Life expectancy Project Report

Litian Zhou

University of Pittsburgh

**1. Introduction**

The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) records health status-related factors for all countries within the United Nations. These datasets are available to the public for the purpose of health data analysis. The dataset used in this project consists of the average life expectancy in a specific year for 193 countries. The dataset also includes other related health factors in these countries collected by the WHO and corresponding socio-economic data collected from the United Nations website. The observations were collected over the past 15 years, and there has been an obvious development in health resulting in an improvement of human life expectancy especially in developing nations. In this project, we analyze data from the year 2000-2015 for 193 countries to make statistical inferences about life expectancy and build a model to make predictions for new emerging health data.

Missing data only occurred in some smaller countries, so we excluded these countries from the final model dataset. The final dataset consists of 22 columns (21 predictors, 1 response) and 2938 rows (observations). All predicting variables fell into several broad categories: Mortality factors, Public health factors, and Socio-economic factors (Figure 1).

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Figure 1. The landscape of the dataset (“Life Expectancy (WHO),” n.d.).

The project aims to answer the following key questions:

1.      Do various initially chosen predicting factors (infant and adult mortality rate, schooling, and immunization coverage) significantly affect the life expectancy? If so, how do each impact life expectancy?

2.     Do other, unexpected variables affect life expectancy? If so, which factors, and how do they impact life expectancy?

3.     Which method and method parameters can build the model which has the best performance while maintaining interpretability; and which predictors does the model select?

4.     What can we conclude from the most interpretable model? What suggestions can we give in terms of both personal and public health across the countries?

Upon the completion of the project, we want to provide insight which can help people make lifestyle changes that can help them to live longer or help governments invest more money in infrastructure facilities or public health programs that increase the average life expectancy of the country’s population.

**2. Methods Overview**

The data file consists of 1649 effective samples, large enough for meaningful interpretation without implementing resampling methods.

The samples are collected from 194 countries between 2000 and 2015. However, as mentioned in the introduction, our major goal is to identify the main factors that might influence the life expectancy and provide guidance for improvement, therefore such exogenous variables like country and time might be reasonable to be removed (as it would not be practical to recommend that people change their birth year, nor ethical to incite people to simply flee from poor countries toward developed countries, etc.).

The predictor variables can be mainly divided into 3 categories, with only one qualitative predictor (development status) and 19 quantitative predictors:

1. Mortality rates: including the number of infant deaths, deaths under age five, and adult deaths (between 15 and 60 years old), measured in deaths per thousand.

2. Public health status: including the immunization coverage rates for hepatitis B, polio and diphtheria (measured in percentages between 0-100%); alcohol consumption per capita (in liters); the occurrence of measles (in reported cases per 1000 population); average BMI for entire population; prevalence (in percentages) of thinness among children (age 5 to 9) and teens (age 10 to 19); and deaths per 1000 live births HIV/AIDS (0-4 years).

3. Socio-economic factors including development status (binary: developed or developing); GDP (in USD); health shares (in percentages) in GDP and government expenditure; income composition of resources (index between 0 and 1); and population.

Our response variable, life expectancy, is also purely continuous with no replicates for each data point, making the problem ill-fitted for multi-level classification. Though it might be a little left skewed, there is no trend suggesting significant outlying cases. Therefore, it makes little sense to divide the response variable into subcategories and conduct classifications on it. Our project is naturally a regression problem and is difficult to convert to a useful classification problem based on the interest of our project.

Additionally, since our major concern is about statistical inference, many non-parametric techniques such as k-nearest neighbors do not seem to be applicable. Instead, we preferred to use parametric regression techniques and model selection techniques such as stepwise selection and best subset selection to find the best model.

