Life Expectancy Prediction Using PCA & Elastic Net

FEM11149 - Introduction to Data Science

Vera Gak Anagrova

26 October, 2025

Introduction

Life expectancy is a widely used measure of population health and overall social development. Countries differ substantially in their health outcomes, reflecting variations in healthcare access, disease prevalence, living conditions, and economic resources. Understanding how these factors relate to life expectancy is important for designing effective public health strategies and allocating resources where they are most needed. This analysis examines cross-country differences in life expectancy and investigates which health, demographic, and economic characteristics are most strongly associated with longer life expectancy.

Data

The data used in this analysis come from the World Bank's Health, Nutrition and Population Statistics database. The main dataset contains 160 countries and 30 variables, capturing health system characteristics (e.g., immunization coverage, healthcare expenditure), demographic indicators (e.g., fertility and population growth), and measures of disease burden (e.g., tuberculosis incidence and diabetes prevalence). A separate dataset provides the life expectancy at birth (years) for the same countries, which serves as the dependent variable in the analysis. A third dataset includes the same 30 predictors for three additional countries, for which life expectancy is not provided; these observations are used for out-of-sample prediction. All variables represent the most recent available values within the last four years, as pre-processed in the assignment data. Prior to modeling, the datasets were merged by country and cleaned to address missing and non-finite values.

Methodology

Before modelling, the datasets were merged by country to ensure that each observation contained both the predictor variables and the corresponding life expectancy value. Missing or non-finite values in the predictors were imputed using the median of each variable to avoid excluding countries while preserving the overall distributional structure of the data. All predictors were treated as numeric, and reproducibility was ensured by setting the random seed to the student ERNA ID `set.seed(772713)`

Because the predictors exhibited substantial multicollinearity, Principal Component Analysis (PCA) was applied to reduce dimensionality. PCA was performed on the correlation matrix so that variables measured on different scales contributed equally to component extraction. To determine the appropriate number of components, multiple criteria were considered: the scree plot, the Kaiser rule (eigenvalues > 1), a permutation test comparing observed eigenvalues to those expected under random structure, and a bootstrap test assessing whether the cumulative variance explained by the selected components exceeded 70%. The retained components were then used as regressors in a Principal Component Regression (PCR) model.

As an alternative modelling strategy, an Elastic Net regression was fitted directly to the original predictors. Elastic Net combines the Lasso and Ridge penalties, enabling both coefficient shrinkage and variable selection, making it suitable for datasets with correlated predictors. The regularization parameter was selected using 10-fold cross-validation, and the model with the lowest cross-validated error was chosen as the final Elastic Net model. Finally, predictive performance of the PCR and Elastic Net models was compared using Root Mean Squared Error (RMSE), and the model with the lower RMSE was selected for forecasting life expectancy in the separate prediction dataset.

Results

The objective of the analysis is to understand how health, demographic, and economic characteristics relate to life expectancy across countries, and to develop a predictive model that generalizes well to new observations. Since many of the predictor variables are strongly correlated—particularly those reflecting healthcare expenditure, immunization coverage, and population size—dimensionality reduction was required to address multicollinearity. Principal Component Analysis (PCA) was therefore applied to summarize the 30 original variables into a smaller set of orthogonal components that capture the main patterns of variation in the dataset.

Figure 1 shows a steep decline in eigenvalues after the first few components, with a noticeable flattening beginning around the fourth and fifth components. Although the scree plot suggests a potential bend near the sixth component as well, only the first five components have eigenvalues greater than one according to the Kaiser criterion, indicating that these components explain more variance than an average standardized variable. Together, the first five components account for approximately 74% of the total variance in the predictors.To verify this choice, a permutation test was conducted (Figure 2) in which the observed eigenvalues were compared to a 95% threshold derived from permuted datasets. The first four components clearly exceed this threshold, indicating strong systematic structure. The fifth component lies very close to the cutoff, but still meets or marginally exceeds the 95% boundary, meaning it likely captures weak but meaningful signal rather than noise.

