Life Expectancy Explanation with Health and Economical Factors

1 Lay Abstract

This work is conducted based on the survey data of 183 countries from the Global Health Observatory (GHO) data repository under World Health Organization (WHO) to explore the relationship between life expectancy and some health and economical factors. Based on the best weight least square model we fitted, it can be concluded that developed countries with less alcohol consumption, longer schooling years, lower under-five deaths and higher GDP tend to larger life expectancy. Keep the other factors the same, there is a 2.59-year decrease of life expectancy for developing countries compared with developed countries; 1 litres increase of alcohol consumption per capita (15+) would lead to 0.39-year decrease of life expectancy; 1-year increase of schooling would leads to a 1.92-year increase of life expectancy; if the number of under five deaths increases to 10-fold, there would be a 1.95-year decrease of life expectancy; if the GDP increases to 10-fold, there would be a 1.43-year decrease of life expectancy. Furthermore, the effects of these significant factors (except status) are the same among developing and developed countries.

2 Introduction and Data Summary

In this report, the scientific goal is to understand how health and economical factors impact the life expectancy of countries. The dataset contains data from the Global Health Observatory (GHO) data repository under World Health Organization (WHO).

The dataset includes life expectancy, health factors and economic data for 183 countries. There are 11 variables in total, and the details of definition, type and number of samples of these variables are shown as Table 1 and 2.

Variable	**Definition**	**Type**	**Number of Samples**
`Life.expectancy`	Life expectancy in years.	Numerical	non-NA - 183
`Status`	Developing status for each country.	Categorical	Developed - 32; Developing - 151.
`infant.deaths`	Number of infant deaths per 1000 population.	Numerical	[0, 1000] - 182; 1200 - 1.
`Alcohol`	Recorded per capita (15+) consumption (in litres of pure alcohol).	Numerical	non-NA - 182; NA - 1.
`Hepatitis.B`	Hepatitis B (HepB) immunization coverage among 1-year-olds (%).	Numerical	non-NA - 168; NA - 15.
`BMI`	Average Body Mass Index of entire population.	Numerical	non-NA - 181; NA - 2.
`under.five.deaths`	Number of under-five deaths per 1000 population	Numerical	[0, 1000] - 182; 1600 - 1.

Table 1: Detailed Information on Variables.

Variable	Definition	Type	Number of Samples
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%).	Numerical	non-NA - 182; NA - 1.
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).	Numerical	non-NA - 182; NA - 1.
GDP	Gross Domestic Product per capita (in USD).	Numerical	non-NA - 156; NA - 27.
Schooling	Number of years of Schooling (in years).	Numerical	non-NA - 173; NA - 10.

Table 2: Detailed Information on Variables.

After removing countries (samples) with NA's and one sample with entry errors (1200 for `infant.deaths` and `under.five.deaths`), there are 141 samples in total.

Besides, there is strong evidence about collinearity between numerical predictors including `infant.deaths`, `Alcohol`, `Hepatitis.B`, `BMI`, `under.five.deaths`, `Polio`, `Diphtheria`, `GDP` and `Schooling`. The condition number is 6149, the variance inflation factor (VIF) of `infant.deaths` and `under.five.deaths` are 66.26 and 67.5 separately, and the correlation between these two predictors are 0.99. To avoid collinearity problem, we chose to use `under.five.deaths` between these two predictors to fit linear models.

3 Data Analysis

We used first 115 samples of the cleaned dataset with 142 samples as the training data and rest as test data. The training data was used to create the model and then we used test data to evaluate different models to select the best one.

