Predicting Life Expectancy Under Multicollinearity: A Comparison of Principal Component Regression and Elastic Net Regularization

FEM11149 - Introduction to Data Science

Maggie Tsou

Oct, 2025

Introduction

Life expectancy is a fundamental indicator of population health and socioeconomic development, reflecting the effects of health services, living conditions, and national resources [Martinez et al., 2021]. The prediction of which is crucial for resource allocation decisions and to help prioritize actions that most effectively improve outcomes across diverse regions, though it is complicated by the fact that many health, economic, and demographic indicators are highly correlated; therefore, standard regression methods may fail to give an accurate prediction. This study explores how predictive modeling techniques can best capture the complex, interdependent drivers of life expectancy in each country. Specifically, it asks whether summarizing correlated health indicators through Principal Component Regression or selectively penalizing them through Elastic Net regularization yields more accurate and interpretable predictions.

Data

The dataset, compiled from the World Health Organization and World Bank, contains observations for 160 countries and 28 indicators describing health, demographic, and socioeconomic conditions, with life expectancy as the dependent variable. Health indicators include total, government, private, and out-of-pocket health expenditure per capita; prevalence of diabetes and hypertension; and tuberculosis incidence, mortality, and treatment success rate. Immunization coverage is measured for DPT, HepB3, Hib3, measles, and Pol3 vaccines, alongside total alcohol consumption per capita. Demographic and nutrition-related indicators comprise fertility rate and prevalence of overweight adults. Population and economic variables include total and urban population, population growth, rural population, labor force participation, unemployment rate, sex ratio at birth, net migration, access to basic drinking water and sanitation, and gross national income per capita. Each variable represents the most recent available observation for each country, and missing data were addressed systematically during preprocessing.

Methodology

To make sure that model performance could be evaluated on unseen data and to mitigate overfitting, this dataset is randomly divided into training (80%, n = 128) and test (20%, n = 32). Imputation with the median (due to its robustness to outliers) is then used to handle missing values that were mostly present in small nations. Next, diagnostic tests were conducted to assess multicollinearity, skewness, and linearity. Logarithmic transformations were applied to continuous right-skewed variables and logit transformations to bounded percentage variables to adjust for the latter two diagnostic tests. Two complementary regression

approaches are implemented to address the remaining multicollinearity. Principal Component Regression (PCR) reduces multicollinearity by constructing uncorrelated linear combinations of predictors that capture the dominant variance structure in the data. The optimal number of components is determined using Kaiser's rule, cumulative variance thresholds, scree-plot analysis, and a permutation test evaluating whether retained components explain more variance than expected by chance, with bootstrap resampling (1,000 iterations) validating component stability. Elastic Net regression is chosen for its ability to combine L1 and L2 penalties, and the tuning parameters alpha and lambda (min and 1se) are optimized to minimize prediction error. Model performance on the test data was evaluated using RMSE, MAE, and R^2, which together quantify overall accuracy, typical deviation, and explained variance. Stability across income levels were also considered in the selection of the final model to predict life expectancy for 3 different countries.

Results

Exploratory analysis was conducted to assess the suitability of ordinary least squares regression. Since the focus is on predictive accuracy, diagnostic assessment prioritized multicollinearity, normality, and linearity on the untransformed training data.