Figure 1: Scree Plot of Principal Components

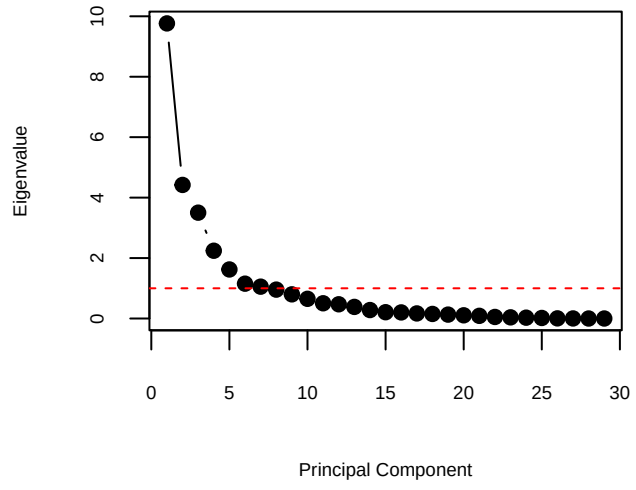
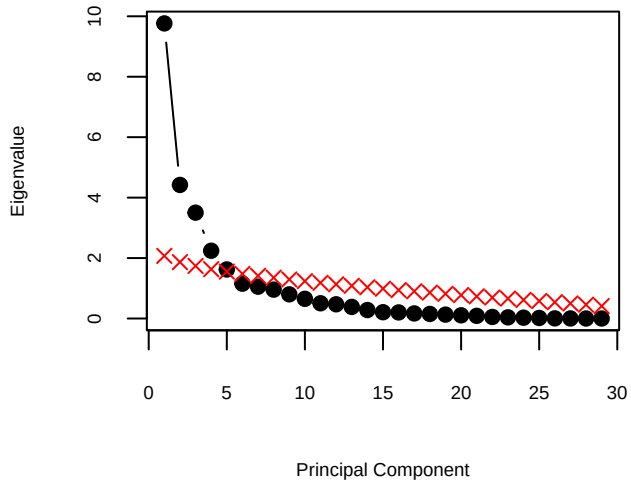


Figure 2: Permutation Test Threshold



To further assess the robustness of the component selection, a bootstrap resampling procedure was conducted

to evaluate whether each component's eigenvalue was significantly greater than one. As shown in Figure 3 (Appendix), the first five components have 95% confidence intervals entirely above the Kaiser threshold, indicating that they each explain more variance than would be expected from an average standardized variable. Although the sixth component's interval approaches the value of one, the permutation test in Figure 2 suggests that only the first five components consistently exceed the 95% random-structure threshold, meaning that the sixth component does not capture stable underlying signal. Figure 4 (Appendix) further supports this conclusion: the cumulative variance explained by the first five components remains above 70% across 500 bootstrap resamples, with the observed value (red line) lying near the center of the distribution. This demonstrates that the five-component solution is stable and not sample-dependent. Based on the combined evidence from the scree plot, permutation test, and bootstrap analyses, five principal components were retained for interpretation and subsequent regression modeling.

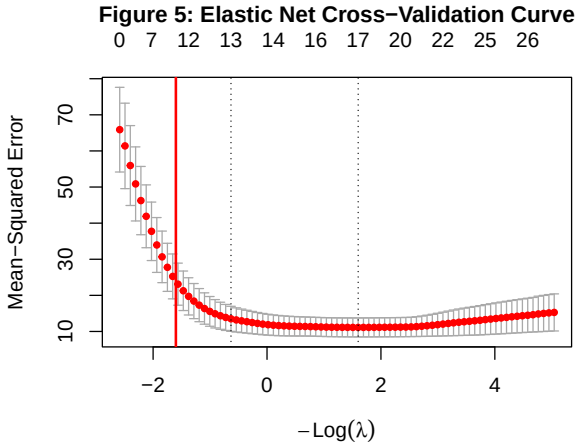
The five retained principal components each capture a distinct structural dimension of global health conditions. PC1 reflects economic development and health system capacity, with high loadings on national income, healthcare spending, and access to sanitation and clean water. PC2 represents population scale and urbanization patterns, distinguishing large, densely populated countries from smaller or more rural ones. PC3 is associated with health financing and the chronic disease environment, indicating variation in resource allocation and the prevalence of noncommunicable diseases such as hypertension. PC4 captures public health and immunization system performance, particularly the effectiveness of preventive care programs. Finally, PC5 reflects lifestyle and behavioral health risk, including obesity and diabetes prevalence. These interpretations are supported by the component loading patterns shown in Table 4 in the Appendix, and together, the components summarize the main axes of variation in the dataset and serve as predictors in the subsequent Principal Component Regression model.

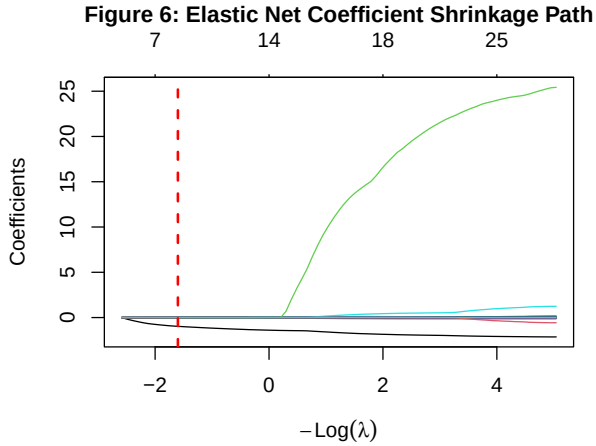
In Table 1, we can see that the PCR model with five components explains about 81% of the variation in life expectancy. PC1 has a strong negative effect, indicating that countries with weaker health and socioeconomic systems tend to have lower life expectancy. PC3 and PC5 show positive and significant associations, suggesting that factors related to healthcare financing, education, and lifestyle contribute to longer lifespans. In contrast, PC2 and PC4 have weaker effects and play a limited role in explaining differences in life expectancy.

Table 1: PCR Model Summary (5 Components)

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	72.082	0.278	259.105	0.000
PC1	-2.298	0.089	-25.809	0.000
PC2	0.042	0.132	0.319	0.750
PC3	0.384	0.149	2.582	0.011
PC4	-0.362	0.186	-1.947	0.053
PC5	0.739	0.218	3.386	0.001
Adjusted R^2	0.811	NA	NA	NA
Cross-Validated RMSE	3.600	NA	NA	NA

A penalized regression model using Elastic Net was also estimated to allow variable selection and shrinkage directly in the original feature space. From the cross-validation curve (Figure 5), the optimal value of λ corresponds to a model that retains approximately ten predictors, while the remaining coefficients are shrunk to zero. This indicates that Elastic Net performs both variable selection and coefficient shrinkage. As shown in the coefficient path plot (Figure 6), only a small subset of predictors remains influential at the optimal penalty level, resulting in a more parsimonious and interpretable model without compromising predictive accuracy.





To ensure a fair comparison, both models were evaluated using 10-fold cross-validation, the Elastic Net model achieves a lower prediction error (RMSE = 3.34) compared to the PCR model (RMSE = 3.60), as seen in Table 2, indicating superior out-of-sample predictive performance. Therefore, the Elastic Net model is selected as the preferred forecasting approach.