3.1 Data Analysis A.1

3.1.1 Simple Linear Model

Firstly, a simple linear model on all predictors (`base model`) was fit to check normal assumptions and unusual points:

$$\begin{aligned}\texttt{Life.expectancy} =& \beta_0 + \beta_1\texttt{I(Status=Developing)} + \beta_2\texttt{Alcohol} + \beta_3\texttt{Hepatitis.B}\\ &+ \beta_4\texttt{BMI} + \beta_5\texttt{under.five.deaths} + \beta_6\texttt{Polio} + \beta_7\texttt{Diphtheria}\\ &+ \beta_7\texttt{GDP} + \beta_8\texttt{Schooling}.\end{aligned}$$

As Figure 1 shows, there is no evidence about heteroskedasticity and no evidence against the normal assumption.

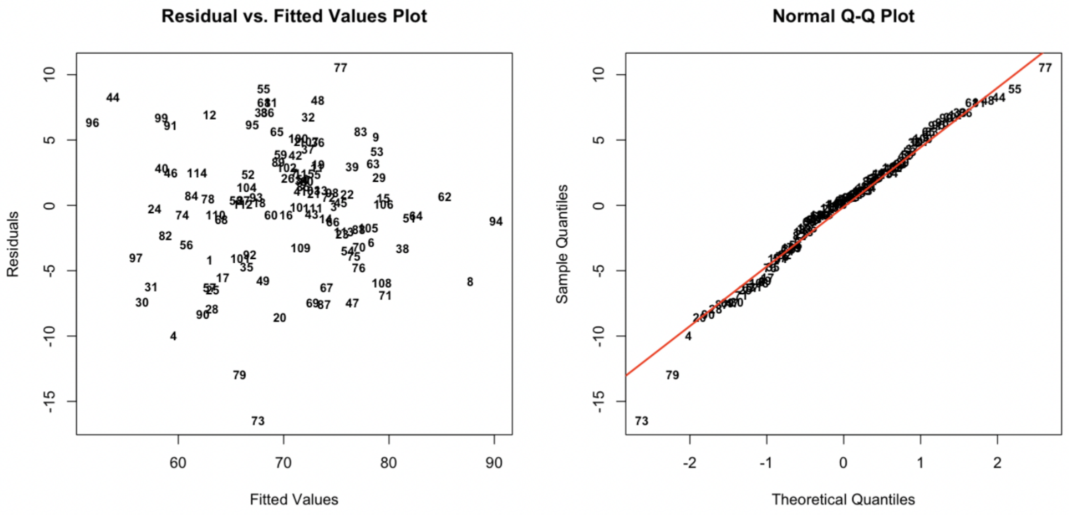

Figure 1: Residual Plot of `base model`.

Moreover, based on `base model`, we removed one outlier with p-value of 0.00085 (smaller than the Bonferroni adjusted $\alpha/n = 0.00087$) and 10 influential points which deviated from most of points in the Cook's distance plot (Figure 2).

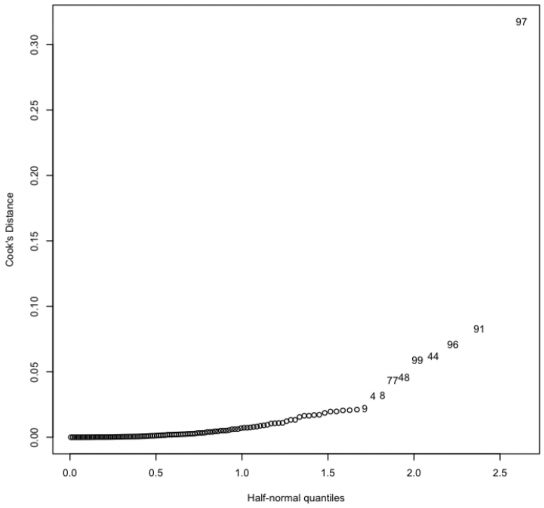

Figure 2: Cook's Distance Plot for `base model`.

