Exploring Factors Affecting Life Expectancy Using Linear Techniques

Spencer Slaton, Leon Weingartner

***Abstract*—This research paper aims to explore the factors that affect life expectancy using linear models. The study utilizes a real-world dataset to derive insights and conclusions based on the techniques learned in class. The analysis goes beyond interpreting coefficient results from a multilinear regression model and provides an in-depth study of the factors that impact life expectancy. We aim to determine not only the direction of influence the covariates have on life expectancy, but also their magnitude; this includes establishing robust confidence intervals for the estimated coefficients.**

I. Introduction

This study begins by exploring the data using descriptive statistics and visualization techniques. Then, various linear models are developed to examine the relationship between life expectancy and several independent variables. The models are evaluated using diagnostic tools such as residual analysis and Cook's distance to ensure their validity and reliability.

In addition to traditional linear models, the study also employs more advanced techniques such as heteroskedastic coefficient estimation, interaction effects, and regularization methods to gain deeper insights into the factors that affect life expectancy. Can linear modeling techniques be used to adequately describe the influence of a complex web of factors on life expectancy across the incredibly heterogenous set of countries, with appropriate care given to statistical rigor? The answer, with a few small caveats, turns out to be yes. The results of the analysis provide a comprehensive understanding of the complex relationship between life expectancy and various socio-economic, demographic, and health-related factors. We aim to highlight adaptable characteristics on a national level to potentially provide valuable insights for policymakers and public health practitioners.

II. The Model

A. Data

We use a dataset that is a combination of life expectancy data provided by GHO (Global Health Observatory) and UNESCO (United Nations Educational Scientific and Culture Organization) *[1]*. The data has life expectancy information about every country for every year from 2000 until 2015. This is our target variable *life_expect*, the life expectancy at birth in years. We are not treating this as a time-series problem, so we will only use a single year for each country, known as that countries' *candidate observation*, leaving us with 183 observations in total. Because the countries are considered to be independent from each other we are able to select candidate observations from different years for different countries. In other words, each potential country and year are considered to be samples from an unknown joint distribution of the covariates, and while these observations are strongly correlated across years when belonging to the same country, they are assumed to be independent across countries. The candidate observation for a given country was taken to be the year for which the fewest number of features have missing values. If there are still missing values in this year, we impute based on the value of the temporally nearest non-missing entry for that country. The resulting set of 183 observations only has missing values for a feature when there is no recorded data for that feature for any year in our dataset. The number of missing values at this stage is low, with the worst features only having 20-30% missing values, and most having less than 4%. We impute the remaining missing values using K-nearest neighbors with parameter k = 2.

There are a few obviously correlated features. For example, *life_exp60*, the life expectancy after the age of 60, and *infant_mort*, the infant mortality rate, are very closely correlated with the target and are therefore dropped. There are also a few features that measure virtually the same thing such as *uni_gni*, the gross national income from the UNESCO source, and *gni_capita*, the gross national income per capita from the GHO source. In this case we only keep the latter as a feature. After this there are around 27 potential predictors, listed below:

- *alcohol*, Recorded alcohol consumption per capita among ages 15+ (in liters of pure alcohol)
- *bmi*, Age-standardized estimate of mean BMI (body mass index) among ages 18+
- *age5*-19thinness, Prevalence of thinness among children and adolescents, BMI < (median - 2 s.d.) (%)
- *age5-19obesity*, Prevalence of obesity among children and adolescents, BMI > (median - 2 s.d.) (%)

