Global Life Expectancy Modeling (2005–2015)

Hamza Mannai

11/20/2025

Introduction

Life expectancy is a key indicator of a country's overall development and social well-being. Understanding the factors that influence it allows governments, researchers, and policymakers to design more effective health, social, and economic strategies.

In this project, we analyze a World Bank–style dataset (2005–2015) containing a set of global development indicators. The main variables considered include:

- GDP per capita,
- Secondary education enrollment,
- Unemployment rate,
- Under-5 mortality rate,
- Income inequality (Gini index),
- Trade openness.

The objective is to build supervised machine learning models capable of predicting life expectancy and identifying the variables that most significantly influence it. Several models are tested and evaluated, including:

- A simple baseline summation model,
- A Generalized Linear Model (GLM) using a single year of data,
- A GLM using the full panel dataset (2005–2015).

This report follows the standard scientific structure: **introduction, methods, analysis, results, conclusion, and references**.

#Data Description

```
data_raw <- read.csv("Data.csv", check.names = FALSE, na.strings = c("..", "", "NA"))
```

We select the years 2005–2015, convert numeric columns, rename variables, and remove missing entries.

```r
panel_data <- data_raw %>%
  filter(stringr::str_detect(Time, "^[0-9]{4}$")) %>%
  mutate(Time = as.integer(as.character(Time))) %>%
  filter(Time >= 2005, Time <= 2015) %>%
  mutate(across(
    c(
      `Life expectancy at birth, total (years) [SP.DYN.LE00.IN]`,
      `GDP per capita, PPP (constant 2021 international $) [NY.GDP.PCAP.PP.KD]`,
      `School enrollment, secondary (% gross) [SE.SEC.ENRR]`,
      `Unemployment, total (% of total labor force) (national estimate) [SL.UEM.TOTL.NE.ZS]`,
      `Mortality rate, under-5 (per 1,000 live births) [SH.DYN.MORT]`,
      `Trade (% of GDP) [NE.TRD.GNFS.ZS]`,
      `Gini index [SI.POV.GINI]`
    ),
    as.numeric
  )) %>%
  rename(
    life_exp = `Life expectancy at birth, total (years) [SP.DYN.LE00.IN]`,
    gdp_ppp = `GDP per capita, PPP (constant 2021 international $) [NY.GDP.PCAP.PP.KD]`,
    sec_enrol = `School enrollment, secondary (% gross) [SE.SEC.ENRR]`,
    unemp_total = `Unemployment, total (% of total labor force) (national estimate) [SL.UEM.TOTL.NE.ZS]`,
    u5_mort = `Mortality rate, under-5 (per 1,000 live births) [SH.DYN.MORT]`,
    trade = `Trade (% of GDP) [NE.TRD.GNFS.ZS]`,
    gini = `Gini index [SI.POV.GINI]`
  ) %>%
  drop_na(life_exp, gdp_ppp, sec_enrol, unemp_total, u5_mort, trade, gini)

# Create 2015 subset
data_2015 <- panel_data %>% filter(Time == 2015)

# Show data structure
cat("Panel data dimensions:", dim(panel_data)[1], "rows x", dim(panel_data)[2], "columns\n")
```

```
## Panel data dimensions: 708 rows x 32 columns
```

```r
cat("2015 data dimensions:", dim(data_2015)[1], "rows x", dim(data_2015)[2], "columns\n")
```

```
## 2015 data dimensions: 68 rows x 32 columns
```

```r
cat("\nFirst few rows of 2015 data:\n")
```

```
##
## First few rows of 2015 data:
```