After removing these unusual points, we refitted `base model` and checked partial regression plots and partial residual residual plots, we found that the `Hepatitis.B`, `Polio` and `Diphtheria` performed highly similar. The PCA of these three predictors suggested that the first principal component is approximately the average of these three predictors and covered around 98% of total explained variance. So, a new predictor called `immunization` was created by taking average of values of `Hepatitis.B`, `Polio` and `Diphtheria` to represent the average immunization coverage level of a country. We used `immunization` to replace `Hepatitis.B`, `Polio` and `Diphtheria` in `base model`, the adjusted R^2 almost did not change. So we applied this replacement in creating all the other models.

Last but not least, we also found that there are some large values (not outliers) for `under.five.deaths` and `GDP` , so we took a log-transformation for these two predictors by setting `log.under.five.deaths = log(under.five.deaths + 1)` and `log.GDP = log(GDP)`.

3.1.2 Final Model

Based on the adjusted R^2 and AIC of predictions for test data, the best model is

$$\begin{aligned}\texttt{Life.expectancy} =& \beta_0 + \beta_1\texttt{I(Status=Developing)} + \beta_2\texttt{Alcohol} + \beta_3\texttt{Schooling}\\ &+ \beta_4\texttt{log.under.five.deaths} + \beta_5\texttt{log.GDP},\end{aligned}$$

which gave an adjusted R^2 of 0.55 and AIC of 104.68 for test data. And then we used the whole cleaned dataset (the combination of training data and test data) to refit the model. The final model is

$$\begin{aligned}\texttt{Life.expectancy} =& 46.59 - 2.64 \times \texttt{I(Status=Developing)} - 0.47 \times \texttt{Alcohol} + 2.03 \times \texttt{Schooling}\\ &- 2.18 \times \texttt{log.under.five.deaths} + 1.34 \times \texttt{log.GDP}.\end{aligned}$$

According to the fitting results of the linear model , controlling the effects of `Status` and `GDP`, the linear associations between life expectancy and alcohol consumption, schooling years, number of under-five deaths are significant at the level of 0.05. If keep the other variables not changing, 1 litres increase of alcohol consumption per capita (15+) would lead to 0.47-year ((95% CI: [0.19, 0.76])) decrease of life expectancy; 1-year increase of schooling would leads to a 2.03-year (95% CI: [1.56, 2.51]) increase of life expectancy; if the number of under five deaths increases to 10-fold, there would be a 2.18-year (95% CI: [0.73, 3.64]) decrease of life expectancy. The adjusted R^2 of this model is 0.67, which implies a not bad fitting of the linear model.

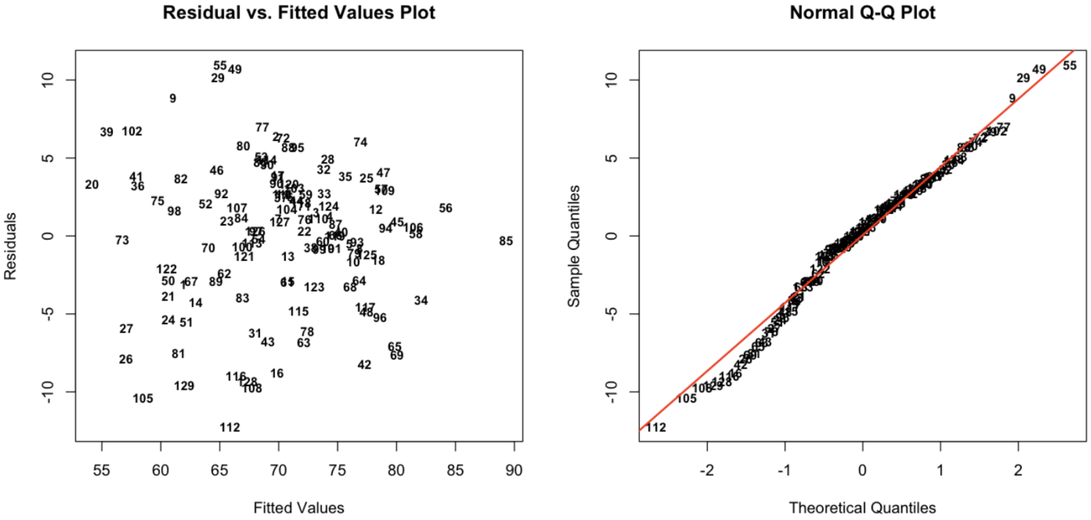

Figure 3: Residual Plots for Refitted Best Linear Model.

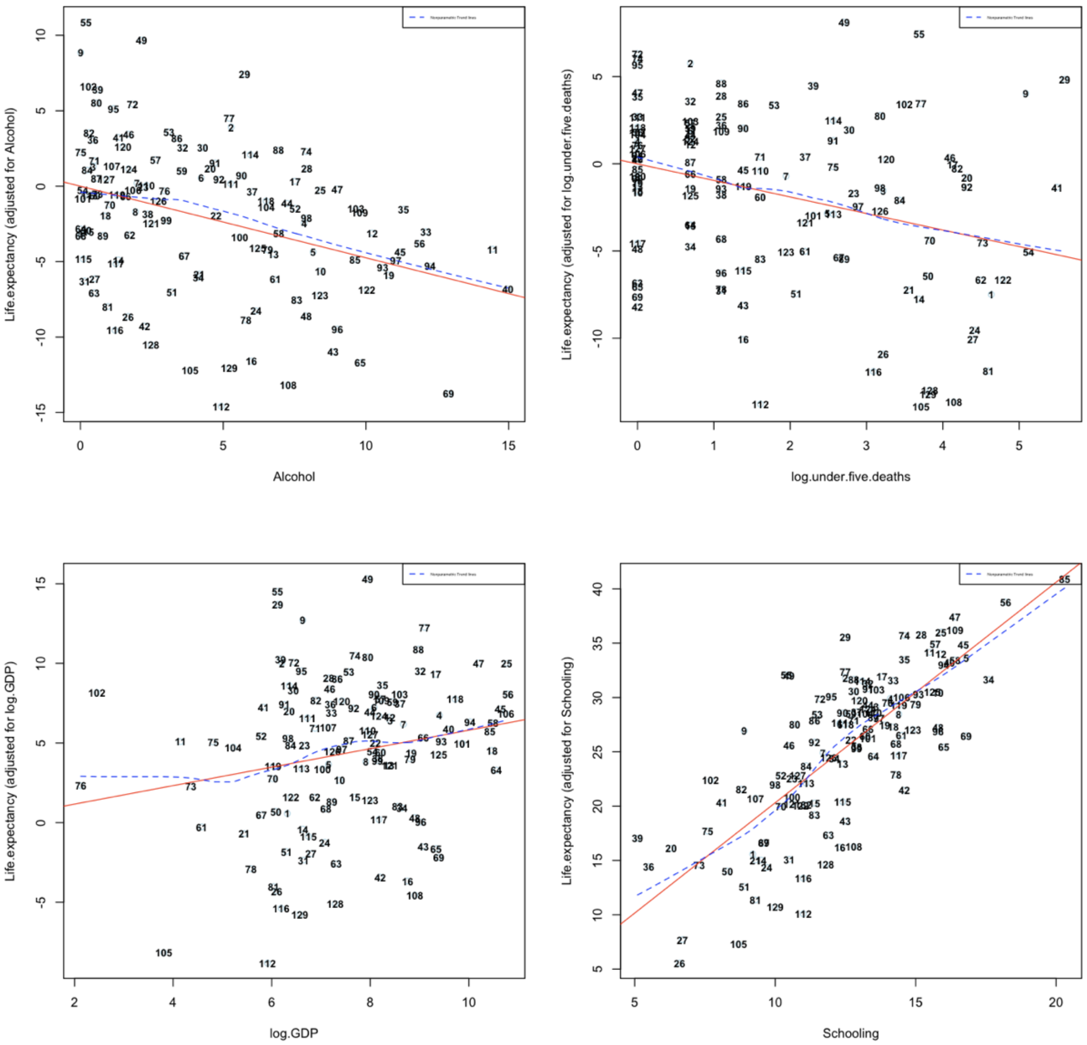

Figure 4: Residual Plots for Refitted Best Linear Model.