Intuitively,  citizens from richer countries would pay more attention to life expectancy, as Jones (2015) mentioned in his paper “The Facts of Economic Growth”(Jones, 2015). With more investment towards public health industry from both private and public sectors, people in the richer countries tend to have higher immunization coverage, lower disease rate, and possibly better lifestyle, which would result in a lower mortality rate. Therefore, there were strong correlations between the predictors, and consequently an issue of collinearity. Ridge regression, lasso regression, PCA and Partial Least Squares methods were tested and compared. Moreover, the Generalized Additive Model was performed to explore the effect of predictors on a more detailed level.

**3.** **Methods Details**

Since the question is inherently a regression problem. We tried the following methods: OLS( Ordinal Least Square), BSS (Best Set Selection), Forward Selection and Backward Selection, Ridge regression and lasso, PCR (Principle Component Regression) and PLS (Partial Least Square), GAM (Generalized Additive Models). Here are the details of those methods.

Before I fit any model, I initialized the validation set for all the model. Thus, we have a standard validation set across multiple methods which gives a stable evaluation.

I first looked at the OLS model. The full model (20 predictors) is fitted by glm() function and tested use both Validation Set approach and LOOCV (Leave One Out Cross Validation). Then, I performed BSS, forward and backward selection using regsubsets() function under the package “leaps.” Under different criteria, the BSS gives model in different size, ranging from 9 to 12. Here is the two examples using Adjusted R Square and BIC as the model selection formula (Figure 2).

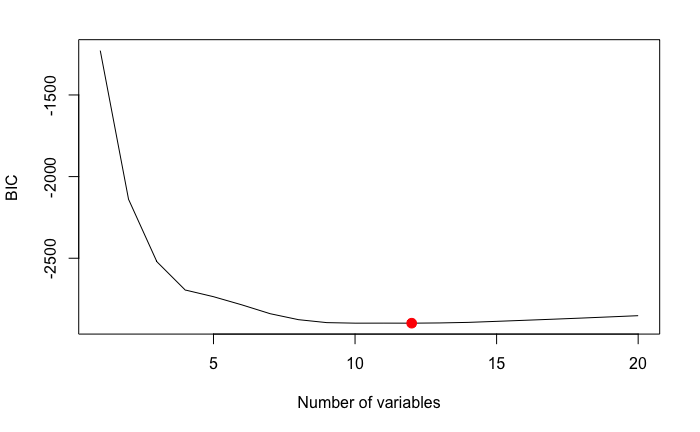
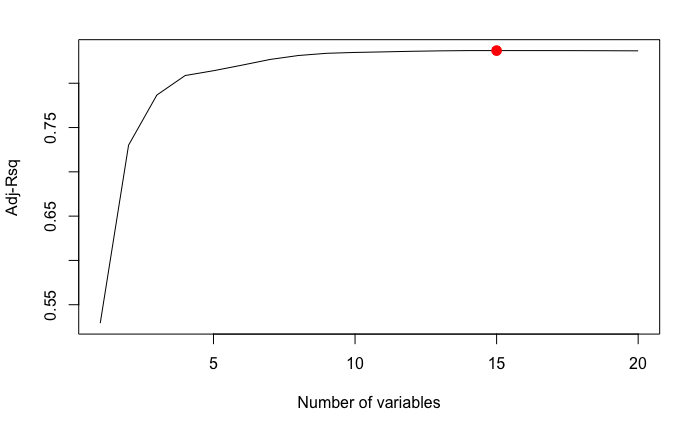


Figure 2. The BSS method using different criteria. Left: Adjusted R Square select the best model with 15 variables. Right: BIC selects the best model with 12 variables. Red dots: the best model selected.

Then, I compared the forward and backward selection method with max variable number 20. They give back the identical models for the 12-predictor models.

Furthermore, I tried the 10-fold Cross Validation to choose the best linear model. It gives back a 14-variable model (Figure 3).