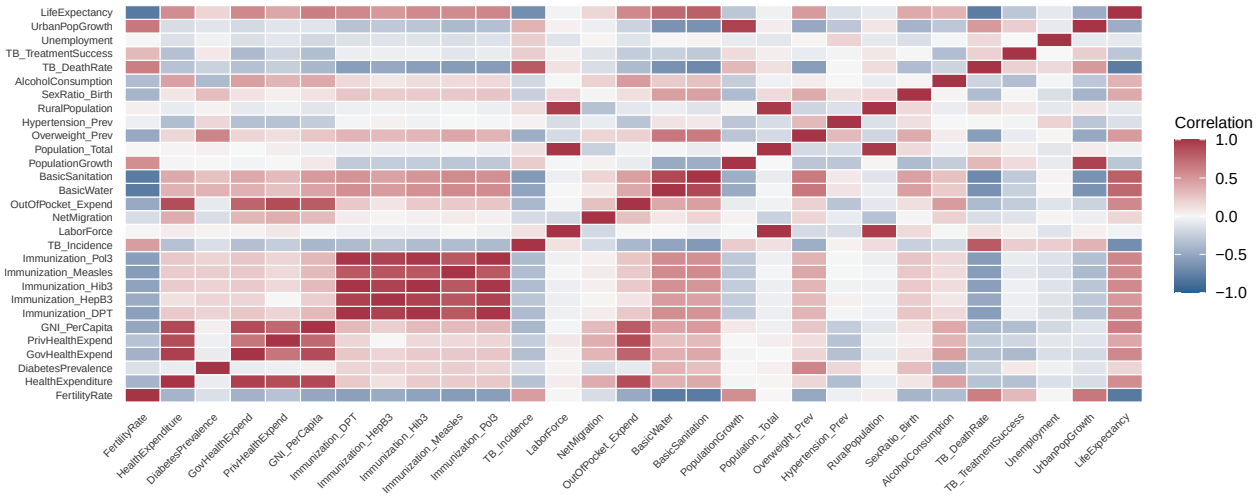

Figure 1: Correlation Plot

As seen in Figure 1, multicollinearity is severe, with strong positive correlations among economic, health expenditure, infrastructure indicators, and clusters of demographic and immunization variables, clearly presenting that many predictors have overlapping information. Normality was violated, with 22 of 28 variables exhibiting $|\text{skewness}| > 1$. (Figure 2). Residual diagnostic plots also reveal violations of linearity. Log transformations were applied, and Table 1 demonstrates the effectiveness of these transformations: mean absolute skewness decreased from 2.32 to 0.40, and linearity improved (Figure 3). However, multicollinearity remained present, and in this moderate-to-high-dimensional setting, specialized methods are needed to prevent overfitting.

Table 1: Data Summary Before and After Transformations

Metric	Before	After
Mean |Skewness|	2.32	0.40
Mean VIF	479.62	17.25
Variables with VIF > 10	14.00	12.00

The Elastic Net regression model estimates coefficients by minimizing the following objective function:

$$\min_{\beta_0,\,\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \left[(1-\alpha) \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \right\}$$

After cross-validation, the optimal alpha is 0.58, which shows a balanced L1 (lasso) and L2 (ridge) penalization, combining variable selection with coefficient shrinkage to handle multicollinearity. Optimal regularization yielded at lambda_min = 0.213, which selected 17 of the 28 predictors. A more conservative penalty (lambda_1se = 1.985) retained only 6 variables, demonstrating the parsimony-accuracy trade-off (Table 4). Both specifications converged on the same core predictors (Table 5): demographic indicators (FertilityRate, SexRatio_Birth), infectious disease burden (TB_DeathRate), basic health infrastructure (BasicSanitation, BasicWater), and government health investment (GovHealthExpend). Increasing penalization from lambda_min to lambda_1se primarily removed secondary variables, giving a simpler representation.

Next, principal component analysis is applied to the 28 transformed predictors. The Principal Component Regression (PCR) model estimates coefficients by first projecting correlated predictors $\mathbf{X}$ onto orthogonal principal components and then regressing the response $\mathbf{y}$ on the first k components, as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{PCR}(k)} = \mathbf{P}_k (\mathbf{T}_k^\top \mathbf{T}_k)^{-1} \mathbf{T}_k^\top \mathbf{y} = \mathbf{P}_k (\mathbf{Z}_k^\top \mathbf{Z}_k)^{-1} \mathbf{Z}_k^\top \mathbf{y}, \quad \text{where } \mathbf{Z}_k = \mathbf{X}\mathbf{P}_k.$$

where $\mathbf{P}_k$ contains the loadings of the first k principal components and $\mathbf{Z}_k$ the corresponding component scores. Figure 4 presents the scree plot, showing an elbow around components 4–5 where the marginal variance explained plateaus. Component selection criteria are consistent: Kaiser's rule (eigenvalue > 1) and the 70% cumulative variance threshold both support retaining five components, which together explain 75.3% of total variance—balancing dimensionality reduction and information retention. A permutation test with 1,000 iterations (Figure 5) confirmed that the first four components capture genuine structure (eigenvalues $>$ 95th%), while the fifth lies near the threshold, indicating marginal but interpretable structure. Bootstrap validation (1,000 resamples; Figure 6) showed stable eigenvalues for PC1–PC4 above Kaiser's threshold, with PC5 generally remaining above it, again supporting its inclusion; PC6 reflected only noise. The bootstrap distribution of cumulative variance explained (Figure 7) indicated that five components account for a mean of 76.7% of total variance (95% CI: 74.5–79.0%), confirming robustness. Finally, sensitivity analysis (Table 6) comparing 3–5 component models demonstrated that the five-component solution achieved the best balance between predictive accuracy (RMSE = 2.669 years), justifying its selection as the final model.

Figure 8 displays the top five variable loadings per component. Component 1 (40.6% of variance) captures a Healthcare Infrastructure and Economic Development dimension, with high loadings for government and total health expenditure, income, and access to sanitation and water, reflecting that wealthier health systems achieve stronger overall outcomes. Component 2 reflects Immunization Coverage with strong negative loadings for vaccination rates across DPT, Polio, Hib3, and HepB3, representing coordinated public health systems. Population Structure is represented by Component 3, which links lower prevalence of hypertension with larger labor forces and total populations, reflecting the effects of workforce composition and demographics. Component 4 captures Demographic Transitions, where declining population growth and urbanization align with economic maturation and rural-to-urban shifts. Finally, Component 5 combines Behavioral Health Risks, with high loadings for diabetes, alcohol consumption, and obesity, indicating lifestyle-related determinants of longevity. Together, these interpretable components confirm that PCR successfully extracted coherent latent dimensions summarizing correlated global health indicators. Notably, SexRatio_Birth—Elastic Net's most influential predictor—was absent from PCR's primary loadings, highlighting a key methodological difference: PCR uncovers structural patterns, whereas Elastic Net isolates specific actionable factors.

