Life Expectancy

Regression Analysis Final Project, Spring 2022

Cole Ryan
Ebise Abdi
Theophilus Anim Bediako

Instructor: Dr. Gemechis Djira

South Dakota
State University

May 5, 2022

Table of contents

1. Introduction
2. Objective
3. Data Description
4. Data Description
5. Fitting the Models
6. Conclusion

Introduction

What’s life expectancy?

★ The number of years a person can expect to live.

★ It is an important indicator for gauging population health.

★ It is based on an estimate of the average age individuals of a particular population will be when they die.

★ In practice, estimating life expectancy entails predicting the likelihood of surviving successive years of life.

Objective

Research problem

★ Data exploratory discussion and interpreting the result.

★ Fitting a models and comparing the models based on their prediction accuracy.

★ What are the factors affecting life expectancy?

Theories

Ordinary Least Squares Regression(OLS) attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The OLS regression is used under the following assumptions: linear relationship,normality, no autocorrelation, homoscedastic (constant variance in residuals), more observations (n) than features(p) and no or little multicollinearity.

Typically, when the assumptions underpinning OLS regression are satisfied, the model coefficients are unbiased and have the smallest variance among all possible linear estimators. Due to the large volume of datasets in todays world, our OLS assumptions are often violated. When these concerns arise, regularized regression (also known as penalized models or shrinkage methods) can be used to manage parameter estimates as an alternative to OLS regression. Regularized regression puts constraints on the magnitude of the coefficients and will progressively shrink them towards zero, reduce the variance and decrease our sample error (Boehmke, 2018).

The common regularization methods are ridge regression, the least absolute shrinkage and selection operator(lasso) regression and the elastic net. (Zou & Hastie, 2005). The objective function of regularized regression methods is very similar to OLS regression; however, we add a penalty parameter (P).

Data Description

Data Description

★ The original data is owned by the World Health Organization (WHO).

★ The data set contains 3111 observations of 31 independent variables.

★ The observations are obtained from 183 countries from 2000 to 2016 (17 years).

★ We generated a random sample of size 51 out of 183 countries from 2016.

★ The sample data is chosen randomly, about 1/3 of the countries from each region.

★ The sample data contains 51 of observations of 17 variables.

Cont. . .

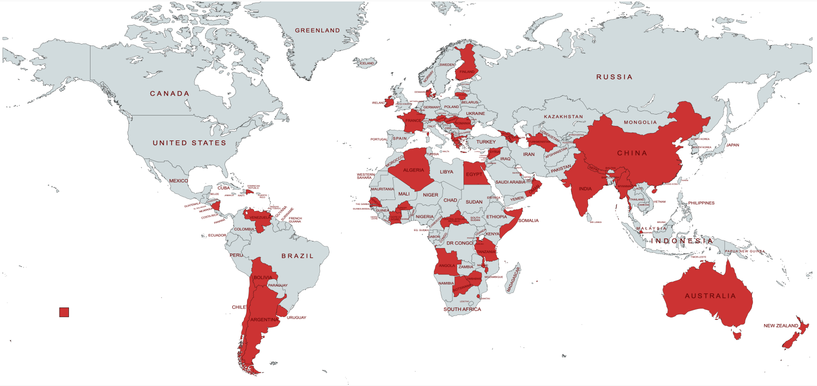

Figure 1: Sampled countries

The visualization of countries included in sample data.

Cont. . .

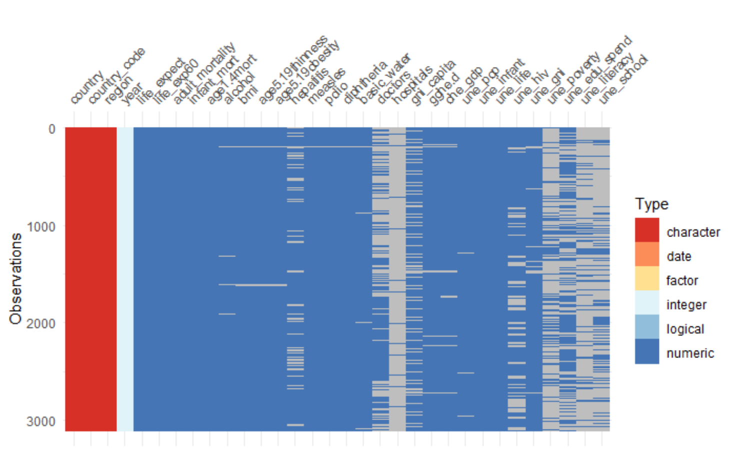

Figure 2: Data Visualization

We've 15.3% of missing values and 84.7 % present of the original data set.

