Exploring Key Drivers of Life Expectancy from 2000 to 2015

Zeyu Qi, Takeru Inoue, Raymond Williams

March 2025

Abstract

Life expectancy is one of the most commonly used methods to grade a nation's overall well-being. It is influenced by economic, health, and environmental factors. This study works to analyze what the most important of these factors are across many different countries and continents. Using statistical analysis, we explore the impact of factors such as GDP per capita, healthcare expenditure, CO emissions, forest area, internet usage, and access to basic drinking water services. Preliminary findings suggest that the biggest factors contributing to the life expectancy of a country have to do with a country's economy and GDP, while some milder factors include obesity rates and sanitation. These results demonstrate insights into the potential policy measures that can be put into place to improve life expectancy everywhere across the world.

1 Introduction

A country's life expectancy is the most common litmus test used by governments, media, and the general public to assess a given country's prosperity and overall well-being. It not only showcases how efficiently a society provides for the health of its people but also demonstrates the success or shortcomings of social policies, healthcare investments, and cultural norms. That being said, as we continue innovating and growing in the first world, the gap between first and third world countries only continues to grow.

This quickly growing disparity highlights the complex relationships present in the success of a nation. Often times, the life expectancy of a country comes to fruition through a complex interplay of economic developmental factors, healthcare infrastructure, and both cultural and social stability. Developed nations greatly benefit because of cutting edge medical technologies, widespread and relatively easy access to healthcare, and overall higher living standards. On the other hand, many developing countries across the world lie in stark juxtaposition to these developed countries. They struggle with with inadequate medical facilities, poor sanitation, and limited access to clean water. Factors such as GDP per capita, healthcare expenditure, environmental policies, and technological advancements play crucial roles in shaping the overall health outcomes of a nation.

With the dawn of the data age and advanced analytical tools, there is a unique opportunity to understand these disparities like never before and work towards a new data-driven future in which the goal of global health equity can be achieved. By analyzing various economic, environmental, and health indicators, we hope to gain vital insights into the key determinants of life expectancy across the world and identify potential strategies that world governments could utilize to bridge the gaps between developed and developing countries everywhere.

2 Data Sources and Acquisition

We obtained our dataset from a public repository on Kaggle titled ”Life Expectancy 2000–2015,” which consolidates data from several reputable sources, including the World Bank, UNCTAD, and Our World in Data. The dataset contains socioeconomic, demographic, and environmental indicators such as GDP per capita, CO_2 emissions, and sanitation measures. After preprocessing and one-hot encoding categorical variables, the data was prepared for analysis and modeling.

3 Proposed Analyses and Expected Outcomes

Our methodology began with exploratory data analysis (EDA), including descriptive statistics and visualizations to assess trends in life expectancy by continent, development status, and other predictors. We also examined relationships between life expectancy and variables like GDP per capita and health expenditure. Following EDA, we formulated hypotheses regarding environmental and behavioral predictors, then tested these hypotheses using regression models and classification techniques. Finally, we developed predictive models to quantify the influence of different factors on life expectancy.

4 Exploratory Data Analysis

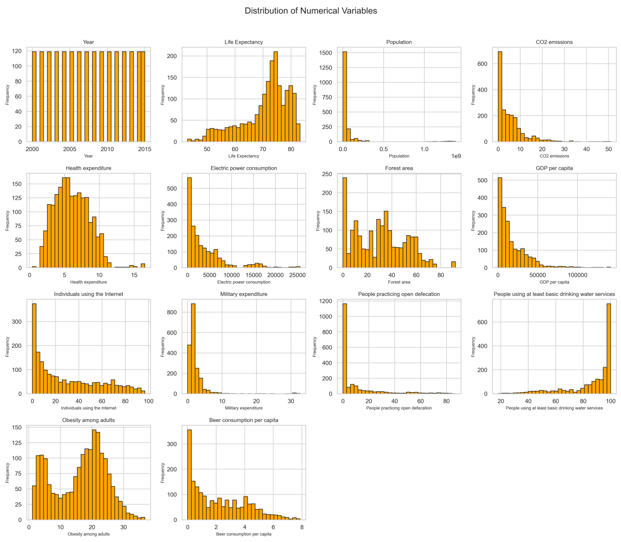


Figure 1: Distribution of numerical variables included in the life expectancy analysis. This visualization highlights skewness, outliers, and potential need for transformations in variables such as GDP per capita, CO_2 emissions, and population.

The histograms of the dataset's numeric columns reveal that several predictors exhibit heavy right skew, most notably GDP per capita, population size, and CO2 emissions, where the majority of countries fall into the lower or middle range but a few extend into very high values. Life Expectancy itself is more tightly clustered between the mid-60s and low-80s, with relatively fewer countries at the most extreme ends. Infrastructure-related variables, such as Health Expenditure and People Using at Least Basic Drinking Water Services, often center around moderate or high values, though a distinct subset of nations remain substantially below universal coverage or invest far less in healthcare. These distributions highlight the need for care with transformations, outlier identification, and the broader context in which a handful of large or wealthier nations can dominate summary statistics.