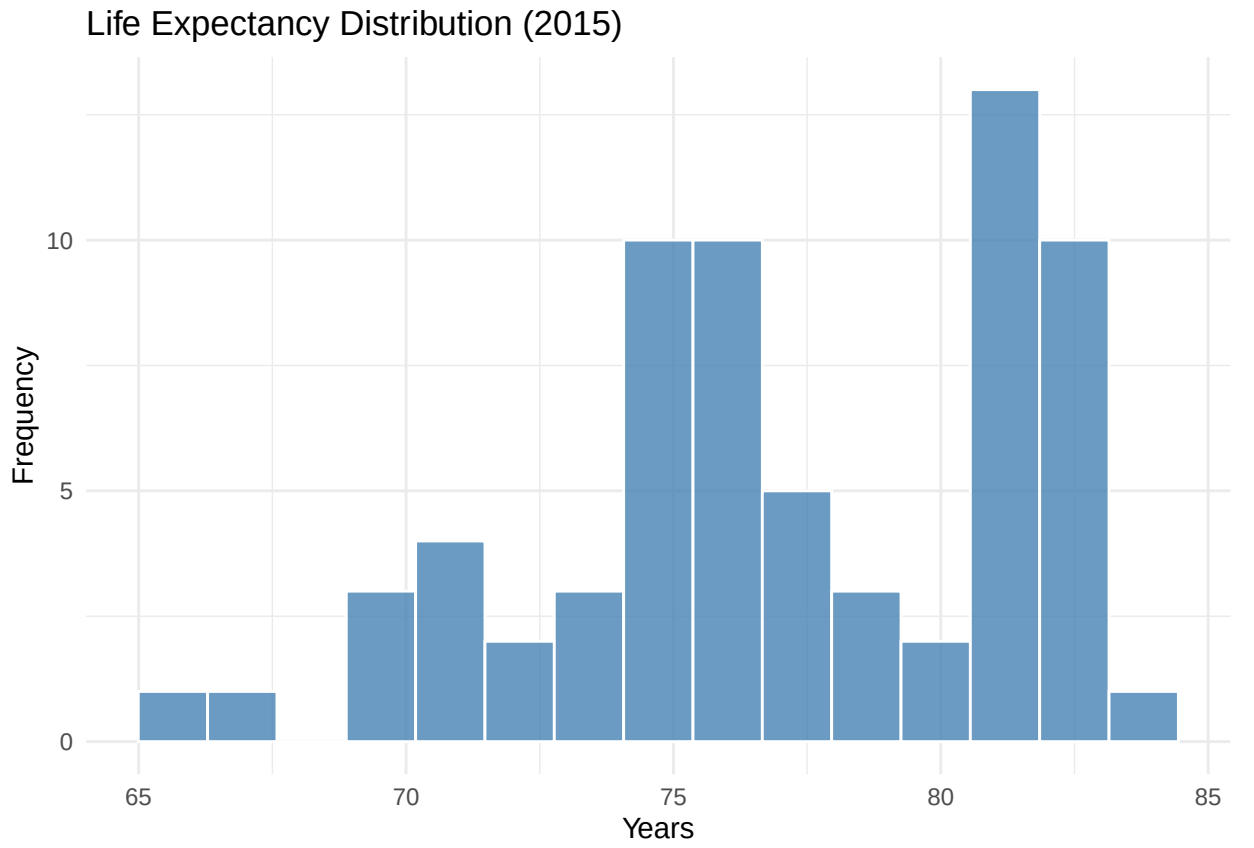
```r
head(data_2015) %>% select(Time, `Country Name`, life_exp, gdp_ppp, sec_enrol, u5_mort)
```

```
##   Time Country Name life_exp   gdp_ppp sec_enrol u5_mort
## 1 2015      Albania 78.35800 13157.089 102.44073     9.6
## 2 2015      Armenia 74.87073 13112.760  86.44214    14.6
## 3 2015      Austria 81.19024 61025.999 101.41158     3.7
## 4 2015      Belarus 73.62439 25513.213 102.89573     4.1
## 5 2015      Belgium 80.99268 57534.397 164.07982     4.1
## 6 2015      Bolivia 67.20500  9141.928  90.61390    31.8
```

The table above provides a quick overview of the cleaned panel dataset from 2005 to 2015. We have numeric variables for life expectancy, GDP, education, unemployment, child mortality, trade, and income inequality. Observing the data types ensures all variables are ready for regression modeling. Missing values have been removed to avoid errors in predictions.