Cont. . .

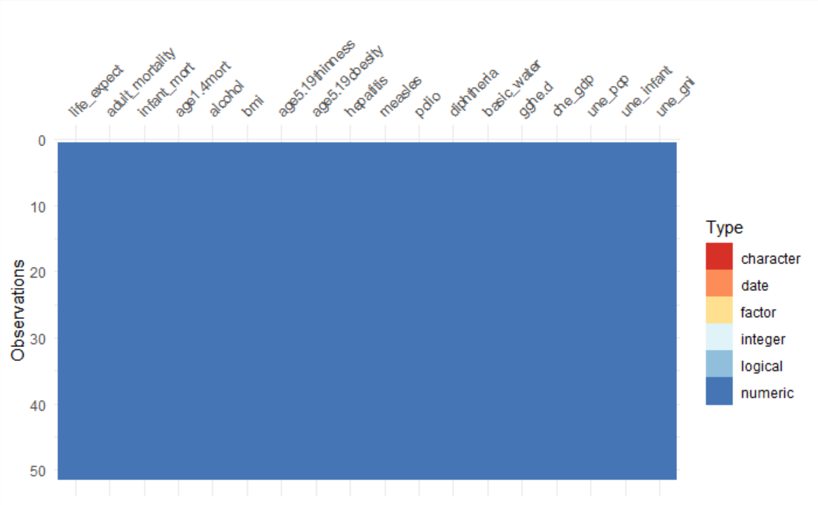

Figure 3: Data visualization after NA values replacement

Correlation plot

The correlation plot indicates that some of the independent variables are highly correlated with each other. The correlation matrix confirms the same result.

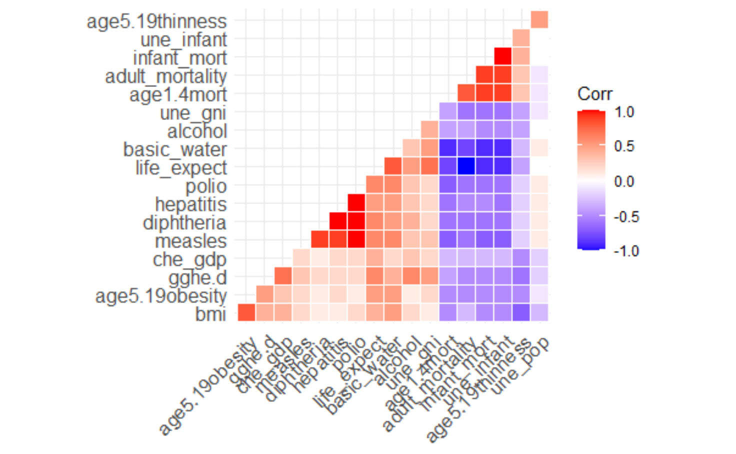

Figure 4: The correlation plot

Fitting the Models

Multiple Regression Model

The objective is to build a multiple linear regression model, in the general form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \quad i = 1, ..., 51 \quad p - 1 = 16$$

where

- ★ Y_i represent the response variable life expectancy,
- ★ $x_{i,p-1}$ are the predictor variables,
- ★ β_0 is the intercept and
- ★ the $\beta_i s$ are the coefficients of the predictor variables.

Cont. . .

As we have seen from the correlation plot 4, some of the variables are highly correlated. We can use the VIF to remove multicollinearity until our model has a low amount of co-linearity.