- *hepatitis*, Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- *measles*, Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds (%)
- *polio*, Polio (Pol3) immunization coverage among 1-year-olds (%)
- *diphtheria*, Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- *diseases*, An engineered feature, the average of polio, measles, and diphtheria
- *basic_water*, Population using at least basic drinking-water services (%)
- *doctors*, Medical doctors (per 10,000)
- *hospitals*, Total density per 100,000 population
- *gghe-d*, Domestic general government health expenditure as percentage of gross domestic product (GDP)
- *che_gdp*, Current health expenditure (CHE) as percentage of gross domestic product (GDP) (%)
- *une_pop*, Population (thousands)
- *une_hiv*, Prevalence of HIV, total (% of population ages 15-49)
- *une_gni*, GNI per capita, PPP (current international $)
- *une_poverty*, Poverty headcount ratio at $1.90 a day (PPP) (%)
- *une_edu_spend*, Government expenditure on education as a percentage of GDP (%)
- *une_literacy*, Adult literacy rate, population 15+ years, both sexes (%)
- *une_school*, Mean years of schooling (ISCED 1 or higher), population 25+ years, both sexes
- *region_X*, 6 binary columns indicating the regions [Africa, Americas, Eastern Mediterranean, Europe, South East Asia, Western Pacific]

After inspecting the distributions and correlations of our feature variables to understand which features should be transformed logarithmically, we decide to replace four features with their log-transformed versions:

- *log_une_poverty*
- *log_une_gni*
- *log_une_hiv*
- *log_une_pop*

B. Final Model

A summary of our final model is shown to the left. It is an OLS regression onto a constant term and 10 predictors, all which have statistically significant coefficients. The model has R-squared 0.896 and AIC 878.7.

The model was found by performing forward variable selection on the processed dataset. The selection criterion was AIC. Selecting based on BIC was considered as well, but the resulting model was virtually identical to the AIC-selected model, this being due to our relatively small n = 183.

The model was chosen in contrast to other candidate models such as models found via backward selection, the LASSO, and models containing various second-order interaction terms mainly due to its simplicity. Its 10 predictors are considerably fewer than other models and it boasts an impressive R-squared. We will expound upon the other candidate models in the *Additional Work* second.

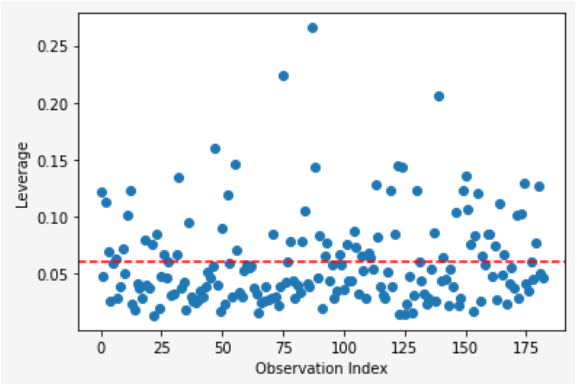
OLS Regression Results

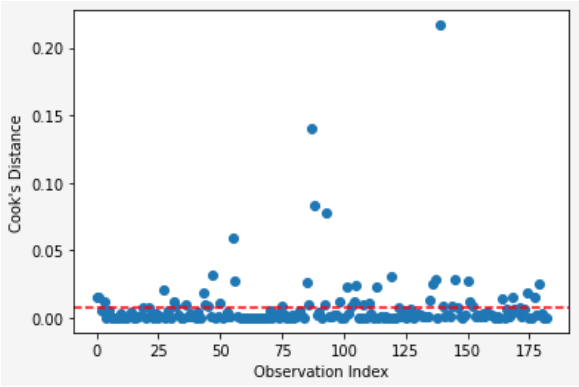
Dep. Variable:	life_expect	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.890
Method:	Least Squares	F-statistic:	143.9
Date:	Sat, 06 May 2023	Prob (F-statistic):	1.22e-76
Time:	23:32:28	Log-Likelihood:	-428.36
No. Observations:	178	AIC:	878.7
Df Residuals:	167	BIC:	913.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.9
const	50.0564	5.971	8.384	0.000	38.269	61.
bmi	-0.6789	0.226	-3.006	0.003	-1.125	-0.
basic_water	0.0723	0.028	2.586	0.011	0.017	0.
region_Eastern Mediterranean	-2.5787	0.853	-3.024	0.003	-4.262	-0.
age5-19obesity	0.4205	0.099	4.241	0.000	0.225	0.
gghe-d	0.4388	0.149	2.942	0.004	0.144	0.
log_une_gni	1.8810	0.357	5.274	0.000	1.177	2.
region_Africa	-3.4454	0.766	-4.498	0.000	-4.958	-1.
log_une_hiv	-1.2026	0.204	-5.894	0.000	-1.605	-0.
age5-19thinness	-0.1471	0.078	-1.890	0.060	-0.301	0.
diphtheria	0.1277	0.020	6.424	0.000	0.088	0.

C. Diagnostics

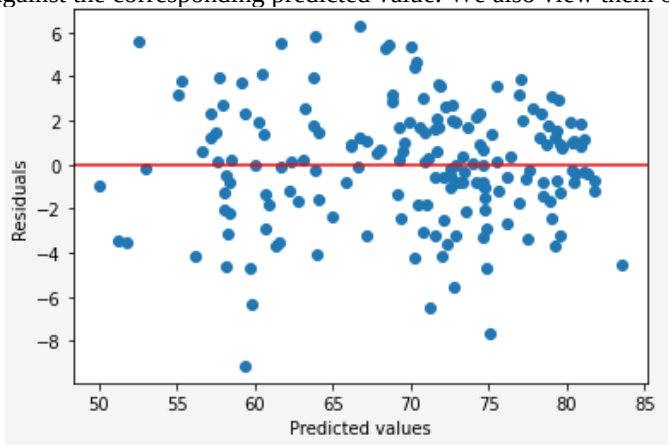
Here we analyze common model diagnostics to ensure the validity of our model. The first diagnostic of interest is to determine how sensitive our model is to perturbation, through leverage and Cook's distance.

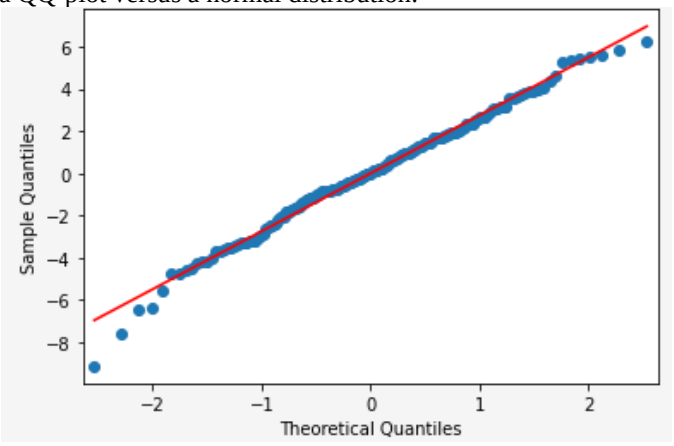




The red line in the above charts indicates the mean observed value. From this, we can see that there are some points with moderate leverage, but even more concerning is that there are some points with extreme Cook's distance. As a measure of how much the model would change if a point was left out, a high Cook's distance indicates our model is being affected greatly by a few individual observations. These five obviously influential points (with Cook's distance > 0.05) were dropped, and the model was refit. The final model described above is the model that was fit on the dataset minus these five observations. While this slightly increases the prediction error on the full dataset, we gain what are potentially more stable coefficient estimates by eliminating these influential points. Since the goal of this project is to explain precisely the magnitude of the effect the predictors have on life expectancy, we consider this to be a worthwhile trade-off.

Another important diagnostic involves checking the residuals for homoskedasticity and normality. For this we plot the residuals against the corresponding predicted value. We also view them on a QQ-plot versus a normal distribution.





The residuals appear to be homoscedastic and follow normal quantiles to a good degree of accuracy. The main concern is that the residuals exhibit a slight parabolic pattern in the residual vs predicted plot. This can sometimes indicate that a second-order term, either a squared term or interaction term, is needed in the model. However, despite extensive testing of including such terms in the model we found that there is no single set of interaction terms that significantly changes the appearance of the residual plot. Since we are unable to capture such a pattern in the data using a linear model on our dataset, we hypothesize that we are simply lacking some signal in our given predictors and that this problem may only be resolved by obtaining novel features. That being said, the parabolic pattern is not strong and could also be attributed to noise. Overall, the unambiguous homoscedasticity and normality of the residuals give us confidence that our model is properly specified.

D. Standard Errors and Confidence Intervals

To evaluate the assumption of homoscedasticity in our linear regression model, we can use the Eicker-Huber-White (EHW) robust standard error estimates for the coefficients. Since our forward selection model assumes homoscedasticity, we can compare these robust standard errors, which account for potential heteroscedasticity, with the standard errors estimated by our model. We can see the model estimates *(Left)* and the robust standard error estimates *(Right)* are very close giving us similar confidence intervals. This result supports the observed residual homoscedasticity and gives us confidence that our error bounds for coefficients are robust even when using the normal standard errors.

coefficient	forward_selection_se	lower	upper	ehw_robust_se	ehw_lower	ehw_upper
const	6.230970	43.511757	67.937158	7.260642	41.493599	69.955316
region_Africa	0.831493	-5.129795	-1.870341	0.892612	-5.249588	-1.750548
log_une_hiv	0.216830	-1.729210	-0.879235	0.247479	-1.789281	-0.819164
region_Eastern Mediterranean	0.920405	-4.392102	-0.784112	0.863685	-4.280930	-0.895285
log_une_gni	0.374320	0.972976	2.440310	0.369160	0.983090	2.430196
bmi	0.230559	-1.269495	-0.365705	0.275976	-1.358512	-0.276688
age5-19thinness	0.081432	-0.393164	-0.073950	0.085905	-0.401931	-0.065183
basic_water	0.028773	0.059180	0.171971	0.032111	0.052638	0.178514
diphtheria	0.020195	0.049009	0.128174	0.028063	0.033589	0.143595
gghe-d	0.155256	0.030659	0.639262	0.154950	0.031259	0.638662
age5-19obesity	0.098835	0.187800	0.575233	0.116980	0.152236	0.610797

III. Discussion

Our purpose is to adequately describe the complex relationships on life expectancy. We have chosen a single model with an impressive balance of goodness of fit and the number of covariates to achieve this goal. Our final covariates are age5-19obesity, age5-19thinness, basic_water, bmi, log_une_hiv, une_gni, diphtheria, and gghe-d, as well as two region indicators for Africa and Eastern Mediterranean. The presence of these indicators in the final model evidences the similarities of countries in these regions with respect to life expectancy, but they do not give specific insights, so we will instead move on to the other eight covariates.

Perhaps surprisingly, covariates such as doctors, hospitals, poverty, schooling, and alcohol consumption do not appear in the final model. The covariates that matter, as it turns out, are more fundamental than that. Increased access to basic water, lack of thinness (can be interpreted as access to food), higher vaccination rates and healthcare spending all point to the primary driver of low life expectancy to not be specifics such as number of doctors or alcohol consumption, but access to basic needs and living in a country with a government able to provide basic necessities. This is further evidenced by the fact that prevalence of obesity is positively correlated with life expectancy, even while bmi is negatively correlated. While it is a fact that increasing BMI past a certain level only increases the likelihood of health complications, the prevalence of obesity can be seen as a proxy for a country affluent enough for people to become obese in the first place.

In future work we would be interested in regressing only on first-world countries to gain actionable political insights for countries with governments stable enough to implement said insights. Another potential extension of this case study would be to include the time component allowing for yearly predictions. This would provide a more comprehensive understanding of how the factors under consideration are changing over time. Earlier, we made the simplifying assumption that the observations are independent across countries, but it is possible that there is some dependence from global events or economic trends such that the time dimension cannot be completely ignored. In addition, time series data could be particularly useful for identifying factors that are driving macroeconomic changes, as macroeconomic trends can often take years or decades to fully develop. We also note from our residual analysis that we suspect the absence of certain features in our original dataset, causing us to be unable to explain the slightly parabolic residual trend. We recommend a follow up investigation on more health-related variables to determine whether they could explain the curvature in the residual plot and improve the overall fit of the model.

IV. Conclusion

Based on the analysis conducted, it can be concluded that linear modeling techniques can be used to adequately describe the influence of various factors on life expectancy across a diverse set of countries. Our initial research goal of obtaining actionable insights appears to have been misspecified, as it is basic needs and national stability that must come first. It is difficult to suggest actionable insights since many of the countries with very low life expectancy are likely lacking policymakers with the ability to take such actions in the first place. However, we are happy with the results of the analysis as it is informative of these issues and our robust confidence intervals assure the validity of our conclusions.

V. Additional Work

A. Additional Data Imputation Techniques

As mentioned above, our goal is to model using a single year for each country. After obtaining the “best” year for each country as described in the *Data* section, we imputed using KNN. However, there are other strategies to try that do not impute data and therefore may be more accurate or preferred in certain contexts. The features with missing entries before imputation fall cleanly into two categories. In the first category the number of missing values is less than 4%. This is likely because there are a few countries that are difficult to get basic data about (think North Korea). In the second category, features have 20-30% missing

values. This is likely because features such as literacy rate or number of schools can be difficult to report. With these two categories in mind, there are two natural ways to remove missing values while avoiding imputation:

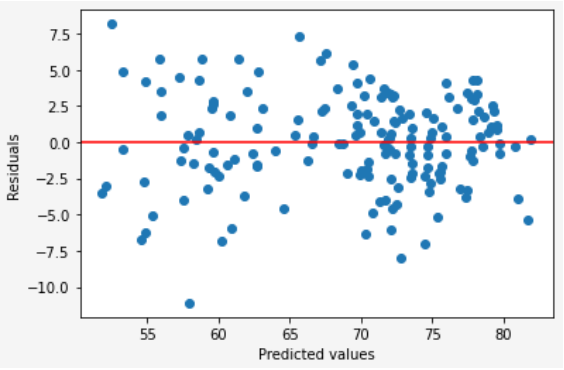
1. Drop all of the countries with missing values for any of the features in the first category. Drop all of the features in the second category completely. This retains most of the features for only slightly fewer observations.
2. Drop all of the countries with missing values for ANY of the features, regardless of category. This retains all of the features, but greatly reduces the number of observations.

With these two data treatments, there is an observation-feature tradeoff for us to balance. The first treatment retains 165 of the 183 countries and 21 of the 27 potential predictors. The second treatment retains only 61 of the 183 countries but all of the potential predictors.

In our initial analysis we used the imputed dataset with all 183 observations. Since this is not a huge number of observations to begin with, we were concerned about the generalizability of a model fit on relatively few observations. Especially if we wanted a decent number of final predictors, we needed enough observations such that n >> p. However, the concern of how imputing may obscure the true data remained, so we went back and repeated our model selection methods on both other treatments.

First treatment:

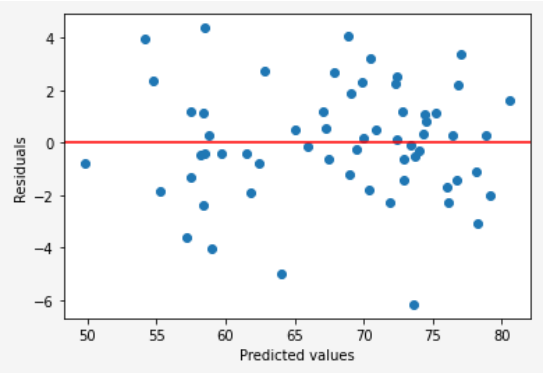
	coef	std err	t	P>|t|	[0.025	0.975]
const	78.4732	6.239	12.578	0.000	66.148	90.798
polio	0.0630	0.030	2.075	0.040	0.003	0.123
age5-19obesity	0.4283	0.119	3.594	0.000	0.193	0.664
che_gdp	-0.3346	0.164	-2.041	0.043	-0.658	-0.011
doctors	0.1105	0.044	2.494	0.014	0.023	0.198
gghe-d	0.9013	0.254	3.550	0.001	0.400	1.403
hepatitis	0.0220	0.014	1.611	0.109	-0.005	0.049
bmi	-1.2999	0.254	-5.112	0.000	-1.802	-0.798
region_Europe	-1.1433	1.073	-1.066	0.288	-3.263	0.976
basic_water	0.1764	0.029	6.175	0.000	0.120	0.233
age5-19thinness	-0.2775	0.091	-3.033	0.003	-0.458	-0.097
region_Africa	-5.5363	0.869	-6.368	0.000	-7.254	-3.819



We see very similar predictors in the forward-selected model. Since this model was trained only on non-imputed data, this is good evidence that our imputation is not obscuring the results. Furthermore, the residual plot exhibits the same parabolic shape as the imputed model.

Second treatment:

	coef	std err	t	P>|t|	[0.025	0.975]
const	85.4190	13.012	6.564	0.000	59.241	111.597
age5-19obesity	0.8798	0.196	4.497	0.000	0.486	1.273
une_literacy	0.0807	0.028	2.913	0.005	0.025	0.137
region_Eastern Mediterranean	-1.8611	1.561	-1.193	0.239	-5.001	1.278
measles	0.1271	0.030	4.305	0.000	0.068	0.187
hospitals	-0.9433	0.297	-3.176	0.003	-1.541	-0.346
bmi	-2.0107	0.521	-3.862	0.000	-3.058	-0.963
log_une_hiv	-1.2486	0.423	-2.953	0.005	-2.099	-0.398
log_une_gni	1.7760	0.720	2.466	0.017	0.327	3.225
basic_water	0.0722	0.045	1.617	0.113	-0.018	0.162
region_Western Pacific	-5.6993	1.746	-3.264	0.002	-9.212	-2.186
region_Africa	-5.2134	1.895	-2.751	0.008	-9.026	-1.400
log_une_pop	-0.7479	0.291	-2.566	0.014	-1.334	-0.161
region_South-East Asia	-4.9321	1.862	-2.649	0.011	-8.678	-1.186



The second treatment has more predictors in the forward selected model, but the common coefficients are similar. In fact, with this many predictors and 61 data points we are quite concerned about overfitting. Again, the parabolic pattern is present in the residuals.

In conclusion, the other potential treatments do not perform significantly differently than the imputed data nor solve the residual issue. As a result, we prefer the imputed data since it uses the most observations and all of the potential predictors.

B. Feature Selection Methods

We experimented with multiple feature selection methods to find a balance of model complexity and goodness of fit. In addition to the favorable forward selection covariates, we performed backwards selection, and Lasso on our observed imputed dataset.

Our target criterion involved examining the value of AIC and R^2. Below shows the features selected by the corresponding methods.

Forward Selection ~ R^2: 0.881, AIC: 937.1

	coef
const	55.7245
region_Africa	-3.5001
log_une_hiv	-1.3042
region_Eastern Mediterranean	-2.5881
log_une_gni	1.7066
bmi	-0.8176
age5-19thinness	-0.2336
basic_water	0.1156
diphtheria	0.0886
gghe-d	0.3350
age5-19obesity	0.3815

Backward Selection ~ R^2: 0.884, AIC: 942.8

	coef
const	44.2470
log_une_hiv	-1.4014
region_Eastern Mediterranean	5.7100
diphtheria	0.0804
age5-19obesity	0.2990
diseases	0.0234
measles	-0.0586
gghe-d	0.3685
region_South-East Asia	8.9900
region_Western Pacific	8.8458
bmi	-0.7126
log_une_gni	1.8670
polio	0.0484
region_Europe	7.4115
age5-19thinness	-0.2822
basic_water	0.1069
region_Africa	4.9395
region_Americas	8.3503