Table 2: Model Performance Comparison

Model	Test_RMSE	Test_MAE	Test_R2	Test_AdjR2	Parameters
PCR (5 PCs)	2.669	2.235	0.892	0.871	5 PCs
Elastic Net (lambda.min)	2.644	2.181	0.894	0.764	17 vars
Elastic Net (lambda.1se)	3.207	2.555	0.844	0.806	6 vars

Table 2 shows test set performance across specifications. PCR (5 components) and Elastic Net (lambda_min) achieved nearly identical accuracy—RMSE of 2.669 vs. 2.644 years, both explaining roughly 89% of variance. The stricter Elastic Net (lambda_1se) specification trades accuracy (RMSE = 3.207) for simplicity. Though Elastic Net edged out PCR by 0.025 years, PCR achieves comparable performance using just 5 dimensions versus 17 separate predictors. To assess robustness across income levels, arguably the most importantly source of global health inequality [World Health Organization, 2025], prediction errors were regressed on log(GNI per capita). Table 7 shows that PCR residuals showed no income relationship ($p = 0.604$), while Elastic Net displayed increasing bias: mild at lambda_min ($p = 0.345$) but significant at lambda_1se ($p = 0.002$). PCR's near-zero slope (-0.17) indicates prediction errors remain constant regardless of national wealth, whereas Elastic Net's slopes (0.31 at lambda_min, 1.17 at lambda_1se) reveal systematic under-prediction for wealthier nations and overprediction for poorer ones—a pattern that intensifies with stronger regularization. This makes PCR more generalizable for cross-country predictions. The five-component PCR model was therefore selected for predictions in the Netherlands, Kenya, and Colombia, spanning the global income spectrum.

Table 3: Life Expectancy Predictions (PCR Model)

Netherlands	Kenya	Colombia
81.81	64.42	74.83

The PCR model predicted life expectancy for the Netherlands (81.81 years), Kenya (64.42 years), and Colombia (74.83 years), spanning a 17-year range across income levels. PCR's income-invariant performance ($p = 0.604$) ensures equally reliable predictions from low-income Kenya to high-income Netherlands, which is essential for global health applications.

Conclusion

This analysis identified high intercorrelations among life expectancy related and addressed two central questions: how to manage multicollinearity in health data and which factors most strongly predict life expectancy. Both principal component regression (PCR) and Elastic Net effectively mitigated severe multicollinearity and achieved comparable predictive accuracy; thus, the optimal method depends on analytical objectives: Elastic Net gives specific variable interpretability, identifying fertility rate, tuberculosis mortality, sanitation access, and government health expenditure as actionable variables for policy intervention. In contrast, PCR reveals broader latent dimensions integrating healthcare infrastructure, economic development, immunization coverage, demographic transition, and behavioral health risks. Although PCR exhibited a marginally higher RMSE, it demonstrated greater robustness across income levels. Elastic Net provides specific targets for within-country interventions, and PCR provides a stable framework for global benchmarking. Future research should expand this comparison to temporal forecasting to evaluate their predictive validity as life expectancy changes over time. The cross-sectional nature of this study restricts inferences to variations between countries, and the 5th principal component exhibited only slight structure, so careful interpretation is needed.