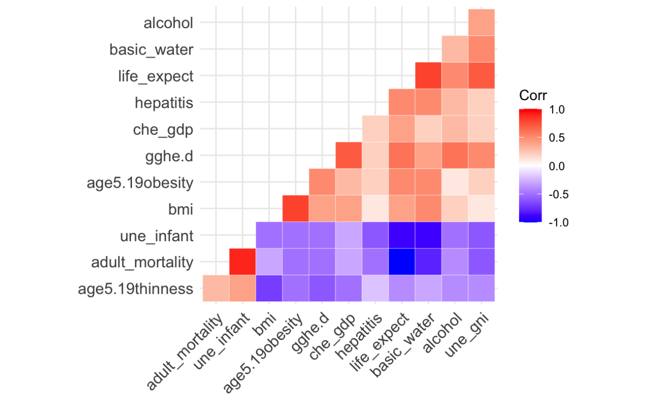

Figure 5: Correlation plot removing highly correlated variables

Cont. . .

The next reduced model is model with reduced co-linearity and a VIF of less than 10. Using this new model with a reduced number of variables, we can use the step function with the AIC criteria to find a new model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \quad i = 1, ..., 51 \quad p - 1 = 6$$

In the next step we used the transformation $Y^{1/2}$ and $Y^{-1/2}$ with the reduced number of variables.

Cont. . .

Response variable

Life expectancy at birth

Independent variable

★ Adult Mortality

★ Infant mortality rate (UNE infant)

★ Gross national income (une-gni)

★ Domestic general government health expenditure as a percentage of gross domestic product (GDP) (gghe-d)

★ Basic water

★ Alcohol

Cont. . .

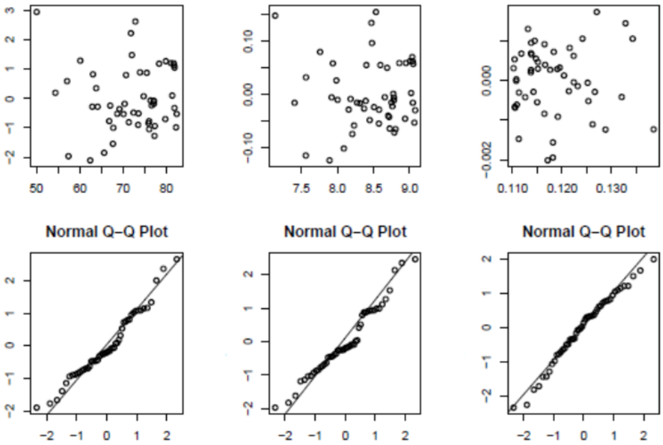

Figure 6: The comparison of full model and transformed models $Y = Y^{1/2}$ and $Y = Y^{-1/2}$

Cont. . .

Comparing the Adjusted R-squared:

Y	$Y = Y^{1/2}$	$Y = Y^{-1/2}$
0.9776771	0.9803632	0.9832662

For adjusted r-squared again, the $Y^{-1/2}$ model performs the best. This model also meets the assumptions best as well, so we will consider it against other models we can make. Other approaches can be used for regression as well. We can use ridge, lasso, and elastic net regression to build other models.

Cont. . .

Full model	Transformed	Ridge	lasso	elastic
0.9797	0.9833	0.9850	0.986	0.9853

This shows that the ridge regression and elastic nets models are providing the best results based solely on adjusted R-squared. When looking at adjusted R-squared what we can see is that the full model is unlikely to be our most suitable option.

Cont. . .

Now that we have built some models, we can compare how they perform when making predictions on new data. We can use the 132 countries that were not part of our original sample.

Full model	Transformed	Ridge	lasso	elastic
2.162	1.462	1.266	1.344	1.257

Conclusion

Conclusion

- These are the MSE for the full model, the model build with AIC criteria, the ridge regression model, the lasso model, and the elastic-net (the last four all built using the $Y = Y^{-1/2}$ transformation) when tested on the non-sample data.

- We can see from this that the elastic-net has the lowest MSE, and is probably the best model that we have built for making new predictions.

Reference

- Boehmke, B.(2018). UC Business Analytics R Programming Guide. University of Cincinnati.

- MMattson. (2020, October 6). Who national life expectancy. Kaggle. Retrieved April 7, 2022, from `https://www.kaggle.com/datasets/mmattson/who-national-life-expectancy/metadata`

- Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie (2013) - "Life Expectancy". Published online at OurWorldInData.org. Retrieved from: `https://ourworldindata.org/life-expectancy`

- Michael H. Kutner, Chris Nachtsheim, John Neter (2004). Applied Linear Regression Models, 4th edition. McGraw-Hill/Irwin. ISBN: