Title: Life expectancy prediction using regression model

Student ID: s3961136 Student Name: Benjaporn Wongmayura email: s3961136@student.rmit.edu.au

| Executive Summary

This report investigates life expectancy prediction using health indicators and socioeconomic factors. We compared Polynomial Ridge Regression and Linear Regression, finding that Polynomial Ridge Regression outperforms in accuracy and robustness, showcasing higher R-squared values and lower mean squared errors. However, challenges include potential overfitting with high-degree polynomials and interpretability issues due to feature transformations. The success of Polynomial Ridge Regression highlights its potential for guiding public health policies and resource allocation by identifying key factors influencing life expectancy. This approach advances beyond traditional analysis, offering precise and actionable insights for improving healthcare outcomes.

| Introduction

Exploring life expectancy through regression analysis offers insights into how health, socioeconomic factors, and other variables affect longevity. This report aims to predict life expectancy, providing valuable data to inform health policies and improve overall health outcomes.

Objective: To predict continuous value for life expectancy(years)

| Methodology

Data

- Data: Global Health and Life Expectancy Dataset from WHO
- Dataset: 2,071 entries with 22 features (2071 rows × 24 columns including ID, Target)
- Types of data: Categorical = 12, Numerical = 12

Model

- Baseline Model: Linear Regression
- Advanced Model: Polynomial Regression

| Table of Contents

| 1. Data Retrieval

First, since given dataset was splitted, we retrieved train and test data and used **.head()** to check the data roughly.

2. Exploratory Data Analysis (EDA)

- Check shape of train_df , test_df and its datatype
- Make sure there is no redundant white space.
- Sanity checks for impossible value in numeric column and found that the numerical number make senses.
- Use .info to check statistics and found that there is no null value. The data type is in the right type. No value is missing since count is 2071 for all features.
- Use **.describe** to check range.

[855] train_df.describe()

	ID	TARGET_LifeExpectancy	Country	Year	Status	AdultMortality	AdultMortality-Male	AdultMortality-Female	SLS	Alcohol	...	Polio	TotalExpenditure	Diphtheria	HIV-AIDS	GDP	Population	Thinness1-19years	Thinness5-9years	IncomeCompositionOfResources	Schooling
count	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	...	2071.000000	2071.000000	2071.000000	2071.000000	2071.000000	2.071000e+03	2071.000000	2071.000000	2071.000000	2071.000000
mean	1036.000000	69.274505	95.360212	2009.518590	0.185418	162.833897	161.908257	163.759536	33.079672	4.696379	...	82.727185	5.883858	82.753259	1.632883	7352.742342	1.203741e+07	4.941284	4.977306	0.609551	3.372453
std	597.990524	9.482281	54.861641	4.614147	0.388730	118.872170	119.442235	118.802292	135.832868	4.205888	...	23.188837	2.554965	23.130969	4.782325	15219.978663	6.391797e+07	4.697830	4.785532	0.216532	0.590832
min	1.000000	37.300000	0.000000	2002.000000	0.000000	1.000000	0.000000	2.000000	0.000000	0.010000	...	3.000000	0.370000	2.000000	0.100000	1.880000	3.400000e+01	0.100000	0.100000	0.000000	0.000000
25%	518.500000	63.000000	50.000000	2006.000000	0.000000	74.000000	74.000000	74.000000	0.000000	0.615000	...	77.000000	4.190000	78.000000	0.100000	413.730000	1.274450e+05	1.600000	1.500000	0.463000	3.065942
50%	1036.000000	71.200000	94.000000	2010.000000	0.000000	144.000000	142.000000	144.000000	3.000000	3.830000	...	93.000000	5.640000	93.000000	0.100000	1410.670000	6.522310e+05	3.200000	3.300000	0.655000	3.449638
75%	1553.500000	76.000000	144.000000	2014.000000	0.000000	228.000000	228.000000	230.000000	22.000000	7.840000	...	97.000000	7.430000	97.000000	0.800000	5811.295000	5.371104e+06	7.400000	7.400000	0.769500	3.741657
max	2071.000000	92.700000	192.000000	2017.000000	1.000000	699.000000	704.000000	722.000000	1800.000000	17.870000	...	99.000000	17.600000	99.000000	50.600000	133473.470000	1.293859e+09	27.700000	28.600000	0.948000	4.381780

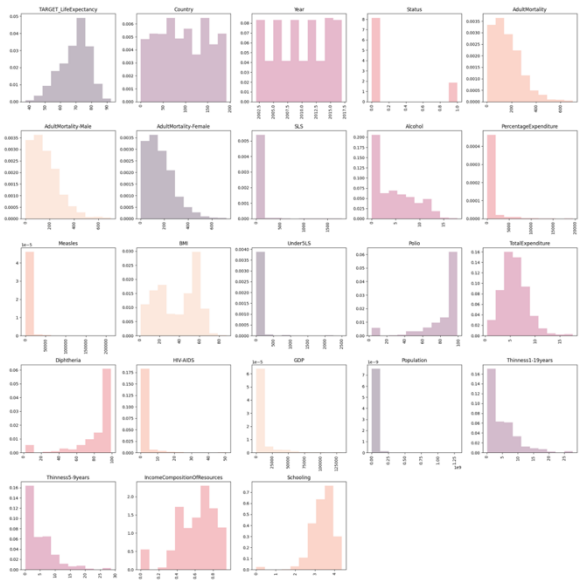
8 rows × 24 columns

Observation

- There is a big difference in numbers for features like GDP (which goes from 1.88 to 133,473.47), Population (from 34 to about 1.29 billion), and other features with smaller ranges. This shows that we **need to adjust the scale** of these numbers.
- We see that the **average of TARGET_LifeExpectancy** is **69** years.
- By checking **count** I see that total count of each feature is the same. Hence, no data is missing

2.1 Data Distribution

2.1.1 Histogram



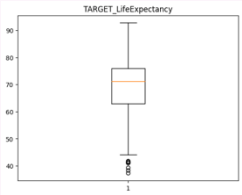
Observation

- The "Status" attribute is categorical, with the majority of data instances belonging to category 0 and only a minority to category 1.

- Many attributes are heavily skewed. e.g. AdultMortality, AdultMortality-Male, AdultMortality - Female, SLS, Alcohol, Polio, Diptheria, Thinness1-19years,Thinness5-9years

2.1.2 Boxplot

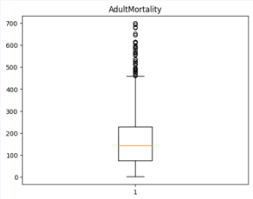
TARGET_LifeExpectancy



Observation: TARGET_LifeExpectancy' shows **a few outliers**, which are the points below the bottom line of the box

Iqr = 13
Upper threshold = 95.5
Lower threshold = 43.5

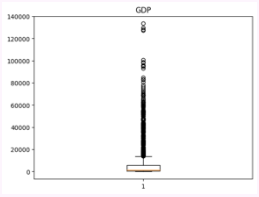
AdultMortality



Observation: The box plot for 'AdultMortality' reveals **several outliers above the upper range**.

Iqr = 154
Upper threshold = 459
Lower threshold = -157

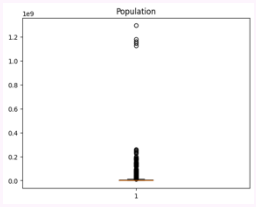
GDP



Observation: This box plot for 'GDP' indicates **a concentration of data near the bottom** with many high-value outliers.

Iqr = 5397.565
Upper threshold = 13907.6425
Lower threshold = -7682.6175

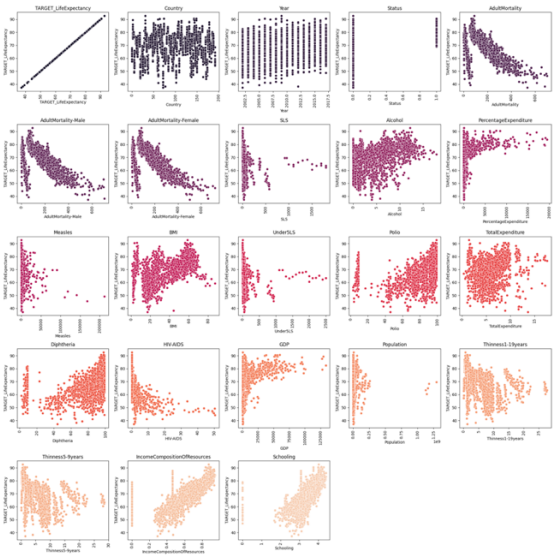
Population



Observation: The 'Population' box plot reveals **a few extremely high outliers** above the main cluster of data.

Iqr = 5243659
Upper threshold = 13236592.5
Lower threshold = -7738043.5

2.2 Relationship between variables



Observation

Variables with a Linear Relationship with TARGET_LifeExpectancy: Schooling,IncomeCompositionOfResources

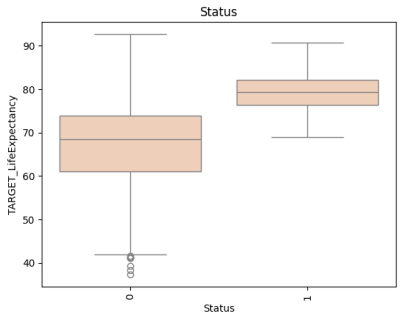
- As life expectancy increases, schooling increases.
- Higher human development, greater life expectancy

Variables with a Non-Linear Relationship with TARGET_LifeExpectancy: BMI, AdultMortality, AdultMortality-Male, AdultMortality-Female, HIV-AIDS, Thinness1-19years, Thinness5-9years

- AdultMortality, AdultMortality-Male, and AdultMortality-Female show a strong, non-linear decrease in 'TARGET_LifeExpectancy' as mortality rates go up, likely following an exponential pattern, where increases in adult mortality are associated with sharp decreases in life expectancy.

Higher HIV/AIDS,lower life expectancy, but this does not seem to follow a straight line

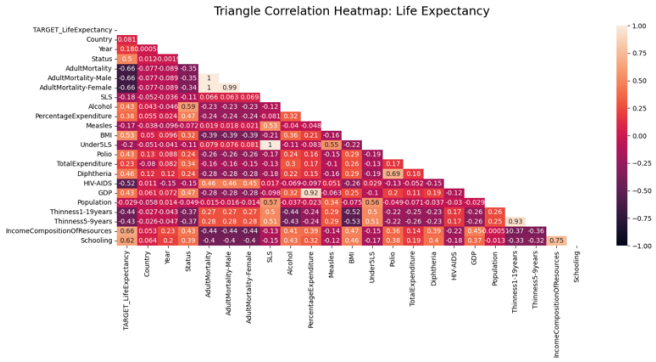
Use Box Plot to find the relationship between categorical attributes and target variable.



Observations

On average, data with 'Status=1' have a higher 'LifeExpectancy' compared to those with 'Status=0'.

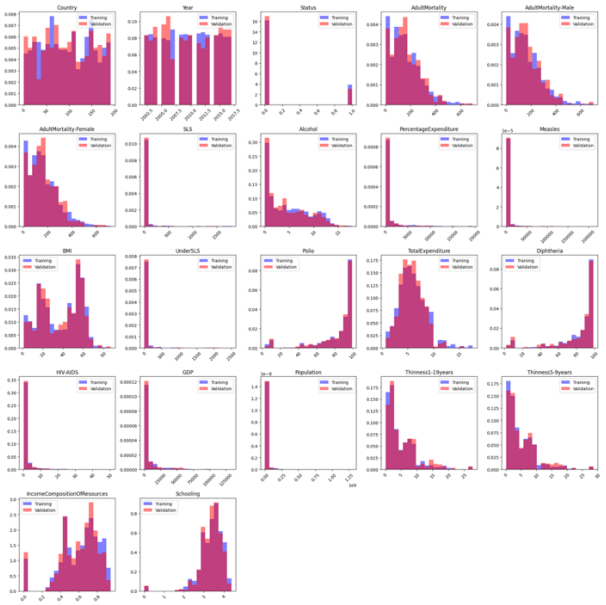
Plot correlation matrix for the numerical data and check for multicollinearity



Observations:

- Higher schooling and income levels are strongly linked to greater life expectancy.
- Higher adult mortality rates and HIV prevalence are strongly linked to lower life expectancy.
- Better BMI and healthcare (immunization rates for diphtheria) are moderately linked to higher life expectancy.
- Thinness in children and adolescents is moderately linked to lower life expectancy.
- Other factors like country, alcohol consumption, and health expenditure show weak links to life expectancy.
- Two variables in the dataset have a correlation coefficient of 1, it means they are perfectly linearly related; in other words, they provide identical information. This can occur due to a data entry error, a duplicate column, or if one variable is a direct calculation from the other e.g. 'AdultMortality and AdultMortality-Male', 'AdultMortality and AdultMortality-Female'.
- A correlation of 1 between SLS and Under5LS means they are perfectly correlated, essentially duplicates. So, we will remove one to avoid redundancy and multicollinearity.

3. Data Splitting



To assess the model, we divide the labeled train_df into two parts:

- training
- validation

Splitting the data into training and validation sets allows me to train my model on one portion of the data and test its performance on another, unseen portion. This helps ensure that the model can generalize well to new data, beyond just memorizing the training set.

We randomly split our data to 80% training set and 20% validation set. Hence, number of instances in the original dataset is 2071. After splitting training has 1656 instances and validation has 415 instances.

Observation
From the histograms, it appears that the distributions of the training and validation sets are similar, suggesting that the random split has been reasonably effective.

4. Data Pre-processing (or Transforming)

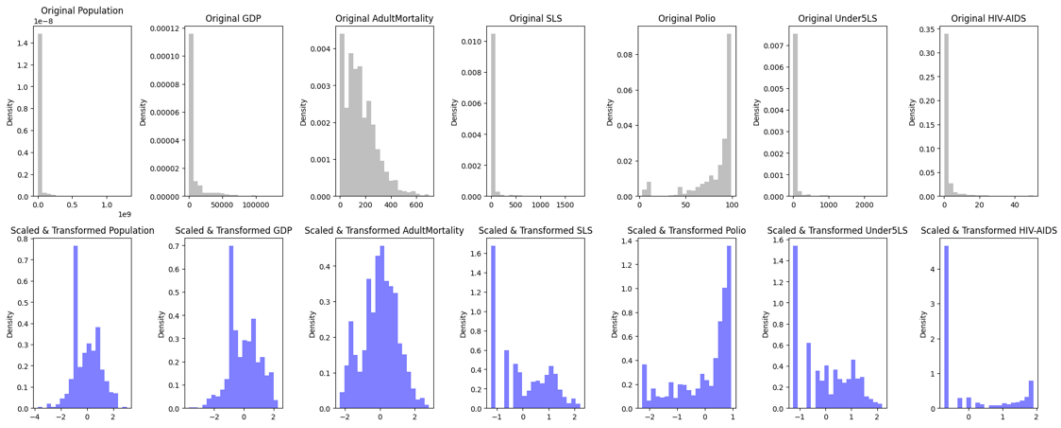
We split data before scaling it to avoid data leakage, which occurs when the model inadvertently learns from information it shouldn't have access to, leading to unrealistically high performance during training that doesn't hold up in actual use. Splitting the data into training and validation sets before applying any transformations is a good practice to ensure that the model is evaluated on unseen data.

We will perform scaling and power transforming on the **training data** to avoid data leakage, which means we will use data that we won't have access to when we assess the performance of the model.
Otherwise, we will overestimate the accuracy of our model

For linear regression using gradient descent, feature scaling aids in faster convergence by ensuring that all features contribute equally to the cost function, allowing the optimization algorithm to proceed more efficiently.