Lasso (5 folds cross validation) ~ R^2: 0.868

Feature	Coefficient
bmi	-0.951555
age5-19thinness	-0.257362
age5-19obesity	0.410026
hepatitis	0.005659
measles	-0.030502
polio	0.033703
diphtheria	0.073439
basic_water	0.152293
doctors	-0.002661
hospitals	-0.020912
gghe-d	0.485722
une_edu_spend	-0.148449
une_literacy	0.035406
region_Africa	-1.242741
region_Western Pacific	0.013534
log_une_poverty	-0.127873
log_une_gni	0.964615
log_une_hiv	-1.336286
log_une_pop	-0.120379

C. *Bootstrap LASSO*

Bootstrap LASSO *[2]* is a method for feature selection in linear regression that combines the LASSO penalty with bootstrap resampling. For our purposes, we use this method to determine which features are significant. Bootstrap LASSO works by fitting LASSO regression models to a large number of bootstrapped samples of the data, with a range of values for the L1 regularization parameter α chosen by 5-fold cross validation. The resulting coefficients are then averaged across the bootstrapped samples to obtain summary statistics. The method determines which coefficients are significant by computing the ratio of the mean coefficient value to its standard deviation, and then comparing this ratio to a threshold value of two. Coefficients with a ratio greater than two are considered significant and are present in all of the feature selection techniques we've investigated. The bootstrap LASSO method is useful for datasets where there are many potential predictor variables, but it is not clear which variables are the most important. Below shows the results with the "Significant" column indicating which predictors pass the threshold.

Feature	Coefficient Mean	Coefficient Std	How Many Std Out	Significant
log_une_hiv	-1.336946	0.275104	4.859787	True
bmi	-0.860639	0.311966	2.758760	True
basic_water	0.126716	0.047386	2.674111	True
age5-19obesity	0.356440	0.137005	2.601654	True
gghe-d	0.496799	0.222349	2.234324	True
log_une_gni	1.401342	0.672540	2.083657	True
age5-19thinness	-0.263585	0.139453	1.890129	False
region_Africa	-1.800007	1.384159	1.300434	False
diphtheria	0.107249	0.089956	1.192238	False
region_Eastern Mediterranean	-1.205189	1.181873	1.019728	False
log_une_pop	-0.129381	0.139152	0.929781	False
une_literacy	0.022888	0.026140	0.875592	False
region_Western Pacific	0.642093	0.767174	0.836959	False

Feature	Coefficient Mean	Coefficient Std	How Many Std Out	Significant
region_Americas	0.488497	0.633313	0.771336	False
log_une_poverty	-0.170274	0.228975	0.743636	False
une_edu_spend	-0.126295	0.173014	0.729970	False
hepatitis	0.009357	0.013502	0.693000	False
measles	-0.035080	0.051483	0.681388	False
region_South-East Asia	0.514420	1.037073	0.496031	False
alcohol	-0.038445	0.086815	0.442833	False
diseases	-0.013923	0.053219	0.261611	False
hospitals	-0.030686	0.125389	0.244724	False
polio	0.019014	0.092976	0.204499	False
region_Europe	-0.044690	0.307677	0.145250	False
doctors	-0.003929	0.035737	0.109942	False
che_gdp	-0.015485	0.188889	0.081978	False
une_school	-0.005025	0.112075	0.044835	False

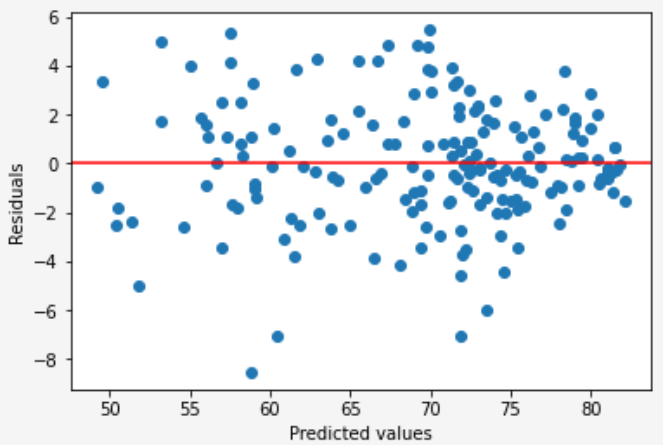
Besides finding significant features, this method can be used as a general feature importance method by comparing the p-values or "How Many Std Out" column. We used this method to understand which features were generally important in the modeling process.

D. Polynomial Features

After discovering the parabolic pattern in the residual plot, we performed extensive testing to see if the plot could be improved by adding second-order terms, either as interactions between two covariates or as squared versions of existing covariates. No combination of these improved the plot, but we will go into detail about some of these tests here.

The initial wide-angle approach was to add every possible second-order term to the dataset and perform feature selection. We attempted feature selection in the same ways as above, with forward and backward selection, as well as LASSO. The results showed many of the interaction terms being selected, but the number of additional covariates was quite large for only a nominal increase in R-squared. Most importantly, the parabolic residual problem was not fixed.

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.2264	5.107	13.360	0.000	58.143	78.310
0	-0.8148	0.244	-3.342	0.001	-1.296	-0.333
2	-9.6991	3.620	-2.679	0.008	-16.848	-2.550
4	0.4960	0.197	2.517	0.013	0.107	0.885
37	-0.0587	0.013	-4.368	0.000	-0.085	-0.032
36	0.0984	0.042	2.346	0.020	0.016	0.181
39	0.0754	0.049	1.526	0.129	-0.022	0.173
42	-0.0717	0.020	-3.525	0.001	-0.112	-0.032
43	-0.0106	0.004	-2.910	0.004	-0.018	-0.003
13	0.0566	0.020	2.818	0.005	0.017	0.096
46	0.9263	0.506	1.831	0.069	-0.073	1.925
47	-0.0376	0.083	-0.452	0.652	-0.202	0.127
54	0.0155	0.003	4.932	0.000	0.009	0.022
22	0.0006	0.005	0.126	0.900	-0.009	0.011
24	0.0081	0.003	2.791	0.006	0.002	0.014
57	-0.5557	0.132	-4.216	0.000	-0.816	-0.295
56	-1.2400	0.496	-2.500	0.013	-2.219	-0.261
59	-0.1973	0.125	-1.575	0.117	-0.445	0.050
60	-0.1192	0.036	-3.347	0.001	-0.190	-0.049



As a sample of the results from these methods, the forward-selected model is shown above. The feature names are indicated only by indices, obscuring the results, however every index greater than 26 is an interaction term of some sort, so the model has 12 interaction terms! The R-squared increased to 0.92 but the residual plot is virtually unchanged, a surprising result. The backward-selected and LASSO model had very similar results.

Clearly, a feature selection approach was not going to work, so we instead tried to engineer features to get at the heart of the problem. From the residual plot, the model tends to sometimes have large negative residuals when the predicted value is either large or small. In other words, the model sometimes predicts middle-of-the-pack countries to have large or small life expectancy. It is conceivable that some features could have a quadratic effect on life expectancy. We experimented with including specific squared terms or interaction terms of predictors in our final model rather than trying to select from the full range of interactions, but no change was observed. Our final conclusion is that no combination of interaction terms can improve the residual plot, and the observed pattern is due to either random noise, a strongly nonlinear relationship, or the lack of important features in our original dataset. Whatever the case, the issue bears further investigating and is outside of the scope of this paper.

VI. References

[1] MMATTSON (2020, October 6). "WHO national life expectancy" URL https://www.kaggle.com/datasets/mmattson/who-national-life-expectancy
[2] Charles Laurin, Dorret Boomsma, Gitta Lubke (Aug 1). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5131926/