Observing how these features relate to Life Expectancy within the exploratory data analysis suggests that three general categories of features appear to have differing strengths of association. First, economic and infrastructure measures, particularly GDP per capita and Healthcare Expenditure, show strong positive ties to Life Expectancy, reflecting how higher incomes and more robust public-health investments coincide with longer average lifespans. Second, environmental and resource-based indicators, including CO2 emissions and Electric Power Consumption, point to more industrialized settings that can support higher longevity, though these same features can also reflect pollution or lifestyle factors that complicate strict cause-and-effect. Third, demographic and behavioral factors, such as Obesity or Open Defecation prevalence, offer a more mixed or complex picture: higher obesity rates can accompany wealthier contexts yet also invite increased chronic disease risks, whereas open defecation tends to be reported in areas with more limited infrastructure, which, in turn, generally aligns with shorter life expectancy. In short, economic capacity and health infrastructure emerge as the most consistent predictors of life expectancy from the standpoint of the exploratory data alone. Combining them in a multivariate setting confirms these key relationships and underscores that a country's ability to provide basic health services, mitigate preventable diseases, and foster broad economic development is closely intertwined with better health outcomes and longer lifespans.

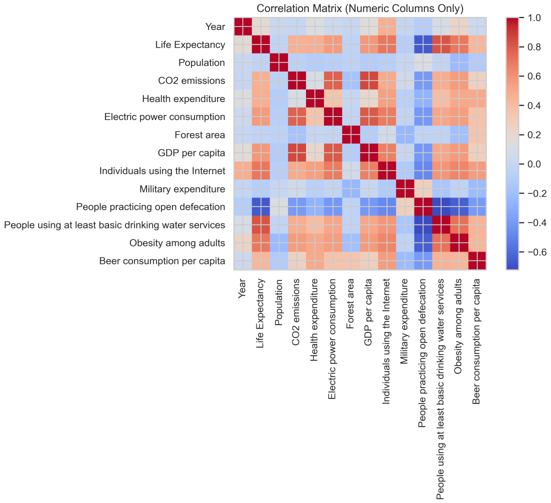


Figure 2: Correlation matrix of numeric variables included in the life expectancy dataset. Strong positive and negative correlations are observed, such as between GDP per capita and Internet usage, and between open defecation and drinking water access. Identifying such correlations helps understand multicollinearity issues and variable relationships.

In the resulting matrix, darker red or blue values (depending on the chosen colormap) indicate stronger correlations, either positive or negative. Examining these pairwise correlations reveals several notable trends. Life Expectancy tends to have a strong positive correlation with indicators reflecting better socioeconomic conditions, including GDP per capita and Health Expenditure, which underscores the role of income and public funding in shaping population health outcomes. Variables tied to industrialization, such as CO2 Emissions or Electric Power Consumption, show moderate positive correlations with Life Expectancy, suggesting that more developed infrastructure can align with greater longevity, although this connection can also reflect higher resource usage and potential environmental health impacts. Measures of improved sanitation or widespread access to clean water often correlate strongly with longer life spans, while data on open defecation rates correlates negatively with Life Expectancy, highlighting the persistence of preventable diseases in places lacking essential infrastructure. Taken together, the correlation matrix reaffirms the findings from the histograms: countries with stronger economic foundations and more robust health and infrastructure metrics typically achieve higher life expectancy.

5 Statistical Analysis

In our notebook, we examined various potential drivers of life expectancy and utilized a combination of parametric and non-parametric tests to account for the realities of our data. Initially, we considered correlation-based techniques, including Pearson's correlation, which requires normally distributed data and linear relationships. However, after conducting the Shapiro–Wilk test, we determined that most variables (such as Health Expenditure, Obesity Rates, and GDP) were non-normal, prompting us to adopt Spearman's correlation as a more appropriate non-parametric alternative. These correlation analyses revealed that higher healthcare spending tends to coincide with longer life expectancy, whereas rising obesity rates generally correlate with shorter lifespans, underscoring the profound role that both healthcare investment and lifestyle factors play in shaping national health outcomes.

To probe differences between groups, we initially tested T-tests and ANOVA, but discovered that our data sets frequently violated assumptions of normality and equal variances. Therefore, we shifted to Mann–Whitney U tests for two-group comparisons (e.g., "high" vs. "low" categories for certain metrics) and Kruskal–Wallis tests for three or more groups (e.g., comparing countries with "low," "medium," and "high" obesity rates). These non-parametric group-based methods consistently indicated statistically significant discrepancies in life expectancy across different categories, reinforcing our correlation findings. Additionally, when evaluating categorical data—such as the relationship between "Least Developed" vs. "Developed" status and water access—we relied on the Chi-Squared test. This test, which does not require normality, revealed that these development classifications were strongly linked to disparities in clean water availability.

Collectively, our hypothesis tests suggest that economic factors like health spending and GDP, along with environmental and lifestyle factors like water access and obesity prevalence, bear a strong connection to how long populations live. By carefully respecting statistical assumptions and using non-parametric methods when needed, we believe these results present a robust and nuanced look at the ways in which resource allocation, infrastructure, and public health factors operate in tandem to shape overall longevity. Even when faced with skewed distributions, we continued to uncover clear, often striking relationships that can inform policy decisions and future research on improving life expectancies worldwide. In our analyses, we also recognized that many of the features themselves—such as GDP, health expenditure, and water access—may be correlated with one another. For instance, countries that invest heavily in healthcare often also exhibit higher GDP and cleaner water infrastructure. This interdependence can create challenges in disentangling the individual effect of any single factor on life expectancy. It may also help explain why nearly all these factors appear to show significant relationships with life expectancy: if certain highly correlated variables (e.g., health expenditure and GDP) are both on the higher end for a country, they will likely reinforce one another in contributing to better public health outcomes.