I decided to use **Standard Scaler** because my data have the below items.
- 1) **Outliers:**
linear regression is vulnerable to outliers. Using StandardScaler to standardize features by their z-score can mitigate this impact, potentially enhancing model performance. However, for extreme outliers, additional outlier handling may be necessary for optimal results.
- 2) **Heavily Skewed Features:** while StandardScaler does not directly transform skewed data into a normal distribution, it ensures each feature has a mean of 0 and a standard deviation of 1, which can be beneficial for linear regression.
- 3) **Plan to Linear Regression:** linear regression assumes that the features are normally distributed. Standardization can help in this regard, especially when combining features that have different scales and distributions.

Check distribution after scaling



Observation:
- StandardScaler doesn't change the shape of the distribution; it simply centers the data by removing the mean and scales it by dividing non-constant features by their standard deviation.

- PowerTransformer transforms the data to make it more Gaussian-like, which can improve the efficacy of linear models and other algorithms that assume normality.

Encoding
Since **Status** attribute is already encoded as 0 for developing countries and 1 for developed country, we will leave it alone.

5. Data Modelling

Training Data:
Country, Year, Status, AdultMortality, AdultMortality-Male, AdultMortality-female,Under5LS, SLS, Alcohol, Measles, BMI, Polio, TotalExpenditure, Diphtheria
, HIV-AIDS, GDP, Population, Thinness1-19years, Thinness5-9years, IncomeCompositionOfResources, Schooling

Target Variable:
TARGET_LifeExpectancy

Selecting Regression Evaluation Metrics

- R-squared is suitable for evaluating life expectancy predictions because it shows the proportion of variance in the life expectancy that our model accounts for. A higher R-squared value means better model performance.

- Mean Squared Error (MSE) works well as an evaluation metric in life expectancy prediction because it measures the average squared difference between the observed actual outcomes and the outcomes predicted by the model, emphasizing the impact of large errors. Lower MSE values are preferable as they indicate higher prediction accuracy.

- Together, they provide a clear picture of the model's explanatory power and predictive accuracy, esse

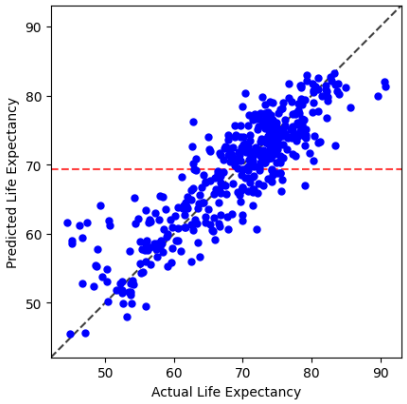
Baseline Model: Linear Regression Model

Reasons to select a linear regression model as a baseline model.

- Our exploratory data analysis revealed linear correlations between TARGET_LifeExpectancy and several variables, notably Schooling and IncomeCompositionOfResources.
- The linear model offers simplicity and intuitive understanding.

Algorithms that use gradient descent as an optimization technique (e.g., linear regression, logistic regression, neural networks) benefit from feature scaling. It ensures a faster convergence because the gradient descent path becomes smoother.

$$\begin{aligned}\textit{Life Expectancy} &= 69.3706 + 0.2248 \times \textit{Country} + 0.2585 \times \textit{Year} + 0.9523 \times \textit{Status} \\ &- 0.8046 \times \textit{AdultMortality} - 0.7400 \times \textit{AdultMortality-Male} \\ &- 0.8661 \times \textit{AdultMortality-Female} + 12.6577 \times \textit{SLS} \\ &+ 0.7237 \times \textit{Alcohol} + 0.2699 \times \textit{PercentageExpenditure} \\ &- 0.0215 \times \textit{Measles} + 0.4497 \times \textit{BMI} - 12.7766 \times \textit{Under5LS} \\ &+ 0.3884 \times \textit{Polio} - 0.0744 \times \textit{TotalExpenditure} + 0.7530 \times \textit{Diphtheria} \\ &- 2.3235 \times \textit{HIV-AIDS} + 0.8491 \times \textit{GDP} - 0.3752 \times \textit{Population} \\ &- 0.5033 \times \textit{Thinness1-19years} - 0.1613 \times \textit{Thinness5-9years} \\ &+ 1.3676 \times \textit{IncomeCompositionOfResources} + 1.2381 \times \textit{Schooling}\end{aligned}$$



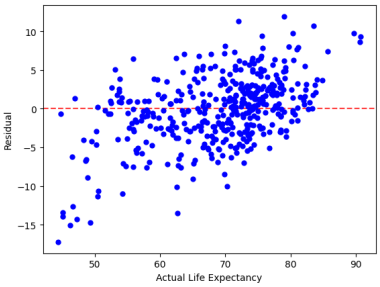
Model 1.1: Linear Regression

Observations:

- The model demonstrates a fair ability to estimate life expectancy for new data.
- The model's predictions are more accurate than a simple model based on the average life span (shown by the red line).
- The model predicts life expectancy with a positive correlation to actual values.
- Prediction accuracy decreases as actual life expectancy increases.
- The model tends to overestimate life expectancy.

Let's obtain quantitative values of performance. The R^2 score for the linear regression model is: 0.762, the Mean Squared Error for the linear regression model is: 19.690. Around 76% of the variance in the target variable is explained by the model, which is quite decent. However, there is still a room to improve the model. The MSE of 19.690 indicates the model's predictions are, on average, approximately 19.690 units away from the true values. lower MSE is typically better, and there's potential to improve the model's accuracy.

Residual plots



Observations:
Model Strengths:

- The residuals are centered around zero, indicating no systematic bias in predictions.
- The residuals are spread on both sides of the zero line, showing a good balance of over and under predictions.

What needs to be changed to improve the model:

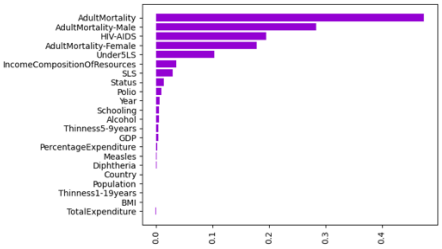
- The presence of patterns in residuals suggests the model may not fully capture all the influential factors or complex relationships.
- Some large negative residuals indicate the model significantly underpredicts in some cases, which may require further investigation and model refinement.

Hence, the model is generally effective but could benefit from non-linear approaches like Polynomial Regression or feature engineering to enhance its predictions.

Tuning Model

Feature importance

- **Permutation Feature**

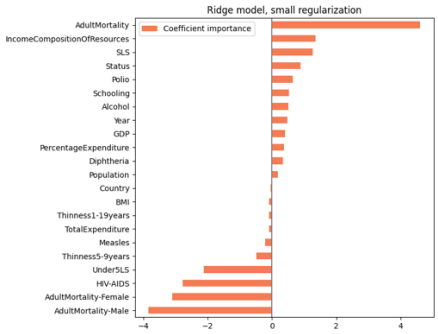


Observations:
High Importance: Features such as "AdultMortality," "AdultMortality-Male," "HIV-AIDS," and "AdultMortality-Female" show the highest importance in the model, suggesting a strong impact on life expectancy predictions.

Moderate Importance: Features like "IncomeCompositionOfResources" and "Under5LS" exhibit moderate importance, indicating a noticeable but lesser influence on life expectancy compared to the top features.

Low Importance: "BMI" and "TotalExpenditure" are among the features with the lowest importance, which may indicate a relatively minor role in the model's life expectancy predictions.

- **Coefficient Importance**



Observations:

- **IncomeCompositionOfResources & Schooling**: These have the highest positive coefficients, suggesting they're strong predictors for higher life expectancy.*Higher education and better resource distribution are linked to longer life expectancy.*
- **Status:** This has a positive coefficient too. *Being from a developed country is positively associated with a longer life.*
- Health indicators such as "Polio" and "Diphtheria" vaccinations, along with "GDP": these variables show smaller positive coefficients, indicating they contribute positively but to a lesser extent to life expectancy. *Greater GDP and immunization coverage modestly increase life expectancy.*
- **PercentageExpenditure & Country:** These have coefficients close to zero, suggesting they have little to no linear predictive power on life expectancy within this model.
- **Predictors like "Population" and "Thinness"** metrics have negative coefficients, suggesting that higher numbers in these categories may be related to lower life expectancy
- **The largest negative coefficients are associated with "HIV-AIDS" and "AdultMortality" metrics.** A higher prevalence or rate in these predictors correlates strongly with lower life expectancy, highlighting their critical impact on health outcomes.

K-Fold Cross Validation

Although we have implemented a form of hold-out validation, Using cross-validation in addition to the current train-validation split can enhance model evaluation, help use all data effectively, and improve the overall robustness and accuracy of the linear regression model

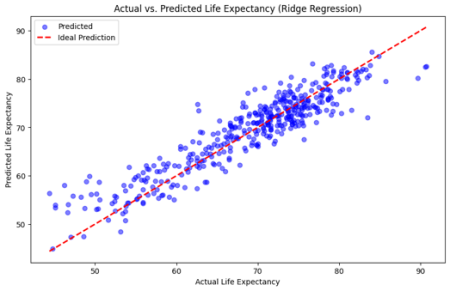
Apply regularisation

Regularisation is the process of adding information to a model in order to prevent overfitting. This is important in order to boost the evaluation metrics. We use regularisation because some of the features in our data are correlated.

After running 5-fold cross-validation to check the performance between Ridge and Lasso Regression. I found that Ridge has better performance when I tried using cross-validation to validate. Ridge model has lower mean MSE, which indicates Ridge Regression's predictions are closer to the actual values. So, I decided to use Ridge in the next model

In my case, **I decided to use Ridge Regression and utilise Grid Search** because Grid Search automates hyperparameter tuning to optimize model performance, ensuring a systematic and reproducible approach to finding the best parameter combination.

Model 1.2: Linear Ridge Regression



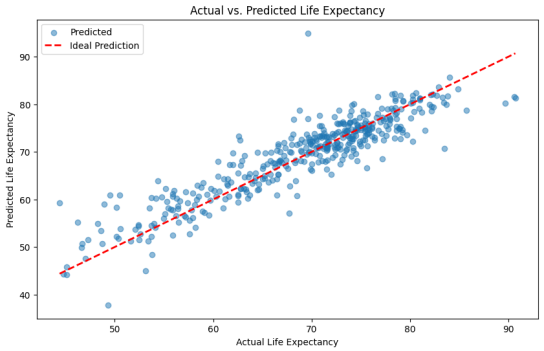
Best alpha value is 1, the R^2 score for the linear regression model is: 0.75838, the Mean Squared Error for the linear regression model is: 21.71403. This model does not improve the performance. Let's try using Polynomial Regression.

Advanced Model : Polynomial Regression

Polynomial regression can capture more complex relationships between the features and the target variable than linear regression by adding polynomial terms, allowing it to model nonlinear phenomena within the data. This makes it a suitable choice for scenarios where the relationship between variables isn't strictly linear, enhancing model flexibility and accuracy in predicting outcomes.

Before running polynomial regression, I have handled outliers by replace the outliers with mean value and removed the redundant features like AdultMortality-Male', 'AdultMortality-Female, Under5LS, Country. These actions help substantially improve the model accuracy.

Model 2.1: Polynomial Regression Model

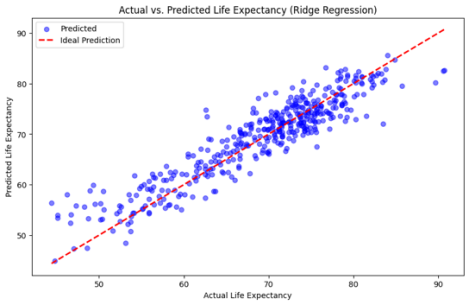


I choose degree=2 because after optimising the Degree of Polynomial Features. I found that **degree = 2 gives the highest R^2 and lowest MSE**, which indicates the best performance.

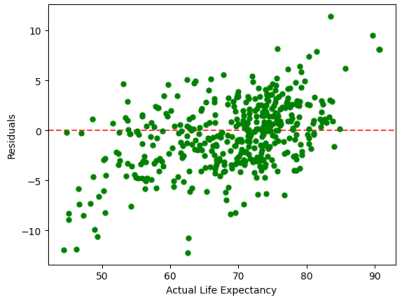
Polynomial Regression R^2: 0.84928, Polynomial Regression MSE: 12.45195

You can see that R^2 and MSE improving indicates better performance.

Model 2.2: Polynomial Ridge Regression Model



After optimization, we found that Best Alpha: 0.0001
Ridge Regression MSE: **12.27503**, Ridge Regression R^2: **0.85142**
Again, you can see that R^2 and MSE improving indicates better performance. Hence, this model has the best performance of four models we have explored.



Observations:

- The residuals are mostly scattered around the zero line, indicating that for many predictions, the model is quite accurate.
- There is no clear pattern in the residuals as they are randomly dispersed, suggesting the model's errors are not systematic across the range of life expectancies.
- However, there are some outliers, particularly for lower life expectancies (around 50 years), where the residuals are more negative, indicating underestimation.
- For higher life expectancies (closer to 90 years), the model also seems to underestimate slightly, but not as severely.
- The variance of the residuals does not appear to increase or decrease with actual life expectancy, which is a good sign that the model is consistent across different values.

Comparing Models

I choose the **Polynomial Ridge Regression Model** because of its **higher R^2 score and lower MSE indicates** it fits the data better and makes more accurate predictions than the Linear Regression Model. A higher R^2 signifies that the model explains a greater proportion of the variance in the target variable, while a lower MSE points to less error in the predictions, making it a more reliable model for forecasting or understanding the relationships within the data.

Model	R^2	MSE	Major Improvement
1.1 Linear Regression	0.762	19.690	Baseline model
1.2 Linear Ridge Regression	0.75838	21.71403	Baseline model with regularisation(Ridge, GridSearch) alpha optimisation
2.1 Polynomial Regression Model	0.84928	12.45195	Handling outliers Drop features Degree Optimisation
2.2 Polynomial Ridge Regression Model	**0.85142**	**12.27503**	Model 2.1 with regularisation (Ridge, GridSearch) and alpha optimisation

6. Conclusion

The choice of Polynomial Ridge Regression is justified by its high R-squared value and low mean squared error, indicating effective variance capture and accuracy in predicting life expectancy. This model adeptly addresses overfitting through Ridge Regression's regularisation, essential for managing the complexity of polynomial features. Its adherence to key regression assumptions—linearity in transformed space, multicollinearity management, and preliminary indications of homoscedasticity and independent residuals—further solidifies its suitability. However, potential challenges include overfitting with high-degree polynomials and reduced interpretability.

Ultimate Judgement: **The best model, in my opinion, is the Polynomial Ridge Regression.** The evidence supporting this judgement includes the highest R-squared value, indicating the model explains a significant portion of the variance in life expectancy from our predictors, and the lowest mean squared error, suggesting the predictions are closely aligned with the actual data. Additionally, the Ridge Regression component helps mitigate overfitting—a common issue with polynomial regression—by penalizing more complex models, which is particularly beneficial given the complexity introduced by polynomial features. The limitation of this ultimate model is its potential to still overfit if the degree of the polynomial is set too high, especially in the presence of limited training data. Moreover, the interpretability of the model can be challenging due to the transformation of features, making it harder to draw direct, intuitive conclusions about the importance and relationships of the original features.

7. Reference

- Analytics Vidhya, (2015). Several ways to improve your machine learning model's performance. Available at: https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/#:~:text=There%20are%20several%20ways%20to,bagging%2C%20boosting%2C%20and%20stacking [3 April 2024].
- Towards Data Science, (Date not provided). Polynomial Regression. Available at: https://towardsdatascience.com/polynomial-regression-bbe8b9d97491 [7 April 2024].
- Haji, S., (Date not provided). Using a Linear Regression Model to Predict Life Expectancy. Available at: https://shanzehhaji.medium.com/using-a-linear-regression-model-to-predict-life-expectancy-de3aef66ac21 [1 April 2024].
- Machine Learning Mastery, (Date not provided). Train-Test Split for Evaluating Machine Learning Algorithms. Available at: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/ [28 Mar 2024].