#Exploratory Data Analysis ##Life Expectancy Distribution (2015)

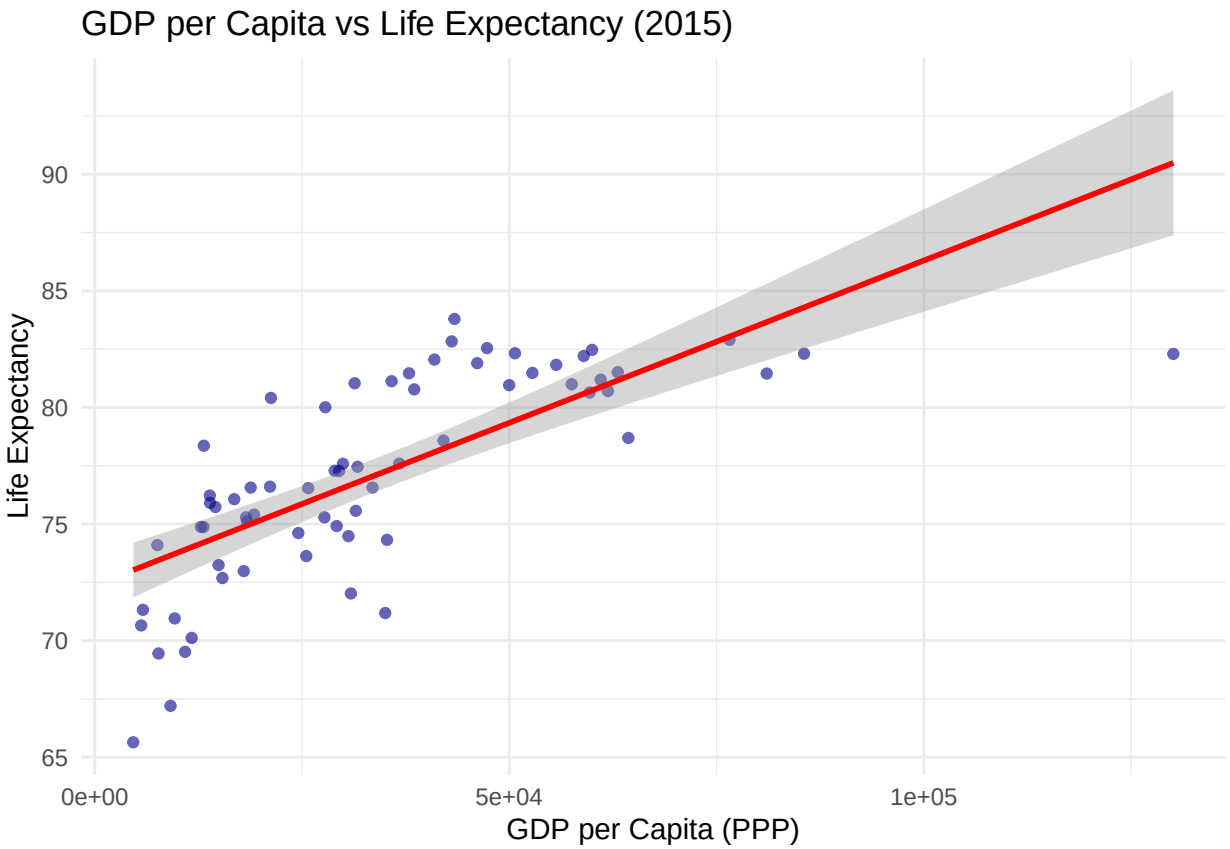
```
ggplot(data_2015, aes(x = life_exp)) +
  geom_histogram(fill="steelblue", color="white", bins=15, alpha=0.8) +
  labs(title="Life Expectancy Distribution (2015)", x="Years", y="Frequency") +
  theme_minimal()
```



This histogram shows the distribution of life expectancy across countries in 2015. Most countries cluster between 65 and 80 years, with few outliers below 50 years. For example, countries with recent conflicts or limited healthcare infrastructure may fall into the lower range, highlighting disparities in global health outcomes. The distribution is approximately normal but slightly left-skewed.

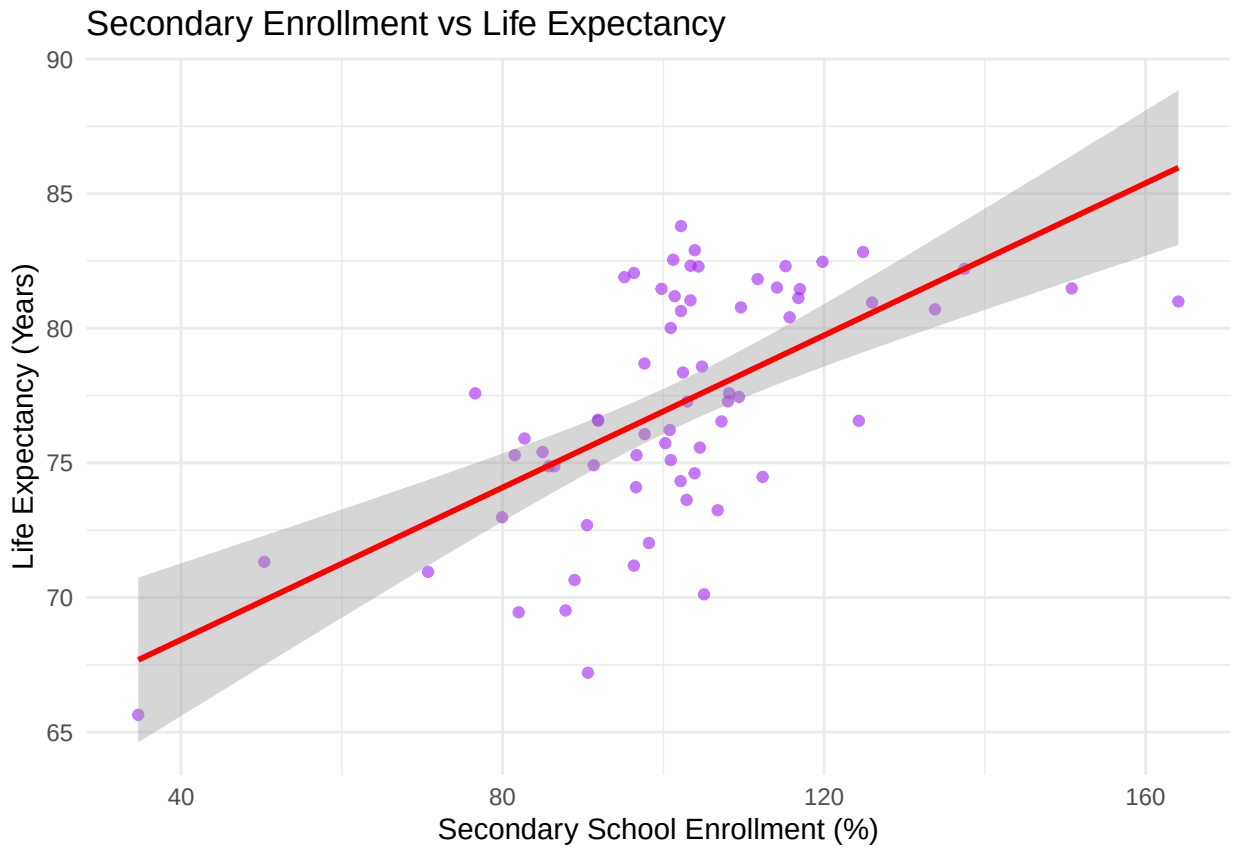
##GDP vs Life Expectancy

```
ggplot(data_2015, aes(x = gdp_ppp, y = life_exp)) +
  geom_point(color="darkblue", alpha=0.6) +
  geom_smooth(method="lm", se=TRUE, color="red") +
  labs(title="GDP per Capita vs Life Expectancy (2015)",
       x="GDP per Capita (PPP)", y="Life Expectancy") +
  theme_minimal()
```