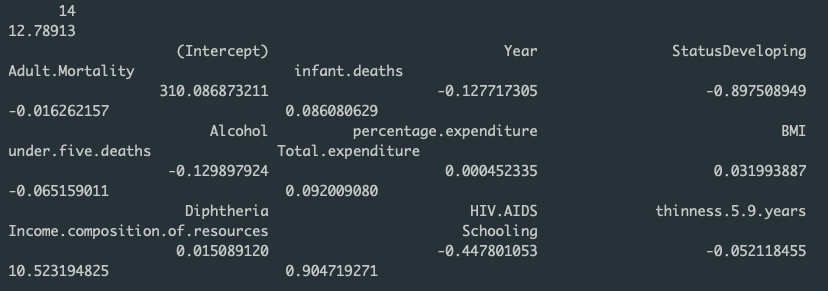
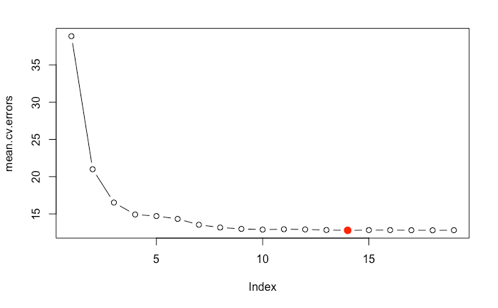


Figure 3. The best model selected by 10-fold Cross Validation. The MSE is 12.78913, while the coefficient of 14 variables are listed. We will use this model for GAM fitting later in this session.

Next, I examined Ridge and lasso methods. For Ridge, the best lambda is around 0.7 with a relative high test MSE: 14.39. For lasso, the lambda is around 0.004 with test MSE of 13.78. The small lambda means all predictors have high significance in the model. Thus, lasso did not force any coefficients to zero, leaving twenty predictors as the full model.

I also tried dimension reduction methods such as PCR and PLS. Using cross validation test with method pcr() and plsr() respectively, PCR chose 10 components and PLS chose 5 (Figure 4).

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 4. Using cross validation to choose number of components in PCR (left) and PLS (right) model. To balance the interpretability and model accuracy, I deliberately chose M=10 in PCR and M=5 in PLS.

Moving beyond linearity, I first tried polynomial regression and splines which reader can check from the appendix file from line 163 to 179. As for the GAM model, I picked up the 14 variables chosen by Cross Validation (See Figure 3) and fit them accumulatively. Note that I remove the following variables: "Adult.Mortality","infant.deaths" , "under.five.deaths" and "percentage.expenditure" which are highly skewed. After fit 10 models, I used ANOVA test to determine which variables significantly explain life expectancy. The result is only the "Total.expenditure" has a relatively low p-value (0.001098), so we remove it and fit the final GAM model with left 9 variables.

In conclusion, I made a bar plot showing the (test) MSE of all the model above (Figure 5), and found all the models have similar MSE while GAM has the best performance (Figure 6). I conclude that GAM, with the highest predict accuracy and outstanding interpretability is the best method in this problem.

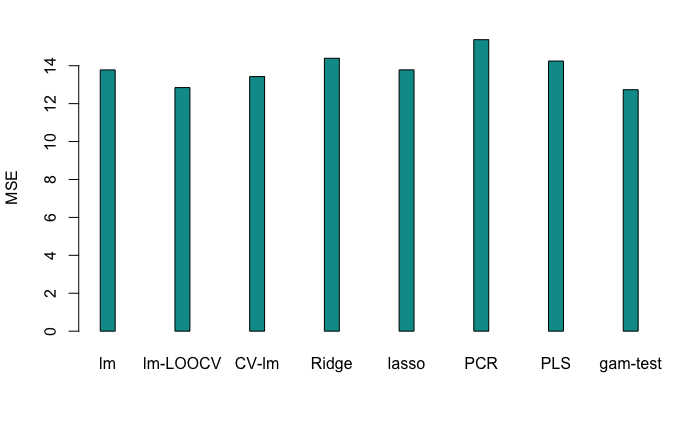


Figure 5. The test or Cross Validation MSE for eight models / methods I used in this project. See details in the appendix file.

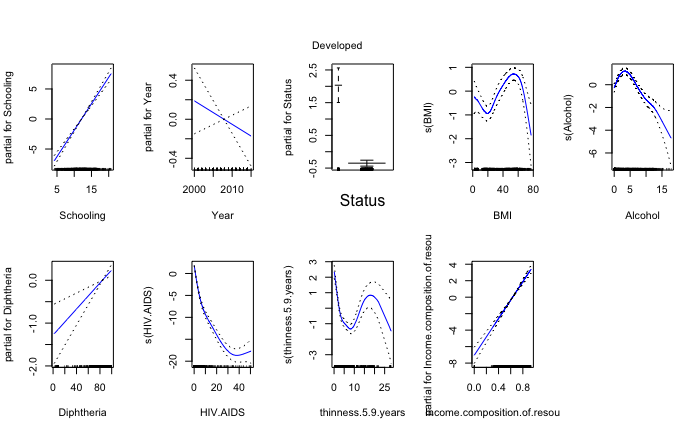


Figure 6. The GAM model with 9 variables selected by ANOVA table.

In the end, we tried the Decision Tree Methods including Regression Trees, Bagging and Random Forest. We chose four variables for the regression tree. Then we used cross validation to decide the best pruning level, which produces nine splits (Figure 7B). The method has a relatively low test MSE of 9.7.

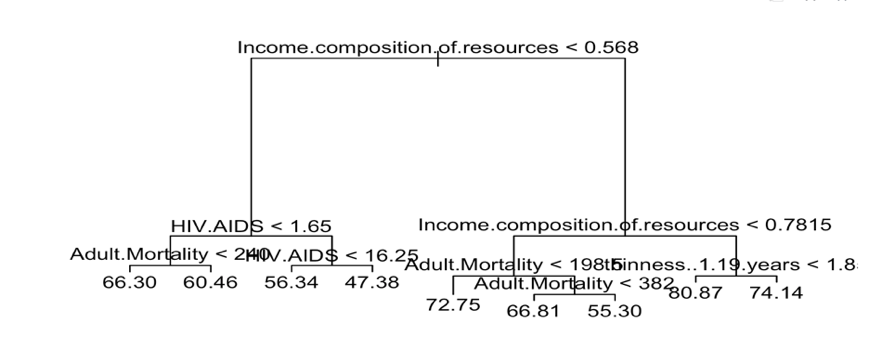
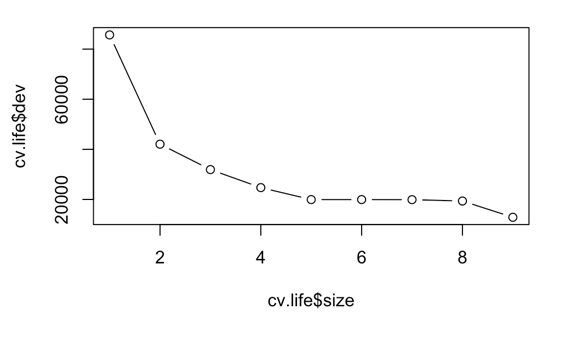


Figure 7. The Decision Tree approach. Left: CV chosen variable number. For better interpretation, we chose four predictor model. Right: After the pruning, a nine-split tree is built.

We then performed Bagging and Random Forest. We bootstrapped 500 trees for bagging, resulting a test MSE of 3.7. As for Random Forest, we picked Mtry=6, resulting a test MSE of 3.67. While the result is similar, I think more trees number are needed as both Bagging and Random Forest will not overfit due to more bootstrapping.

**4.**  **Summary of** **Results**

Coming together as a group, we fortunately had very similar findings. The following tables indicate that our test MSE values (Table 1) and R^2 (Table 2) are all quite similar, and the methods that appeared to perform best are consistent throughout. Even though forward and backward stepwise selection gave the same results for all of us, different group members still chose different subsets of predictors based on the penalty term they used for model selection (BIC, AIC, CP, Adjusted R^2 or cross validation). This resulted in very slight differences for values of model performance metrics, but the trends between methods remain consistent for all of us. For the linear cases, the models are all consistent, especially between OLS, ridge regression, and lasso. The three models deemed the same variables as important, which is reasonable as the shrinkage values chosen by cross-validation on ridge and lasso are very low, making them nearly the same as OLS. This also explains why no particular method does far better than the other–among the three, test MSE and test R^2 are extremely close; however, these three methods have much better performance than PCR. PCR (as well as PLS) of course creates a linear combination of all predictors, so we cannot compare which variables the method deems most important, but we can see that PCR produces an inferior fit to the data with higher MSE and lower R^2 values. Contrarily, PLS produces a fit quite similar to the first three methods; therefore, the unsupervised aspect of PCR may be impairing its performance.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test MSE** | OLS Full | OLS Reduced | Ridge (lambda) | Lasso  (lambda) | PCR  (# of comp) | PLS  (# of comp) | GAM |
| Sam | 13.13 | 14.55 | 14.14 (0.7) | 13.83 (0.003) | 13.83 (5) | 13.83 (2) | 7.85 |
| Aaron | 13.72 | 13.95 | 13.71 (.01) | 13.73 (.01) | 15.9 (5) | 14 (3) | 8.85 |
| Tianke | 12.79 | 12.71 | 12.79 (0.001) | 12.80  (0.004) | 14.88  (7) | 13.81(3) | 7.21 |
| Litian | 13.78 | 13.43 | 14.39 (0.71) | 13.79  (0.006) | 16.62  (7) | 16.48  (2) | 12.73 |
| Alex | 13.78 | 14.94 | 13.83 (0.01) | 13.84 (0.011) | 14.75 (5) | 13.78 (3) | 13.80 |

Table 1. Group members test MSE values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test R squared** | OLS Full | OLS Reduced | Ridge | Lasso | PCR | PLS | GAM |
| Sam | 81.56% | 81.91% | 80.67% | 81.56% | 77.85% | 80.09% | 89.8% |
| Aaron | 80.4 | 80.24% | 80.6% | 80.6% | 77.5% | 80% | 88.5% |
| Tianke | 84.39% | 84.48% | 84.39% | 84.37% | 81.83% | 83.13% | 91.2% |
| Litian | 83.66% | 85.52% | 83.33% | 82.23% | 81.91% | 81.89% | 88.88% |
| Alex | 81.63% | 80.48% | 81.56% | 81.54% | 80.34% | 81.63% | 81.97% |

Table 2. Group members R square values.

Lastly, the non-linear method appears to do significantly better than all other attempted methods. Although the linear models fit very well, the GAM model we built using the same predictors as from those linear models performed with nearly half the test MSE, and the test R^2 increased by at least 8%. Although it produces the most complex model, it is still interpretable and represents some non-linear trends that the previous models do not capture. We are concerned that this complex model may be overfitting our population, and we would like to see if the test error would still be as low if we had thousands of more observations as opposed to 1,649. For this reason, we still choose to continue with the OLS model that requires less data, has very interpretable coefficients, and still produces a very adequately fit model.

**5.**  **Conclusion and Takeaway:**

As a group, we discussed and made the following conclusions:

* All models have very similar test MSEs, while the non-linear models have a potential to make more accuracy predictions.
* BSS, forward selection, backward selection do not produce small models as we expect. Luckily, they all agree with each other by choosing similar variables. Meanwhile, Ridge Regression and Lass also have a small lambda by Cross Validation, which do not provide large shrinkage either. As a result, the models are always relatively big with nine to fifteen predictors.
* The above conclusion makes senses that the life expectancy of people around the world is a complex problem influenced by many factors.
* GAM has the lowest test MSE, and also good interpretability. Thus we can make prediction and inference accurately and easily.
* The good performance of the Tree Methods implies that there are different groups clustering together with similar life expectancy. As a result, it is easy for decision trees to make accurate prediction.

