Inference on Global **Life Expectancy**

POLIMI GRADUATE SCHOOL OF MANAGEMENT

Research Question

What are the main factors influencing global life expectancy, and how can they be studied to formulate more effective policies and forecasts?

Our Mission

Through the usage of different statistical models perform a data driven analysis on the main variables impacting life expectancy across the world.

Our Purpose

Generate statistically significant inference, and predictive tools to aid the WHO in crafting measures which improve life expectancy globally in the most efficient manner.

World Health Organization

Methodology

A. Initial Data Processing and Analysis

1. Data importing
2. Data cleaning
3. Descriptive analysis
4. Data transformations
5. Tecnical approach

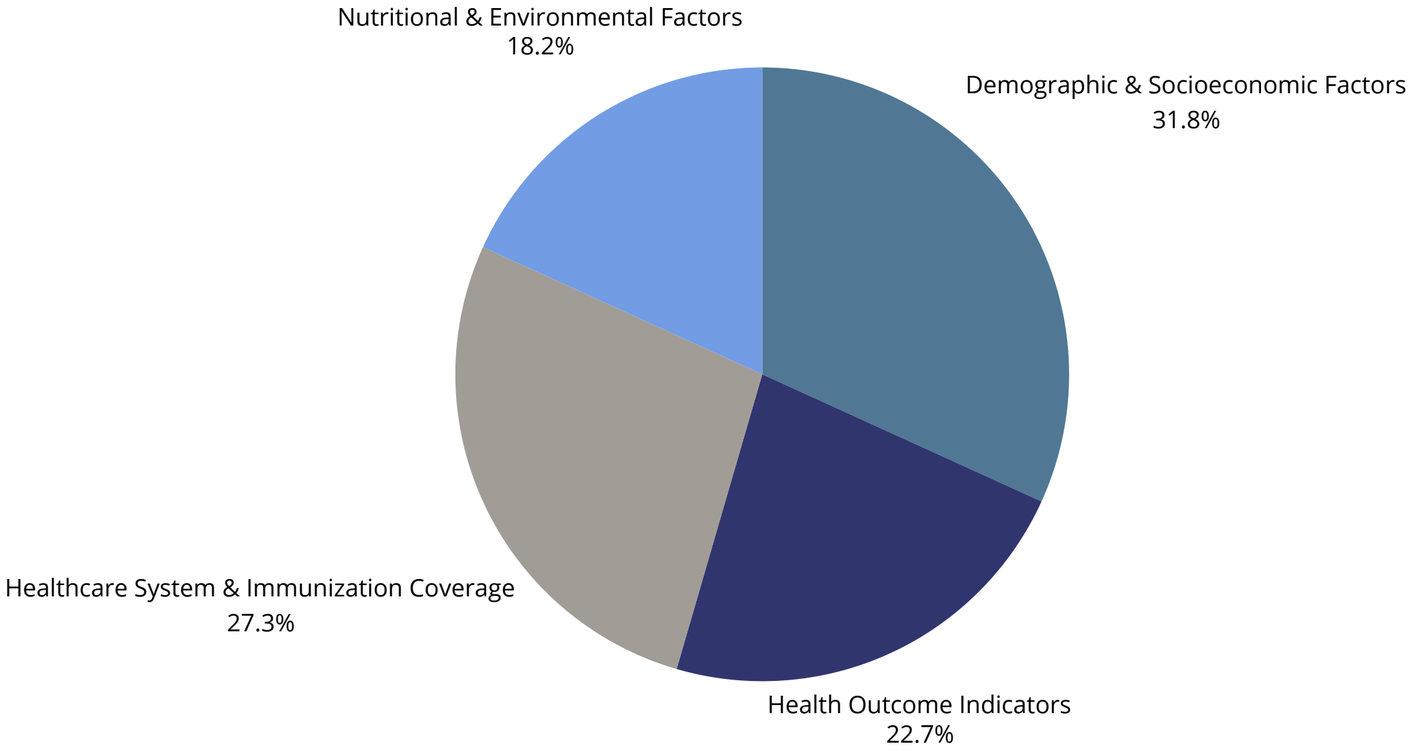
B. Modeling and Evaluation

1. Linear assumptions testing (Considerations)
2. MLR fitting and RMSE calculation
3. PCA analysis
4. Lasso regression analysis
5. Comparison
6. Findings
7. Conclusions
8. Next Steps

Data Card :

Life Expectancy (WHO): Statistical Analysis on factors influencing Life Expectancy by KumarRajarshi

Data set composition:



Source: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Key characteristics:

- Time Range: 2000-2015
- Total Files: 1 CSV dataset

Demographic & Socioeconomic Factors (7 variables)

- Country
- Year
- Status
- GDP
- Population
- Income composition of resources
- Schooling

Health Outcome Indicators (5 variables)

- Life expectancy
- Adult Mortality
- under-five deaths
- infant deaths
- HIV/AIDS

Healthcare System & Immunization Coverage (6 variables)

- Total expenditure
- percentage expenditure
- Hepatitis B
- Measles
- Polio
- Diphtheria

Nutritional & Environmental Factors (4 variables)

- BMI
- thinness 1-19 years
- thinness 5-9 years
- Alcohol

A. Initial Data Processing and Analysis

Data Importing

21 Different Features

Consisting of 19 numerical features and 2 categorical ones

3000 Observations

Not a lot of null values, about 150 only for specific countries/years are they present

Normal Distributions

All of the variables present a normal distribution with varying degrees of skewness

A. Initial Data Processing and Analysis

Data Cleaning

- Filling null values with median
- Create Dummy variable for Status and Year
- Outliers removal: used Tukey's depth to detect and remove outliers. 176 rows dropped

Descriptive Analysis

- Checking Histograms and Box Plots for each variable
- Generation of Covariance and Correlation Matrices
- Plot Feature Scatterplot

Data Transformation

- Multicollinearity Test with VIF > 10: dropped **Infant Deaths**
- Residual Transformation with Box-Cox to get Normality
- Log-Transformation for Residuals

A. Initial Data Processing and Analysis

Tecnical approach

Start with a Multiple Linear Regression (MLR) since the level of strictness of this model allows us to perform a better inference than other approaches

Select the most important variables for predicting and bettering life expectancy in countries

Perform alternative models (PCA and Lasso regression), which increases credibility and robustness of the investigation

Validate the performance of our models through K-Folds to see which fits best

B. Modeling and Evaluation

MLR Key Assumptions

MLR allows to run an accurate model when the following predictions hold:

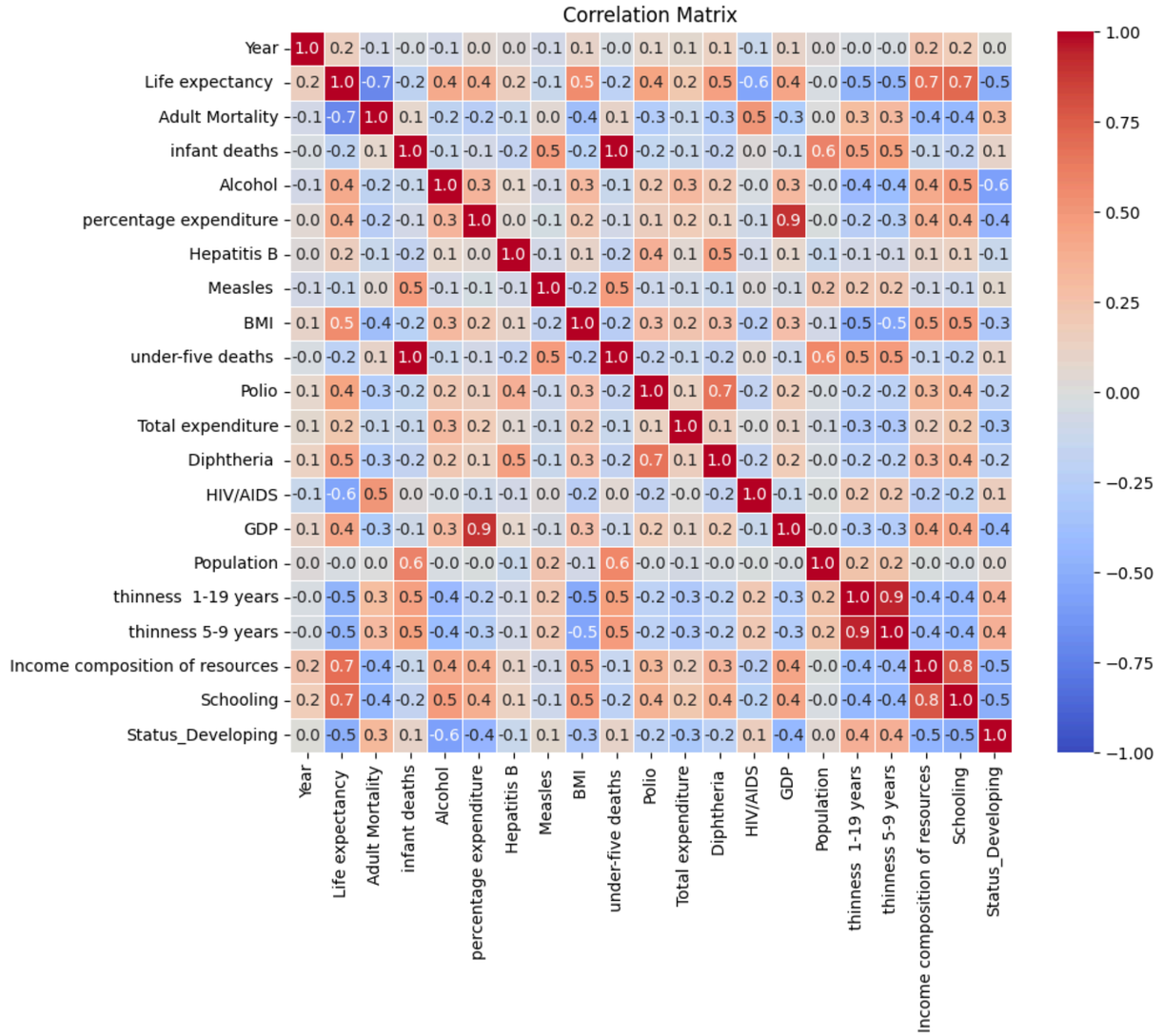
- Linearity
- Homoscedasticity of residuals(residual plots)
- Independence of residuals (Durbin-Watson test)
- Normality of residuals (Shapiro-Wilk test, Q-Q plots)
- No Perfect Multicollinearity (VIF analysis)

In the case these assumptions do not hold it is possible to run the analysis, however estimates might be biased and inefficient

B. Modeling and Evaluation

Linearity

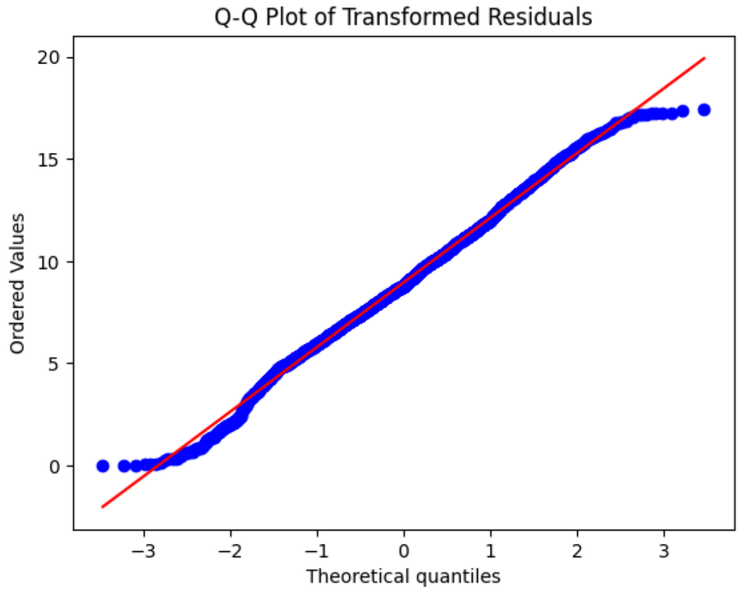
- Plotting the correlation matrix after removing outliers was used to check linearity
- Considering the correlation coefficients between the independent variable (life expectancy) and the dependent variables
- Inspecting the coefficients one by one, we can assume that there is a linear relation

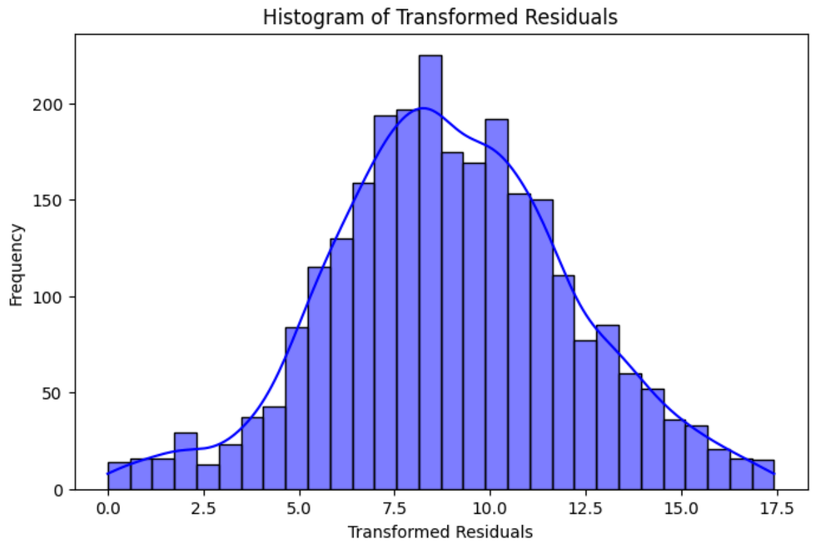


B. Modeling and Evaluation

Normality of Residuals

- Normality of residuals is key for confidence intervals and hypothesis testing
- The QQ plot of log transformed residuals displays a diagonal and the histogram resembles a normally distributed bell curve
- The data shows a normal distribution of its residuals post log transformation





B. Modeling and Evaluation

Multicolinearity

- Running VIF on all predictors helps spot collinearity
- An iterative loop removing variables until all the variables but constant have a VIF < 10

Drop multi-correlated predictors

	Variable	VIF
0	const	82.776253
1	Year	1.147502
2	Adult Mortality	1.699367
3	infant deaths	195.864398
4	Alcohol	1.857860
5	percentage expenditure	5.846138
6	Hepatitis B	1.319398
7	Measles	1.351197
8	BMI	1.686109
9	under-five deaths	191.957638
10	Polio	1.907991
11	Total expenditure	1.242026
12	Diphtheria	2.157951
13	HIV/AIDS	1.394974
14	GDP	6.030345
15	Population	1.596671
16	thinness 1-19 years	8.688270
17	thinness 5-9 years	8.819863
18	Income composition of resources	2.935252
19	Schooling	3.193970
20	Status_Developing	1.871214

Dropped infant deaths due to high VIF.

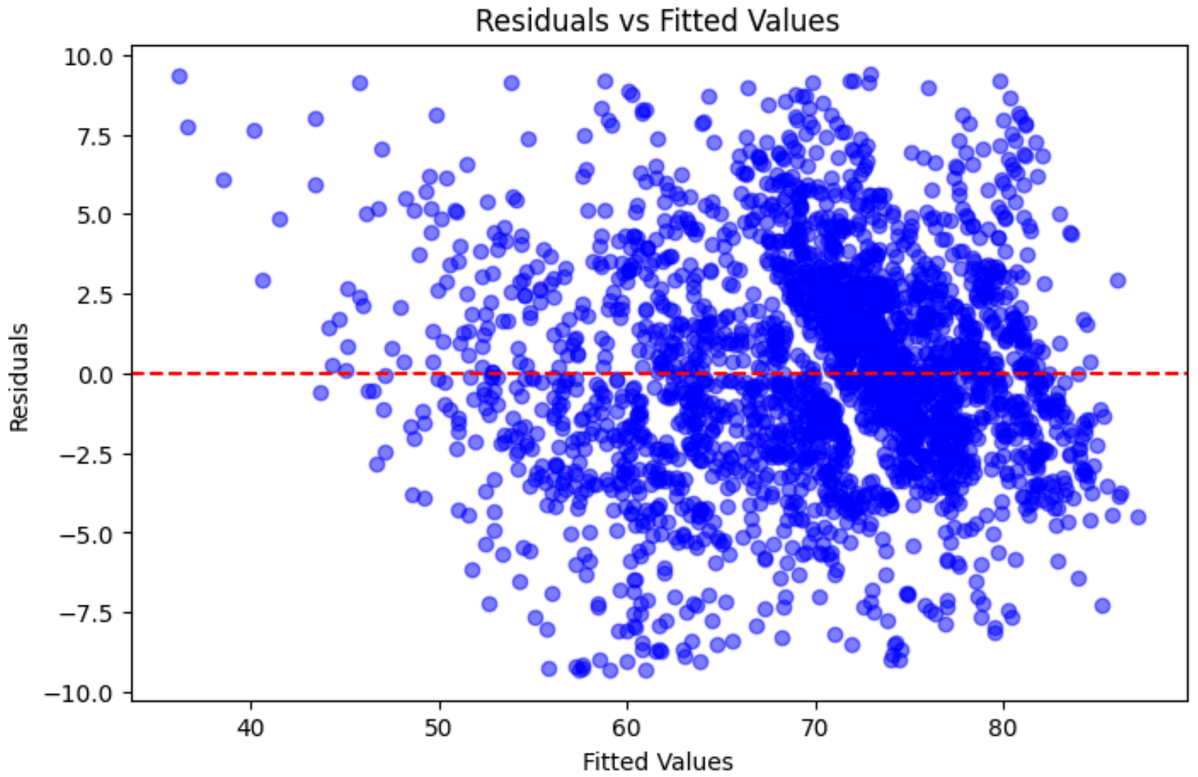
Current VIF Data:

	Variable	VIF
0	const	80.977565
1	Year	1.147069
2	Adult Mortality	1.690078
3	Alcohol	1.845417
4	percentage expenditure	5.844993
5	Hepatitis B	1.312008
6	Measles	1.336858
7	BMI	1.683635
8	under-five deaths	2.270740
9	Polio	1.904254
10	Total expenditure	1.241773
11	Diphtheria	2.117663
12	HIV/AIDS	1.390018
13	GDP	6.022672
14	Population	1.544291
15	thinness 1-19 years	8.688200
16	thinness 5-9 years	8.803137
17	Income composition of resources	2.918852
18	Schooling	3.189128
19	Status_Developing	1.870225

B. Modeling and Evaluation

Homoscedacity

- Plot Residual vs fitted values to check the distribution of the erros
- Visible pattern, therefore the assumption do not hold, there is slight Heteroscedasticity in the model
- Inference may be slightly affected, but the overall model remains useful for decision-making given it's not heteroscedastic everywhere.



B. Modeling and Evaluation

Residual Independence

- The Durbin-Watson statistic is lower than the ideal value of 2, indicating some autocorrelation.
- While this suggests a minor pattern in the residuals, it does not severely impact predictions.
- Inference may be slightly affected, but the overall model remains useful for decision-making.
- Further adjustments (e.g., adding lagged variables or using robust standard errors) could refine the model if needed.

Consider more details on future work.

B. Modeling and Evaluation

MLR Fitted Model

- After testing the OLS assumption, we fit the model
- Considering the P values (significance), we performed backward selection to improve our model
- Backward selection improved predictor significance but reduced Adjusted R^2 from 0.826 to 0.803, indicating a small loss of explanatory power.
- The refined model is more interpretable, but some removed variables likely contributed some information (to be addressed later)

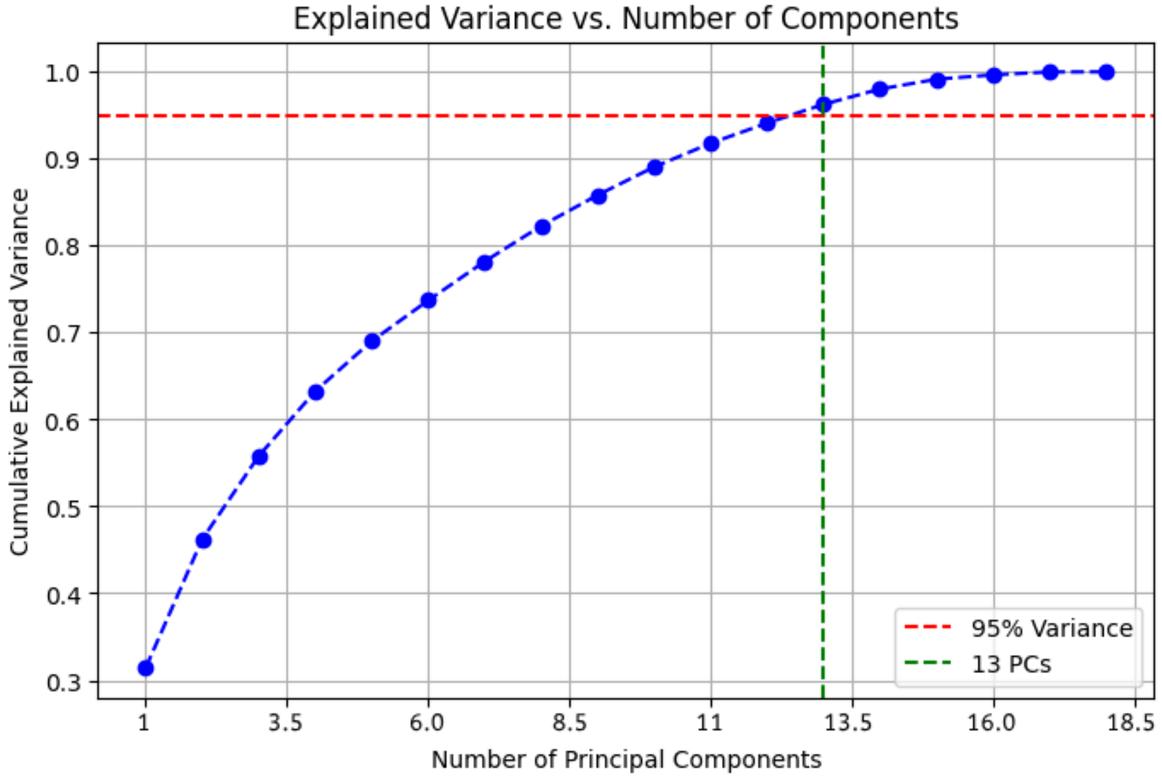
```
Removed 'percentage expenditure' (p = 0.6950)
Removed 'Year' (p = 0.4753)
Removed ' thinness 5-9 years' (p = 0.3248)
Removed 'Measles ' (p = 0.1751)
Removed 'Alcohol' (p = 0.0719)

Final Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:       Life expectancy   R-squared:                       0.804
Model:                          OLS   Adj. R-squared:                  0.803
Method:               Least Squares   F-statistic:                     801.2
Date:              Thu, 06 Feb 2025   Prob (F-statistic):               0.00
Time:                      13:49:30   Log-Likelihood:                -7811.7
No. Observations:              2754   AIC:                         1.565e+04
Df Residuals:                  2739   BIC:                         1.574e+04
Df Model:                        14
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           56.5709      0.702     80.612      0.000      55.195      57.947
Adult Mortality                 -0.0210      0.001    -24.927      0.000      -0.023      -0.019
Hepatitis B                     -0.0187      0.004     -4.773      0.000      -0.026      -0.011
 BMI                             0.0441      0.005      8.694      0.000       0.034       0.054
under-five deaths               -0.0032      0.001     -4.490      0.000      -0.005      -0.002
Polio                            0.0295      0.005      6.294      0.000       0.020       0.039
Total expenditure                0.0965      0.036      2.701      0.007       0.026       0.167
Diphtheria                       0.0459      0.005      9.389      0.000       0.036       0.055
 HIV/AIDS                       -0.4874      0.019    -25.912      0.000      -0.524      -0.451
GDP                          4.299e-05   6.83e-06      6.295      0.000    2.96e-05    5.64e-05
Population                   4.816e-09   1.76e-09      2.740      0.006    1.37e-09    8.26e-09
 thinness  1-19 years          -0.0665      0.025     -2.641      0.008      -0.116      -0.017
Income composition of resources  5.8199      0.650      8.951      0.000       4.545       7.095
Schooling                        0.6715      0.043     15.761      0.000       0.588       0.755
Status_Developing               -1.8138      0.260     -6.974      0.000      -2.324      -1.304
==============================================================================
Omnibus:                      136.314   Durbin-Watson:                   0.731
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              408.950
Skew:                          -0.192   Prob(JB):                     1.58e-89
Kurtosis:                       4.848   Cond. No.                     5.31e+08
==============================================================================
```

B. Modeling and Evaluation

PCA

- We also ran a PCA analysis to later compare performance
- We first standardised the data and then the PCA
- Using the graph we established that the 95% of variance explanation was at 13 PC
- We have a useful setting for unsupervised learning but little interpretability can be extracted.



Change in the size of the PC's

```
Shape of original dataset: (2754, 18)
Shape of PCA-transformed dataset: (2754, 13)
```

B. Modeling and Evaluation

Lasso

- Additionally we ran a Lasso regression as third and last model to test and compare
- We applied standardization and then proceeded
- The penalty was 0.0035 and it kept all the variables
- This can be interpreted as no predictor was weak enough to be shrunk to 0

Features kept by Lasso:

	Feature	Coefficient
0	Year	-0.035608
1	Adult Mortality	-2.477058
2	infant deaths	11.026263
3	Alcohol	0.261748
4	percentage expenditure	0.049585
5	Hepatitis B	-0.350191
6	Measles	-0.216974
7	BMI	0.840877
8	under-five deaths	-11.282617
9	Polio	0.628307
10	Total expenditure	0.198476
11	Diphtheria	0.908932
12	HIV/AIDS	-2.348437
13	GDP	0.583691
14	Population	0.071816
15	thinness 1-19 years	-0.390440
16	thinness 5-9 years	0.062504
17	Income composition of resources	1.105029
18	Schooling	2.084505
19	Status_Developing	-0.633624

B. Modeling and Evaluation

K-Folds Validation

- Lasso and PCA performed equally well (Mean RMSE = 4.0708), indicating similar generalization performance despite their different approaches.
- MLR had a slightly higher RMSE (4.1616), which may be attributed to residual independence violations (autocorrelation), affecting its ability to generalise effectively.
- Lasso's penalty (λ = 0.0035) was too weak to remove any predictors, reaffirming that all variables contribute meaningfully to the model, and there was not overfitting in the MLR.

B. Modeling and Evaluation

Findings

Multivariable Linear Regression Model refinement process was carried out to improve its accuracy in estimating life expectancy. We compared the initial model and the final model after removing non-significant variable: Initial Model: R^2= 0.826 vs Final Model: R^2 = 0.803

Demographic and Socioeconomic

Variable	Direction	Interpretation
Status	Negative	Developing countries tend to have less robust health systems
GDP	Positive	Higher income per capita imply better quality of life
Population	Positive	Greater infrastructure, economies of scale and diversified economies
Income Composition Resources	Positive	Higher ICOR indicates optimal utilization of available resources
Schooling	Positive	More education increases access to health, income and healthy habits

Health Outcome indicators

Variable	Direction	Interpretation
Adult Mortality	Negative	Higher rates decrease life expectancy
Under-Five Deaths	Negative	Higher infant mortality means less prolongation of life
HIV/AIDS	Negative	Reduces immunity, causes serious diseases and increases early mortality.

B. Modeling and Evaluation

Findings

Healthcare System and Immunization

Variable	Direction	Interpretation
Total Expenditure	Positive	Greater spending on health improves access, treatments and prevention, increasing longevity.
Hepatitis B	Negative	Causes chronic liver disease and cancer
Polio and Diphteria	Both Positive	(*) Quality of the Healthcare system: better epidemiological surveillance and access to medical diagnosis (**) Data bias: In countries with better health records, polio is more documented

Nutrional and Environmental Factors

Variable	Direction	Interpretation
BMI	Positive	Associated with good nutrition and higher standards of living
Thinness 1-19 years	Negative	Malnutrition and increased risk of contracting diseases

B. Modeling and Evaluation

Conclusions

1) **Unexpected Findings & Interpretative Challenges**

- **Polio** and **Diphteria** *are statistically significant and positively correlated with life expectancy, which is counterintuitive*. This could be due to data collinearity, vaccination programs, or data collection biases. Further investigation is needed.

2) **Refined Model Performance & Variable Selection**

- The final model eliminated insignificant predictors (e.g., Alcohol, Measles) while maintaining key drivers, improving robustness.
- R-squared decreased slightly (0.826 → 0.803), but the model is now more interpretable and avoids overfitting.
- Assumption violation: Durbin-Watson statistic showed some autocorrelation, whilst lower values also displayed heteroscedasticity, slightly hindering the inference capabilities of the model.

3) **Comparison of Model Approaches: PCA & Lasso**

- Lasso's penalty (λ = 0.0035) was too weak to remove any predictors, reaffirming that **all variables contribute meaningfully to the model**.
- Overall, **Lasso and PCA generalized equally well**, while MLR showed slightly poorer performance due to potential assumption violations.
- The **range** of the data was from **30 to 90 years**, so an **RMSE of +/- 4** deviates by **only 6.67%** of the total range. This suggests that our model **captures key patterns** while leaving room for improvement."

B. Modeling and Evaluation

Next Steps

1) **Regarding the current model:**

- Further validation of unexpected results (Polio and Diphteria).
- Consider how to fix the autocorrelation and heteroscedacity of the residuals -> improve quality of the inference
- Explore interaction terms or non-linear models for more complex relationships.

2) **Considering Development of new approaches:**

- Extend regression analysis to initially non-significant predictors over significant variables, for example:

$$Adult\ Mortality_i = \beta_0 + \beta_1 Alcohol_i + \beta_2 GDP_i + \varepsilon_i$$

Evaluating the significant variable Adult Mortality and its relationship with factors such as Alcohol (which cause illness, accidents and mental disorders) while controlling with GDP Variable (as a proxy of healthcare system).

- Use findings from some regression models, like the one above, to **propose specific one-shot interventions** such as health communication initiatives, vaccination campaigns or nutrition improvement programs, based on empirical evidence.

Thank you!