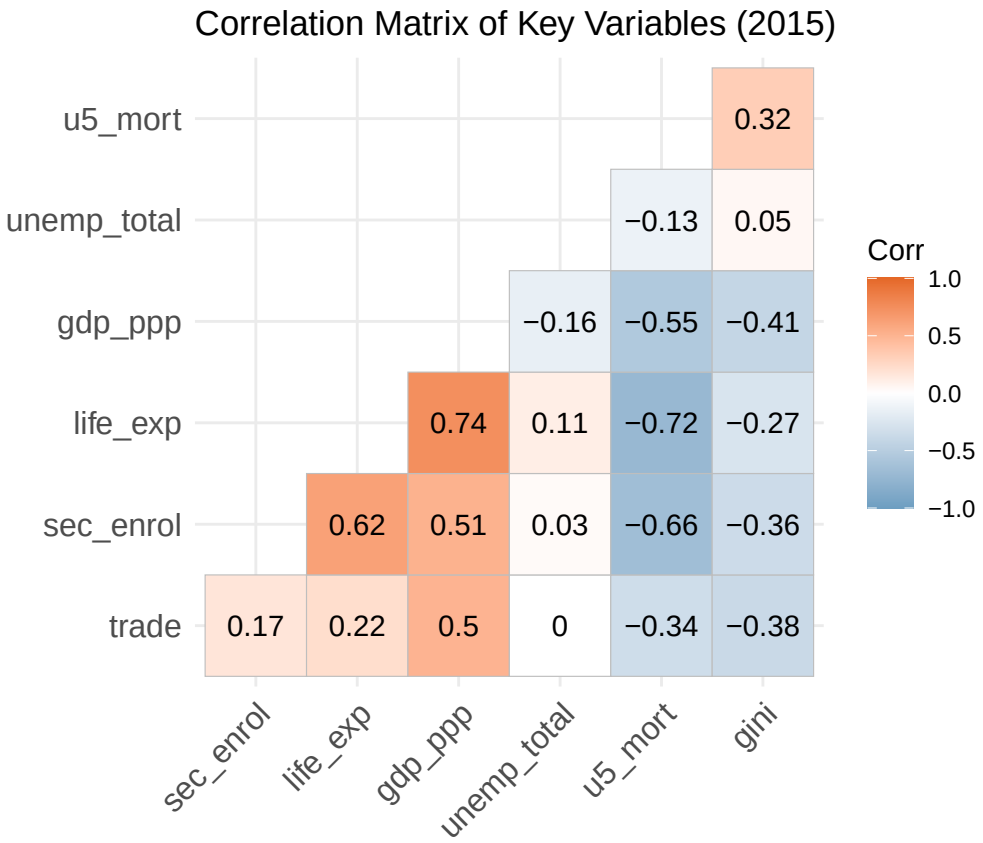
The scatter plot indicates a positive correlation between GDP per capita and life expectancy. Countries with higher wealth tend to have better health infrastructure, nutrition, and access to education, which contributes to longer lifespans. For instance, high-GDP countries like Norway or Switzerland are above 80 years, while lower-GDP countries may be below 60 years. The linear trend line confirms this strong relationship.

##Education vs Life Expectancy



This plot shows that countries with higher secondary school enrollment tend to have higher life expectancy. Education improves health awareness, employment opportunities, and economic development. For example, countries with >90% secondary enrollment usually enjoy life expectancy above 75 years, while countries with lower enrollment struggle to reach that level.

##Correlation Heatmap (2015)



The correlation heatmap identifies relationships between variables. Life expectancy is strongly positively correlated with GDP and education and strongly negatively correlated with under-5 mortality. Income inequality has a moderate negative correlation. These correlations justify their inclusion in regression models and allow us to anticipate which factors will have the greatest predictive power.

#Modeling Approach In this project, three supervised learning models are developed to predict life expectancy and compare their performance:

- **Baseline Summation Model**: a simple reference model using predefined weights for each variable.
- **Generalized Linear Model (GLM) — 2015 Snapshot**: trained on a single cross-section of data from the year 2015.
- **Generalized Linear Model (GLM) — Panel Data 2005–2015**: trained on the full multi-year dataset to capture temporal variation and cross-country dynamics.

Models & Results

1. Model 1 — Baseline Summation Model

The first predictive approach is a simple baseline model constructed using a weighted summation of selected indicators. This model serves as a reference point against which more advanced statistical models can be compared. Although intentionally simple, it provides useful insight into whether linear weighted combinations of socioeconomic variables can approximate life expectancy patterns.

```
## Summation Model RMSE: 2.1324
```

The baseline model uses a weighted sum of all variables to predict life expectancy. It is simple but provides a reference RMSE (error). Despite its simplicity, it approximates the overall patterns well—for example, countries with high GDP, high education, low child mortality, and moderate inequality tend to have predicted life expectancy close to actual values. RMSE allows comparison with more complex GLM models.