As takeaway for the general readers, there are many interesting interpretation from the GAM model (Figure 6). As I said, the GAM model not only has a low test MSE, but provides good interpretability.

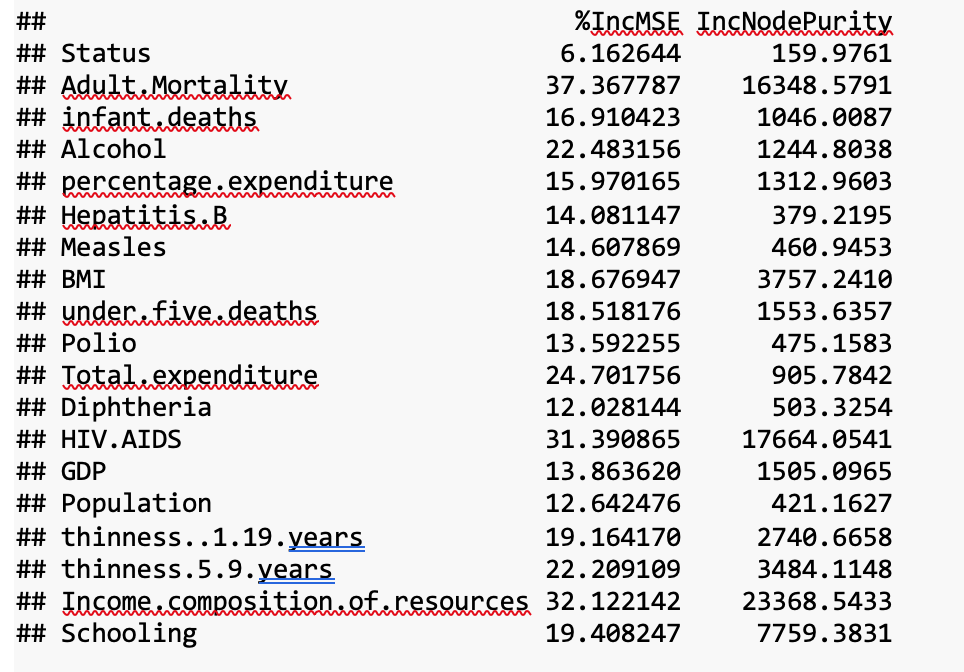
1. Only the variable measures the government expenditure on health shows relatively low significance in the initial ten-variable model, which means life expectancy is more of a personal matter than government control.
2. The scholars have relatively longer life expectancy as they increases the length of schooling.
3. Individual should keep their BMI in the standard range and take vaccine such as DTP3.
4. Economics is still one of the most important factor of longer life expectancy. Since better education, public health system and hospitals are all associated with GDP.
5. Both government and individual should control spread and take precaution against AIDS.
6. For analysts, we also provide a table of Variable importance from Tree model(Table 8). See code and details in appendix.

Table 8. Variable Importance Table After Using Bagging. Importance ranges from low to high.

1. Conflicted with current medical knowledge (GBD 2016 Alcohol Collaborators, 2018), a small amount of alcohol intake is positively related with longer life span. However, there could be some confounding factors like financial condition.

Finally, I hope our project can bring insights for how life expectancy is affected by the complex world, which will provide some suggestion to both individual and governmental sections for healthier and longer life.

References

GBD 2016 Alcohol Collaborators. (2018). Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet (London, England)*, *392*(10152), 1015–1035. https://doi.org/10.1016/S0140-6736(18)31310-2

Jones, C. (2015). *The Facts of Economic Growth*. Retrieved from National Bureau of Economic Research, Inc website: https://EconPapers.repec.org/RePEc:nbr:nberwo:21142

Life Expectancy (WHO). (n.d.). Retrieved April 7, 2019, from https://kaggle.com/kumarajarshi/life-expectancy-who