

Section 2.1. Data Characteristics

The data was compiled from 2 sources – data related to life expectancy and health factors were compiled from the Global Health Observatory (GHO) data repository under World Health Organization (WHO), while economic data was collected from the United Nations website. The data covers 193 countries over the years 2000 to 2015, and includes 19 predicting factors as shown in Figure 1. Descriptions of the variables can be found in A2. In total, we have 2938 cases.

	Variable Name
X_1	Status
X_2	Adult Mortality
X_3	Infant Deaths
X_4	Alcohol
X_5	Percentage Expenditure
X_6	Hepatitis B
X_7	Measles
X_8	BMI
X_9	Under-five Deaths
X_{10}	Polio
X_{11}	Total Expenditure
X_{12}	Diphtheria
X_{13}	HIV/AIDS
X_{14}	GDP
X_{15}	Population
X_{16}	Thinness 10-19 Years
X_{17}	Thinness 5-9 Years
X_{18}	Income Composition Of Resources
X_{19}	Schooling

Figure 1: List of Predicting Variables

First, we explore the distribution of our dependent variable – life expectancy. We observe that life expectancy ranges from 36 to 89 years, with an average lifespan of 69 years.

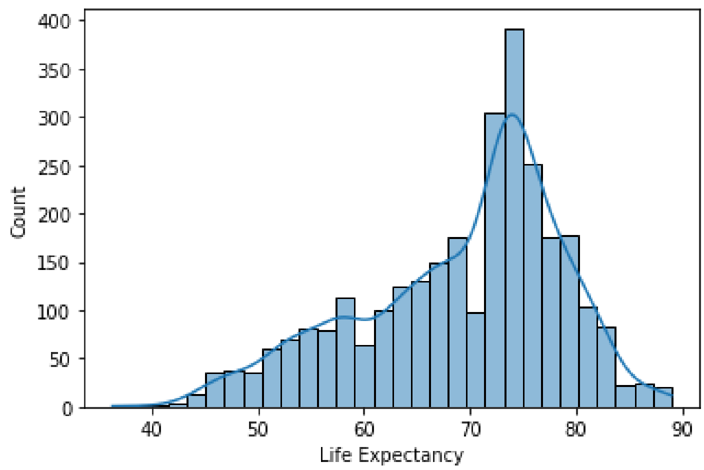


Figure 2: Histogram of Life Expectancy

We also use a correlation heat map to visualise and explore the correlation between different variables. We observe that life expectancy has high correlation with variables such as adult mortality, BMI, schooling, HIV/AIDS, income composition of resources and GDP.

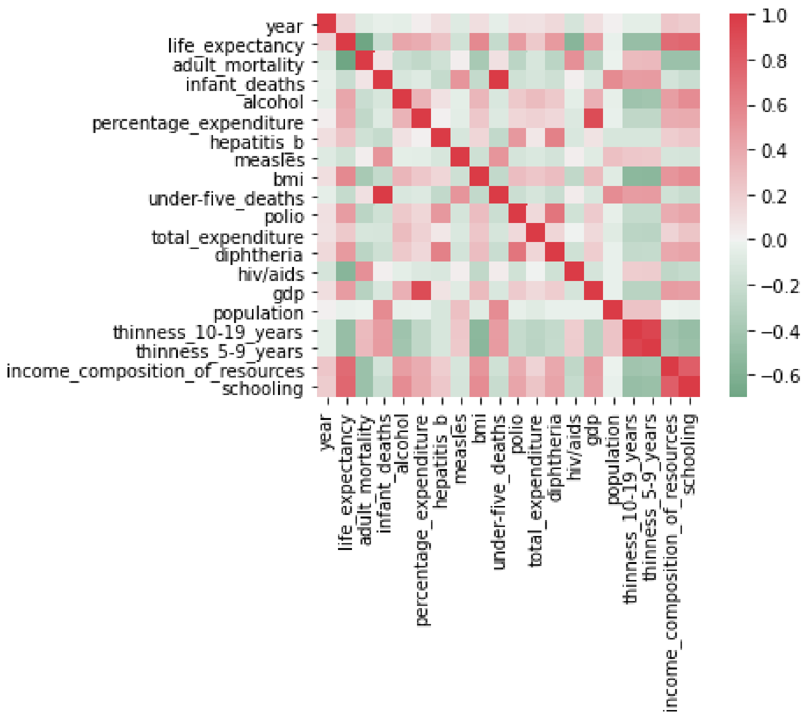


Figure 3: Correlation Heatmap of All Variables

We also plotted some scatterplots to visualise the relationship between life expectancy and its predictor variables. For instance, schooling appears to have a direct relationship with life expectancy as shown in Figure 4. This is a logical observation as higher education typically leads to a higher standard of living, which results in longer life expectancy. On the contrary, HIV/AIDS appears to have an inverse relationship with life expectancy as shown in Figure 5. This is also a logical observation as the number of deaths caused by HIV/AIDS is inversely related to life expectancy.

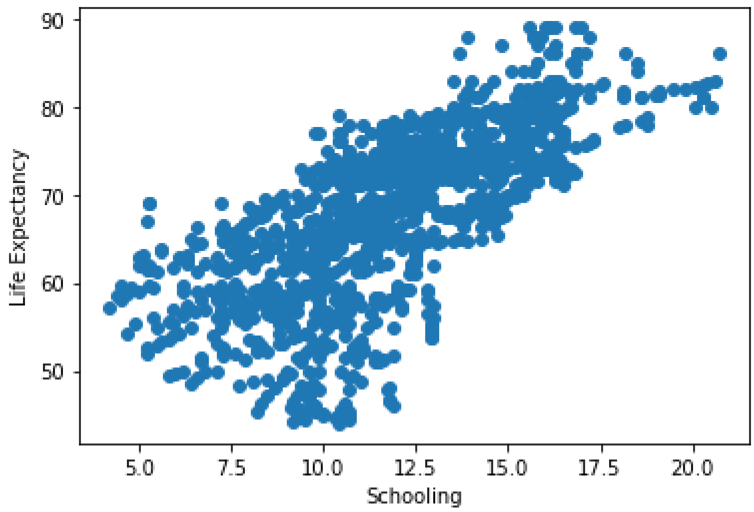


Figure 4: Scatterplot of Life Expectancy against Schooling

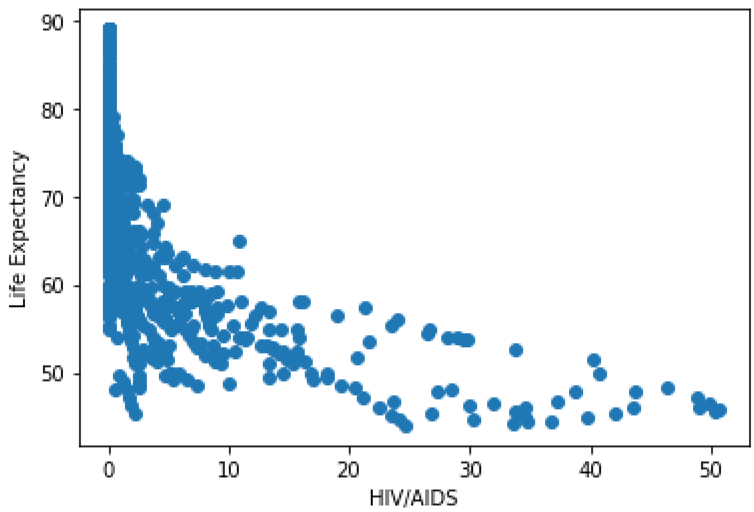


Figure 5: Scatterplot of Life Expectancy against HIV/AIDS

Section 2.2. Data Preparation

After data exploration, we start to prepare our data for model selection.

First, we noticed that 1289 out of 2938 cases contained null values. Considering our data volume is large enough, we decided to drop all cases that contained null values. After elimination, our dataset has 1649 cases. We also decided against interpolating the missing values as we wanted to retain all original data characteristics and avoid introducing more biasness to the model.

Second, one of our predictor variables, status, is a qualitative variable. Hence, we created a dummy variable, using 0 to represent a developing country and 1 to represent a developed country.

Lastly, we split our cases between training set and validation set using a ratio of 70:30. Hence, we have 1154 cases for the training set and 495 cases for the validation set.

Section 3. Model Refinement and Selection

Section 3.1. Full Model

First, we examine the full model. When we regress life expectancy against all the 19 variables, the adjusted R squared stands at around 0.830. Moreover, when using the overall significance test, the F test, we find out that the set of variables collectively have some explanation power on our dependent variable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.913[a]	.833	.830	3.5999

a. Predictors: (Constant), schooling, population, hivaids, hepatitis_b, total_expenditure, measles, percentage_expenditure, polio, status, thinness_59_years, adult_mortality, bmi, diphtheria, alcohol, income_composition_of_resources, underfive_deaths, thinness_1019_years, gdp, infant_deaths

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	73218.169	19	3853.588	297.367	.000[b]
	Residual	14695.564	1134	12.959		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), schooling, population, hivaids, hepatitis_b, total_expenditure, measles, percentage_expenditure, polio, status, thinness_59_years, adult_mortality, bmi, diphtheria, alcohol, income_composition_of_resources, underfive_deaths, thinness_1019_years, gdp, infant_deaths

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	53.383	.891		59.938	.000
	status	.971	.401	.040	2.420	.016
	adult_mortality	-.016	.001	-.229	-13.927	<.001
	infant_deaths	.099	.015	1.380	6.784	<.001
	alcohol	-.120	.040	-.056	-3.020	.003
	percentage_expenditure	.001	.000	.116	2.285	.022
	hepatitis_b	-.004	.005	-.011	-.692	.489
	measles	-1.047E-5	.000	-.013	-.828	.408
	bmi	.032	.007	.071	4.317	<.001
	underfive_deaths	-.072	.010	-1.367	-7.049	<.001
	polio	.012	.006	.032	2.028	.043
	total_expenditure	.083	.050	.021	1.655	.098
	diphtheria	.012	.007	.031	1.784	.075
	hivaids	-.440	.021	-.309	-21.078	<.001
	gdp	-1.521E-5	.000	-.021	-.404	.686
	population	-2.628E-9	.000	-.025	-1.212	.226
	thinness_1019_years	-.042	.061	-.021	-.689	.491
	thinness_59_years	-.052	.060	-.027	-.864	.388
	income_composition_of_resources	9.242	.954	.197	9.684	<.001
	schooling	.892	.073	.280	12.284	<.001

a. Dependent Variable: life_expectancy

Figure 6: Linear Regression Results for Full Model

Next, we conducted an informal residual analysis to examine the aptness of the linear regression model. We need to ensure that the data is inline with the initial assumptions of a linear regression model. We regressed life expectancy on all 19 predictor variables and produced 3 plots as seen in Figures 7 and 8.

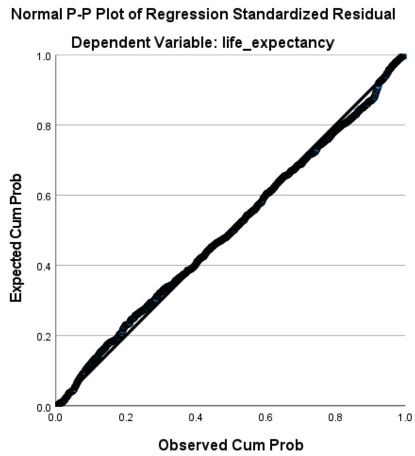


Figure 7: P-P Plot of Full Model

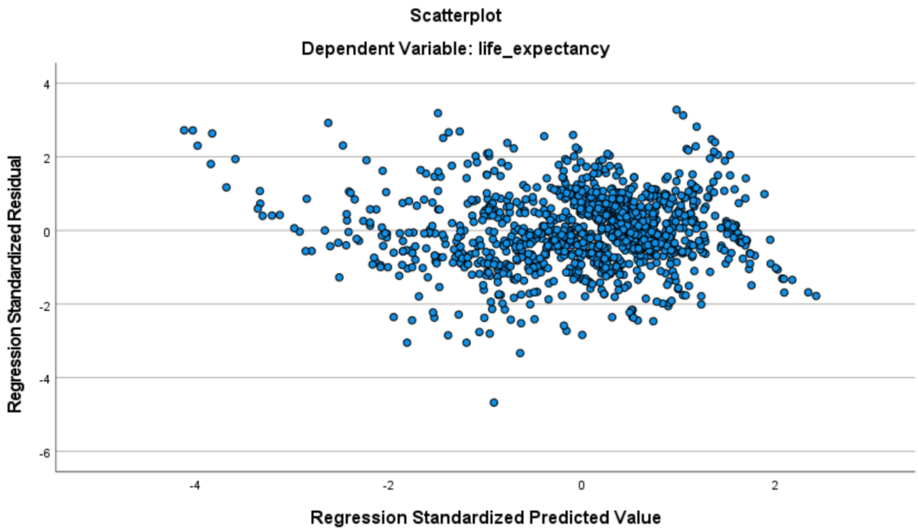


Figure 8: Residual Plot of Full Model

From Figure 7, we can infer that the CDF of the standardized residual quite matched up with the CDF of the normal distribution, suggesting that the standardized residual of the all variables model matches the assumption of normality of residual.

From Figure 8, we observe that residuals are randomly distributed around the horizontal line through zero. This is a sign that our all factors model meets the constant variance assumption. Hence, the above plots show that the life expectancy and predictor variables follow a linear relation.

Section 3.2. Model Selection

First, we used the automatic method to narrow down the optimal number of predictor variables to include in our final model. As seen in Figure 10, both forward and stepwise method suggested 11 variables, while backward method suggested 13 variables. We decided to go with 11 variables as our starting point as forward and stepwise method both suggested 11 variables. Furthermore, stepwise method is the most widely used technique.

	R^2	$R_{adjusted}^2$	Number of variables
Full Model	0.833	0.830	19
Forward	0.824	0.822	11
Backward	0.832	0.830	13
Stepwise	0.824	0.822	11

Figure 9: Results from Automatic Method

Next, we need to identify the optimal pool of variables to include in our final model. We used the all-possible-regression procedure to test for all combinations of 10, 11 and 12 variables, and evaluated the models based on four criteria: Mallow's C_p, $PRESS_p$, $R_{adjusted}^2$ and AIC_p. Based on the results, we shortlisted the top 3 models as seen in Figure 11.

	Number of variables	C_p	$PRESS_p$	$R_{adjusted}^2$	AIC_p
1	12	34.677257	15200.926468	0.829813	2968.780273
2	12	34.697796	15200.825245	0.829810	2968.800658
3	12	35.190159	15206.618220	0.829738	2969.289235

Figure 10: Results of All-Possible-Regression Procedure for Top 3 Models

	Variable Name	Model 1	Model 2	Model 3
X1	Status	1	1	1
X2	Adult Mortality	1	1	1
X3	Infant Deaths	1	1	1
X4	Alcohol	1	1	1
X5	Percentage Expenditure	1	1	1
X6	Hepatitis B	0	0	0
X7	Measles	0	0	0
X8	BMI	1	1	1
X9	Under-five Deaths	1	1	1
X10	Polio	1	1	0
X11	Total Expenditure	0	0	0
X12	Diphtheria	0	0	1
X13	HIV/AIDS	1	1	1
X14	GDP	0	0	0
X15	Population	0	0	0
X16	Thinness 10-19 Years	0	1	0
X17	Thinness 5-9 Years	1	0	1
X18	Income Composition Of Resources	1	1	1
X19	Schooling	1	1	1

Figure 11: Top 3 Models

From Figure 11, we observe that all 3 models contain 12 predictor variables each and are quite similar to each other, differing only by 1 or 2 variables.

Section 3.3. Model Refinement

We analyse each model to determine if we need to take any extra steps to refine the model. For Model 1, $R_{adjusted}^2$ and F-test show that a significant linear relationship exists as seen in Figure 12. However, the collinearity diagnostics variance inflation factor (VIF) shows that infant deaths and under-5 deaths have high multicollinearity. Hence, we would need to decide which variable to drop.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.912[a]	.832	.830	3.6023	2.050

a. Predictors: (Constant), schooling, hivaids, infant_deaths, polio, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources, underfive_deaths

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	73107.668	12	6092.306	469.491	.000[b]
	Residual	14806.064	1141	12.976		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), schooling, hivaids, infant_deaths, polio, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources, underfive_deaths

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	54.016	.817		66.115	.000		
	status	.934	.398	.038	2.345	.019	.559	1.790
	adult_mortality	-.016	.001	-.230	-14.158	<.001	.557	1.794
	infant_deaths	.091	.012	1.272	7.565	<.001	.005	191.476
	alcohol	-.112	.039	-.052	-2.843	.005	.442	2.261
	percentage_expenditure	.000	.000	.099	7.018	<.001	.741	1.350
	bmi	.032	.007	.071	4.354	<.001	.554	1.804
	underfive_deaths	-.068	.009	-1.288	-7.651	<.001	.005	192.156
	polio	.016	.005	.042	3.127	.002	.824	1.213
	hivaids	-.437	.021	-.308	-21.094	<.001	.694	1.442
	thinness_59_years	-.088	.033	-.045	-2.696	.007	.527	1.897
	income_composition_of_resources	9.400	.946	.200	9.934	<.001	.364	2.746
	schooling	.903	.072	.283	12.584	<.001	.291	3.435

a. Dependent Variable: life_expectancy

Figure 12: Linear Regression Results for Model 1

We tried dropping each variable and compared the results. From Figure 13, it is evident that the removal of either variable would have a similar effect on our model. In both cases, dropping either variable resolved the issue of multicollinearity. In fact, when looking at those

2 revised models, the F-test is significant. However, both $R_{adjusted}^2$ decreased to 0.821. We decided to drop infant deaths as it is a subset of under-5 deaths. We prefer under-5 deaths as it is more encompassing and allows our model to be applicable to wider scenarios.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.821	3.6899	2.059

a. Predictors: (Constant), schooling, hivaids, underfive_deaths, polio, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources

b. Dependent Variable: life_expectancy

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.821	3.6919	2.058

a. Predictors: (Constant), infant_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_59_years, income_composition_of_resources, schooling

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72365.037	11	6578.640	483.179	.000[b]
	Residual	15548.696	1142	13.615		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), schooling, hivaids, underfive_deaths, polio, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72348.120	11	6577.102	482.541	.000[b]
	Residual	15565.612	1142	13.630		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), infant_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_59_years, income_composition_of_resources, schooling

Coefficients[a]

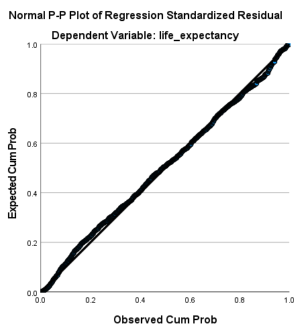
Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	53.177	.829		64.136	.000		
	status	.868	.408	.035	2.127	.034	.559	1.789
	adult_mortality	-.017	.001	-.249	-15.090	<.001	.570	1.754
	alcohol	-.154	.040	-.072	-3.875	<.001	.452	2.214
	percentage_expenditure	.000	.000	.096	6.653	<.001	.741	1.349
	bmi	.031	.007	.069	4.158	<.001	.554	1.804
	underfive_deaths	-.001	.001	-.019	-1.317	.188	.761	1.315
	polio	.022	.005	.057	4.182	<.001	.842	1.187
	hivaids	-.439	.021	-.309	-20.652	<.001	.694	1.442
	thinness_59_years	-.074	.033	-.038	-2.217	.027	.529	1.891
	income_composition_of_resources	10.051	.965	.214	10.413	<.001	.367	2.724
	schooling	.930	.073	.292	12.680	<.001	.292	3.426

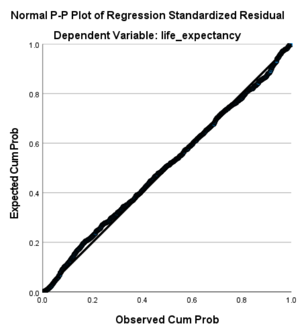
a. Dependent Variable: life_expectancy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	53.128	.829		64.100	.000		
	status	.872	.408	.036	2.137	.033	.559	1.790
	adult_mortality	-.017	.001	-.248	-15.043	<.001	.569	1.757
	alcohol	-.158	.040	-.073	-3.958	<.001	.453	2.208
	percentage_expenditure	.000	.000	.096	6.649	<.001	.741	1.349
	bmi	.031	.007	.070	4.159	<.001	.554	1.804
	polio	.022	.005	.058	4.287	<.001	.845	1.183
	hivaids	-.438	.021	-.308	-20.607	<.001	.694	1.442
	thinness_59_years	-.082	.034	-.042	-2.454	.014	.527	1.896
	income_composition_of_resources	10.010	.966	.213	10.359	<.001	.367	2.727
	schooling	.935	.073	.293	12.740	<.001	.292	3.423
	infant_deaths	-.001	.001	-.010	-.701	.483	.763	1.310

a. Dependent Variable: life_expectancy





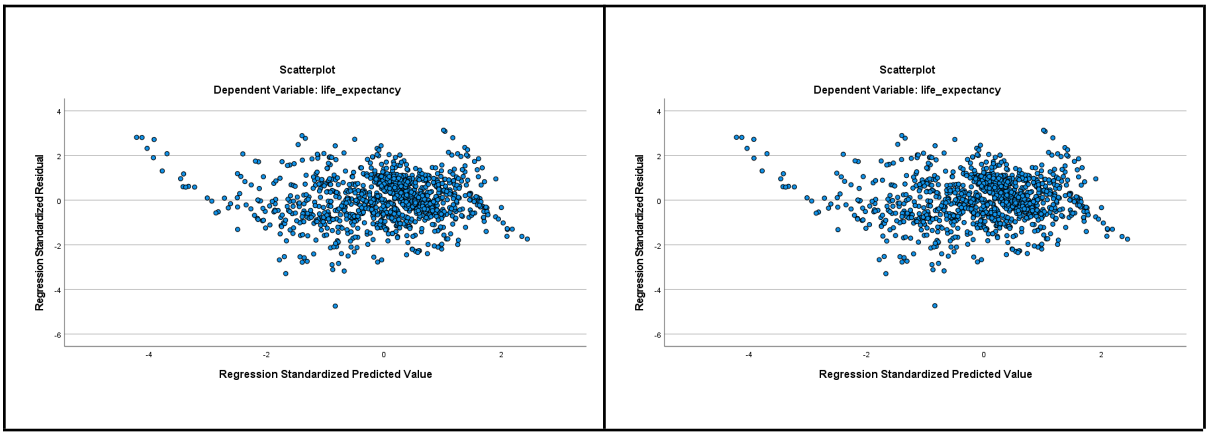


Figure 13: Linear Regression Results for Model 1 After Dropping Infant Deaths (left) and Under-5 Deaths (right)

We tested for multicollinearity for Model 2 and 3 and faced the same problem. We also dropped infant deaths in these models. The revised top 3 models can be seen in Figure 14. Refer to A4 and A5 for the refinement process for Models 2 and 3.

	Variable Name	Model 1	Model 2	Model 3
X1	Status	1	1	1
X2	Adult Mortality	1	1	1
X3	Infant Deaths	0	0	0
X4	Alcohol	1	1	1
X5	Percentage Expenditure	1	1	1
X6	Hepatitis B	0	0	0
X7	Measles	0	0	0
X8	BMI	1	1	1
X9	Under-five Deaths	1	1	1
X10	Polio	1	1	0
X11	Total Expenditure	0	0	0
X12	Diphtheria	0	0	1
X13	HIV/AIDS	1	1	1
X14	GDP	0	0	0
X15	Population	0	0	0
X16	Thinness 10-19 Years	0	1	0
X17	Thinness 5-9 Years	1	0	1
X18	Income Composition Of Resources	1	1	1
X19	Schooling	1	1	1

Figure 14: Top 3 Models (revised)

Section 4. Residual Analysis

We perform residual analysis on each of our models to examine the aptness of the model, ensuring that the assumptions of linear regression models are met. We used a mix of formal and informal methods.

Section 4.1. Test for Normality

For Model 1, we plotted the residuals against predicted life expectancy. From Figure 15, there is no observable pattern. Residuals are randomly distributed around the horizontal line through zero. This suggests that a linear relationship exists.

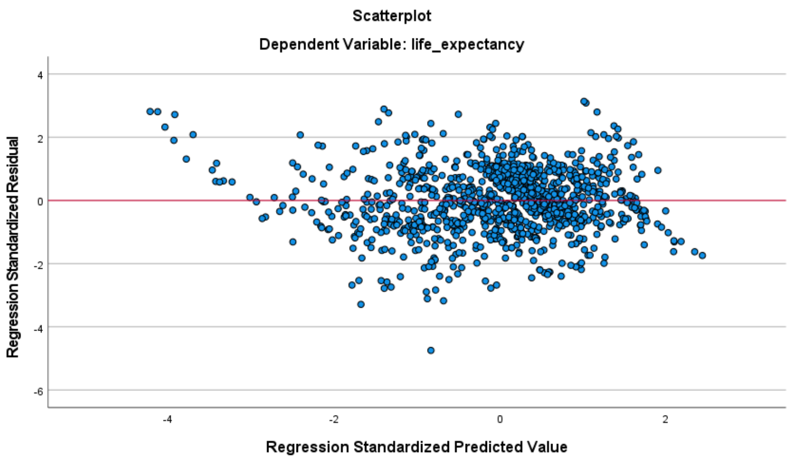


Figure 15: Residual Plot for Model 1

We repeated this process for Model 2 and Model 3, and obtained similar results. Refer to A6 for the results.

Section 4.2. Test for Variance Homogeneity

To test the variance constancy of residuals, we performed two tests.

Section 4.2.1 Levene Test

For all 3 models, p values based on mean, median are both high, which means we cannot reject the hypothesis that variance of residuals are equal. That is to say, residual variance has homogeneity.

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
ZRE_1	Based on Mean	.100	1	1152	.752
	Based on Median	.098	1	1152	.755
	Based on Median and with adjusted df	.098	1	1148.827	.755
	Based on trimmed mean	.101	1	1152	.750
ZRE_2	Based on Mean	.100	1	1152	.751
	Based on Median	.096	1	1152	.757
	Based on Median and with adjusted df	.096	1	1148.868	.757
	Based on trimmed mean	.100	1	1152	.752
ZRE_3	Based on Mean	.035	1	1152	.851
	Based on Median	.033	1	1152	.856
	Based on Median and with adjusted df	.033	1	1149.194	.856
	Based on trimmed mean	.034	1	1152	.853

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
ZRE_1	Between Groups	.078	1	.078	.079	.779
	Within Groups	1141.921	1152	.991		
	Total	1142.000	1153			
ZRE_2	Between Groups	.089	1	.089	.090	.764
	Within Groups	1141.911	1152	.991		
	Total	1142.000	1153			
ZRE_3	Between Groups	.084	1	.084	.085	.771
	Within Groups	1141.916	1152	.991		
	Total	1142.000	1153			

Figure 16: Levene Test Results for Models 1, 2 and 3

Section 4.2.2 Brown Forsythe Test

We also conduct the Brown Forsythe test. We divide the standardized residual into two groups and test the variance. p values for the tests are high, which indicate we cannot reject the hypothesis of constant variance. Accordingly, the three models have constancy of error variance.

ANOVA

ZRE_D1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.040	1	.040	.100	.752
Within Groups	461.219	1152	.400		
Total	461.259	1153			

ANOVA

ZRE_D2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.040	1	.040	.100	.751
Within Groups	458.853	1152	.398		
Total	458.893	1153			

ANOVA

ZRE_D3

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.014	1	.014	.035	.851
Within Groups	461.899	1152	.401		
Total	461.913	1153			

Figure 17: Brown Forsythe Test Results for Models 1, 2 and 3

Section 5. Outlier Analysis

We have to identify outliers and determine whether the outlier is influential. We will elaborate on the analysis of Model 1. We obtained similar results for Model 2 and Model 3. Refer to A8 and A9 for the analysis.

Section 5.1. Identifying Outlying Observations

When we use the boxplot to plot the standardized residual of our model, we find out that there are some cases that can potentially be an outlier for our model.

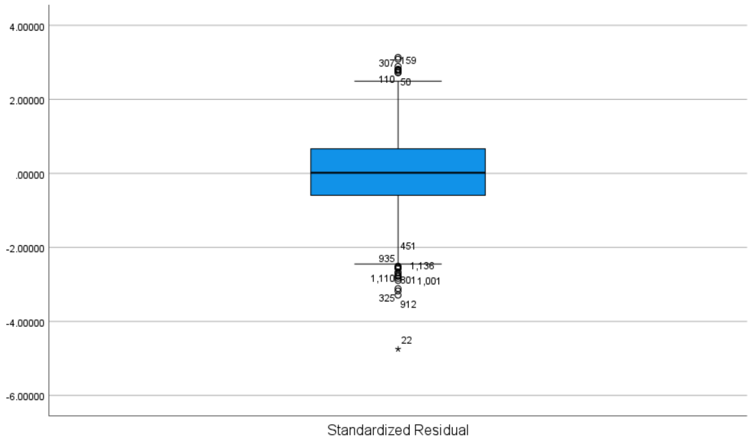


Figure 18: Boxplot of Standardized Residual of Model 1

Section 5.2. Identifying Influential Cases

We need to further determine if these outliers are influential. We have to consider both the influence on a single fitted value and influence on regression coefficient.

Section 5.2.1. Cook's Distance

To analyse the influence on single fitted values, we decided to use Cook's Distance as it is more conservative than DFFITS. From Figure 19, we see that the maximum value is smaller than 1. This means that the Cook's Distance of all cases is below 1, implying that there is no influential outlier within our model.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	36.014	88.727	69.413	7.9223	1154
Std. Predicted Value	-4.216	2.438	.000	1.000	1154
Standard Error of Predicted Value	.159	1.389	.343	.155	1154
Adjusted Predicted Value	35.390	88.974	69.408	7.9363	1154
Residual	-17.5159	11.5724	.0000	3.6723	1154
Std. Residual	-4.747	3.136	.000	.995	1154
Stud. Residual	-4.772	3.150	.001	1.002	1154
Deleted Residual	-17.7039	11.6729	.0050	3.7227	1154
Stud. Deleted Residual	-4.819	3.162	.001	1.003	1154
Mahal. Distance	1.141	162.439	10.990	14.623	1154
Cook's Distance	.000	.042	.001	.003	1154
Centered Leverage Value	.001	.141	.010	.013	1154

a. Dependent Variable: life_expectancy

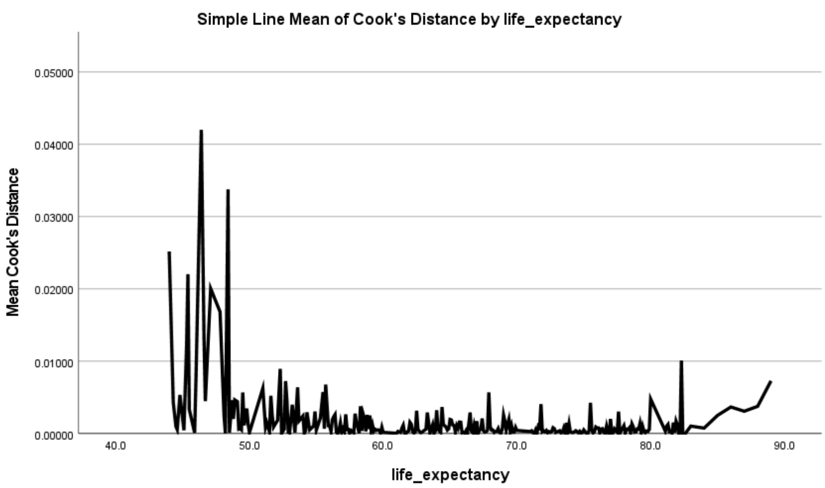


Figure 19: Cook's Distance Results for Model 1

Section 5.2.2 DFBETAS

To analyse the influence on the regression coefficient, we use DFBETAS. We sorted the results in descending order and found that the largest value is still smaller than 1. Hence, we conclude that there are no influential outliers.

DFB0_1	DFB1_1	DFB2_1	DFB3_1	DFB4_1	DFB5_1	DFB6_1	DFB7_1
-.03235	-.03253	-.00002	-.01084	.00324	.00000	.00018	.00830
-.02484	.00567	-.00004	.00089	.00019	.00000	.00041	-.00066
-.01750	-.00113	-.00006	.00090	.00177	.00000	.00041	-.00068
-.01242	-.00429	-.00007	.00082	.00250	.00000	.00037	-.00063
-.07482	.01562	.00001	-.00016	-.00019	-.00004	-.00033	.00013
-.09307	-.00315	.00040	.00133	.00278	.00000	.00068	-.00100
-.08338	.02091	.00008	.00025	-.00378	.00000	.00042	-.00019
-.13578	.01310	.00043	.00053	-.00291	.00000	-.00009	-.00036
-.28688	.01491	.00038	.00057	-.00741	.00000	.00145	-.00034
.08165	-.01693	-.00012	.00037	.00322	.00000	.00056	-.00025
-.00280	.08254	-.00003	-.00021	-.00198	-.00001	-.00249	.00017
.05067	-.03123	-.00006	.00050	.00677	.00000	-.00004	-.00038
-.01989	.00101	.00000	.00060	-.00006	.00000	.00020	-.00044
-.12324	.04445	.00029	.00173	-.00812	.00000	.00079	-.00130
-.02181	.01796	.00010	.00042	-.00383	.00000	-.00007	-.00026
.06507	-.01016	-.00015	.00009	-.00001	.00000	-.00089	-.00006
-.10303	.00786	.00032	.00043	-.00174	.00000	-.00008	-.00030
-.02678	.00344	-.00002	.00008	.00002	.00000	-.00026	-.00011
.05283	-.01032	-.00013	.00008	-.00009	.00000	-.00086	-.00005

Figure 20: DFBETAS Results for Model 1

Section 6. Model Validation

From Figure 21, we conclude that Model 2 is the best model.

	Model 1 Training Data Set	Model 1 Validation Data Set	Model 2 Training Data Set	Model 2 Validation Data Set	Model 3 Training Data Set	Model 3 Validation Data Set
p	12	12	12	12	12	12
b0	53.176524	52.303914	53.329375	52.150106	53.119528	51.600691
s[b0]	0.829116	1.255296	0.839376	1.255415	0.831206	1.274214
b1	0.867766	1.220452	0.885223	1.218722	0.872781	1.207842
s[b1]	0.407999	0.650101	0.407923	0.649485	0.407847	0.647333
b2	-0.017307	-0.018877	-0.017406	-0.018957	-0.017535	-0.018645
s[b2]	0.001147	0.001801	0.001144	0.001802	0.001144	0.001797
b4	-0.154327	-0.087942	-0.157699	-0.083788	-0.151832	-0.090225
s[b4]	0.039824	0.062392	0.03991	0.062466	0.039776	0.062131
b5	0.000452	0.000306	0.000453	0.000304	0.000452	0.000311
s[b5]	0.000068	0.000135	0.000068	0.000135	0.000068	0.000135
b8	0.030996	0.04137	0.030608	0.042136	0.032102	0.042097
s[b8]	0.007455	0.010746	0.007404	0.01063	0.007459	0.010695
b9	-0.000996	-0.003608	-0.000918	-0.003841	-0.000996	-0.003396
s[b9]	0.000756	0.001244	0.000755	0.00126	0.000755	0.001243
b10	0.021869	0.010147	0.022393	0.01014	-	-
s[b10]	0.005229	0.008096	0.005242	0.008086	-	-
b12	-	-	-	-	0.023261	0.020384
s[b12]	-	-	-	-	0.005435	0.008511
b13	-0.438672	-0.426734	-0.438369	-0.427023	-0.435682	-0.428208
s[b13]	0.021241	0.035439	0.021226	0.035411	0.021249	0.035283
b16	-	-	-0.08462	0.058364	-	-
s[b16]	-	-	0.033925	0.047966	-	-
b17	-0.074242	0.041787	-	-	-0.069642	0.038667
s[b17]	0.033482	0.046608	-	-	0.03339	0.046356
b18	10.051414	13.156548	9.992438	13.203582	9.74138	12.88322
s[b18]	0.96527	1.761906	0.965832	1.760682	0.970419	1.754947
b19	0.930441	0.83162	0.924254	0.83331	0.936919	0.828183
s[b19]	0.073381	0.108691	0.073435	0.108573	0.073096	0.108097
SSEp	15548.69561	6529.274629	15531.0247	6520.155169	15537.60896	6473.629919
PRESSp	15979.23673	6997.378324	15961.18123	6986.693527	15970.79875	6933.212278
Cp	32.677257	50.531486	32.697796	49.28955	33.190159	50.012274
MSEp	13.61532	13.518167	13.599846	13.499286	13.605612	13.40296
MSPR	13.4737401	-	13.45842761	-	13.46413251	

Figure 21: Summary of Data Validation

To evaluate and choose the best model, we look at 2 aspects – predictive and descriptive power. All 3 models performed similarly.

Among the three models, Model 2 has lowest MSPR, which indicates it has the most predictive power.

When we compare the beta coefficients in the training and validation sets, we notice that it is relatively stable and the changes in magnitude are generally insignificant. However, we noticed that the coefficients of X16 and X17 (Thinness 5-9 and Thinness 10-19) are negative in the training sets but positive in the validation sets. Using a 99% confidence level, we notice that the thinness predictor is not that significant in the model using t-test. Therefore, we conclude the descriptive power of these models are not affected by this single insignificant variable.

Since all three models have either Thinness 5-9 or Thinness 10-19 as a predictor variable inside, we choose model 2 as the champion model based on its relatively strong predictive power.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.822	3.6878	2.062

a. Predictors: (Constant), underfive_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72382.708	11	6580.246	483.847	.000[b]
	Residual	15531.025	1142	13.600		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), underfive_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	53.329	.839		63.535	.000		
	status	.885	.408	.036	2.170	.030	.558	1.791
	adult_mortality	-.017	.001	-.250	-15.210	<.001	.572	1.749
	alcohol	-.158	.040	-.073	-3.951	<.001	.449	2.227
	percentage_expenditure	.000	.000	.096	6.675	<.001	.742	1.348
	bmi	.031	.007	.069	4.134	<.001	.561	1.781
	polio	.022	.005	.058	4.272	<.001	.837	1.195
	hivaids	-.438	.021	-.308	-20.652	<.001	.694	1.441
	income_composition_of_resources	9.992	.966	.213	10.346	<.001	.366	2.730
	schooling	.924	.073	.290	12.586	<.001	.291	3.435
	thinness_1019_years	-.085	.034	-.043	-2.494	.013	.524	1.909
	underfive_deaths	-.001	.001	-.017	-1.216	.224	.762	1.312

a. Dependent Variable: life_expectancy

Figure 22: Linear Regression Results for Global Optimal Model – Model 2

Section 7. Summary and Concluding Remarks

Based on the automatic method and the all-possible-regression method, we build a linear regression model with life expectancy as a dependent variable and 11 predictor variables. We conclude that life expectancy can be explained by the status of the country, adult mortality, alcohol, percentage expenditure, BMI, under-five deaths, polio, HIV/AIDS, thinness 10-19 years, income composition of resources and schooling. From the standardised beta coefficient, we know that HIV/AIDS, schooling and adult mortality are the most important factors in predicting values of life expectancy. Governments can place more emphasis on these aspects to improve life expectancy for the country. We also acknowledge that our model can possibly be improved. One way is to dive deep into the data analysis with professional domain knowledge to filter out highly correlated factors and insignificant variables during the preliminary data preparation stage. By doing this, it is likely that we do not need remedial measures for multicollinearity after the globally optimised models have been selected. This could also ensure that all predictors are significant in the model.

Appendix

Appendix Table of Contents

A1. References

Applied Linear Statistical Models 5th (fifth) Edition by Kutner, Michael H, Neter, John, Nachtsheim, Christopher J., published by McGraw-Hill Higher Education (2004) (5th ed.). (2004). McGraw-Hill Higher Education.

Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2013, May 23). Life Expectancy. Our World in Data. https://ourworldindata.org/life-expectancy

Life Expectancy (WHO). (2018, February 10). Kaggle. https://www.kaggle.com/kumarajarshi/life-expectancy-who

Life Expectancy: Exploratory Data Analysis. (2019, November 2). Kaggle. https://www.kaggle.com/philbowman212/life-expectancy-exploratory-data-analysis

“Life Expectancy” – What does this actually mean? (n.d.). Our World in Data. Retrieved August 28, 2017, from https://ourworldindata.org/life-expectancy-how-is-it-calculated-and-how-should-it-be-interpreted

GHE: Life expectancy and healthy life expectancy. (2020b, December 1). THE GLOBAL HEALTH OBSERVATORY. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy

Roser, M. (2013, May 23). Life Expectancy. Our World in Data. https://ourworldindata.org/life-expectancy#twice-as-long-life-expectancy-around-the-world

A2. Variable Definition

Variable Name	Description
Life expectancy	Life expectancy of people in years for a particular country and year
Status	Whether a country is considered to be 'Developing' or 'Developed' by WHO standards
Adult mortality	Number of people dying between 15 and 60 years per 1000 population
Infant deaths	Number of infant deaths per 1000 population
Alcohol	The country's alcohol consumption rate measured as litres of pure alcohol consumption per capita
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product (GDP)
Hepatitis B	Number of 1-year-olds with Hepatitis B immunization over all 1-year-olds in population
Measles	Number of reported Measles cases per 1000 population
BMI	Average Body Mass Index (BMI) of a country's total population
Under-five deaths	Number of people under the age of five deaths per 1000 populations
Polio	Number of 1-year-olds with Polio immunization over the number of all 1-year-olds in population
Total expenditure	Government expenditure on health as a percentage of total government expenditure
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1-year-olds
HIV/AIDS	Deaths per 1000 live births caused by HIV/AIDS for people under 5
GDP	GDP per capita
Population	Population of a country
Thinness 10-19 years	Rate of thinness among people aged 10-19
Thinness 5-9 years	Rate of thinness among people aged 5-9
Income composition of resources	Human Development Index in terms of income composition of resources, ranging from 0 to 1
Schooling	Average number of years of schooling of a population

A3. Python Codes for All-Possible-Regression Procedure

We used python to conduct the all-possible-regression procedure as our dataset contained too many variables to do so manually or in SPSS. We made references to this site:

https://github.com/Superbblue2021/Linear-Regression.

A4. Refinement Process for Model 2

<u>Before dropping collinearity variable</u>

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.912[a]	.832	.830	3.6023	2.054

a. Predictors: (Constant), underfive_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling, infant_deaths

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	73107.407	12	6092.284	469.481	.000[b]
	Residual	14806.326	1141	12.977		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), underfive_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling, infant_deaths

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	54.083	.826		65.468	.000		
	status	.950	.399	.039	2.383	.017	.558	1.792
	adult_mortality	-.016	.001	-.232	-14.312	<.001	.560	1.787
	alcohol	-.114	.039	-.053	-2.891	.004	.439	2.277
	percentage_expenditure	.000	.000	.099	7.045	<.001	.741	1.349
	bmi	.032	.007	.072	4.429	<.001	.561	1.782
	polio	.017	.005	.043	3.212	.001	.818	1.222
	hivaids	-.438	.021	-.308	-21.103	<.001	.694	1.441
	income_composition_of_resources	9.362	.947	.199	9.883	<.001	.363	2.752
	schooling	.897	.072	.282	12.490	<.001	.290	3.443
	infant_deaths	.090	.012	1.255	7.473	<.001	.005	190.928
	thinness_1019_years	-.089	.033	-.045	-2.692	.007	.524	1.910
	underfive_deaths	-.067	.009	-1.271	-7.550	<.001	.005	192.086

a. Dependent Variable: life_expectancy

<u>Results if we drop infant deaths</u>

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.822	3.6878	2.062

a. Predictors: (Constant), underfive_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72382.708	11	6580.246	483.847	.000[b]
	Residual	15531.025	1142	13.600		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), underfive_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	53.329	.839		63.535	.000		
	status	.885	.408	.036	2.170	.030	.558	1.791
	adult_mortality	-.017	.001	-.250	-15.210	<.001	.572	1.749
	alcohol	-.158	.040	-.073	-3.951	<.001	.449	2.227
	percentage_expenditure	.000	.000	.096	6.675	<.001	.742	1.348
	bmi	.031	.007	.069	4.134	<.001	.561	1.781
	polio	.022	.005	.058	4.272	<.001	.837	1.195
	hivaids	-.438	.021	-.308	-20.652	<.001	.694	1.441
	income_composition_of_resources	9.992	.966	.213	10.346	<.001	.366	2.730
	schooling	.924	.073	.290	12.586	<.001	.291	3.435
	thinness_1019_years	-.085	.034	-.043	-2.494	.013	.524	1.909
	underfive_deaths	-.001	.001	-.017	-1.216	.224	.762	1.312

a. Dependent Variable: life_expectancy

Results if we drop under-5 deaths

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.821	3.6896	2.061

a. Predictors: (Constant), infant_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72367.635	11	6578.876	483.277	.000[b]
	Residual	15546.098	1142	13.613		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), infant_deaths, hivaids, alcohol, polio, percentage_expenditure, bmi, adult_mortality, status, thinness_1019_years, income_composition_of_resources, schooling

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	53.289	.839		63.498	.000		
	status	.891	.408	.036	2.183	.029	.558	1.791
	adult_mortality	-.017	.001	-.250	-15.175	<.001	.571	1.751
	alcohol	-.161	.040	-.075	-4.037	<.001	.450	2.220
	percentage_expenditure	.000	.000	.096	6.673	<.001	.742	1.348
	bmi	.031	.007	.069	4.139	<.001	.561	1.781
	polio	.023	.005	.059	4.382	<.001	.840	1.190
	hivaids	-.438	.021	-.308	-20.609	<.001	.694	1.441
	income_composition_of_resources	9.947	.967	.212	10.288	<.001	.366	2.733
	schooling	.928	.073	.291	12.638	<.001	.291	3.432
	thinness_1019_years	-.093	.034	-.047	-2.732	.006	.524	1.910
	infant_deaths	-.001	.001	-.009	-.608	.543	.767	1.304

a. Dependent Variable: life_expectancy

A5. Refinement Process for Model 3

<u>Before dropping collinearity variable</u>

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.912[a]	.832	.830	3.6031	2.057

a. Predictors: (Constant), schooling, hivaids, infant_deaths, diphtheria, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources, underfive_deaths

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	73101.137	12	6091.761	469.243	.000[b]
	Residual	14812.596	1141	12.982		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), schooling, hivaids, infant_deaths, diphtheria, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources, underfive_deaths

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	54.008	.821		65.816	.000		
	status	.938	.398	.038	2.353	.019	.559	1.790
	adult_mortality	-.016	.001	-.233	-14.326	<.001	.558	1.791
	infant_deaths	.090	.012	1.261	7.473	<.001	.005	192.947
	alcohol	-.110	.039	-.051	-2.800	.005	.443	2.257
	percentage_expenditure	.000	.000	.099	7.003	<.001	.741	1.349
	bmi	.032	.007	.073	4.454	<.001	.553	1.807
	underfive_deaths	-.068	.009	-1.278	-7.559	<.001	.005	193.714
	diphtheria	.016	.005	.041	3.044	.002	.807	1.240
	hivaids	-.435	.021	-.306	-20.974	<.001	.693	1.444
	thinness_59_years	-.084	.033	-.043	-2.585	.010	.529	1.889
	income_composition_of_resources	9.193	.951	.196	9.669	<.001	.361	2.771
	schooling	.909	.071	.285	12.716	<.001	.293	3.411

a. Dependent Variable: life_expectancy

<u>Results if we drop infant deaths</u>

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.822	3.6886	2.069

a. Predictors: (Constant), schooling, hivaids, underfive_deaths, diphtheria, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72376.123	11	6579.648	483.598	.000[b]
	Residual	15537.609	1142	13.606		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), schooling, hivaids, underfive_deaths, diphtheria, percentage_expenditure, status, bmi, adult_mortality, thinness_59_years, alcohol, income_composition_of_resources

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	53.120	.831		63.907	.000		
	status	.873	.408	.036	2.140	.033	.559	1.789
	adult_mortality	-.018	.001	-.252	-15.323	<.001	.572	1.748
	alcohol	-.152	.040	-.071	-3.817	<.001	.452	2.211
	percentage_expenditure	.000	.000	.096	6.647	<.001	.742	1.348
	bmi	.032	.007	.072	4.304	<.001	.553	1.807
	underfive_deaths	-.001	.001	-.019	-1.319	.188	.761	1.313
	diphtheria	.023	.005	.058	4.280	<.001	.831	1.204
	hivaids	-.436	.021	-.306	-20.503	<.001	.693	1.444
	thinness_59_years	-.070	.033	-.036	-2.086	.037	.531	1.882
	income_composition_of_resources	9.741	.970	.207	10.038	<.001	.363	2.755
	schooling	.937	.073	.294	12.818	<.001	.294	3.401

a. Dependent Variable: life_expectancy

<u>Results if we drop under-5 deaths</u>

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.907[a]	.823	.821	3.6906	2.068

a. Predictors: (Constant), infant_deaths, hivaids, alcohol, diphtheria, percentage_expenditure, bmi, adult_mortality, status, thinness_59_years, income_composition_of_resources, schooling

b. Dependent Variable: life_expectancy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72359.384	11	6578.126	482.966	.000[b]
	Residual	15554.348	1142	13.620		
	Total	87913.732	1153			

a. Dependent Variable: life_expectancy

b. Predictors: (Constant), infant_deaths, hivaids, alcohol, diphtheria, percentage_expenditure, bmi, adult_mortality, status, thinness_59_years, income_composition_of_resources, schooling

Coefficients[a]

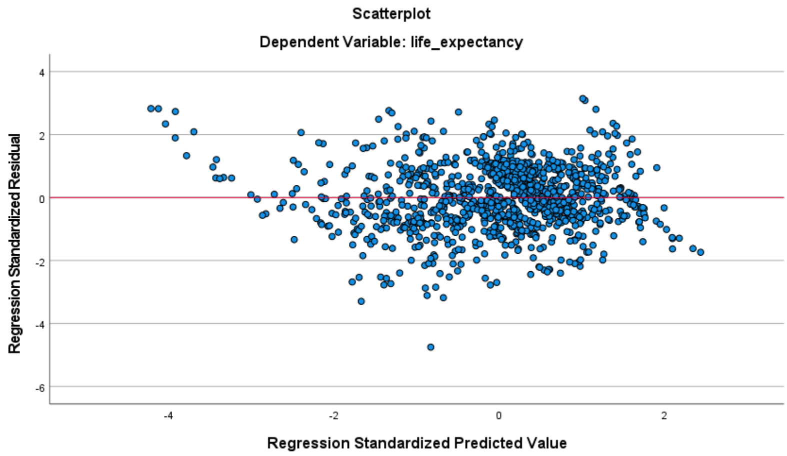
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	53.071	.831		63.874	.000		
	status	.877	.408	.036	2.150	.032	.559	1.790
	adult_mortality	-.018	.001	-.252	-15.281	<.001	.571	1.750
	alcohol	-.155	.040	-.072	-3.898	<.001	.454	2.205
	percentage_expenditure	.000	.000	.096	6.643	<.001	.742	1.348
	bmi	.032	.007	.072	4.308	<.001	.553	1.807
	diphtheria	.024	.005	.060	4.384	<.001	.834	1.199
	hivaids	-.435	.021	-.306	-20.456	<.001	.693	1.444
	thinness_59_years	-.077	.033	-.040	-2.316	.021	.530	1.887
	income_composition_of_resources	9.694	.971	.206	9.979	<.001	.363	2.758
	schooling	.942	.073	.296	12.881	<.001	.294	3.398
	infant_deaths	-.001	.001	-.010	-.713	.476	.764	1.308

a. Dependent Variable: life_expectancy

A6. Residual Plot for Model 2 and Model 3

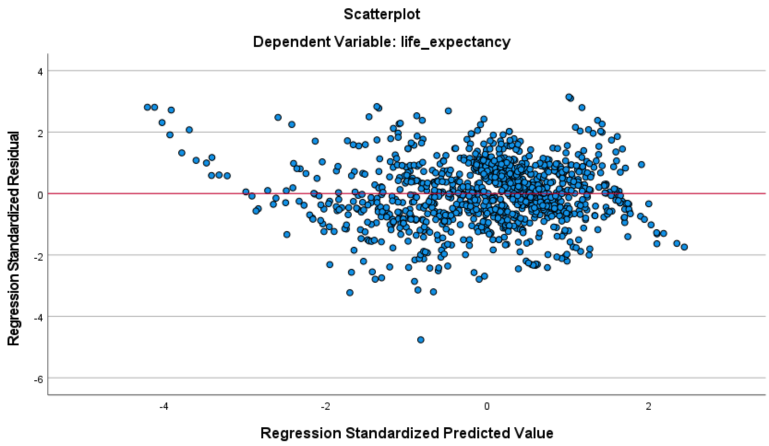
<u>Model 2</u>

No observable pattern, residuals are randomly distributed around the horizontal line through zero. This suggests that a linear relationship exists.



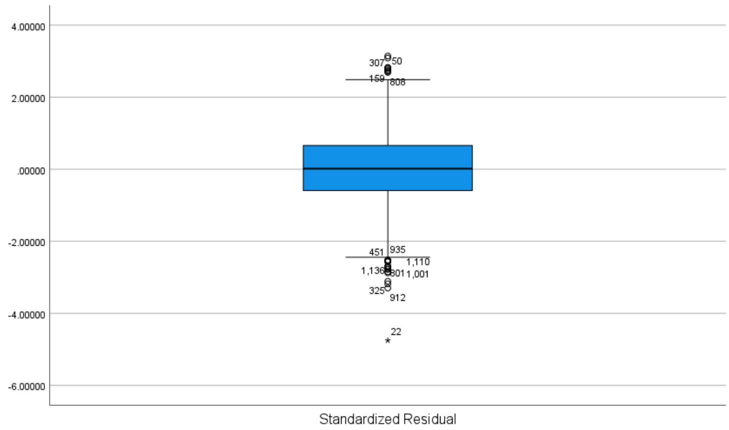
<u>Model 3</u>

No observable pattern, residuals are randomly distributed around the horizontal line through zero. This suggests that a linear relationship exists.



A7. Outlier Analysis for Model 2

Boxplot

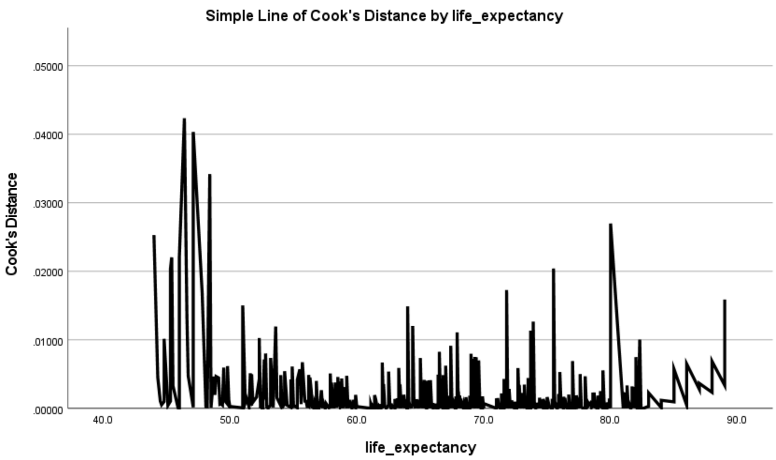


Cook's Distance

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	35.980	88.708	69.413	7.9232	1154
Std. Predicted Value	-4.220	2.435	.000	1.000	1154
Standard Error of Predicted Value	.157	1.388	.343	.155	1154
Adjusted Predicted Value	35.354	88.954	69.408	7.9371	1154
Residual	-17.5235	11.5973	.0000	3.6702	1154
Std. Residual	-4.752	3.145	.000	.995	1154
Stud. Residual	-4.777	3.158	.001	1.002	1154
Deleted Residual	-17.7110	11.6980	.0048	3.7206	1154
Stud. Deleted Residual	-4.823	3.171	.001	1.003	1154
Mahal. Distance	1.103	162.232	10.990	14.574	1154
Cook's Distance	.000	.042	.001	.003	1154
Centered Leverage Value	.001	.141	.010	.013	1154

a. Dependent Variable: life_expectancy

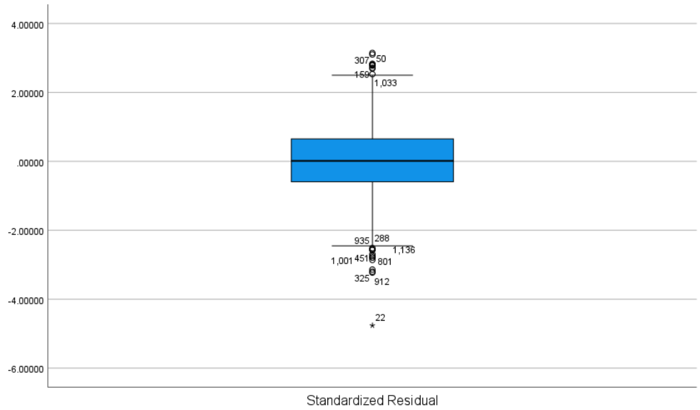


DFBETAS

DFB0_2	DFB1_2	DFB2_2	DFB3_2	DFB4_2	DFB5_2	DFB6_2	DFB7_2
-.03126	-.03259	-.00002	-.01080	.00322	.00000	.00016	.00828
-.02889	.00552	-.00004	.00089	.00028	.00000	.00045	-.00067
-.02207	-.00142	-.00006	.00091	.00188	.00000	.00044	-.00069
-.01704	-.00467	-.00007	.00084	.00262	.00000	.00040	-.00064
-.07750	.01573	.00001	-.00017	-.00015	-.00004	-.00031	.00013
-.09588	-.00341	.00040	.00135	.00285	.00000	.00069	-.00101
-.08839	.02067	.00008	.00027	-.00369	.00000	.00044	-.00020
-.13722	.01331	.00042	.00050	-.00291	.00000	-.00006	-.00035
-.28928	.01506	.00037	.00055	-.00739	.00000	.00148	-.00032
.08522	-.01655	-.00012	.00035	.00312	.00000	.00054	-.00024
-.00350	.08346	-.00004	-.00028	-.00201	-.00001	-.00245	.00022
.05311	-.03105	-.00005	.00051	.00668	.00000	-.00007	-.00038
-.02203	.00103	.00000	.00060	-.00002	.00000	.00022	-.00044
-.12364	.04447	.00029	.00173	-.00812	.00000	.00079	-.00130
-.02152	.01874	.00009	.00036	-.00391	.00000	-.00003	-.00022
-.10408	.00800	.00031	.00042	-.00173	.00000	-.00006	-.00029
.06708	-.00980	-.00015	.00006	-.00008	.00000	-.00089	-.00004
.22438	.02580	-.00003	-.00158	-.00137	.00000	-.00041	.00121
.05449	-.01001	-.00014	.00005	-.00014	.00000	-.00086	-.00003

A8. Outlier Analysis for Model 3

Boxplot

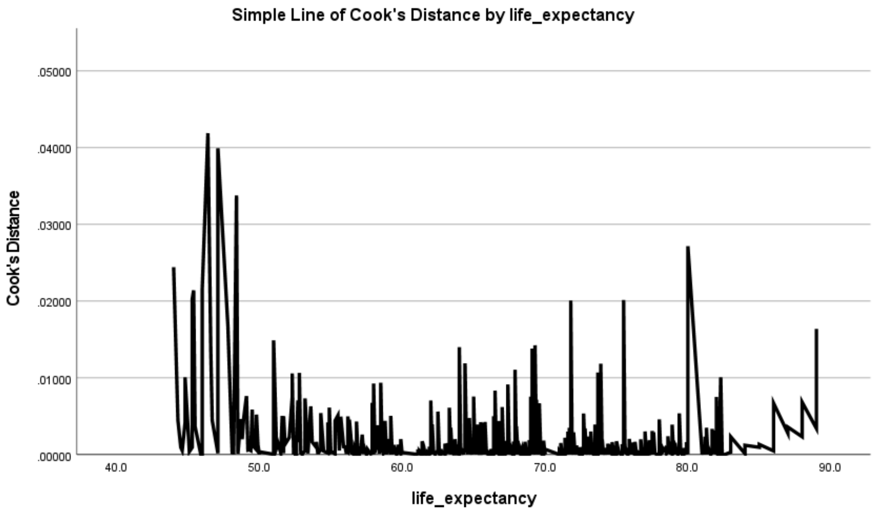


Cook's Distance

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	36.035	88.724	69.413	7.9229	1154
Std. Predicted Value	-4.213	2.437	.000	1.000	1154
Standard Error of Predicted Value	.157	1.391	.342	.157	1154
Adjusted Predicted Value	35.412	88.971	69.408	7.9371	1154
Residual	-17.5616	11.6032	.0000	3.6709	1154
Std. Residual	-4.761	3.146	.000	.995	1154
Stud. Residual	-4.786	3.159	.001	1.002	1154
Deleted Residual	-17.7478	11.7039	.0050	3.7218	1154
Stud. Deleted Residual	-4.833	3.172	.001	1.003	1154
Mahal. Distance	1.096	162.892	10.990	14.752	1154
Cook's Distance	.000	.042	.001	.003	1154
Centered Leverage Value	.001	.141	.010	.013	1154

a. Dependent Variable: life_expectancy



DFBETAS

DFB0_3	DFB1_3	DFB2_3	DFB3_3	DFB4_3	DFB5_3	DFB6_3	DFB7_3
-.08918	-.03075	-.00002	-.01041	.00227	.00000	.00018	.00799
-.03018	.00553	-.00005	.00087	.00030	.00000	.00046	-.00066
-.02245	-.00140	-.00006	.00090	.00191	.00000	.00045	-.00069
-.01748	-.00466	-.00007	.00083	.00265	.00000	.00041	-.00064
-.07575	.01578	.00001	-.00015	-.00016	-.00004	-.00032	.00012
-.10708	-.00348	.00040	.00124	.00267	.00000	.00069	-.00092
-.18486	.01159	.00038	.00001	-.00322	.00000	.00001	.00004
-.09043	.02058	.00008	.00024	-.00369	.00000	.00045	-.00018
-.28671	.01519	.00037	.00056	-.00730	.00000	.00151	-.00034
.07831	-.01646	-.00012	.00030	.00305	.00000	.00055	-.00020
.15525	-.00742	-.00008	.00091	.00281	.00000	.00035	-.00068
-.00005	.08380	-.00004	-.00026	-.00198	-.00001	-.00246	.00020
.04361	-.03068	-.00006	.00043	.00654	.00000	-.00004	-.00032
-.03674	-.00112	-.00008	-.00153	-.00099	.00000	.00023	.00137
-.02317	.00104	-.00001	.00058	.00000	.00000	.00023	-.00043
-.01987	.01892	.00009	.00037	-.00390	.00000	-.00003	-.00023
-.11959	.04503	.00029	.00177	-.00807	.00000	.00082	-.00133
-.09970	.00814	.00032	.00046	-.00168	.00000	-.00006	-.00033
.15255	.00397	-.00003	.00079	.00077	.00000	.00023	-.00058
.05818	-.00963	-.00014	.00001	-.00017	.00000	-.00087	.00000
.15114	.00540	-.00003	.00077	.00049	.00000	.00022	-.00057