Appendix

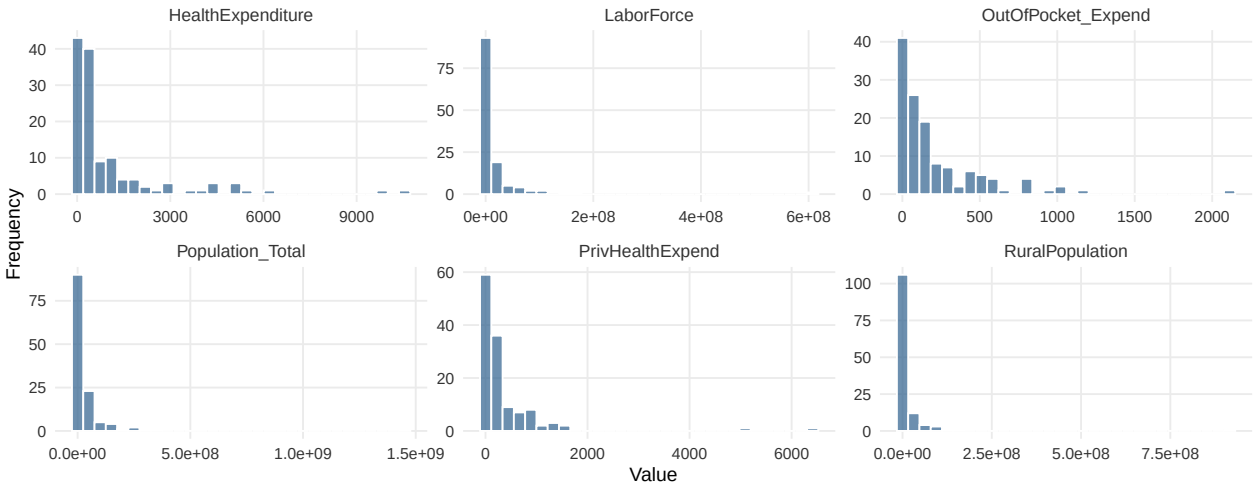

Figure 2: Distribution of Most Skewed Variables

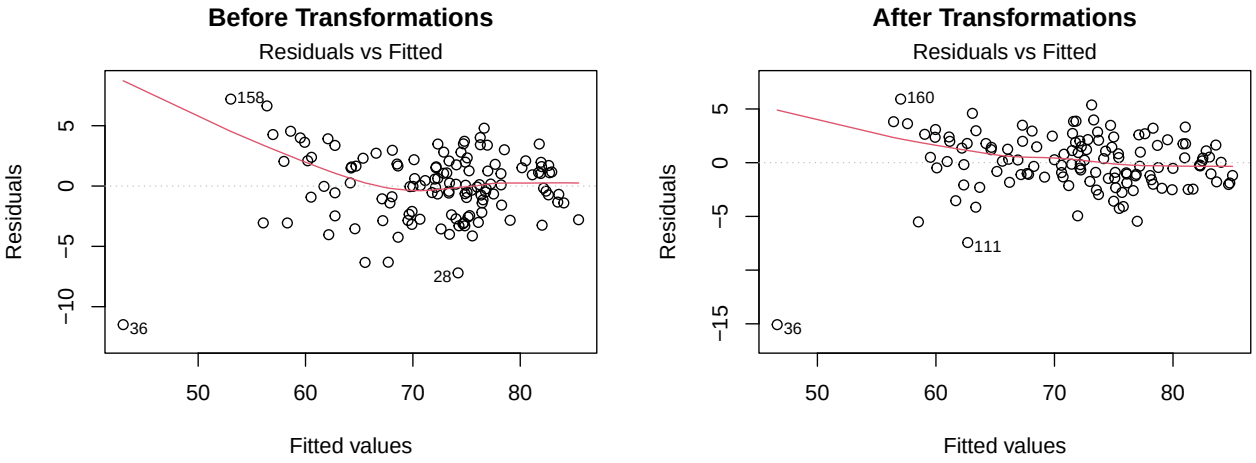

Figure 3: Linearity Before and After Transformations

Table 4: Elastic Net Cross-Validation Results

Parameter	Value
Alpha	0.580
Lambda (min)	0.213
Lambda (1-SE)	1.985

Table 5: Elastic Net Model Comparison

Model	Lambda	Predictors	TopPredictors
Elastic Net (lambda_min)	0.21	17	SexRatio_Birth, FertilityRate, TB_DeathRate_logit, GovHealthExpend_log, BasicSanitation_logit
Elastic Net (lambda_1se)	1.99	6	FertilityRate, TB_DeathRate_logit, BasicSanitation_logit, GovHealthExpend_log, BasicWater_logit

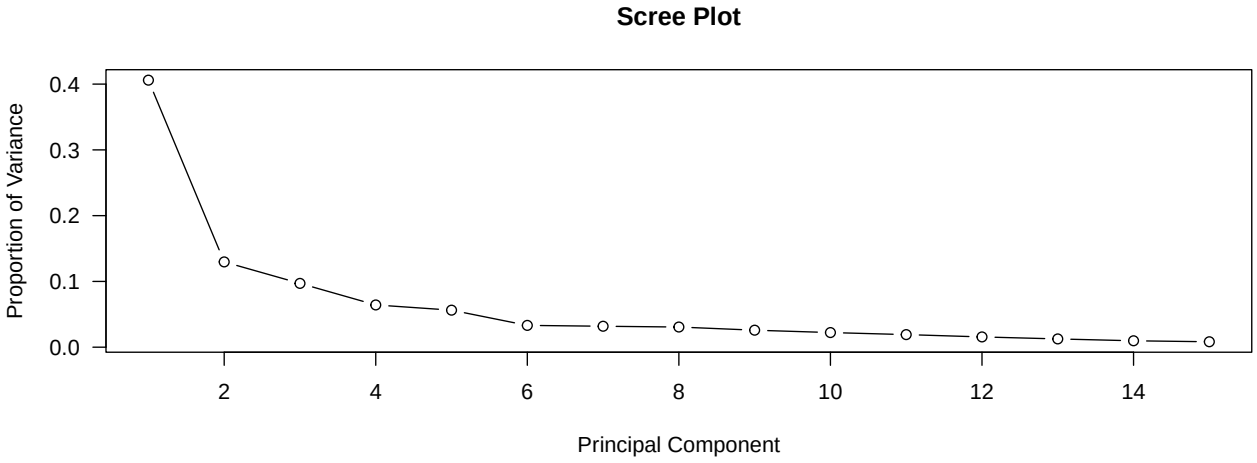

Figure 4: Eigenvalue Scree Plot of PCs

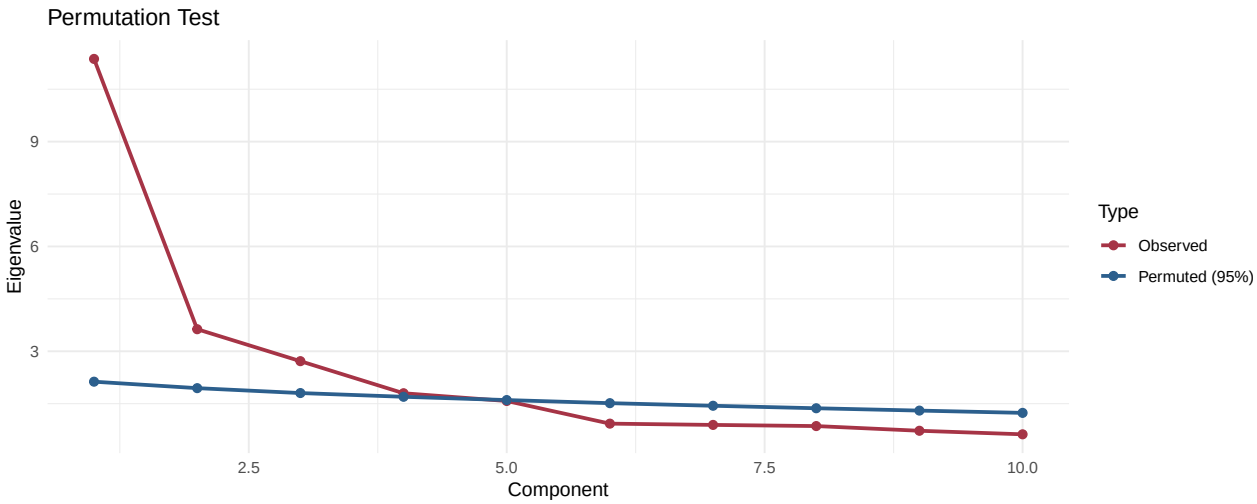

Figure 5: Permutation Test for Component Significance

Table 6: Sensitivity analysis of PCR performance across different numbers of components (k)

Number.of.Components..k.	Test.RMSE	Test.MAE	Variance.Explained. . . .
3	2.725	2.171	63.3
4	2.807	2.226	69.7
5	2.669	2.235	75.3

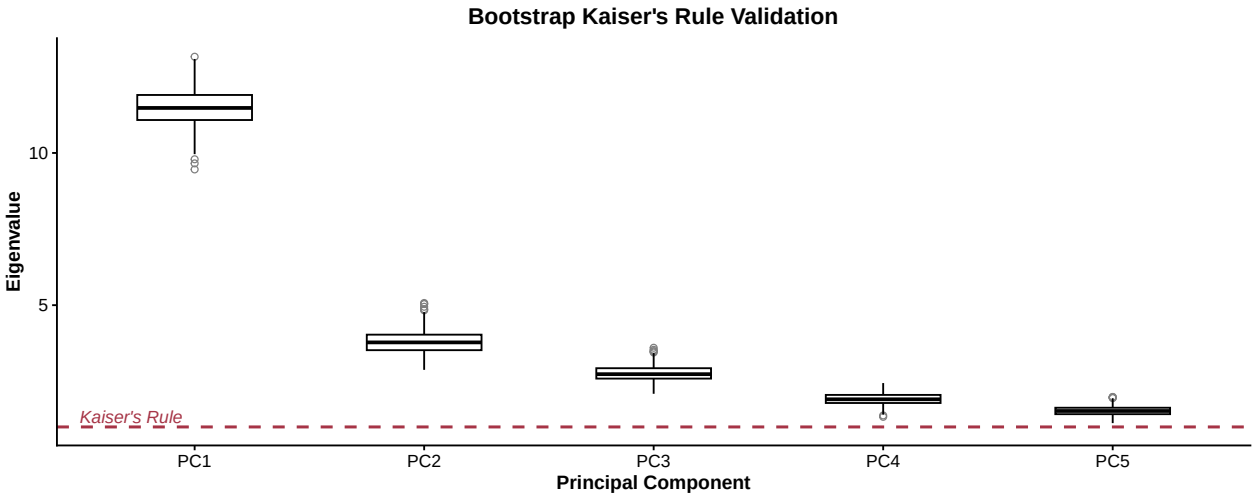

Figure 6: Bootstrap Validation of Kaiser's Rule

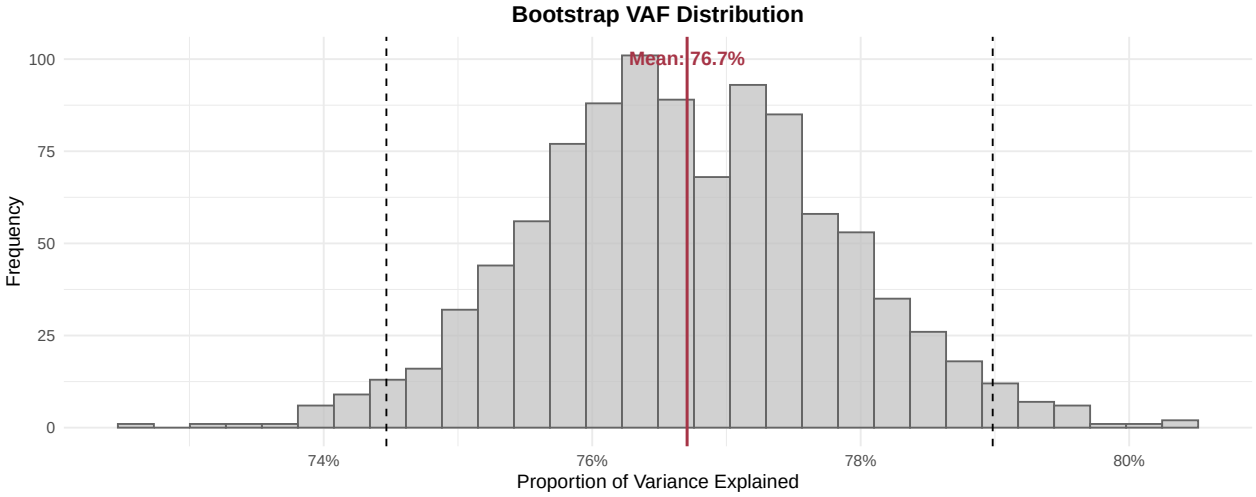

Figure 7: Bootstrap Distribution of Cumulative Variance Explained

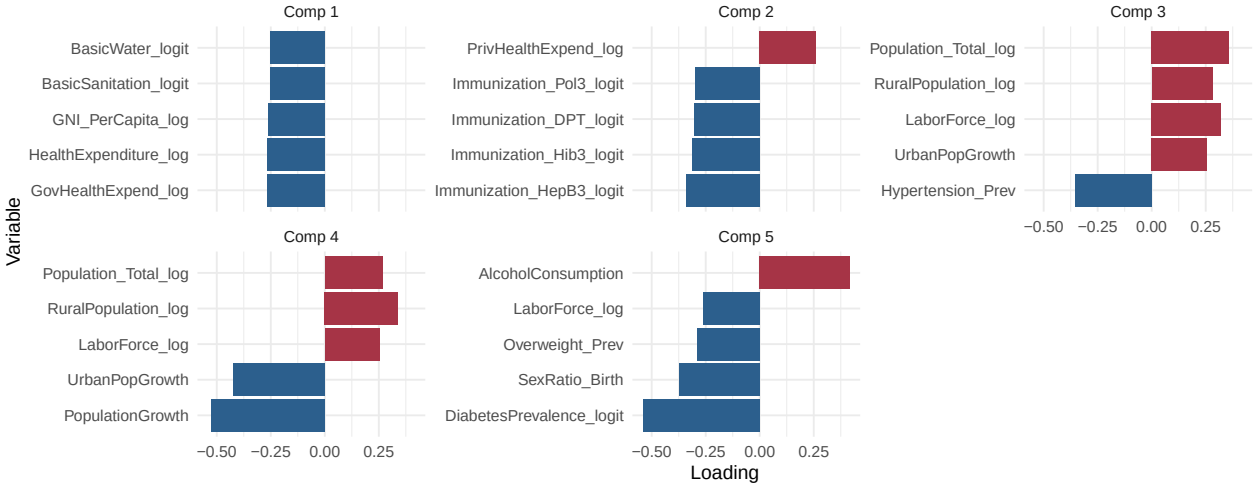

Figure 8: Top 5 Variable Loadings per Principal Component

Table 7: Stability Test of Prediction Errors Across Income Levels

Model	Slope	p.value
PCR Regression	-0.1686	6.04e-01
Elastic Net (lambda_min)	0.3121	3.45e-01
Elastic Net (lambda_1se)	1.1734	1.80e-03

```
invisible(lapply(c("knitr","dplyr","glmnet","tidyr","ggplot2","corrplot","naniar",
"car","lmtest","pls","reshape2","e1071"),require,character.only=TRUE))
health_data<-read.csv("health_nutrition_population.csv",stringsAsFactors=FALSE)
life_exp<-read.csv("life_expectancy_data-1.csv",stringsAsFactors=FALSE)
predictions_data<-read.csv("predictions-1.csv",stringsAsFactors=FALSE)
set.seed(100)
train_index<-sample(1:nrow(full_data),0.8*nrow(full_data))
train_data<-full_data[train_index,]; test_data<-full_data[-train_index,]
numeric_cols<-sapply(train_data,is.numeric)
numeric_cols["LifeExpectancy"]<-FALSE
train_medians<-list()
for(col in names(train_data)[numeric_cols]){
 if(sum(is.na(train_data[[col]]))>0){
  train_medians[[col]]<-median(train_data[[col]],na.rm=TRUE)}}
for(col in names(train_medians)){
 train_data[[col]][is.na(train_data[[col]])]<-train_medians[[col]]}
for(col in names(train_medians)){
 if(col%in%colnames(test_data)){
  test_data[[col]][is.na(test_data[[col]])]<-train_medians[[col]]}}
train_data<-train_data%>%select(-UrbanPopulation)
test_data <-test_data %>%select(-UrbanPopulation)
train_numeric<-train_data%>%select(-Country)%>%select(where(is.numeric))
lm_baseline<-lm(LifeExpectancy~.-Country,data=train_data)
vif_values<-vif(lm_baseline)
skewness_values<-sapply(train_numeric,skewness,na.rm=TRUE)
skewed_vars<-names(skewness_values[abs(skewness_values)>1])
logit_transform<-function(x){
 if(max(x,na.rm=TRUE)>1)x<-x/100
 e<-0.001; x<-pmax(pmin(x,1-e),e); log(x/(1-x))}
# Lists of percentage and continuous skewed variables used for log/logit transformations
# are omitted here for brevity; transformations followed the same procedure as in training.
train_x<-train_transformed%>%select(-Country,-LifeExpectancy)
train_y<-train_transformed$LifeExpectancy
test_x <-test_transformed %>%select(-Country,-LifeExpectancy)
test_y <-test_transformed $LifeExpectancy
train_x_matrix<-as.matrix(train_x); train_y_vector<-as.numeric(train_y)
test_x_matrix <-as.matrix(test_x);  test_y_vector <-as.numeric(test_y)
train_pcr_data<-data.frame(life_expectancy=train_y,train_x)
test_pcr_data <-data.frame(life_expectancy=test_y,test_x)
lm_transformed<-lm(LifeExpectancy~.-Country,data=train_transformed)