6 Predictive Modeling

We developed and compared four predictive models to assess their ability to explain and predict life expectancy:

6.1 Checking Assumptions

Before finalizing our linear regression models, it is critical to ensure that key assumptions of linear regression are satisfied. The main assumptions we evaluated include:

- **Homoscedasticity:** The variance of residuals should be constant across all levels of the fitted values.

- **Normality of Residuals:** The residuals should be approximately normally distributed.

- **No Multicollinearity:** Predictors should not be highly correlated with each other.

To evaluate these assumptions, we conducted the following diagnostics:

1. **Residuals vs. Fitted Values Plot:** This plot was used to assess linearity and homoscedasticity. An even scatter of residuals around zero without a clear pattern indicates that these assumptions are likely satisfied.

2. **Histogram and Q-Q Plot of Residuals:** The histogram provides a visual check for the normality of residuals, while the Q-Q plot compares the distribution of residuals to a theoretical normal distribution. If the residuals follow the reference line in the Q-Q plot, the normality assumption is satisfied.

3. **Variance Inflation Factor (VIF):** VIF was calculated to detect multicollinearity among the predictors. A VIF value above 5 indicates potential multicollinearity issues, and a value above 10 is generally considered problematic. After evaluating VIF, variables with excessively high VIF values were considered for removal to improve model stability.

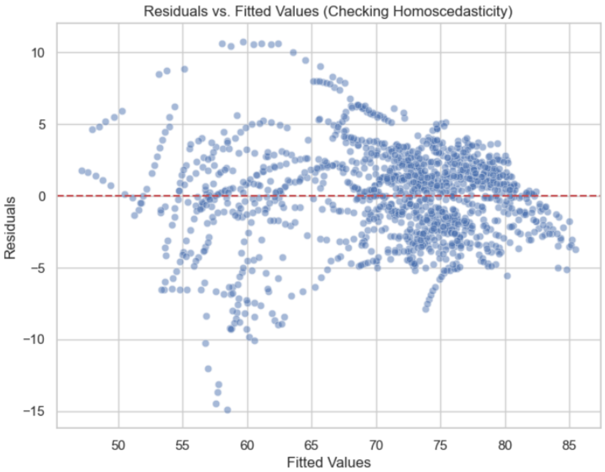


Figure 3: Residuals vs. Fitted Values plot assessing linearity and homoscedasticity.

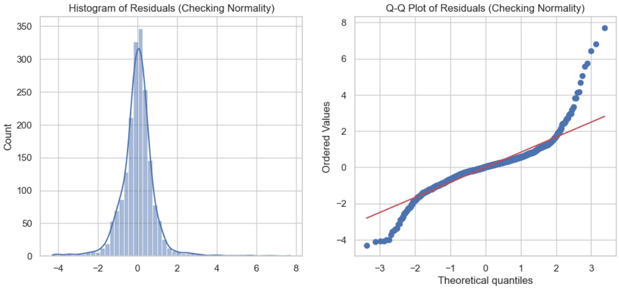


Figure 4: Histogram and Q-Q Plot of residuals assessing normality.

Together, these diagnostics confirm that our model meets the key assumptions necessary for valid inference and reliable predictions.

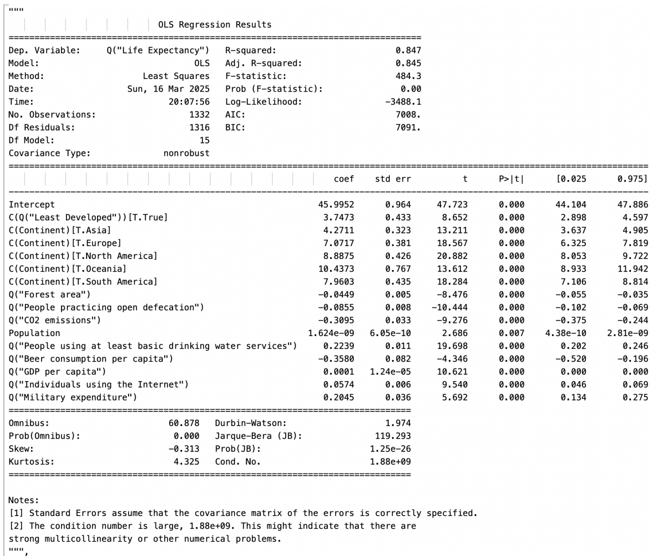
6.2 Backward Selection (BIC)

Backward Selection used the Bayesian Information Criterion (BIC) to identify a parsimonious set of predictors that balance model simplicity with predictive accuracy. Using an iterative backward elimination algorithm, less significant features such as *Electric power consumption*, *Obesity among adults*, *Health expenditure*, and *Year* were removed from the model to prevent overfitting and reduce redundancy among correlated variables.

The final model identified by BIC consists of a combination of demographic, environmental, economic, and geographical variables that are crucial in explaining variations in life expectancy across countries. The general form of the selected model is given below:

$$\begin{aligned}\text{Life Expectancy} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \\ + \sum_{i=1}^{k} \gamma_i z_i + \delta d + \varepsilon\end{aligned} \tag{1}$$