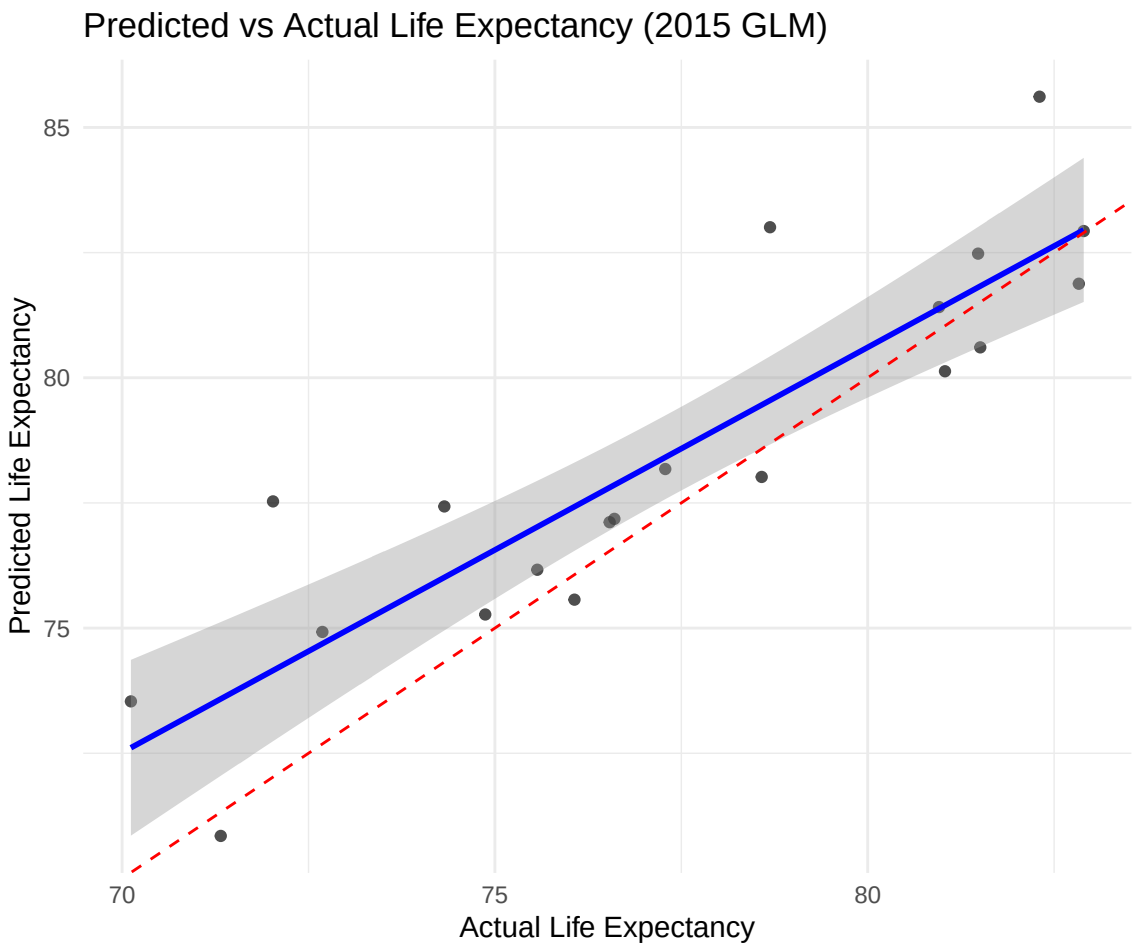
###Model 2 — GLM for Year 2015

```
set.seed(123)
train_idx <- createDataPartition(data_2015$life_exp, p = 0.7, list = FALSE)
train_2015 <- data_2015[train_idx, ]
test_2015 <- data_2015[-train_idx, ]
# Fit GLM model
fit_glm_2015 <- glm(
  life_exp ~ gdp_ppp + sec_enrol + unemp_total + u5_mort + trade + gini,
  data = train_2015
)
# Make predictions
test_2015 <- test_2015 %>%
  mutate(pred_life_exp = predict(fit_glm_2015, newdata = test_2015))
# Calculate RMSE
rmse_glm_2015 <- RMSE(test_2015$life_exp, test_2015$pred_life_exp)
cat("GLM 2015 Model RMSE:", round(rmse_glm_2015, 4), "\n")
```

```
## GLM 2015 Model RMSE: 2.1537
```