Table 2: Model Performance Comparison

Model	RMSE
Elastic Net (CV)	3.335
PCR (CV)	3.600

The selected Elastic Net model was applied to the three new countries in the prediction dataset. The results indicate notable differences in life expectancy: the Netherlands has the highest predicted life expectancy ($\approx$ 81.9 years), Colombia has a moderate level ($\approx$ 76.7 years), and Kenya has the lowest ($\approx$ 64.2 years). These differences align with broader disparities in healthcare investment, disease burden, and economic development, suggesting that structural health and economic conditions play a central role in shaping population longevity.

Table 3: Predicted Life Expectancy for New Countries

Country	Predicted_Life_Expectancy
Netherlands	81.92
Kenya	64.20
Colombia	76.67

Conclusion

This study examined international differences in life expectancy using global health, demographic, and socioeconomic indicators. Principal Component Analysis reduced the data to five meaningful dimensions capturing health system quality, disease environment, population structure, and lifestyle-related risks. A Principal Component Regression model explained a substantial share of the variation (Adjusted $R^2 \approx 0.81$), but showed moderate predictive error. The Elastic Net model improved predictive accuracy and selected only the most relevant variables, resulting in a simpler and more interpretable approach for forecasting life expectancy across countries.

Appendix

Figure 3: Bootstrap 95% CI for PCA Eigenvalues (Kaiser Test)

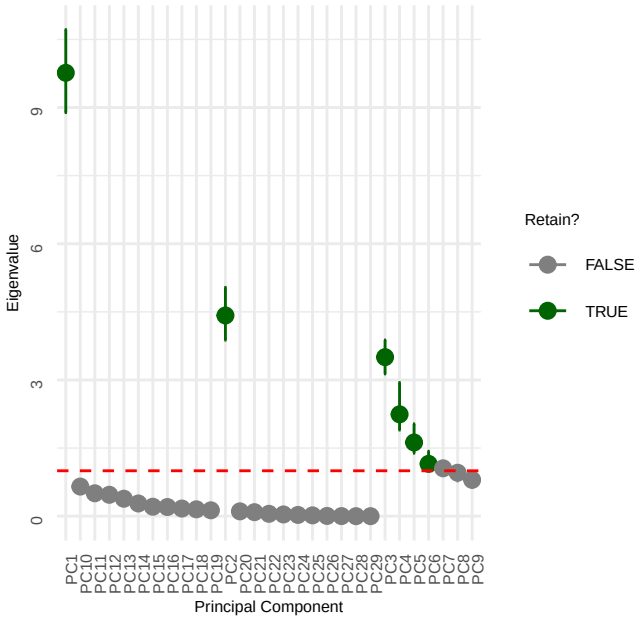


Figure 4: Bootstrap Distribution of Cumulative Variance Explained

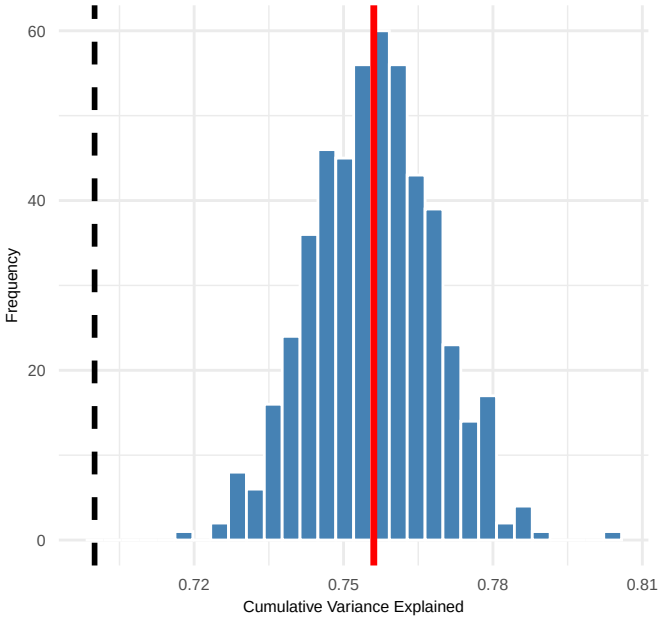


Table 4: Principal Component Loadings

Variable	Component	Loading
population_total	Comp.2	-0.439
labor_force_total	Comp.2	-0.435
urban_population	Comp.2	-0.420
rural_population	Comp.2	-0.418
current_health_expenditure_per_capita_current_us	Comp.3	0.330
domestic_general_government_health_expenditure_per_capita_current_us	Comp.3	0.325
domestic_private_health_expenditure_per_capita_current_us	Comp.3	0.309
prevalence_of_hypertension_of_adults_ages_30_79	Comp.3	-0.303
immunization_hep_b3_of_one_year_old_children	Comp.4	0.334
immunization_hib3_of_children_ages_12_23_months	Comp.4	0.333
urban_population_growth_annual	Comp.4	0.319
immunization_dpt_of_children_ages_12_23_months	Comp.4	0.317
diabetes_prevalence_of_population_ages_20_to_79	Comp.5	0.503
total_alcohol_consumption_per_capita_liters_of_pure_alcohol_projected_estimates_15_years_of_age	Comp.5	-0.431
population_growth_annual	Comp.5	0.422
urban_population_growth_annual	Comp.5	0.317
prevalence_of_overweight_of_adults	Comp.5	0.300

Code

```
# --- Merge & prepare data ---
df_model <- df_nutrition %>% left_join(df_life_expectancy, by="country")
df_pred  <- df_prediction
X <- df_model %>% select(-country, -life_expectancy_at_birth_total_years) %>%
  mutate(across(everything(), ~ ifelse(is.na(.)|is.infinite(.), median(.,na.rm=TRUE), .))) %>%
  mutate(across(everything(), as.numeric))
# Check for countries that failed to merge
df_model %>% filter(is.na(life_expectancy_at_birth_total_years)) %>% select(country)
# --- PCA (correlation matrix standardizes variables) ---
pca_fit <- princomp(X, cor=TRUE, scores=TRUE)

# Figure 1: Scree Plot
plot(pca_fit$sdev^2, type="b", pch=19,
     main="Figure 1: Scree Plot of Principal Components", xlab="Principal Component", ylab="Eigenvalue")
abline(h=1, col="red", lty=2)

# --- Figure 2: Permutation Test ---
set.seed(772713)
eig_obs <- eigen(cor(X))$values
eig_perm <- replicate(500, eigen(cor(as.data.frame(lapply(X, sample))))$values)
eig_cutoff <- apply(eig_perm,1,quantile,0.95)
plot(eig_obs, type="b", pch=19, col="black", main="Figure 2: Permutation Test Threshold",
     xlab="Principal Component", ylab="Eigenvalue")
lines(eig_cutoff, type="b", pch=4, col="red", lty=2)

# --- Figure 3: Bootstrap CI for eigenvalues ---
boot_kaiser <- function(X,B=500){
  eig_obs <- eigen(cor(X))$values
  eig_boot <- replicate(B, eigen(cor(X[sample(nrow(X),replace=TRUE),]))$values)
  ci <- apply(eig_boot,1,quantile,c(.025,.975))
  data.frame(PC=paste0("PC",1:length(eig_obs)),
             eig_obs, ci_low=ci[1,], ci_high=ci[2,], retain=ci[1,]>1)
}
bk <- boot_kaiser(X)
ggplot(bk, aes(PC, eig_obs, color=retain)) + geom_point(size=2.5) +
  geom_errorbar(aes(ymin=ci_low,ymax=ci_high), width=.15) +
  geom_hline(yintercept=1, lty=2, col="red") +
  labs(title="Figure 3: Bootstrap 95% CI for PCA Eigenvalues (Kaiser Test)")+ theme_minimal()

# --- Figure 4: Bootstrap cumulative variance (PC1-PC5) ---
set.seed(772713)
vaf_boot <- replicate(500, sum(princomp(X[sample(nrow(X),replace=TRUE),],
                                        cor=TRUE)$sdev[1:5]^2)/ncol(X))
ggplot(data.frame(VAF=vaf_boot), aes(VAF)) +
  geom_histogram(fill="steelblue", color="white", bins=30) +
  geom_vline(xintercept=mean(vaf_boot), col="red") +
  geom_vline(xintercept=0.70, lty=2) +
  labs(title="Figure 4: Bootstrap Distribution of Cumulative Variance Explained (PC1-PC5)") +
  theme_minimal()

# --- Table 1: PCA Loadings (|Loading|  0.30) ---
L_clean <- as.data.frame(unclass(pca_fit$loadings))[,1:5] %>%
```

```r
  tibble::rownames_to_column("Variable") %>%
  pivot_longer(starts_with("Comp"), names_to="Component", values_to="Loading") %>%
  filter(abs(Loading)>=0.30) %>% mutate(Loading=round(Loading,3))
kable(L_clean, caption="Table 1: PCR Model Summary") %>%
  kable_styling(full_width=FALSE, font_size=8)

# --- PCR with 10-fold CV ---
PC_scores <- as.data.frame(pca_fit$scores[,1:5]) %>%
  mutate(life_expectancy = df_model$life_expectancy_at_birth_total_years)

set.seed(772713)
fold_id <- sample(rep(1:10, length.out = nrow(PC_scores)))
Xsc <- as.matrix(PC_scores[,1:5]); y <- PC_scores$life_expectancy

rmse_k <- sapply(1:5, function(k){
  pred <- rep(NA, length(y))
  for(f in 1:10){
    tr <- fold_id != f; te <- !tr
    fit <- lm(y[tr] ~ Xsc[tr,1:k,drop=FALSE])
    pred[te] <- cbind(1, Xsc[te,1:k,drop=FALSE]) %*% coef(fit)
  }
  sqrt(mean((y - pred)^2))
})
select_n <- which.min(rmse_k)
pcr_final_lm <- lm(y ~ Xsc[,1:select_n,drop=FALSE])
summary(pcr_final_lm)

# --- Elastic Net ---
X_en <- as.matrix(X); y_en <- df_model$life_expectancy_at_birth_total_years
set.seed(772713)
cv_en <- cv.glmnet(X_en,y_en,alpha=.5,standardize=TRUE)
en_final <- glmnet(X_en,y_en,alpha=.5,lambda=cv_en$lambda.min)
rmse_en <- sqrt(mean((y_en - predict(en_final, X_en))^2))
# Figure 5 & 6
plot(cv_en); abline(v=log(cv_en$lambda.min),col="red",lwd=2)
mtext("Figure 5: Elastic Net Cross-Validation Curve",side=3,line=2,font=2)
plot(glmnet(X_en,y_en,alpha=.5), xvar="lambda",label=FALSE)
abline(v=log(cv_en$lambda.min),col="red",lty=2)
mtext("Figure 6: Elastic Net Coefficient Shrinkage Path",side=3,line=2,font=2)

# --- Model Performance Comparison ---
comparison_table <- data.frame(Model = c("Elastic Net (CV)", "PCR (CV)"),
  RMSE = c(round(rmse_en_cv, 3), round(rmse_pcr, 3)))
kable(comparison_table, caption = "Model Performance Comparison")

# --- Predict New Countries ---
df_pred_clean <- df_pred %>% select(all_of(colnames(X_en))) %>%
  mutate(across(everything(), ~ ifelse(is.na(.) | is.infinite(.), median(., na.rm = TRUE), .))) %>%
  as.matrix()
prediction_table <- data.frame(Country = df_pred$country,
  Predicted_Life_Expectancy = round(as.numeric(predict(en_final, newx = df_pred_clean)), 2)
)
kable(prediction_table, caption = "Predicted Life Expectancy for New Countries")
```