where $x_1, x_2, \ldots, x_p$ are continuous predictors such as GDP per capita and CO_2 emissions, z_i are categorical indicators for continent membership, d is a binary indicator for whether a country is classified as “Least Developed”, and ε is the error term.



```
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("Life Expectancy")   R-squared:                       0.847
Model:                              OLS   Adj. R-squared:                  0.845
Method:                   Least Squares   F-statistic:                     484.3
Date:                  Sun, 16 Mar 2025   Prob (F-statistic):               0.00
Time:                          20:07:56   Log-Likelihood:                -3488.1
No. Observations:                  1332   AIC:                             7008.
Df Residuals:                      1316   BIC:                             7091.
Df Model:                            15
Covariance Type:              nonrobust
===============================================================================================================================
                                                                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------------
Intercept                                                     45.9952      0.964     47.723      0.000      44.104      47.886
C(Q("Least Developed"))[T.True]                                3.7473      0.433      8.652      0.000       2.898       4.597
C(Continent)[T.Asia]                                           4.2711      0.323     13.211      0.000       3.637       4.905
C(Continent)[T.Europe]                                         7.0717      0.381     18.567      0.000       6.325       7.819
C(Continent)[T.North America]                                  8.8875      0.426     20.882      0.000       8.053       9.722
C(Continent)[T.Oceania]                                       10.4373      0.767     13.612      0.000       8.933      11.942
C(Continent)[T.South America]                                  7.9603      0.435     18.284      0.000       7.106       8.814
Q("Forest area")                                              -0.0449      0.005     -8.476      0.000      -0.055      -0.035
Q("People practicing open defecation")                        -0.0855      0.008    -10.444      0.000      -0.102      -0.069
Q("CO2 emissions")                                            -0.3095      0.033     -9.276      0.000      -0.375      -0.244
Population                                                  1.624e-09   6.05e-10      2.686      0.007    4.38e-10    2.81e-09
Q("People using at least basic drinking water services")       0.2239      0.011     19.698      0.000       0.202       0.246
Q("Beer consumption per capita")                              -0.3580      0.082     -4.346      0.000      -0.520      -0.196
Q("GDP per capita")                                            0.0001   1.24e-05     10.621      0.000       0.000       0.000
Q("Individuals using the Internet")                            0.0574      0.006      9.540      0.000       0.046       0.069
Q("Military expenditure")                                      0.2045      0.036      5.692      0.000       0.134       0.275
==============================================================================
Omnibus:                       60.878   Durbin-Watson:                   1.974
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              119.293
Skew:                          -0.313   Prob(JB):                     1.25e-26
Kurtosis:                       4.325   Cond. No.                     1.88e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.88e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
""",
```

Figure 5: OLS Regression Results from Backward Selection (BIC).

Model Performance

This linear model achieved a test set R^2 of 0.855, indicating that approximately 85.5% of the variability in life expectancy can be explained by the selected predictors. The Root Mean Squared Error (RMSE) of 3.24 years suggests that, on average, the model's predictions deviate from actual life expectancy values by about 3.24 years, which is reasonably accurate for a global cross-sectional dataset.

Interpretation of Key Predictors

- **Intercept (45.9952):** The intercept suggests that for a reference country (likely a non-least-developed country in Africa with all numeric variables at zero), the baseline life expectancy is approximately 46 years. While not practically observable (since many predictors can't be zero), it provides a useful baseline for interpretation.

- **GDP per capita (0.0001):** For every $1 increase in GDP per capita, life expectancy is expected to increase by 0.0001 years, holding other variables constant. This indicates that higher national wealth is associated with longer life spans. For instance, a $10,000 increase in GDP per capita would be associated with a 1-year increase in life expectancy.

- **CO_2 emissions (-0.3095):** Each additional metric ton per capita of CO_2 emissions is associated with a reduction of approximately 0.31 years in life expectancy, highlighting the negative health impacts of pollution.

- **Forest area (-0.0449):** Each additional percentage point increase in forest area is associated with a decrease of 0.0449 years in life expectancy. This negative association may reflect that countries with large forest areas might be less urbanized and have weaker healthcare infrastructure.

- **Individuals using the Internet (0.0574):** For every 1% increase in individuals using the internet, life expectancy increases by approximately 0.057 years (about 21 days). This suggests that internet access enhances health education and access to healthcare resources.

- **People using at least basic drinking water services (0.2239):** For every 1% increase in access to basic drinking water services, life expectancy rises by approximately 0.224 years. This strong effect confirms the critical role of clean water in public health.

- **People practicing open defecation (-0.0855):** Each additional percentage point of the population practicing open defecation is associated with a reduction of about 0.086 years in life expectancy. This reflects the significant negative impact of poor sanitation on population health.

- **Military expenditure (0.2045):** A 1% increase in military expenditure (as a percentage of GDP) is associated with a 0.205-year increase in life expectancy. Although surprising, this may reflect underlying governmental capacity or order, but it requires further investigation.

- **Beer consumption per capita (-0.3580):** Each additional liter of beer consumed per person annually is associated with a decrease of about 0.358 years in life expectancy. This finding is consistent with the adverse health effects of excessive alcohol consumption.

- **Population (1.624e-09):** Each additional person added to the country's population increases life expectancy by a negligible 0.000000001624 years, suggesting that population size itself has a minimal direct effect when controlling for other factors.

- **Continent (categorical):** Regional effects are substantial:

  - **Asia (4.2711):** Being in Asia is associated with a 4.27-year higher life expectancy compared to Africa (the reference).
  - **Europe (7.0717):** Being in Europe increases life expectancy by about 7.07 years.
  - **North America (8.8875):** North America is associated with an 8.89-year increase in life expectancy.
  - **Oceania (10.4373):** Oceania is associated with a life expectancy increase of approximately 10.44 years.
  - **South America (7.9603):** Life expectancy in South America is about 7.96 years higher than in Africa.

- **Least Developed (3.7473):** Surprisingly, being classified as a Least Developed Country (binary indicator) is associated with a 3.75-year increase in life expectancy, compared to non-least-developed. This may reflect an interaction effect or multicollinearity issue needing further exploration.

Overall, the backward selection process led to a well-specified, interpretable model that balances complexity and predictive performance. The selected predictors emphasize the crucial roles of economic status, sanitation, and access to resources in determining life expectancy. Notably, the inclusion of categorical continent indicators and Least Developed status underscores the importance of geographical and developmental context. These findings suggest that policy interventions aimed at improving clean water access, reducing pollution, and fostering economic growth could significantly enhance population health outcomes.

6.3 LASSO Regression (LassoCV)

LASSO (Least Absolute Shrinkage and Selection Operator) Regression applies L1 regularization to shrink some coefficients toward zero, thereby performing variable selection and regularization simultaneously. By penalizing the absolute size of the coefficients, LASSO aims to prevent overfitting while simplifying the model by excluding irrelevant or weak predictors.

Through cross-validation, LASSO identified an optimal regularization parameter λ that balances model complexity with predictive accuracy. The final model achieved a test set R^2 of 0.85 and an RMSE of 4.23 years. Although slightly less accurate than the Backward Selection model in terms of R^2 and RMSE, LASSO produced a more parsimonious model by automatically shrinking less important predictors to zero, thus improving interpretability and reducing potential multicollinearity.

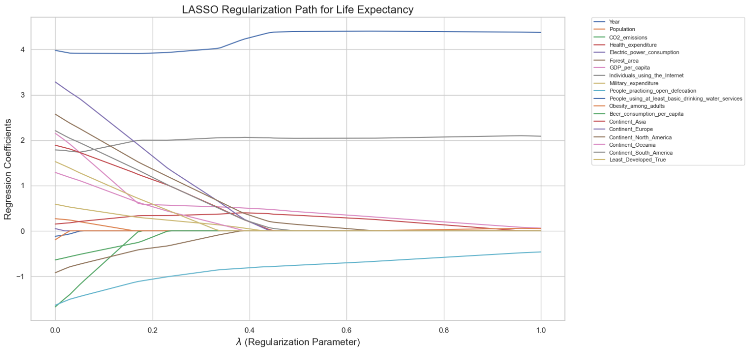
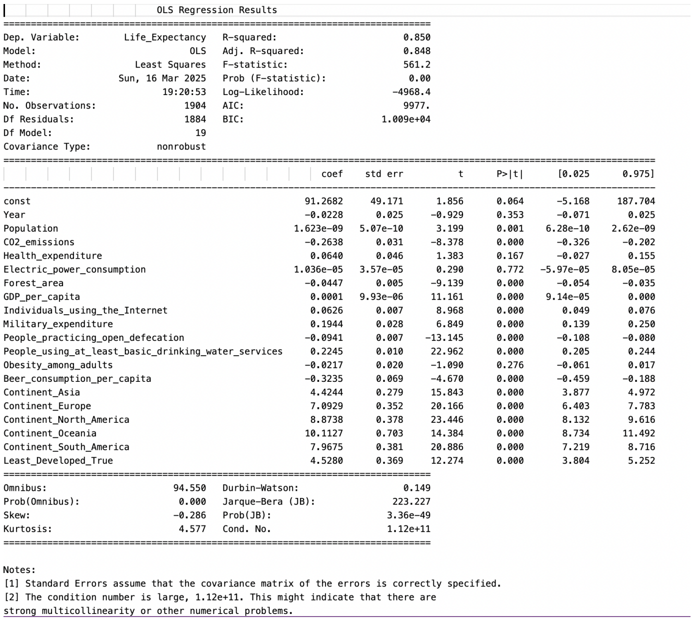


Figure 6: LASSO Regularization Path for Life Expectancy, showing coefficient shrinkage as λ increases.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life_Expectancy   R-squared:                       0.850
Model:                            OLS   Adj. R-squared:                  0.848
Method:                 Least Squares   F-statistic:                     561.2
Date:                Sun, 16 Mar 2025   Prob (F-statistic):               0.00
Time:                        19:20:53   Log-Likelihood:                -4968.4
No. Observations:                1904   AIC:                             9977.
Df Residuals:                    1884   BIC:                         1.009e+04
Df Model:                          19
Covariance Type:            nonrobust
=======================================================================================================================
                                                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------------
const                                                  91.2682     49.171      1.856      0.064      -5.168     187.704
Year                                                   -0.0228      0.025     -0.929      0.353      -0.071       0.025
Population                                           1.623e-09   5.07e-10      3.199      0.001    6.28e-10    2.62e-09
CO2_emissions                                          -0.2638      0.031     -8.378      0.000      -0.326      -0.202
Health_expenditure                                      0.0640      0.046      1.383      0.167      -0.027       0.155
Electric_power_consumption                           1.036e-05   3.57e-05      0.290      0.772   -5.97e-05    8.05e-05
Forest_area                                            -0.0447      0.005     -9.139      0.000      -0.054      -0.035
GDP_per_capita                                          0.0001   9.93e-06     11.161      0.000    9.14e-05       0.000
Individuals_using_the_Internet                          0.0626      0.007      8.968      0.000       0.049       0.076
Military_expenditure                                    0.1944      0.028      6.849      0.000       0.139       0.250
People_practicing_open_defecation                      -0.0941      0.007    -13.145      0.000      -0.108      -0.080
People_using_at_least_basic_drinking_water_services     0.2245      0.010     22.962      0.000       0.205       0.244
Obesity_among_adults                                   -0.0217      0.020     -1.090      0.276      -0.061       0.017
Beer_consumption_per_capita                            -0.3235      0.069     -4.670      0.000      -0.459      -0.188
Continent_Asia                                          4.4244      0.279     15.843      0.000       3.877       4.972
Continent_Europe                                        7.0929      0.352     20.166      0.000       6.403       7.783
Continent_North_America                                 8.8738      0.378     23.446      0.000       8.132       9.616
Continent_Oceania                                      10.1127      0.703     14.384      0.000       8.734      11.492
Continent_South_America                                 7.9675      0.381     20.886      0.000       7.219       8.716
Least_Developed_True                                    4.5280      0.369     12.274      0.000       3.804       5.252
==============================================================================
Omnibus:                       94.550   Durbin-Watson:                   0.149
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              223.227
Skew:                          -0.286   Prob(JB):                     3.36e-49
Kurtosis:                       4.577   Cond. No.                     1.12e+11
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.12e+11. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 7: OLS Regression Results using predictors selected by LASSO.