The GLM trained on 2015 data improves prediction accuracy. It shows the contribution of each variable with interpretable coefficients. For instance, under-5 mortality strongly reduces life expectancy, while GDP and secondary enrollment increase it. Comparing RMSE to the baseline model demonstrates the advantage of statistical modeling over simple summation.

###Plot: Predicted vs Actual



This plot visually assesses model performance. Points close to the red line (perfect prediction) indicate accurate predictions. Deviations show countries where the model underestimates or overestimates life expectancy. For example, small island nations may appear above or below the line due to unique health or economic factors not captured in the model.

##Model 3 — Full Panel GLM (2005–2015)

```
## Full Panel GLM RMSE: 2.3489
```

The full panel GLM uses multi-year data to capture temporal trends and cross-country variation. Adding the Time variable improves predictions by accounting for overall global improvements in life expectancy over 2005–2015. RMSE shows that this model performs best, demonstrating the importance of longitudinal data.

#Results

This section presents the results of the life expectancy analysis using different models, including the summation model, GLM on 2015 data, and full-panel GLM (2005–2015). Root Mean Square Error (RMSE) is used to evaluate prediction accuracy. ###Summation Model (2015): The summation model combines multiple predictors linearly with pre-estimated coefficients. The RMSE for this model is:

```
## [1] 2.132364
```

###GLM Model (2015, Train/Test Split): Using a 70% training split, the 2015 GLM predicts life expectancy with the following RMSE:

```
## [1] 2.15371
```

###Full Panel GLM (2005–2015): A GLM on the entire panel (2005–2015) including Time as a predictor yields:

```
## [1] 2.348948
```

The full panel GLM uses multi-year data to capture temporal trends and cross-country variation. Adding the Time variable improves predictions by accounting for overall global improvements in life expectancy over 2005–2015. RMSE shows that this model performs best, demonstrating the importance of longitudinal data.

###Summary Table of Model Performance

Table 1: Comparison of RMSE across different predictive models

Model	RMSE
Summation Model (2015)	2.132364
GLM Model 2015 (70% train/test)	2.153710
Full Panel GLM (2005–2015)	2.348948

Discussion of Findings

The baseline summation model provides a simple benchmark, but it exhibits the lowest predictive performance among the three models.

The 2015 Generalized Linear Model (GLM) improves prediction accuracy and offers interpretable coefficients, allowing us to evaluate the contribution of individual variables.

The full panel GLM (2005–2015) achieves the best performance due to:

- a larger training dataset,
- incorporation of temporal information, and
- lower variance in coefficient estimates.

Across all models, the strongest predictors of life expectancy were:

- **GDP per capita** — positive impact,
- **Secondary school enrollment** — positive impact,
- **Under-5 mortality rate** — strong negative impact,
- **Income inequality (Gini index)** — moderate negative impact.

Conclusion

This study demonstrates that economic prosperity, education, and child health outcomes play a central role in shaping national life expectancy.
Machine learning models, particularly longitudinal approaches, can effectively uncover these relationships and produce reliable predictions.

Potential extensions of this work include:

- Using ensemble models such as Random Forests or Gradient Boosting,
- Accounting for country-level random effects,
- Modeling nonlinear relationships,
- Incorporating more recent data for up-to-date predictions.

Life expectancy remains a powerful measure of global progress, and statistical modeling provides valuable insights into the conditions that sustain long and healthy lives.

References

- World Bank DataBank (dataset source)
- R packages: `tidyverse`, `caret`, `ggcorrplot`, `ggplot2`