STATS513
Final Project 2 Life Expectancy

junhuuu@umich.edu

April 28, 2022

1 Lay Abstract

This work is conducted based on the the dataset includes $Life.expectancy$, health factors and economic data for 183 countries. It has been observed health factors and economic status may affect the lift expectancy. From this work, we can conclude there are five factors influence $Life.expectancy$ which are $Status, Alcohol, BMI, GDP$ and $Schooling$ respectively. However, $Status$ and $Alcohol$ has a very slightly influence. There is 758.8 increase in $Life.expectancy^3$ for each one increase in BMI, There is 1.18 increase in $Life.expectancy^3$ for each one increase in GDP, And there is 27090 increase in $Life.expectancy^3$ for each one increase in BMI .From our prior knowledge, we may think that the $Status$ of a country maybe have some inner relationship with other health and economic factors so that the variables will influence the $Life.expectancy$ with respect to developed and developing countries differently, however after analysis we find the effects are the same among developed and developing countries.

2 Introduction and Data Summary

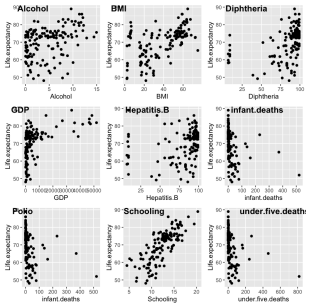

Figure 1: scatterplot

Firstly, we check the missing values, then we find that there are 41 rows in our dataframe that contains missing value, and we remove these rows.According to the description, we know that $infant.deaths$ and $under.five.deaths$ should be equal or less than 1000. However the maximum of $infant.deaths$ is 1200.00 and the maximum of $under.five.deaths$ is 1600.00. Therefor there are some entry errors for

the two variables, we are supposed to filter these errors.
Secondly, no matter the histogram or boxplot, we can see that the life expectancy in developed countries are much longer than develoing countries. Also, the range in developing countries is larger than developed countries. The developed countries have a relatively more stable distribution in life expectancy compared with developing countries. From the summary result above, we can notice that there are some "abnormal values". For example, the maximum values for $infant.deaths$ and $under.five.deaths$ are 521 and 817, respectively, which are not convincing to some extent.

Thirdly, The $Life.expectancy$ as dependent variable has somewhat strong positive correlation with

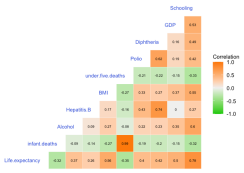

Figure 2: correlations between numerical variables

$Schooling$, we are going to see it further on the model analysis. On the other hand, it has negative correlation with $Infant.death$, it is valid since $infant.death$ usually happens at a very young age. The number of Developing Countries on this observations are way bigger than the Developed Countries. On the Development Status, it was clearly that distribution of higher $Life.expectancy$ lies on the Developed Countries, with a significant Median distance. As the p-value ANOVA Analysis is less than the significance level 0.05, we can conclude that there are significant differences of Life Expectancy between the Developed and Developing Countries.

3 Data Analysis

3.1 Data Analysis A.1

3.1.1 Identify and deal with unusual points

For leverage points, we have the 24 and 97, And we have 91,97 and 99 to be influential points. Therefore, I decide to delete the 97 row in our training set so that to avoid extreme values distracting model.

3.1.2 Variable Transformation

Carried out the Box-Cox transform analysis and generated plot of the likelihood function with the maximum-likelihood estimate and 95% confidence intervals shown on the plot. It looks like a reasonable transformation might be the degree of 3, so we utilize that power law of y^3 and re-run least squares regression to get new least-squares estimates with this transformed model.

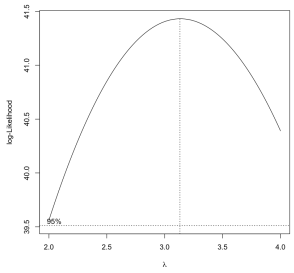

Figure 3: Box-Cox analysis

3.1.3 Variable selection

After dealing with unusual points and doing variable transformation, we fit a new linear model:

$$Life.expectancy^3 \sim Status + infant.deaths + Alcohol + Hepatitis.B + BMI + Polio + Diphtheria + GDP + Schooling \quad (1)$$

Here, we have training RMSE and testing RMSE are 4.88508755036417 and 6.81139090761671 respectively. Then we try to use stepwise AIC ,exhaustive AIc and BIC for model selection. No matter which method we use, we have training RMSE around 4.5 and testing RMSE around 6.5. And all the three method choose 5 variables which are $Status$, $Alcohol$, BMI, GDP and $Schooling$.

3.1.4 Robust method

Robust regression is an alternative to least squares regression when data are contaminated with outliers or influential observations, and it can also be used for the purpose of detecting influential observations. From results above, we can see that given our dataset is a small dataset with somehow large variables, our model is easily to be infected by outliers or influential observations, so here we use robust method.I apply two robust methods (i.e. Huber's method and Least absolute deviations) on the regression with selected five variables. For the two robust model, the training RMSE are still around 4.5 and the testing RMSE are still around 6.5. So we can see that the differences between different methods are really small.The difference between different models are really small, which is approximately 0.001.

3.1.5 conclusion

Based on the analysis above, I did the following steps:
1. Two rows of data were removed from the training dataset because they are leverage/influential points.
2. A transformation was performed on the response variable, i.e. $Life.expectancy \rightarrow Life.expectancy^3$.
3. Several criterion-based methods were used for predictor selection. Eventually 5 predictors were selected.
4. Two robust methods were performed (Huber's method and Least absolute deviations), and it turned out the result differences are really small.

In conclusion, I choose my final model as follows:

$$Life.expectancy^3 \sim Status + Alcohol + BMI + GDP + Schooling$$

So from the two plots, we can see that the model follows the assumption pretty well.

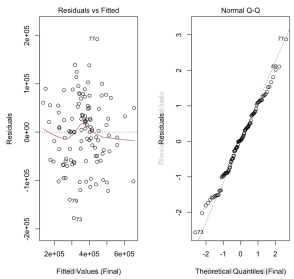

Figure 4: final selected model

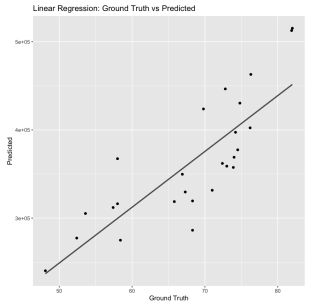

Figure 5: OLS

3.2 Data Analysis A.2

From A1, the final model has 5 predictors. From the model summary, there are 2 significant predictors, which are BMI and $Schooling$. Therefore, we need to make an analysis these 2 predictors to see whether they differently effects between developed and developing countries. Then I separately induce the interaction term i.e other factors interact with Status to do summary and ANOVA analysis. Then we find that p-values in the anova analysis are larger than 0.05, so we cannot conclude that the factors have different effects between developed and developing countries. Also, for interaction with $Alcohol$ and GDP, we also find that p-values in the anova analysis are larger than 0.05, so also we cannot conclude that the factors have different effects between developed and developing countries.

4 Discussion

At the begining, in my work I try to add some features like interaction between the variables, the second, third and even fourth degree of the variables. Since this is a small dataset, after adding new features,

the RMSE exploded even reached 1000+, so finally I gave up adding new features. With many features and small observations, the model is easily effected with influntial points and outliers.
There are several limitations in this work. Firstly, in this work, we exclude the missing data when we analyze this dataset. However, there maybe some underlying missing patterns and our result may be biased since we simply handle missing data by deleting the observation. Secondly, we divide the traning dataset and testing data set with specific division. For more accurate and in further work, we can use k-fold method for model selection so that maybe the traning RMSE will somehow increase, but it will increase accuracy when deal with unseen datasets. In our work, we have a much larger traning dataset compared with our testing dataset, there is high potential that overfitting happens.