Determining Key Predictors of Global Life Expectancy

STAT 311-50 FINAL PROJECT

Course: STAT 311 – Regression Analysis
Names: Michael Slather, Kathryn Stoven
Instructor: Dr. Iresha Premarathna
Date: Fall 2025

Introduction

Life expectancy is one of the most important global health indicators used by the World Health Organization to assess a population's overall health and development. Nations differ widely in income, education, disease burden, and access to healthcare, all of which may influence longevity.

The purpose of this study is to build a multiple linear regression model to identify which factors are most strongly associated with life expectancy across countries. Understanding these relationships can help inform public health policies, economic investments, and global development programs.

This report uses multiple regression modeling, model selection techniques (stepwise, all possible models, k-folds cross-validation), and transformations of the response variable to address the following research questions:

Research Questions

1. Which socioeconomic and health predictors are most strongly associated with life expectancy across countries in 2015?
2. Do interaction or quadratic (second order) terms substantially improve model adequacy and predictive accuracy?
3. Does transforming the response variable improve the normality of residuals and satisfy regression assumptions?
4. Which final model provides the best balance of interpretability, predictive accuracy, and assumption validity, and what is its prediction equation?

The overall goal is to develop a statistically significant regression model that explains life expectancy effectively while satisfying standard regression assumptions.

Data Description

Dataset and Missing Values

This report used a dataset from the WHO Global Health Observatory containing annual observations from 2000-2015. The WHO collects yearly nationwide health indicators from government agencies, healthcare reports, and data agencies, Estimates are standardized before releasing the report.

The dataset was filtered to only include 2015 due to uneven data collection and numerous null values. Observations with missing values were removed from the analysis. Thus, 10 national observations were excluded from the final, cleaned dataset.

Observations and Variables

The cleaned dataset contains:

- 183 observations (data from 183 nations out of 193)
- 7 first-order predictors: **GDP** (per capita), **Schooling** (mean years), **Adult Mortality** (probability of dying between ages 15-60), **Infant Deaths** (per 1,000 live births), **HIV/AIDS prevalence** (deaths per 1,000 population), **BMI** (average), **Alcohol** (consumption per capita)

Disclaimer: The alcohol consumption predictor displayed extreme inconsistencies and was mostly disqualified from this report, though used in exploratory analysis.

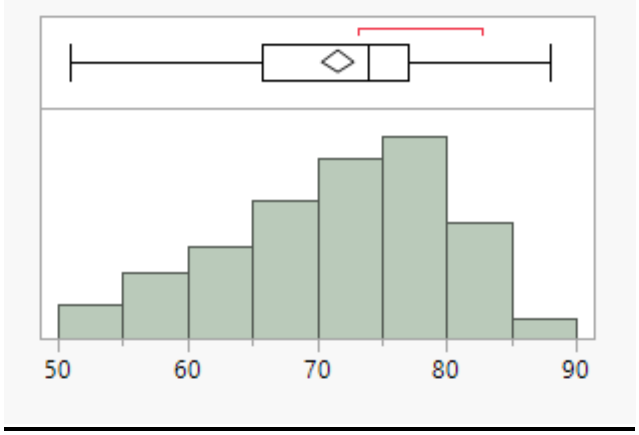
Second-Order Terms and Interactions

Second-order terms and predictor interactions were determined as significant through stepwise regression and all-possible-models techniques. Interactions and second-order terms in models with favorable metrics (AICc, BIC, Mallow's Cp) were considered significant.

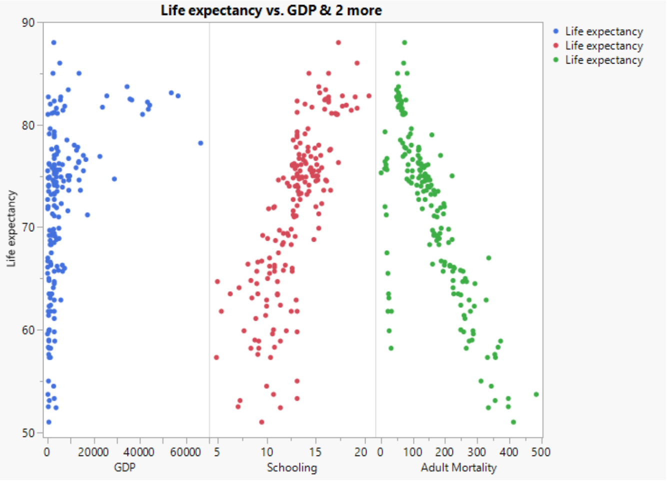
Exploratory Data Analysis (EDA)

Before model building, exploratory data analysis was conducted to visualize relationships, assess curvature, and identify potential outliers/influential observations.

Response Variable (life expectancy) Distribution



The histogram shows a left-skewed distribution with values concentrated between 60-85 years. This suggests potential outliers or influential observations requiring investigation during model building.

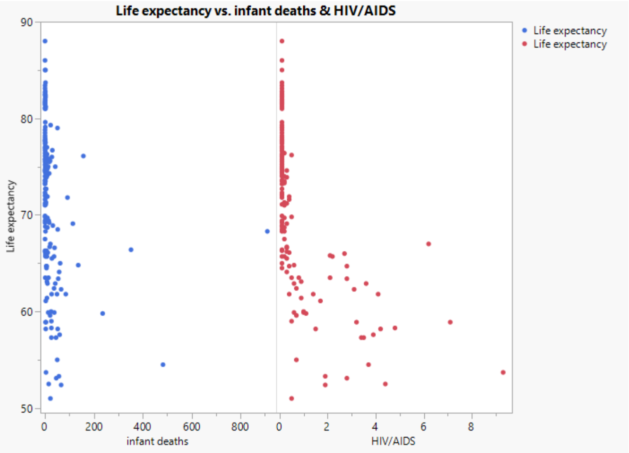


This plot visualizes the relationship between life expectancy and GDP, Schooling, and Adult Mortality respectively.

Life Expectancy vs GDP: Positive curvilinear relationship; life expectancy plateaus as GDP increases, suggesting need for a second-order term.

Life Expectancy vs Schooling: Strong positive linear relationship.

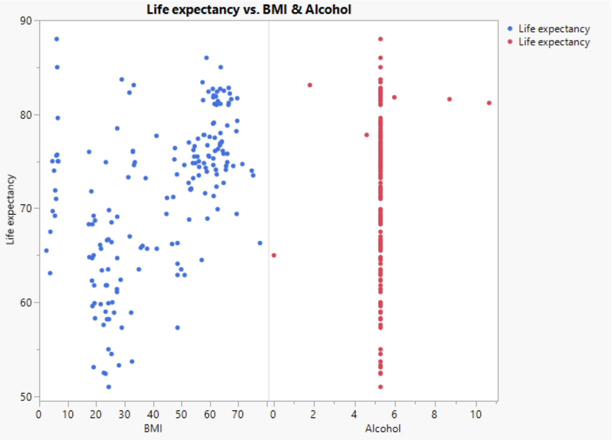
Life Expectancy vs Adult Mortality: Negative curvilinear relationship with rapid increases at lower adult mortality values.



This next plot visualizes the relationships between Life Expectancy vs Infant Deaths, and Life Expectancy vs HIV/AIDS.

Life Expectancy vs HIV/AIDS: Possible curvilinear relationship requiring a second-order term.

Life Expectancy vs Infant Deaths: High variability with little linear relationship.



This plot displays the relationships between Life Expectancy and both Average Body Mass Index and Alcohol Consumption.

Life Expectancy vs BMI: U-shaped curvilinear relationship; life expectancy decreases toward median BMI levels then increases at higher BMI

Life Expectancy vs Alcohol: No relationship; data unreliable

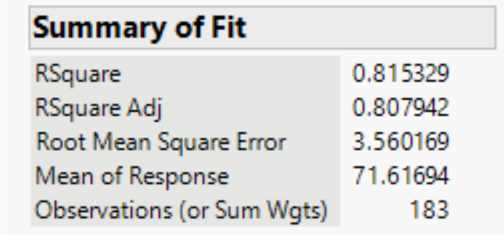
First Order Model Exploration with all First Order Terms

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

Where:

y = Life Expectancy, x_1 = GDP, x_2 = Schooling, x_3= Adult Mortality, x_4 = Infant Deaths, x_5 = HIV/AIDS, x_6 = BMI, x_7 = Alcohol

Testing Fit:



Summary of Fit	
RSquare	0.815329
RSquare Adj	0.807942
Root Mean Square Error	3.560169
Mean of Response	71.61694
Observations (or Sum Wgts)	183

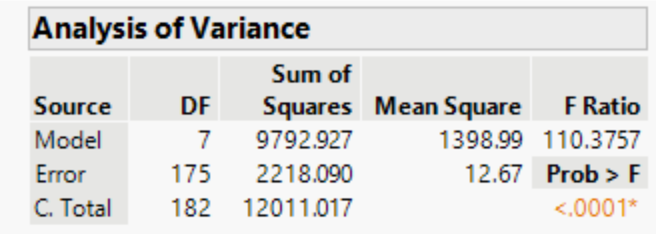
R_a^2 = 0.807942 indicates strong fit. Approximately 80.8% of variance explained by the model.

Testing Model Utility

$$H_o: \beta_i = 0, where\ i = 1,2,3,4,5,6,7$$

$$H_a: at\ least\ one\ \beta_i \neq 0, where\ i = 1,2,3,4,5,6,7$$

$$\alpha = 0.05$$

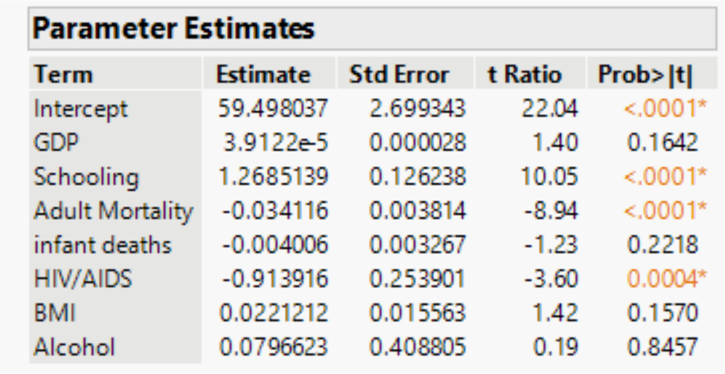


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	9792.927	1398.99	110.3757
Error	175	2218.090	12.67	Prob > F
C. Total	182	12011.017		<.0001*

F-Ratio of 110.3757, p-value < 0.0001. We reject the null hypothesis. At least one β_i is significantly different from zero at 95% confidence level.

Beta Testing

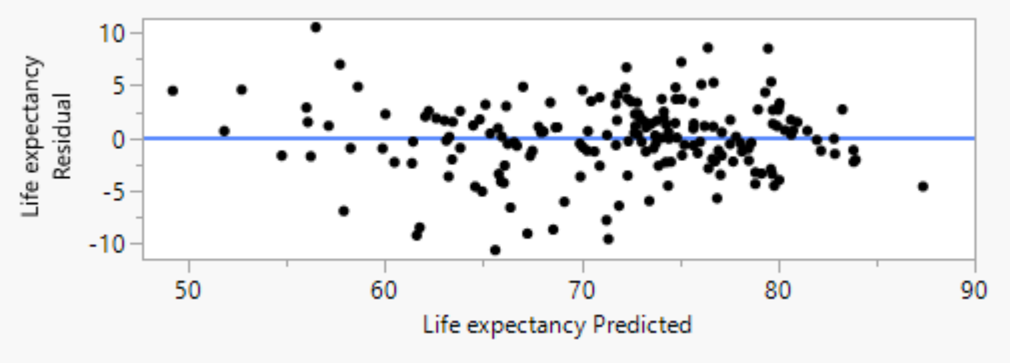


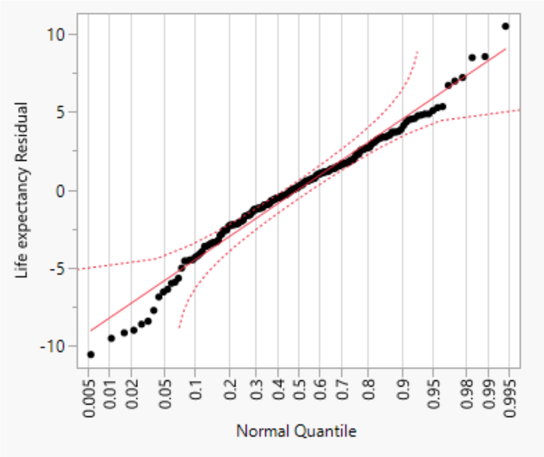
Parameter Estimates

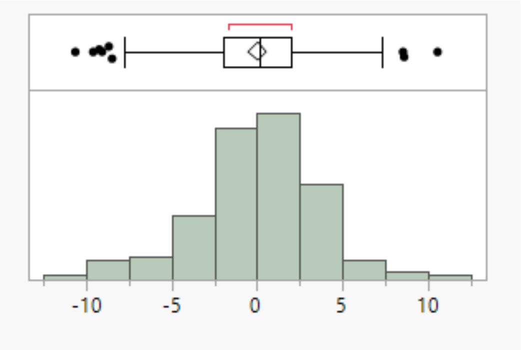
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	59.498037	2.699343	22.04	<.0001*
GDP	3.9122e-5	0.000028	1.40	0.1642
Schooling	1.2685139	0.126238	10.05	<.0001*
Adult Mortality	-0.034116	0.003814	-8.94	<.0001*
infant deaths	-0.004006	0.003267	-1.23	0.2218
HIV/AIDS	-0.913916	0.253901	-3.60	0.0004*
BMI	0.0221212	0.015563	1.42	0.1570
Alcohol	0.0796623	0.408805	0.19	0.8457

Schooling, Adult Mortality, and HIV/AIDS exhibit statistically significant p-values.

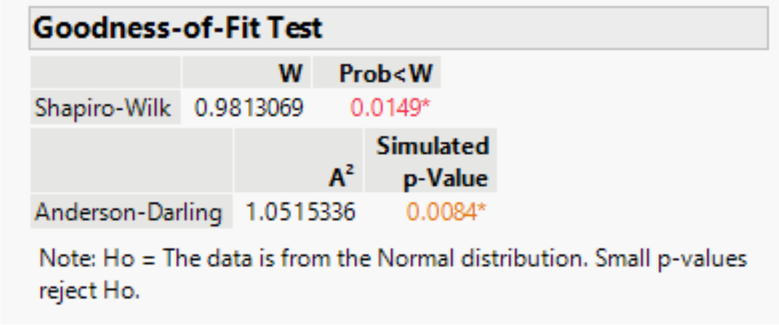
Residual/Normality Testing







Both residual vs $\hat{y}$ and Q-Q plots indicate non-normal distribution with slight bell-like form and S-shaped pattern, respectively. However, the histogram of the residuals for the full first-order model model shows a reasonably normal distribution.



Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9813069	0.0149*

	A^2	Simulated p-Value
Anderson-Darling	1.0515336	0.0084*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

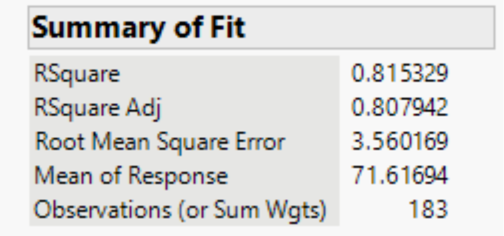
$$H_o\text{: Normal Distribution}$$

$$H_a\text{: Distribution is not normal}$$

$$\alpha = 0.05$$

With Shapiro-Wilk test statistic $W = 0.9813069, \mathrm{p-value} = 0.0149 < \alpha = 0.05$, we reject normality. The residual distribution is not normal at 95% confidence.

Outlier/Influence Testing

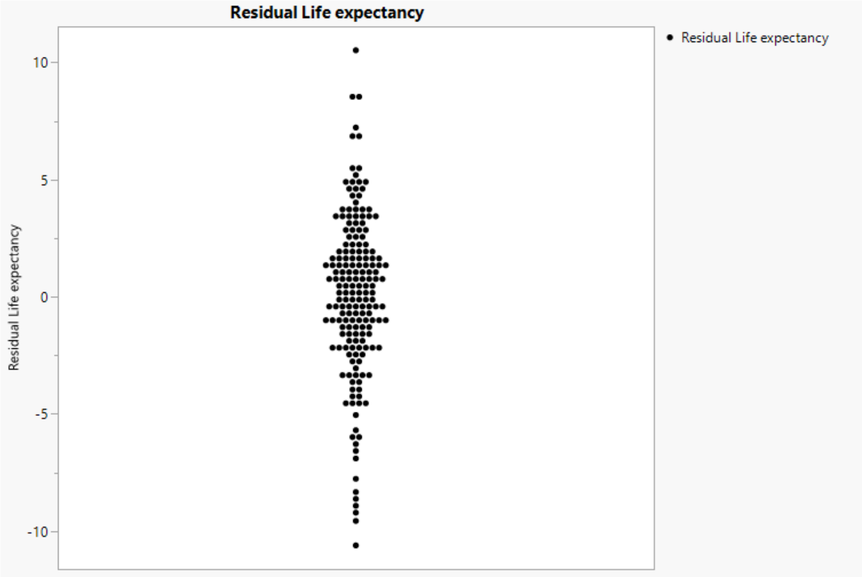


Summary of Fit

RSquare	0.815329
RSquare Adj	0.807942
Root Mean Square Error	3.560169
Mean of Response	71.61694
Observations (or Sum Wgts)	183

$$RMSE = 3.560169$$

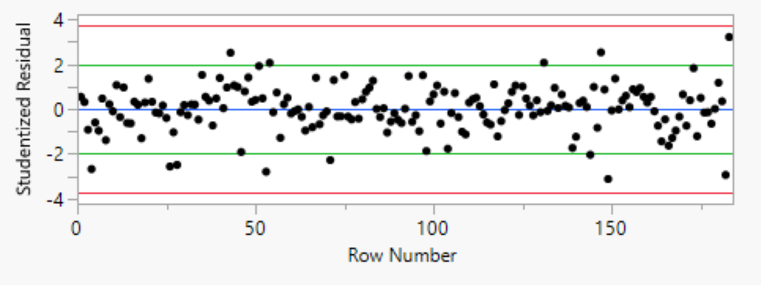
Using $3 * RMSE = 10.68$ as threshold:



From this residual plot, two observations could be outliers.

Country	Residual Life expectancy
Zimbabwe	10.524525048
Somalia	-10.6066211

However, Zimbabwe (10.52) and Somalia (-10.61) approach but don't exceed threshold.



Country	**Externally Studentized Residuals**	**h**	**Cook's D Influence**
Zimbabwe	**3.21**	**0.105**	**0.144**
Somalia	**-3.11**	**0.0389**	**0.047**

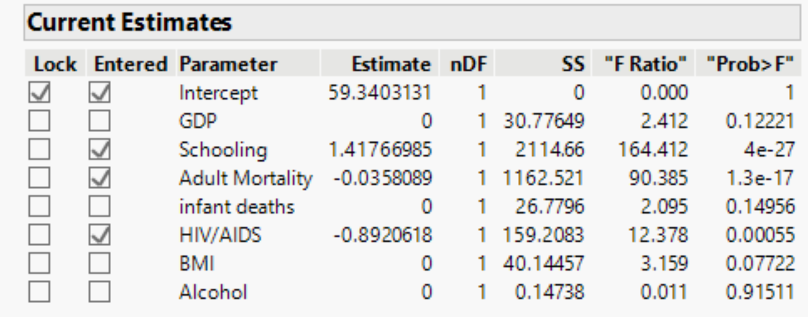
The hat threshold for the model is ~0.098 (2*7 variables/183 observations). Both observations retained due to inconsistent influence metrics and Cook's D values are below exclusion threshold (D > 1).

Methods

Model Selection

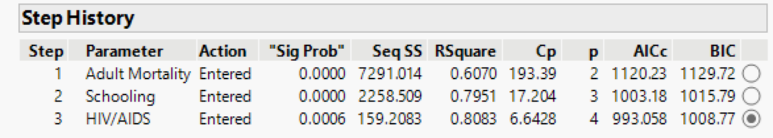
This report used stepwise regression, all possible models, and k-folds cross-validation within JMP Student Edition. Metrics compared: R-squared adjusted (R_a^2), Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC), Mallow's C_p, and Root Mean Square Error (RMSE). All tests used $\alpha = 0.05$.

Stepwise test 1 (Forward): All First Order terms



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	59.3403131	1	0	0.000	1
□	□	GDP	0	1	30.77649	2.412	0.12221
□	☑	Schooling	1.41766985	1	2114.66	164.412	4e-27
□	☑	Adult Mortality	-0.0358089	1	1162.521	90.385	1.3e-17
□	□	infant deaths	0	1	26.7796	2.095	0.14956
□	☑	HIV/AIDS	-0.8920618	1	159.2083	12.378	0.00055
□	□	BMI	0	1	40.14457	3.159	0.07722
□	□	Alcohol	0	1	0.14738	0.011	0.91511



Step History

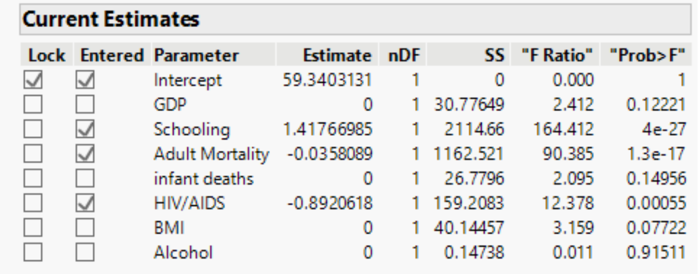
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Adult Mortality	Entered	0.0000	7291.014	0.6070	193.39	2	1120.23	1129.72	○
2	Schooling	Entered	0.0000	2258.509	0.7951	17.204	3	1003.18	1015.79	○
3	HIV/AIDS	Entered	0.0006	159.2083	0.8083	6.6428	4	993.058	1008.77	●

Model selected:

$$E(y) = \beta_0 + \beta_3 x_3 + \beta_2 x_2 + \beta_5 x_5$$

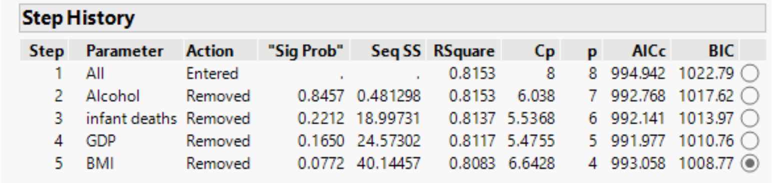
Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS

Stepwise test 2 (backward): All First Order terms



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	59.3403131	1	0	0.000	1
☐	☐	GDP	0	1	30.77649	2.412	0.12221
☐	☑	Schooling	1.41766985	1	2114.66	164.412	4e-27
☐	☑	Adult Mortality	-0.0358089	1	1162.521	90.385	1.3e-17
☐	☐	infant deaths	0	1	26.7796	2.095	0.14956
☐	☑	HIV/AIDS	-0.8920618	1	159.2083	12.378	0.00055
☐	☐	BMI	0	1	40.14457	3.159	0.07722
☐	☐	Alcohol	0	1	0.14738	0.011	0.91511



Step History

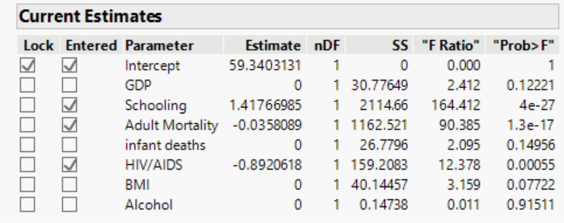
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	All	Entered	.	.	0.8153	8	8	994.942	1022.79	○
2	Alcohol	Removed	0.8457	0.481298	0.8153	6.038	7	992.768	1017.62	○
3	infant deaths	Removed	0.2212	18.99731	0.8137	5.5368	6	992.141	1013.97	○
4	GDP	Removed	0.1650	24.57302	0.8117	5.4755	5	991.977	1010.76	○
5	BMI	Removed	0.0772	40.14457	0.8083	6.6428	4	993.058	1008.77	●

Model selected:

$$E(y) = \beta_0 + \beta_3 x_3 + \beta_2 x_2 + \beta_5 x_5$$

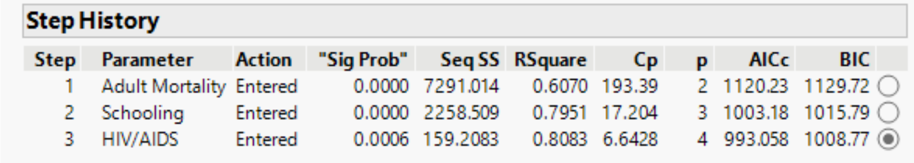
Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS

Stepwise test 3 (Mixed Stepwise): First Order Terms



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	59.3403131	1	0	0.000	1
☐	☐	GDP	0	1	30.77649	2.412	0.12221
☐	☑	Schooling	1.41766985	1	2114.66	164.412	4e-27
☐	☑	Adult Mortality	-0.0358089	1	1162.521	90.385	1.3e-17
☐	☐	infant deaths	0	1	26.7796	2.095	0.14956
☐	☑	HIV/AIDS	-0.8920618	1	159.2083	12.378	0.00055
☐	☐	BMI	0	1	40.14457	3.159	0.07722
☐	☐	Alcohol	0	1	0.14738	0.011	0.91511



Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Adult Mortality	Entered	0.0000	7291.014	0.6070	193.39	2	1120.23	1129.72	○
2	Schooling	Entered	0.0000	2258.509	0.7951	17.204	3	1003.18	1015.79	○
3	HIV/AIDS	Entered	0.0006	159.2083	0.8083	6.6428	4	993.058	1008.77	●

Model selected:

$$E(y) = \beta_0 + \beta_3 x_3 + \beta_2 x_2 + \beta_5 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS

All three stepwise methods (forward, backward, mixed) selected the same first- order model:

$$E(y) = \beta_0 + \beta_3 x_3 + \beta_2 x_2 + \beta_5 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS

All-Possible-Models testing for all First Order Predictors

Model Predictors	Adj R-sqr	RMSE	AICc	BIC	Cp
1	0.605	5.1066	1120.23	1129.72	193.3927
2	0.793	3.698	1003.18	1015.79	17.2038
3	0.805	3.5864	993.058	1008.77	6.6428
4	0.807	3.5649	991.977	1010.76	5.4755
5	0.808	3.5555	992.141	1013.97	5.5368
6	0.809	3.5504	992.768	1017.62	6.038
7	0.808	3.5602	994.942	1022.79	8

Models with more than 4 predictors exhibit diminishing returns in R_a^2 and less-preferable values of AICc and BIC. Although the models with more terms possess preferable C_p values closer to the number of terms in the model and have lower RMSE values, the lowest C_p value of the 4-predictor model indicates the best overall fit, and the lower BIC value coupled with diminishing improvements in R_a^2 signal that the added terms don't significantly improve the model. Also, the lowest AICc value indicates that the 4-predictor model performs the best when predicting Life Expectancy. Thus, the best first order model going forward is:

$$E(y) = \beta_0 + \beta_3 x_3 + \beta_2 x_2 + \beta_5 x_5 + \beta_6 x_6$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Stepwise and All Models Tests: Full Second Order Model.

Second-order and interaction terms were considered because the EDA revealed curvature in the relationships with GDP, Adult Mortality, and BMI.

Model being tested:

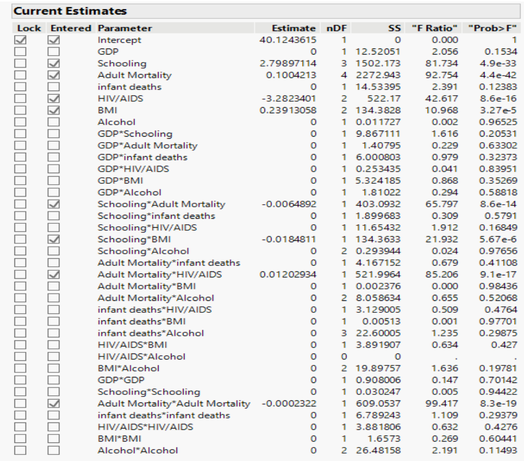
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

$$+\beta_8 x_1^2 + \beta_9 x_2^2 + \beta_{10} x_3^2 + \beta_{11} x_4^2 + \beta_{12} x_5^2 + \beta_{13} x_6^2 + \beta_{14} x_7^2$$

$$\begin{aligned}+\beta_{15} x_1 x_2 + \beta_{16} x_1 x_3 + \beta_{17} x_1 x_4 + \beta_{18} x_1 x_5 + \beta_{19} x_1 x_6 + \beta_{20} x_1 x_7 + \beta_{21} x_2 x_3 + \beta_{22} x_2 x_4 \\ + \beta_{23} x_2 x_5 + \beta_{24} x_2 x_6 + \beta_{25} x_2 x_7 + \beta_{26} x_3 x_4 + \beta_{27} x_3 x_5 + \beta_{28} x_3 x_6 + \beta_{29} x_3 x_7 \\ + \beta_{30} x_4 x_5 + \beta_{31} x_4 x_6 + \beta_{32} x_4 x_7 + \beta_{33} x_5 x_6 + \beta_{34} x_5 x_7 + \beta_{35} x_6 x_7\end{aligned}$$

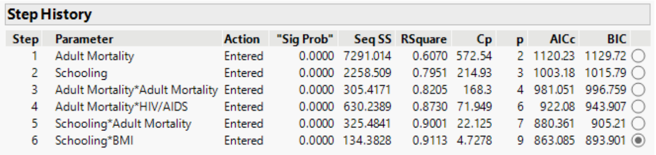
Where y = Life Expectancy, x_1 = GDP, x_2 = Schooling, x_3= Adult Mortality, x_4 = Infant Deaths, x_5 = HIV/AIDS, x_6 = BMI, x_7 = Alcohol

Forward Stepwise Regression:



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	40.1243615	1	0	0.000	1
□	□	GDP	0	1	12.52051	2.056	0.1534
□	☑	Schooling	2.79897114	3	1502.173	81.734	4.9e-33
□	☑	Adult Mortality	0.1004213	4	2272.943	92.754	4.4e-42
□	□	infant deaths	0	1	14.53395	2.391	0.12383
□	☑	HIV/AIDS	-3.2823401	2	522.17	42.617	8.6e-16
□	☑	BMI	0.23913058	2	134.3828	10.968	3.27e-5
□	□	Alcohol	0	1	0.011727	0.002	0.96525
□	□	GDP*Schooling	0	1	9.867111	1.616	0.20531
□	□	GDP*Adult Mortality	0	1	1.40795	0.229	0.63302
□	□	GDP*infant deaths	0	1	6.000803	0.979	0.32373
□	□	GDP*HIV/AIDS	0	1	0.253435	0.041	0.83951
□	□	GDP*BMI	0	1	5.324185	0.868	0.35269
□	□	GDP*Alcohol	0	1	1.81022	0.294	0.58818
□	☑	Schooling*Adult Mortality	-0.0064892	1	403.0932	65.797	8.6e-14
□	□	Schooling*infant deaths	0	1	1.899683	0.309	0.5791
□	□	Schooling*HIV/AIDS	0	1	11.65432	1.912	0.16849
□	☑	Schooling*BMI	-0.0184811	1	134.3633	21.932	5.67e-6
□	□	Schooling*Alcohol	0	2	0.293944	0.024	0.97656
□	□	Adult Mortality*infant deaths	0	1	4.167152	0.679	0.41108
□	☑	Adult Mortality*HIV/AIDS	0.01202934	1	521.9964	85.206	9.1e-17
□	□	Adult Mortality*BMI	0	1	0.002376	0.000	0.98436
□	□	Adult Mortality*Alcohol	0	2	8.058634	0.655	0.52068
□	□	infant deaths*HIV/AIDS	0	1	3.129005	0.509	0.4764
□	□	infant deaths*BMI	0	1	0.00513	0.001	0.97701
□	□	infant deaths*Alcohol	0	3	22.60005	1.235	0.29875
□	□	HIV/AIDS*BMI	0	1	3.891907	0.634	0.427
□	□	HIV/AIDS*Alcohol	0	0	0	.	.
□	□	BMI*Alcohol	0	2	19.89757	1.636	0.19781
□	□	GDP*GDP	0	1	0.908006	0.147	0.70142
□	□	Schooling*Schooling	0	1	0.030247	0.005	0.94422
□	☑	Adult Mortality*Adult Mortality	-0.0002322	1	609.0537	99.417	8.3e-19
□	□	infant deaths*infant deaths	0	1	6.789243	1.109	0.29379
□	□	HIV/AIDS*HIV/AIDS	0	1	3.881806	0.632	0.4276
□	□	BMI*BMI	0	1	1.6573	0.269	0.60441
□	□	Alcohol*Alcohol	0	2	26.48158	2.191	0.11493



Step History

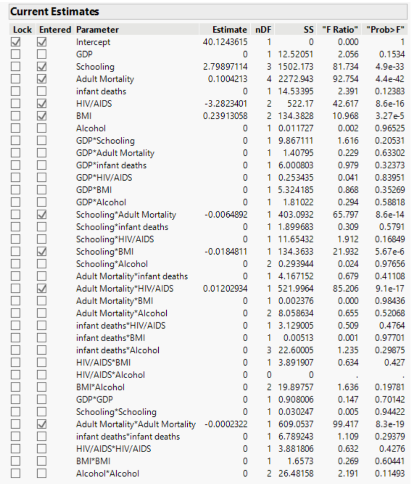
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Adult Mortality	Entered	0.0000	7291.014	0.6070	572.54	2	1120.23	1129.72	○
2	Schooling	Entered	0.0000	2258.509	0.7951	214.93	3	1003.18	1015.79	○
3	Adult Mortality*Adult Mortality	Entered	0.0000	305.4171	0.8205	168.3	4	981.051	996.759	○
4	Adult Mortality*HIV/AIDS	Entered	0.0000	630.2389	0.8730	71.949	6	922.08	943.907	○
5	Schooling*Adult Mortality	Entered	0.0000	325.4841	0.9001	22.125	7	880.361	905.21	○
6	Schooling*BMI	Entered	0.0000	134.3828	0.9113	4.7278	9	863.085	893.901	●

Model Selected:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

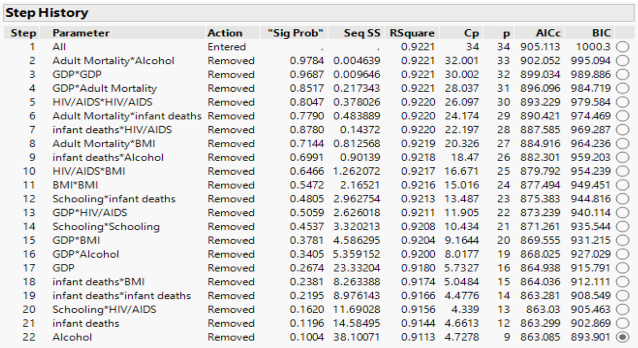
Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Backward Stepwise Regression:



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	40.1243615	1	0	0.000	1
□	□	GDP	0	1	12.52051	2.056	0.1534
□	☑	Schooling	2.79897114	3	1502.173	81.734	4.9e-33
□	☑	Adult Mortality	0.1004213	4	2272.943	92.754	4.4e-42
□	□	infant deaths	0	1	14.53395	2.391	0.12383
□	☑	HIV/AIDS	-3.2823401	2	522.17	42.617	8.6e-16
□	☑	BMI	0.23913058	2	134.3828	10.968	3.27e-5
□	□	Alcohol	0	1	0.011727	0.002	0.96525
□	□	GDP*Schooling	0	1	9.867111	1.616	0.20531
□	□	GDP*Adult Mortality	0	1	1.40795	0.229	0.63302
□	□	GDP*infant deaths	0	1	6.000803	0.979	0.32373
□	□	GDP*HIV/AIDS	0	1	0.253435	0.041	0.83951
□	□	GDP*BMI	0	1	5.324185	0.868	0.35269
□	□	GDP*Alcohol	0	1	1.81022	0.294	0.58818
□	☑	Schooling*Adult Mortality	-0.0064892	1	403.0932	65.797	8.6e-14
□	□	Schooling*infant deaths	0	1	1.899683	0.309	0.5791
□	□	Schooling*HIV/AIDS	0	1	11.65432	1.912	0.16849
□	☑	Schooling*BMI	-0.0184811	1	134.3633	21.932	5.67e-6
□	□	Schooling*Alcohol	0	2	0.293944	0.024	0.97656
□	□	Adult Mortality*infant deaths	0	1	4.167152	0.679	0.41108
□	☑	Adult Mortality*HIV/AIDS	0.01202934	1	521.9964	85.206	9.1e-17
□	□	Adult Mortality*BMI	0	1	0.002376	0.000	0.98436
□	□	Adult Mortality*Alcohol	0	2	8.058634	0.655	0.52068
□	□	infant deaths*HIV/AIDS	0	1	3.129005	0.509	0.4764
□	□	infant deaths*BMI	0	1	0.00513	0.001	0.97701
□	□	infant deaths*Alcohol	0	3	22.60005	1.235	0.29875
□	□	HIV/AIDS*BMI	0	1	3.891907	0.634	0.427
□	□	HIV/AIDS*Alcohol	0	0	0	.	.
□	□	BMI*Alcohol	0	2	19.89757	1.636	0.19781
□	□	GDP*GDP	0	1	0.908006	0.147	0.70142
□	□	Schooling*Schooling	0	1	0.030247	0.005	0.94422
□	☑	Adult Mortality*Adult Mortality	-0.0002322	1	609.0537	99.417	8.3e-19
□	□	infant deaths*infant deaths	0	1	6.789243	1.109	0.29379
□	□	HIV/AIDS*HIV/AIDS	0	1	3.881806	0.632	0.4276
□	□	BMI*BMI	0	1	1.6573	0.269	0.60441
□	□	Alcohol*Alcohol	0	2	26.48158	2.191	0.11493



Step History

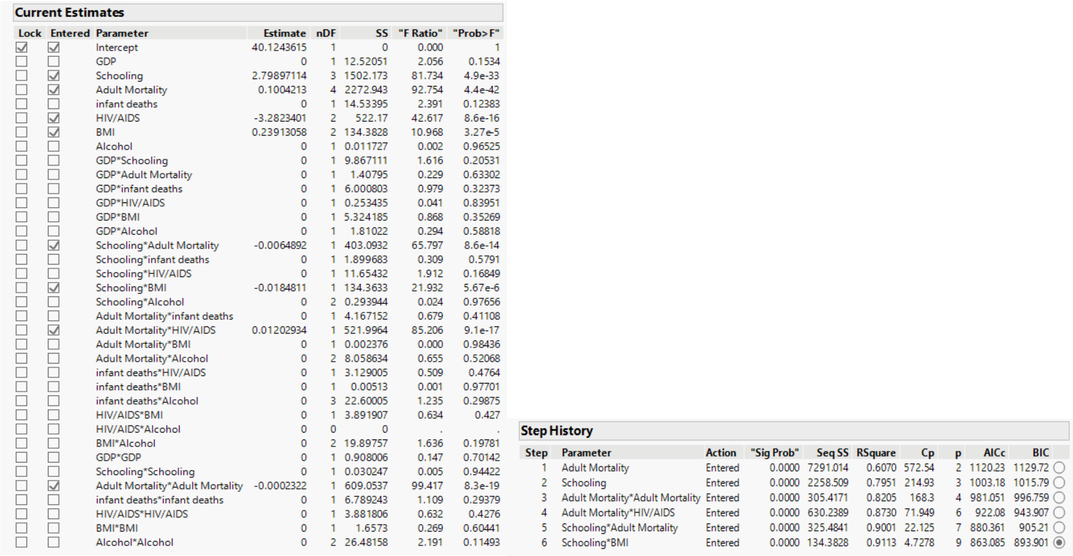
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	All	Entered	.	.	0.9221	34	34	905.113	1000.3	○
2	Adult Mortality*Alcohol	Removed	0.9784	0.004639	0.9221	32.001	33	902.052	995.094	○
3	GDP*GDP	Removed	0.9687	0.009646	0.9221	30.002	32	899.034	989.886	○
4	GDP*Adult Mortality	Removed	0.8517	0.217343	0.9221	28.037	31	896.096	984.719	○
5	HIV/AIDS*HIV/AIDS	Removed	0.8047	0.378026	0.9220	26.097	30	893.229	979.584	○
6	Adult Mortality*infant deaths	Removed	0.7790	0.483889	0.9220	24.174	29	890.421	974.469	○
7	infant deaths*HIV/AIDS	Removed	0.8780	0.14372	0.9220	22.197	28	887.585	969.287	○
8	Adult Mortality*BMI	Removed	0.7144	0.812568	0.9219	20.326	27	884.916	964.236	○
9	infant deaths*Alcohol	Removed	0.6991	0.90139	0.9218	18.47	26	882.301	959.203	○
10	HIV/AIDS*BMI	Removed	0.6466	1.262072	0.9217	16.671	25	879.792	954.239	○
11	BMI*BMI	Removed	0.5472	2.16521	0.9216	15.016	24	877.494	949.451	○
12	Schooling*infant deaths	Removed	0.4805	2.962754	0.9213	13.487	23	875.383	944.816	○
13	GDP*HIV/AIDS	Removed	0.5059	2.626018	0.9211	11.905	22	873.239	940.114	○
14	Schooling*Schooling	Removed	0.4537	3.320213	0.9208	10.434	21	871.261	935.544	○
15	GDP*BMI	Removed	0.3781	4.586295	0.9204	9.1644	20	869.555	931.215	○
16	GDP*Alcohol	Removed	0.3405	5.359152	0.9200	8.0177	19	868.025	927.029	○
17	GDP	Removed	0.2674	23.33204	0.9180	5.7327	16	864.938	915.791	○
18	infant deaths*BMI	Removed	0.2381	8.263388	0.9174	5.0484	15	864.036	912.111	○
19	infant deaths*infant deaths	Removed	0.2195	8.976143	0.9166	4.4776	14	863.281	908.549	○
20	Schooling*HIV/AIDS	Removed	0.1620	11.69028	0.9156	4.339	13	863.03	905.463	○
21	infant deaths	Removed	0.1196	14.58495	0.9144	4.6613	12	863.299	902.869	○
22	Alcohol	Removed	0.1004	38.10071	0.9113	4.7278	9	863.085	893.901	●

Model Selected:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Mixed Stepwise Regression:



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	40.1243615	1	0	0.000	1
□	□	GDP	0	1	12.52051	2.056	0.1534
□	☑	Schooling	2.79897114	3	1502.173	81.734	4.9e-33
□	☑	Adult Mortality	0.1004213	4	2272.943	92.754	4.4e-42
□	□	infant deaths	0	1	14.53395	2.391	0.12383
□	☑	HIV/AIDS	-3.2823401	2	522.17	42.617	8.6e-16
□	☑	BMI	0.23913058	2	134.3828	10.968	3.27e-5
□	□	Alcohol	0	1	0.011727	0.002	0.96525
□	□	GDP*Schooling	0	1	9.867111	1.616	0.20531
□	□	GDP*Adult Mortality	0	1	1.40795	0.229	0.63302
□	□	GDP*infant deaths	0	1	6.000803	0.979	0.32373
□	□	GDP*HIV/AIDS	0	1	0.253435	0.041	0.83951
□	□	GDP*BMI	0	1	5.324185	0.868	0.35269
□	□	GDP*Alcohol	0	1	1.81022	0.294	0.58818
□	☑	Schooling*Adult Mortality	-0.0064892	1	403.0932	65.797	8.6e-14
□	□	Schooling*infant deaths	0	1	1.899683	0.309	0.5791
□	□	Schooling*HIV/AIDS	0	1	11.65432	1.912	0.16849
□	☑	Schooling*BMI	-0.0184811	1	134.3633	21.932	5.67e-6
□	□	Schooling*Alcohol	0	2	0.293944	0.024	0.97656
□	□	Adult Mortality*infant deaths	0	1	4.167152	0.679	0.41108
□	☑	Adult Mortality*HIV/AIDS	0.01202934	1	521.9964	85.206	9.1e-17
□	□	Adult Mortality*BMI	0	1	0.002376	0.000	0.98436
□	□	Adult Mortality*Alcohol	0	2	8.058634	0.655	0.52068
□	□	infant deaths*HIV/AIDS	0	1	3.129005	0.509	0.4764
□	□	infant deaths*BMI	0	1	0.00513	0.001	0.97701
□	□	infant deaths*Alcohol	0	3	22.60005	1.235	0.29875
□	□	HIV/AIDS*BMI	0	1	3.891907	0.634	0.427
□	□	HIV/AIDS*Alcohol	0	0	0	.	.
□	□	BMI*Alcohol	0	2	19.89757	1.636	0.19781
□	□	GDP*GDP	0	1	0.908006	0.147	0.70142
□	□	Schooling*Schooling	0	1	0.030247	0.005	0.94422
□	☑	Adult Mortality*Adult Mortality	-0.0002322	1	609.0537	99.417	8.3e-19
□	□	infant deaths*infant deaths	0	1	6.789243	1.109	0.29379
□	□	HIV/AIDS*HIV/AIDS	0	1	3.881806	0.632	0.4276
□	□	BMI*BMI	0	1	1.6573	0.269	0.60441
□	□	Alcohol*Alcohol	0	2	26.48158	2.191	0.11493

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Adult Mortality	Entered	0.0000	7291.014	0.6070	572.54	2	1120.23	1129.72	○
2	Schooling	Entered	0.0000	2258.509	0.7951	214.93	3	1003.18	1015.79	○
3	Adult Mortality*Adult Mortality	Entered	0.0000	305.4171	0.8205	168.3	4	981.051	996.759	○
4	Adult Mortality*HIV/AIDS	Entered	0.0000	630.2389	0.8730	71.949	6	922.08	943.907	○
5	Schooling*Adult Mortality	Entered	0.0000	325.4841	0.9001	22.125	7	880.361	905.21	○
6	Schooling*BMI	Entered	0.0000	134.3828	0.9113	4.7278	9	863.085	893.901	●

Model Selected:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Like the stepwise testing for the first order terms, all 3 stepwise tests for the second order terms and interactions yielded the same result as a recommended model:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

All Possible Models: Full Second order model

Model Predictors	Adj R-Sqr	RMSE	AICc	BIC	Cp
1	0.604856	4.91	1105.86	1115.36	521.5139
2	0.793	3.4616	979.004	991.617	169.2511
3	0.805	3.3736	970.68	986.388	152.0538
4	0.807	2.908	917.434	936.213	68.647
5	0.822	2.8765	914.58	936.406	64.1129
6	0.841	2.6116	880.361	905.21	23.6981
7	0.855	2.5729	876.068	903.913	18.9678
8	0.907	2.4751	863.085	893.901	6.125
9	0.907	2.4687	863.338	897.099	6.2541
10	0.907	2.4729	865.196	901.874	7.8534
11	0.907	2.4766	867.001	906.571	9.3806
12	0.907	2.4818	869.03	911.463	11.089

13	0.906	2.4886	871.333	916.601	13.023
14	0.905	2.4958	873.71	921.785	15

All Possible Models Testing indicates the model with 8 predictors has the best overall metrics for selection: High R_a^2, second lowest RSME, lowest both AICc and BIC, and low C_p close to the number of predictors. Conveniently, the same 8 predictors in all-possible-models testing and stepwise testing were selected, thus the model going forward is:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where, y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

K-Folds Cross Validation: Full Second-Order Model

The full second-order model was tested:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

$$+\beta_8 x_1^2 + \beta_9 x_2^2 + \beta_{10} x_3^2 + \beta_{11} x_4^2 + \beta_{12} x_5^2 + \beta_{13} x_6^2 + \beta_{14} x_7^2$$

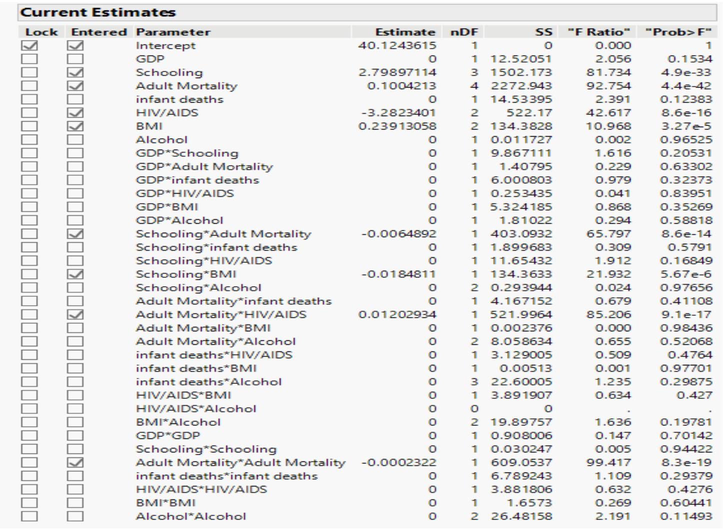
$$+\beta_{15} x_1 x_2 + \beta_{16} x_1 x_3 + \beta_{17} x_1 x_4 + \beta_{18} x_1 x_5 + \beta_{19} x_1 x_6 + \beta_{20} x_1 x_7 + \beta_{21} x_2 x_3 + \beta_{22} x_2 x_4$$

$$+ \beta_{23} x_2 x_5 + \beta_{24} x_2 x_6 + \beta_{25} x_2 x_7 + \beta_{26} x_3 x_4 + \beta_{27} x_3 x_5 + \beta_{28} x_3 x_6 + \beta_{29} x_3 x_7$$

$$+ \beta_{30} x_4 x_5 + \beta_{31} x_4 x_6 + \beta_{32} x_4 x_7 + \beta_{33} x_5 x_6 + \beta_{34} x_5 x_7 + \beta_{35} x_6 x_7$$

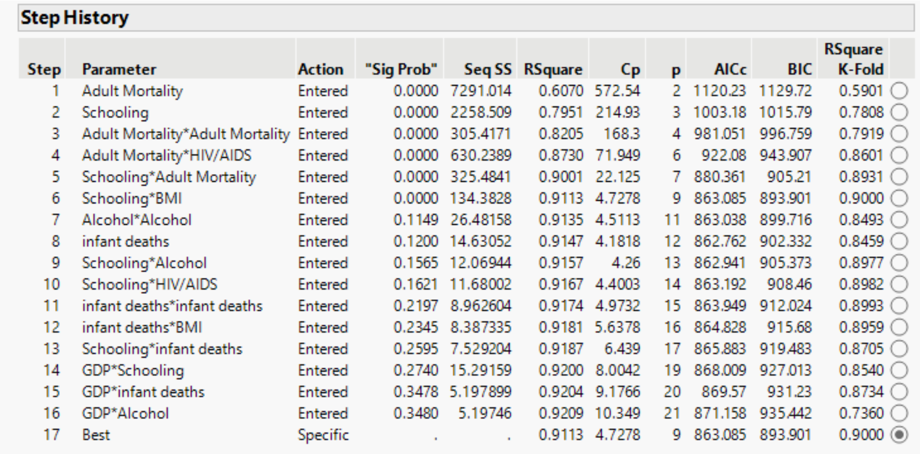
Where y = Life Expectancy, x_1 = GDP, x_2 = Schooling, x_3= Adult Mortality, x_4 = Infant Deaths, x_5 = HIV/AIDS, x_6 = BMI, x_7 = Alcohol

The results:



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	40.1243615	1	0	0.000	1
□	□	GDP	0	1	12.52051	2.056	0.1534
□	☑	Schooling	2.79897114	3	1502.173	81.734	4.9e-33
□	☑	Adult Mortality	0.1004213	4	2272.943	92.754	4.4e-42
□	□	infant deaths	0	1	14.53395	2.391	0.12383
□	☑	HIV/AIDS	-3.2823401	2	522.17	42.617	8.6e-16
□	☑	BMI	0.23913058	2	134.3828	10.968	3.27e-5
□	□	Alcohol	0	1	0.011727	0.002	0.96525
□	□	GDP*Schooling	0	1	9.867111	1.616	0.20531
□	□	GDP*Adult Mortality	0	1	1.40795	0.229	0.63302
□	□	GDP*infant deaths	0	1	6.000803	0.979	0.32373
□	□	GDP*HIV/AIDS	0	1	0.253435	0.041	0.83951
□	□	GDP*BMI	0	1	5.324185	0.868	0.35269
□	□	GDP*Alcohol	0	1	1.81022	0.294	0.58818
□	☑	Schooling*Adult Mortality	-0.0064892	1	403.0932	65.797	8.6e-14
□	□	Schooling*infant deaths	0	1	1.899683	0.309	0.5791
□	□	Schooling*HIV/AIDS	0	1	11.65432	1.912	0.16849
□	☑	Schooling*BMI	-0.0184811	1	134.3633	21.932	5.67e-6
□	□	Schooling*Alcohol	0	2	0.293944	0.024	0.97656
□	□	Adult Mortality*infant deaths	0	1	4.167152	0.679	0.41108
□	☑	Adult Mortality*HIV/AIDS	0.01202934	1	521.9964	85.206	9.1e-17
□	□	Adult Mortality*BMI	0	1	0.002376	0.000	0.98436
□	□	Adult Mortality*Alcohol	0	2	8.058634	0.655	0.52068
□	□	infant deaths*HIV/AIDS	0	1	3.129005	0.509	0.4764
□	□	infant deaths*BMI	0	1	0.00513	0.001	0.97701
□	□	infant deaths*Alcohol	0	3	22.60005	1.235	0.29875
□	□	HIV/AIDS*BMI	0	1	3.891907	0.634	0.427
□	□	HIV/AIDS*Alcohol	0	0	0	.	.
□	□	BMI*Alcohol	0	2	19.89757	1.636	0.19781
□	□	GDP*GDP	0	1	0.908006	0.147	0.70142
□	□	Schooling*Schooling	0	1	0.030247	0.005	0.94422
□	☑	Adult Mortality*Adult Mortality	-0.0002322	1	609.0537	99.417	8.3e-19
□	□	infant deaths*infant deaths	0	1	6.789243	1.109	0.29379
□	□	HIV/AIDS*HIV/AIDS	0	1	3.881806	0.632	0.4276
□	□	BMI*BMI	0	1	1.6573	0.269	0.60441
□	□	Alcohol*Alcohol	0	2	26.48158	2.191	0.11493



Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	RSquare K-Fold	
1	Adult Mortality	Entered	0.0000	7291.014	0.6070	572.54	2	1120.23	1129.72	0.5901	○
2	Schooling	Entered	0.0000	2258.509	0.7951	214.93	3	1003.18	1015.79	0.7808	○
3	Adult Mortality*Adult Mortality	Entered	0.0000	305.4171	0.8205	168.3	4	981.051	996.759	0.7919	○
4	Adult Mortality*HIV/AIDS	Entered	0.0000	630.2389	0.8730	71.949	6	922.08	943.907	0.8601	○
5	Schooling*Adult Mortality	Entered	0.0000	325.4841	0.9001	22.125	7	880.361	905.21	0.8931	○
6	Schooling*BMI	Entered	0.0000	134.3828	0.9113	4.7278	9	863.085	893.901	0.9000	○
7	Alcohol*Alcohol	Entered	0.1149	26.48158	0.9135	4.5113	11	863.038	899.716	0.8493	○
8	infant deaths	Entered	0.1200	14.63052	0.9147	4.1818	12	862.762	902.332	0.8459	○
9	Schooling*Alcohol	Entered	0.1565	12.06944	0.9157	4.26	13	862.941	905.373	0.8977	○
10	Schooling*HIV/AIDS	Entered	0.1621	11.68002	0.9167	4.4003	14	863.192	908.46	0.8982	○
11	infant deaths*infant deaths	Entered	0.2197	8.962604	0.9174	4.9732	15	863.949	912.024	0.8993	○
12	infant deaths*BMI	Entered	0.2345	8.387335	0.9181	5.6378	16	864.828	915.68	0.8959	○
13	Schooling*infant deaths	Entered	0.2595	7.529204	0.9187	6.439	17	865.883	919.483	0.8705	○
14	GDP*Schooling	Entered	0.2740	15.29159	0.9200	8.0042	19	868.009	927.013	0.8540	○
15	GDP*infant deaths	Entered	0.3478	5.197899	0.9204	9.1766	20	869.57	931.23	0.8734	○
16	GDP*Alcohol	Entered	0.3480	5.19746	0.9209	10.349	21	871.158	935.442	0.7360	○
17	Best	Specific	.	.	0.9113	4.7278	9	863.085	893.901	0.9000	●

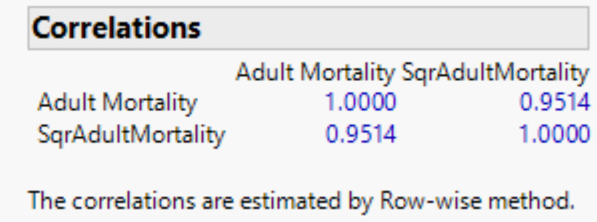
After including the full second order model in k-folds cross validation test, the same second-order model that was selected by both the stepwise and all-possible-models tests has been selected to best represent the data for prediction. This model is identified as:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Comparing the difference between the training and validation R^2 values yield $0.9113 - 0.9000 = 0.0113$, indicating good predictive power.

Multicollinearity (x_3^2)



Correlations

	Adult Mortality	SqrAdultMortality
Adult Mortality	1.0000	0.9514
SqrAdultMortality	0.9514	1.0000

The correlations are estimated by Row-wise method.

Although the second-order predictor has high multicollinearity with its first-order counterpart, no values were coded due to high significance of β estimates.

Analysis: Selected Model

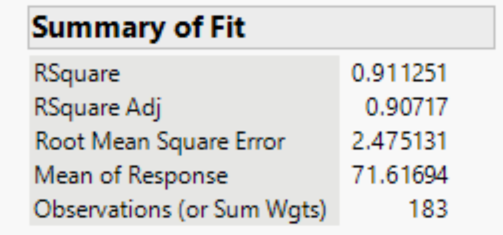
Model 1: Selected Second-Order Model without Transformation

The model being tested is:

$$E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Testing Fit

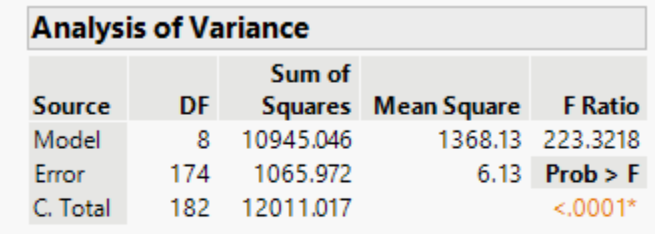


Summary of Fit

RSquare	0.911251
RSquare Adj	0.90717
Root Mean Square Error	2.475131
Mean of Response	71.61694
Observations (or Sum Wgts)	183

R_a^2 = 0.90717. Approximately 90.7% of variance in the data explained by the model. Overall strong fit.

Testing Model Utility



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	10945.046	1368.13	223.3218
Error	174	1065.972	6.13	**Prob > F**
C. Total	182	12011.017		<.0001*

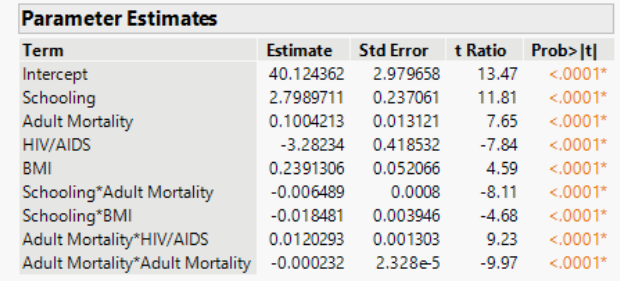
$H_o\!: \beta_i = 0, where\ i = 2, 3, 5, 6, 10, 21, 24, 27$

$H_a\!: at\ least\ one\ \beta_i \neq 0, where\ i = 2, 3, 5, 6, 10, 21, 24, 27$

$\alpha = 0.05$

F-Ratio of 223.3218 and p-value < 0.001< α, we reject the null hypothesis. At least one $\beta_i, where\ i = 2, 3, 5, 6, 10, 21, 24, 27$ does not equal zero. This model has overall good model utility.

Beta Significance Testing



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	40.124362	2.979658	13.47	<.0001*
Schooling	2.7989711	0.237061	11.81	<.0001*
Adult Mortality	0.1004213	0.013121	7.65	<.0001*
HIV/AIDS	-3.28234	0.418532	-7.84	<.0001*
BMI	0.2391306	0.052066	4.59	<.0001*
Schooling*Adult Mortality	-0.006489	0.0008	-8.11	<.0001*
Schooling*BMI	-0.018481	0.003946	-4.68	<.0001*
Adult Mortality*HIV/AIDS	0.0120293	0.001303	9.23	<.0001*
Adult Mortality*Adult Mortality	-0.000232	2.328e-5	-9.97	<.0001*

All β estimates are significant at any reasonable α level. Prediction equation:

$$\hat{y} = 40.124 + 2.799x_2 + 0.1x_3 - 3.282x_5 + 0.239x_6 - 0.0002x_3^2 - 0.006x_2x_3 - 0.018x_2x_6 + 0.012x_3x_5$$

Testing for Outliers

Country	**Residual Life expectancy**
Zimbabwe	6.7134184127
Sierra Leone	5.4789374768
Eritrea	5.7007546432
Haiti	-4.997859978
Democratic Republic of the Congo	-5.252433352

Outlier threshold: ±7.44 (3*RMSE). No outliers found.

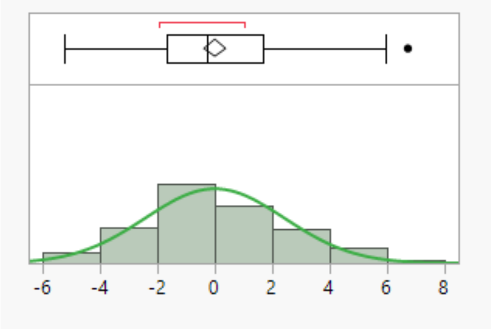
Testing for Influential Observations

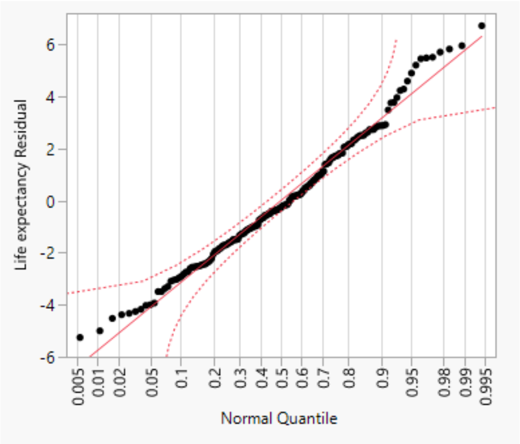
Influence thresholds: Hat = 0.0984, Cook's D = 0.19 for identification (D > 1 for exclusion)

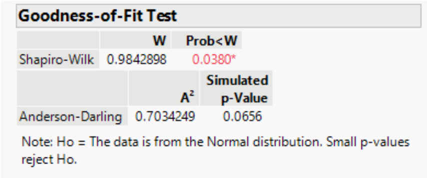
Country	**External Studentized Residuals**	**Hat Values**	**Cook's D**
Zimbabwe	2.973	0.13	0.14
Sierra Leone	2.82	0.36	0.478
Eritrea	2.543	0.154	0.127
Haiti	-2.14	0.092	0.05
D.R Congo	-2.18	0.032	0.017

No observations fail all three tests or exceed Cook's D exclusion threshold. All observations will be used going forward.

Testing Normal Distribution Assumptions







Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9842898	0.0380*

	A^2	Simulated p-Value
Anderson-Darling	0.7034249	0.0656

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

H_o: Normal Distribution

H_a: Distribution is not normal

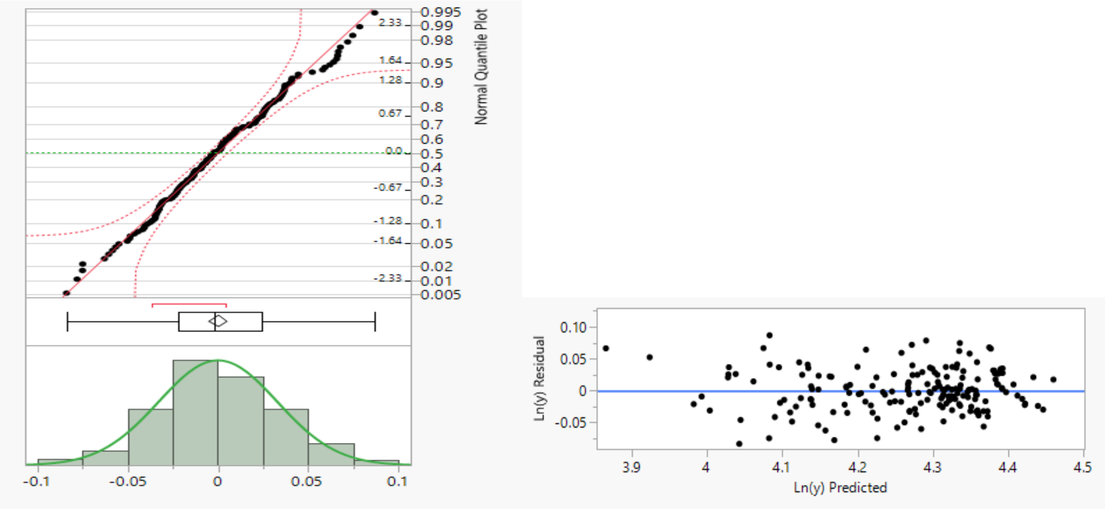
$\alpha = 0.05$

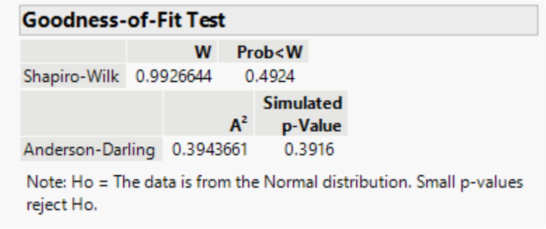
Shapiro-Wilk p-value = 0.038 < 0.05. Normality assumption not satisfied. Residual distribution shows skewed form. Possible transformation needed.

Applying Transformations to the Response Variable

The lack of significant normality in the previous model indicates that a transformation may need to be applied to the response variable to achieve normality in the model. Three transformations were tested: ln(y), $\sqrt{(y)}$, and Box Cox ($\lambda = 0.242$).

Model 2: ln(y) Transformation Distribution





Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9926644	0.4924

	A^2	Simulated p-Value
Anderson-Darling	0.3943661	0.3916

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

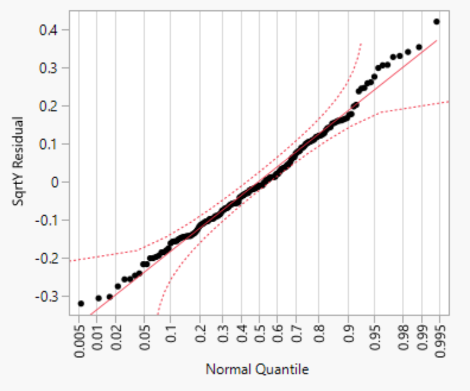
H_o: Normal Distribution

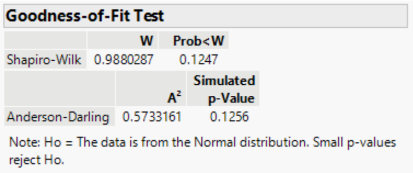
H_a: Distribution is not normal

$\alpha = 0.05$

All residual plots appear normal. Shapiro-Wilk p-value = 0.4924 > 0.05. We fail to reject null hypothesis. Normal residual distribution achieved through natural log transformation.

Model 3: $\sqrt(y)$ Transformation Distribution





Goodness-of-Fit Test

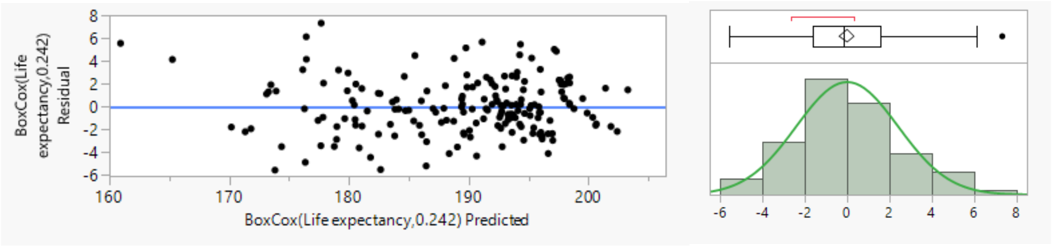
	W	Prob<W
Shapiro-Wilk	0.9880287	0.1247

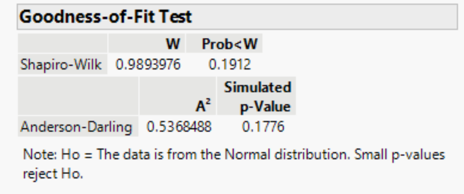
	A^2	Simulated p-Value
Anderson-Darling	0.5733161	0.1256

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Same outcome for $\sqrt(y)$ transformation.

Model 4: Box Cox Transformation ($\lambda = 0.242$)





Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9893976	0.1912

	A^2	Simulated p-Value
Anderson-Darling	0.5368488	0.1776

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

The residual vs predicted, residual histogram, and Shapiro-Wilk test all indicate a normal distribution for the Box Cox transformation.

Results: Final Model Selection

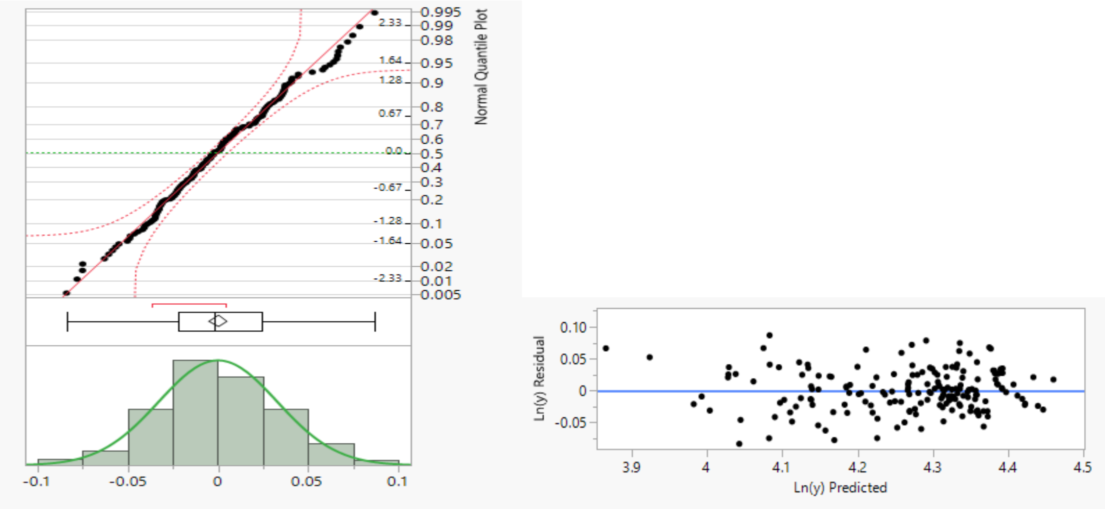
Model	Adj R-sqr	F-Ratio	Prob>F	Shapiro-Wilk W	Prob < W
1 (Untransformed)	0.90717	223.3218	<0.0001	0.9843	0.038
2 (ln(y))	0.92	260.5119	<0.0001	0.993	0.4924
3 (sqrt(y))	0.911	234.6943	<0.0001	0.988	0.1247
4(Box Cox, lambda = 0.242	0.913	240.1232	<0.0001	0.989	0.1912

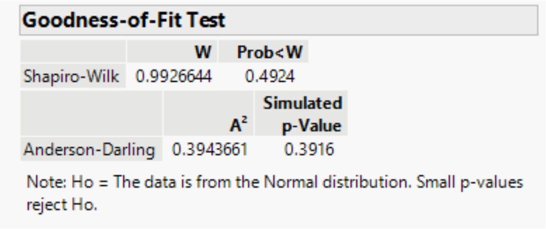
The ln(y) transformation yields the most significant results. It explains the most variance in the data, has good model utility, and has the strongest evidence of normal residual distribution as indicated in the previous section. Thus, the final model used in this data to explain Life Expectancy is:

$$E(\ln(y)) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6 + \beta_{10} x_3^2 + \beta_{21} x_2 x_3 + \beta_{24} x_2 x_6 + \beta_{27} x_3 x_5$$

Where y = Life Expectancy, x_2 = Schooling, x_3= Adult Mortality, x_5 = HIV/AIDS, x_6 = BMI

Confirming Normality in Final Model





Goodness-of-Fit Test

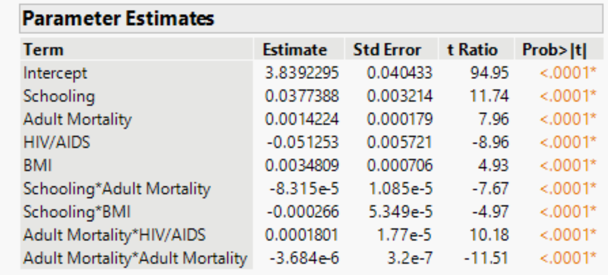
	W	Prob<W
Shapiro-Wilk	0.9926644	0.4924

	A^2	Simulated p-Value
Anderson-Darling	0.3943661	0.3916

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

All distribution assessments yield strong indications of normal residual distribution in the final model. The Q-Q plot shows even distribution along the center line, the residual vs predicted plot shows a negation of the fan-shaped distribution in the untransformed model, and the Shapiro-Wilk test resulted in a significant p-value at any reasonable confidence level, indicating evidence of normality.

Prediction Equation: Final Model



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.8392295	0.040433	94.95	<.0001*
Schooling	0.0377388	0.003214	11.74	<.0001*
Adult Mortality	0.0014224	0.000179	7.96	<.0001*
HIV/AIDS	-0.051253	0.005721	-8.96	<.0001*
BMI	0.0034809	0.000706	4.93	<.0001*
Schooling*Adult Mortality	-8.315e-5	1.085e-5	-7.67	<.0001*
Schooling*BMI	-0.000266	5.349e-5	-4.97	<.0001*
Adult Mortality*HIV/AIDS	0.0001801	1.77e-5	10.18	<.0001*
Adult Mortality*Adult Mortality	-3.684e-6	3.2e-7	-11.51	<.0001*

$$\ln(\hat{y}) = 3.83 + 0.038x_2 + 0.0014x_3 - 0.051x_5 + 0.0034x_6 - 0.0000037x_3^2 \\ - 0.000083x_2x_3 - 0.00027x_2x_6 + 0.00018x_3x_5$$

Research Questions Answered

1. Which factors are most strongly associated with life expectancy across countries?

Our final model shows that **Schooling**, **Adult Mortality**, **HIV/AIDS prevalence**, and **BMI** are the primary predictors that best explain life expectancy. Schooling has a strong positive association with longevity, while Adult Mortality and HIV/AIDS prevalence have strong negative associations. BMI maintains a smaller but significant positive effect. These four predictors consistently appeared in all model-selection procedures and remained statistically significant in the final ln(y) model, indicating that they explain most of the variation in global life expectancy.

2. Do higher-order or interaction effects improve our ability to explain life expectancy?

Yes. Including **one quadratic term** (Adult Mortality^2) and **three interaction terms** (Schooling×Adult Mortality, Schooling×BMI, and Adult Mortality×HIV/AIDS) significantly improved model fit. These additions captured important curvilinear and interactions that first-order models could not account for. As a result, R_a^2 increased to **0.92**, and residual normality was greatly improved after applying the ln(y) transformation.

3. Which model best explains variation in life expectancy?

After comparing first-order, second-order, and transformed models, the **ln(y) second-order model** performed best. It achieved the highest adjusted R^2, strong model utility (F-ratio), and the best evidence of normal residual distribution (Shapiro-Wilk p = 0.4924). This model demonstrated both high explanatory power and predictive validity, with a training–validation R^2 difference of only **0.0113** in k-fold cross-validation.

4. How well does the final model satisfy regression assumptions?

The final ln(y) model satisfies all major regression assumptions:

Linearity/functional form: Improved through inclusion of quadratic and interaction terms.

Normality: Satisfied after transformation (Shapiro-Wilk p = 0.4924).

Homoscedasticity: Residual vs. predicted plot showed no systematic pattern.

Influence/outliers: No observations exceeded all influence thresholds; Sierra Leone showed some influence but was not sufficient for removal.

Predictive stability: K-fold validation confirmed strong predictive power.

Conclusion

Overall, the model meets the necessary statistical assumptions and provides a reliable explanation of life expectancy.

This analysis successfully identified a statistically significant regression model explaining life expectancy across 183 nations. The final model with ln(y) transformation explains 92% of variance in life expectancy through schooling, adult mortality, HIV/AIDS prevalence, BMI, and their interactions. The model satisfies all regression assumptions including normality of residuals and demonstrates strong predictive accuracy (training-validation R^2 difference = 0.0113). Key predictors (schooling, adult mortality, and HIV/AIDS) emerged as consistently significant across all selection methods, confirming their crucial role in determining life expectancy worldwide.

References

Mendenhall, W., & Sincich, T. (2016). *A Second Course in Statistics: Regression Analysis* (8th ed.). Pearson.

World Health Organization. (2015). *WHO Life Expectancy Dataset*. Retrieved from https://www.who.int/data