vif_transformed<-vif(lm_transformed)
set.seed(100)
alpha_grid<-seq(0,1,by=0.01)
cv_list<-vector("list",length(alpha_grid))
cv_min <-numeric(length(alpha_grid))
for(i in seq_along(alpha_grid)){
```

```
 cv_list[[i]]<-cv.glmnet(train_x_matrix,train_y_vector,
  family="gaussian",alpha=alpha_grid[i],nfolds=10,
  type.measure="mse",standardize=TRUE)
 cv_min[i]<-min(cv_list[[i]]$cvm)}
best_i<-which.min(cv_min)
best_alpha<-alpha_grid[best_i]
best_lambda_min <-cv_list[[best_i]]$lambda.min
best_lambda_1se <-cv_list[[best_i]]$lambda.1se
enet_min<-glmnet(train_x_matrix,train_y_vector,family="gaussian",
 alpha=best_alpha,lambda=best_lambda_min,standardize=TRUE)
enet_1se<-glmnet(train_x_matrix,train_y_vector,family="gaussian",
 alpha=best_alpha,lambda=best_lambda_1se,standardize=TRUE)
coef_min<-coef(enet_min,s=best_lambda_min)
coef_1se<-coef(enet_1se,s=best_lambda_1se)
nz_min_idx<-which(coef_min!=0); nz_1se_idx<-which(coef_1se!=0)
nz_min_coef<-coef_min[nz_min_idx]; nz_1se_coef<-coef_1se[nz_1se_idx]
vars_min<-rownames(coef_min)[nz_min_idx]
vars_1se<-rownames(coef_1se)[nz_1se_idx]
num_vars_min<-length(vars_min)-1
num_vars_1se<-length(vars_1se)-1
coef_min_no_int<-nz_min_coef[-1]; names(coef_min_no_int)<-vars_min[-1]
coef_1se_no_int<-nz_1se_coef[-1]; names(coef_1se_no_int)<-vars_1se[-1]
pca_result<-prcomp(train_x,center=TRUE,scale.=TRUE)
eigenvalues<-pca_result$sdev^2
prop_var<-eigenvalues/sum(eigenvalues)
cumvar<-cumsum(prop_var)
kaiser_components<-sum(eigenvalues>1)
n_comp_70<-which(cumvar>=0.7)[1]
set.seed(100)
n_perm<-1000
perm_eigenvalues<-matrix(0,nrow=ncol(train_x),ncol=n_perm)
for(i in 1:n_perm){
 perm_data<-apply(train_x,2,sample)
 perm_pca<-prcomp(perm_data,center=TRUE,scale.=TRUE)
 perm_eigenvalues[,i]<-perm_pca$sdev^2}
perm_threshold<-apply(perm_eigenvalues,1,quantile,probs=0.95)
perm_components<-sum(eigenvalues>perm_threshold)
set.seed(100)
B<-1000; n<-nrow(train_x); m<-ncol(train_x); k<-5
eigs.boot<-matrix(0,m,B); boot_var<-numeric(B)
for(i in 1:B){
 samp_index<-sample(1:n,size=n,replace=TRUE)
 boot_sample<-train_x[samp_index,]
 boot_pca<-prcomp(boot_sample,center=TRUE,scale.=TRUE)
 eigs.boot[,i]<-boot_pca$sdev^2
 boot_var[i]<-sum((boot_pca$sdev[1:k]^2)/sum(boot_pca$sdev^2))}
k_values<-3:5
results<-data.frame(k=k_values,Test_RMSE=NA,Test_MAE=NA,VAF=NA)
for(i in 1:length(k_values)){
 k<-k_values[i]
 pcr_mod<-pcr(life_expectancy~.,data=train_pcr_data,ncomp=k,scale=TRUE)
 pred<-predict(pcr_mod,newdata=test_pcr_data,ncomp=k)
 results$Test_RMSE[i]<-sqrt(mean((test_y-pred)^2))
```

```
 results$Test_MAE[i] <-mean(abs(test_y-pred))
 results$VAF[i]      <-sum(pca_result$sdev[1:k]^2)/sum(pca_result$sdev^2)}
pcr_final<-pcr(life_expectancy~.,data=train_pcr_data,ncomp=5,scale=TRUE)
final_k<-5
loadings_matrix<-pcr_final$loadings[,1:final_k]
for(i in 1:final_k){
 comp_loadings<-loadings_matrix[,i]
 top_vars<-sort(abs(comp_loadings),decreasing=TRUE)[1:5]
 top_names<-names(top_vars)
 for(v in top_names){
  tmp<-comp_loadings[v]}}
train_pred_pcr<-as.numeric(predict(pcr_final,newdata=train_pcr_data,
 ncomp=final_k))
test_pred_pcr <-as.numeric(predict(pcr_final,newdata=test_pcr_data,
 ncomp=final_k))
test_rmse_pcr<-sqrt(mean((test_y-test_pred_pcr)^2))
test_mae_pcr <-mean(abs(test_y-test_pred_pcr))
test_r2_pcr  <-1-sum((test_y-test_pred_pcr)^2)/
 sum((test_y-mean(test_y))^2)
n_test<-length(test_y)
test_adjr2_pcr<-1-(1-test_r2_pcr)*(n_test-1)/(n_test-final_k-1)
test_pred_min <-as.numeric(predict(enet_min,newx=test_x_matrix,
 s=best_lambda_min))
test_pred_1se <-as.numeric(predict(enet_1se,newx=test_x_matrix,
 s=best_lambda_1se))
test_rmse_min<-sqrt(mean((test_y_vector-test_pred_min)^2))
test_mae_min <-mean(abs(test_y_vector-test_pred_min))
test_r2_min  <-1-sum((test_y_vector-test_pred_min)^2)/
 sum((test_y_vector-mean(test_y_vector))^2)
test_rmse_1se<-sqrt(mean((test_y_vector-test_pred_1se)^2))
test_mae_1se <-mean(abs(test_y_vector-test_pred_1se))
test_r2_1se  <-1-sum((test_y_vector-test_pred_1se)^2)/
 sum((test_y_vector-mean(test_y_vector))^2)
n_test<-length(test_y_vector)
adjr2_min<-1-(1-test_r2_min)*(n_test-1)/(n_test-num_vars_min-1)
adjr2_1se<-1-(1-test_r2_1se)*(n_test-1)/(n_test-num_vars_1se-1)
pcr_stability   <-lm(pcr_resid~log(GNI_PerCapita),data=resid_df)
enet_min_stab   <-lm(enet_min_resid~log(GNI_PerCapita),data=resid_df)
enet_1se_stab   <-lm(enet_1se_resid~log(GNI_PerCapita),data=resid_df)
resid_df<-data.frame(
  GNI_PerCapita = exp(test_x$GNI_PerCapita_log)-1,
  pcr_resid     = test_y - test_pred_pcr,
  enet_min_resid= test_y - test_pred_min,
  enet_1se_resid= test_y - test_pred_1se
)
pred_transformed<-pred_transformed[,colnames(train_x)]
pred_final<-predict(pcr_final,newdata=pred_transformed,ncomp=final_k)
pred_values<-round(as.numeric(pred_final),2)
prediction_table<-data.frame(t(pred_values))
colnames(prediction_table)<-pred_countries
```

References

Rodrigo Martinez, Paula Morsch, Patricia Soliz, Cornelia Hommes, Pedro Ordunez, and Enrique Vega. Life expectancy, healthy life expectancy, and burden of disease in older people in the americas, 1990–2019: a population-based study. *Revista Panamericana de Salud Pública*, 45:e114, 2021. doi: 10.26633/rpsp.2021.114. URL https://doi.org/10.26633/rpsp.2021.114.

World Health Organization. Health inequities are shortening lives by decades. https://www.who.int/news/item/06-05-2025-health-inequities-are-shortening-lives-by-decades, May 2025. News release.