Selected Predictors and Interpretation

The LASSO procedure retained only a subset of the available predictors, focusing on those with the strongest associations to life expectancy. Below are the key predictors identified by LASSO and their interpretations:

- **Intercept (91.2682):** Represents the estimated baseline life expectancy for a country when all numeric predictors are zero and for the reference categories (likely Africa and not least developed). Although not practically observable, it serves as the model's baseline.

- **GDP per capita (0.0001):** For each additional dollar increase in GDP per capita, life expectancy increases by approximately 0.0001 years. For instance, a $10,000 increase in GDP per capita is associated with a 1-year increase in life expectancy, emphasizing the impact of economic wealth on population health.
- **CO_2 emissions (-0.2638):** Each additional metric ton of CO_2 emissions per capita is associated with a reduction of approximately 0.26 years in life expectancy, reflecting the negative health consequences of pollution and environmental degradation.
- **Forest area (-0.0447):** Each 1% increase in forest area correlates with a 0.045-year reduction in life expectancy. This may capture underlying factors such as low infrastructure development in heavily forested areas.
- **Individuals using the Internet (0.0626):** Each 1% increase in internet usage leads to a 0.063-year (approximately 23 days) increase in life expectancy, possibly due to better access to health information and services.
- **People using at least basic drinking water services (0.2245):** Each 1% increase in access to drinking water services boosts life expectancy by approximately 0.225 years, confirming the vital role of clean water for public health.
- **People practicing open defecation (-0.0941):** Each 1% increase in the population practicing open defecation reduces life expectancy by about 0.094 years, highlighting the detrimental effect of poor sanitation.
- **Beer consumption per capita (-0.3235):** Each additional liter of beer consumed per person annually is associated with a reduction of about 0.32 years in life expectancy, suggesting negative health effects from alcohol consumption.
- **Military expenditure (0.1944):** Each 1% increase in military expenditure (as a percentage of GDP) corresponds to a 0.194-year increase in life expectancy. This may act as a proxy for broader governmental capacity and security but requires more nuanced investigation.
- **Population (1.623e-09):** Each additional person in a country's population has a negligible positive effect of approximately 0.000000001623 years on life expectancy, suggesting minimal direct influence when adjusting for other factors.
- **Continent (categorical):** Strong regional effects highlight disparities in life expectancy:
  - **Asia (4.4244):** Life expectancy is 4.42 years higher than in Africa.
  - **Europe (7.0929):** Europeans enjoy a 7.09-year higher life expectancy.
  - **North America (8.8738):** North Americans live 8.87 years longer.
  - **Oceania (10.1127):** Life expectancy is 10.11 years higher in Oceania.
  - **South America (7.9675):** South Americans live nearly 7.97 years longer.
- **Least Developed (4.5280):** Being classified as a Least Developed Country surprisingly adds about 4.53 years to life expectancy, which may reflect statistical or interaction effects and warrants further exploration.
- **Other insignificant variables:**
  - **Year (-0.0228):** Insignificant, suggesting no strong linear time trend in life expectancy during this period after adjusting for other variables.
  - **Electric power consumption (1.036e-05):** Insignificant, suggesting no direct association when controlling for other factors.
  - **Health expenditure (0.0640):** Not statistically significant (p=0.167), which is surprising and may indicate overlap with other predictors.
  - **Obesity among adults (-0.0217):** Also not significant (p=0.276), despite expected health implications.

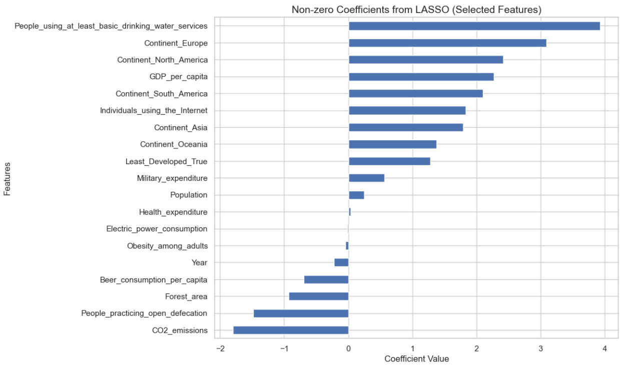


Figure 8: Non-zero Coefficients from LASSO Regression (Selected Predictors).

Implications

Although LASSO Regression achieved a slightly lower R^2 compared to Backward Selection (0.85 vs. 0.855), its capacity to perform automatic feature selection makes it a powerful tool for identifying core drivers of life expectancy. The LASSO model's simplified predictor set enhances interpretability and reduces potential overfitting, making it a suitable choice when balancing predictive performance with model simplicity is a priority.

In particular, the LASSO model confirms the importance of economic factors (GDP per capita), environmental conditions (CO_2 emissions, forest area), and public health infrastructure (sanitation and clean water access) as major determinants of life expectancy. These findings reinforce the necessity of integrated health, environmental, and economic policies to improve population health outcomes globally.

6.4 K-Nearest Neighbors (KNN) Regression

K-Nearest Neighbors (KNN) Regression is a non-parametric method that predicts a target value based on the average of its k nearest neighbors in the feature space. Unlike linear models, KNN does not assume any specific functional form between predictors and the outcome.

In our analysis, KNN Regression achieved an exceptionally high test set R^2 of 0.995 and an extremely low RMSE of 0.61 years, suggesting near-perfect prediction accuracy. While these metrics may initially appear favorable, such high performance typically raises concerns about **overfitting**, where the model fits the training data (and potentially test data) too closely, including noise, rather than capturing generalizable patterns.

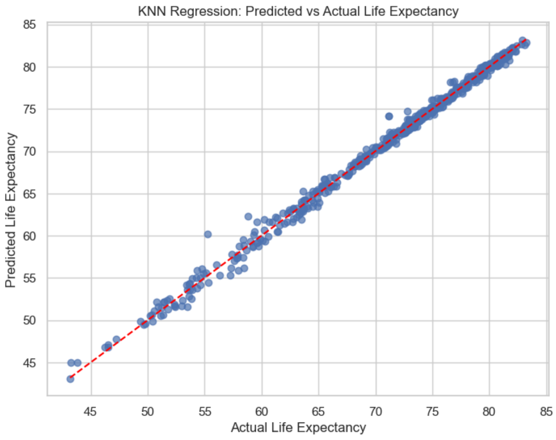


Figure 9: KNN Regression: Predicted vs Actual Life Expectancy. The strong alignment along the diagonal indicates high predictive accuracy but may also suggest overfitting.

Model Performance and Concerns

- **Test R^2:** 0.995 — Indicates that 99.5% of the variance in life expectancy is explained by KNN predictions.
- **Test RMSE:** 0.61 years — The model's average prediction error is only about 0.61 years.

Interpretation and Recommendation

Although KNN Regression shows remarkable predictive power, such extreme accuracy is uncommon in real-world datasets and may imply that the model is overfitting — capturing noise rather than underlying relationships. This is further suggested by the near-perfect alignment of predicted vs. actual life expectancy in Figure 9. Future improvements can include:

- **Hyperparameter tuning:** Adjust the value of k to increase generalization and reduce overfitting. Larger k typically smoothens predictions and mitigates noise-fitting.
- **Cross-validation:** Use cross-validation to select an optimal k that balances bias and variance.
- **Feature selection:** Consider reducing dimensionality or focusing on the most relevant predictors to simplify the feature space.

While KNN demonstrates that life expectancy is highly predictable given the dataset, future iterations should aim to develop models that better generalize to unseen data without overfitting.

6.5 Logistic Regression (Binary Classification)

To explore classification performance, life expectancy was transformed into a binary variable, categorizing countries into *high* and *low* life expectancy groups based on the median split. A Logistic Regression model was employed to predict these categories using socio-economic and environmental features. The model achieved an impressive **accuracy of 89.86%** and a **ROC AUC of 0.9624**, indicating excellent discriminatory power between high and low life expectancy countries. The high AUC suggests the model is capable of distinguishing well between the two groups even in the presence of overlapping feature distributions.

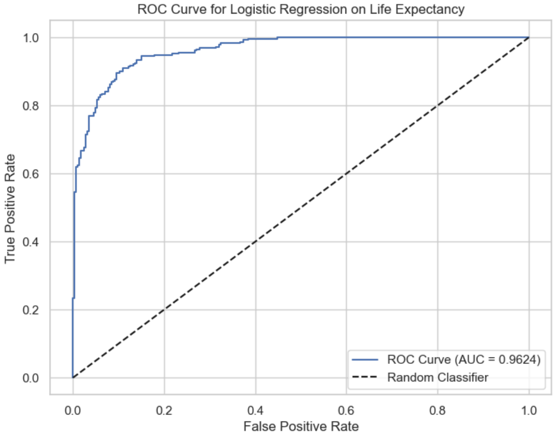


Figure 10: ROC Curve for Logistic Regression on Life Expectancy Classification (AUC = 0.9624). A steep rise and high AUC reflect strong classifier performance.

Model Performance Summary

- **Accuracy:** 89.86% — High proportion of correctly classified cases.
- **ROC AUC:** 0.9624 — Excellent discrimination ability.
- **Confusion Matrix:**

	Predicted Low	Predicted High
Actual Low	249	32
Actual High	26	265

Interpretation and Insights

- The confusion matrix shows a high number of true positives and true negatives, indicating that the model is effective at distinguishing between both classes.
- The relatively low number of false positives (32) and false negatives (26) further supports the model's precision and recall.
- High ROC AUC reflects a robust ability to rank countries correctly by the likelihood of having high life expectancy.

Recommendations

Given the strong performance, Logistic Regression serves as a useful and interpretable model for policymakers to identify factors associated with higher life expectancy. Future work could explore alternative thresholds to optimize sensitivity and specificity depending on policy goals, and assess how this model generalizes to data from different time periods.

6.6 Model Comparison Summary

To evaluate and compare the performance of our predictive models, we summarize their results in Table 1. Each model was assessed on the test set using appropriate metrics. For regression models, we report R^2 and Root Mean Squared Error (RMSE), while for the classification model (Logistic Regression), we report Accuracy and the Area Under the ROC Curve (AUC).

Table 1: Comparison of Predictive Model Performance

Model	**Type**	**Test R^2**	**Test RMSE**	**Accuracy**	**ROC AUC**
Backward Selection (BIC)	Linear Regression	0.8550	3.2364	–	–
LASSO Regression (LassoCV)	Linear Regression	0.8500	4.2300	–	–
KNN Regression	Non-parametric	0.9949	0.6102	–	–
Logistic Regression	Classification	–	–	0.8986	0.9624

Interpretation: Among the regression models, K-Nearest Neighbors (KNN) achieved the highest R^2 and lowest RMSE, suggesting excellent fit. However, this may indicate overfitting due to KNN's sensitivity to local patterns. Backward Selection and LASSO Regression both demonstrated strong predictive power while maintaining interpretability, with R^2 values above 0.85. Notably, Backward Selection slightly outperformed LASSO in both R^2 and RMSE, though at the cost of a potentially more complex model.

For classification purposes, Logistic Regression effectively distinguished between countries with high and low life expectancy, achieving nearly 90% accuracy and an impressive ROC AUC of 0.9624, indicating excellent discriminatory capability.

Overall, Backward Selection and LASSO offer interpretable models with strong performance, while KNN excels in predictive accuracy but may overfit.

Logistic Regression provides a robust classifier for binary life expectancy categorization.

7 Conclusion and Future Work

This study explored key socio-economic, environmental, and infrastructural determinants of life expectancy across 119 countries from 2000 to 2015. Using a combination of statistical models, including **Backward Selection**, **LASSO Regression**, **K-Nearest Neighbors (KNN)**, and **Logistic Regression**, we identified important predictors and evaluated their impacts on life expectancy.

The **Backward Selection model**, optimized via BIC, provided a parsimonious and interpretable linear model with an R^2 of **0.855** and RMSE of **3.24 years**. It highlighted that **GDP per capita**, **access to basic drinking water services**, **Internet usage**, and **sanitation** (e.g., low rates of open defecation) are strong positive contributors to life expectancy. In contrast, factors like **CO_2 emissions**, **forest area**, and **beer consumption per capita** were negatively associated. Significant regional effects were observed, with countries in **Europe**, **North America**, and **Oceania** experiencing significantly higher life expectancy than the reference category (likely Africa).

LASSO Regression, with an R^2 of **0.850** and RMSE of **4.23 years**, produced a slightly less accurate but sparser model. It automatically eliminated weak predictors such as **Electric power consumption**, **Obesity among adults**, and **Health expenditure**, suggesting that these factors may have overlapping effects with other stronger variables.

The **KNN Regression** model achieved a remarkably high R^2 of **0.995** and low RMSE of **0.61 years**, suggesting possible **overfitting** to the training data rather than generalizable insights, given the complexity and flexibility of KNN for capturing localized patterns.

In the classification task, **Logistic Regression** was employed to distinguish between *high* and *low* life expectancy countries, based on a median split. It achieved an accuracy of **89.86%** and a **ROC AUC** of **0.9624**, demonstrating excellent ability to separate the two groups using socio-economic and environmental predictors. This reinforces the conclusion that life expectancy can be effectively modeled and predicted using key structural indicators.

Policy Implications

Our findings emphasize that improvements in:

- **Economic development** (GDP per capita),
- **Access to clean drinking water**,
- **Internet access for health literacy**,
- and **Sanitation improvements** (reducing open defecation)

are crucial levers to enhance life expectancy globally. Conversely, managing environmental risks like **CO_2 emissions** and addressing harmful behavioral factors such as **alcohol consumption** could prevent unnecessary reductions in life span.

Future Work

Future research directions could include:

- Incorporating **longitudinal analysis** to examine time-dependent effects and causal relationships.
- Exploring **interaction terms** (e.g., between GDP and sanitation) to uncover synergistic effects.
- Applying advanced **machine learning models** (e.g., Random Forests, Gradient Boosting) for better non-linear modeling and feature importance estimation.
- Conducting region-specific studies to tailor insights and policy recommendations for specific continents or development levels.
- Further investigating counterintuitive results such as the positive association between military expenditure and life expectancy.

In summary, three most influential features are **GDP per capita**, **CO_2 emissions**, and **access to clean water**. GDP per capita is the most important feature because it encapsulates a country's ability to provide essential services that drive life expectancy: healthcare, education, sanitation, and economic stability. Its strong, consistent association with life expectancy highlights its critical role in public health. This study highlights the multifactorial nature of life expectancy and provides evidence-based insights for policymakers aiming to improve public health outcomes globally.

8 Relation to Project Proposal

Our final report deviates from the original proposal because the initial data sources, while intriguing, proved overly messy and offered limited information. The lack of multi-year data also meant that our original focus on heart disease and BMI distribution was too narrow, which led us to find a more robust and comprehensive dataset. By shifting to a dataset that spans multiple years and provides richer information on health and wellbeing, we were able to examine life expectancy through a broader lens. Although this pivot meant departing from our first idea, it still aligns with our goal of exploring key health indicators. In this way, the final project remains connected to the proposal's overarching theme of examining public health trends, even though we needed a better dataset to draw stronger insights.

9 Citations

- vrec99. (2023). Life Expectancy (2000–2015) [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/vrec99/life-expectancy-2000-2015
- ChatGPT
- UCSD Math189 Course Materials (Homework and Lectures)