From the model diagnostics (Figure 4), all predictors have an approximately linear relationship to the response, but as Figure 3 shows, it can be seen that the model has obvious heteroskedasticity problem (the p-value of studentized Breusch-Pagan test is 0.02) and the normal assumption of residuals is violated.

To control the heteroskedasticity and follow the normal distribution consumption, a weighted least square model with square-transformation on the response `Life.expectancy` was created by calculating the weights from the linear model `|residual|=`β_0 `+` $\beta_1\times$`fitted value of linear model` and finding reasonable transformation (any λ lay inside the 95% confidence interval) by box-cox plot (Figure 5).

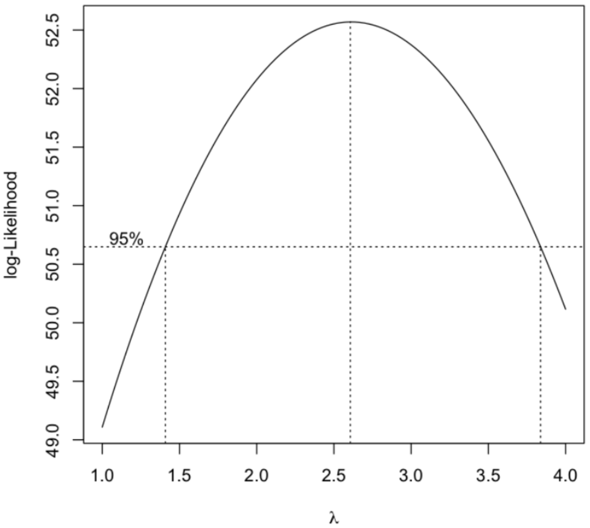

Figure 5: Box-cox Plot for Linear Model.

The WLS with square-transformation on response was fitted as

$$\begin{aligned}\texttt{Life.expectancy}^2 =& 1638.51 - 464.05 \times \texttt{I(Status=Developing)} - 57.84 \times \texttt{Alcohol} \\ &+ 282.17 \times \texttt{Schooling} - 228.82 \times \texttt{log.under.five.deaths} + 204.28 \times \texttt{log.GDP}.\end{aligned}$$

This model perfectly solved the heteroskedasticity. However, the right-skewness of residuals was still not be removed after applying square-transformation on the response. Since the square-transformation of response made it harder to interpret the results, we chose to not use transformation on the response and only use the WLS.

After fitting the WLS model, we got the final model as

$$\begin{aligned}\texttt{Life.expectancy} =& 47.11 - 2.59 \times \texttt{I(Status=Developing)} - 0.39 \times \texttt{Alcohol} + 1.92 \times \texttt{Schooling} \\ &- 1.95 \times \texttt{log.under.five.deaths} + 1.43 \times \texttt{log.GDP}.\end{aligned}$$

The WLS model improves both the R^2 (from 0.68 to 0.75) and adjusted R^2 (from 0.67 to 0.74) a lot, and under the WLS, all predictors (`Status`, `Alcohol`, `Schooling`, `log.under.five.death` and `log.GDP`) become significant at the level of 0.05.

Besides, from the fitting results of WLS, we found that if keep the other variables not changing, there would be 2.59-year (95% CI: [0.28, 4.90]) decrease of life expectancy for developing countries compared with developed countries; 1 litres increase of alcohol consumption per capita (15+) would lead to 0.39-year ((95% CI: [0.14, 0.64])) decrease of life expectancy; 1-year increase of schooling would leads to a 1.92-year (95% CI: [1.33, 2.50]) increase of life expectancy; if the number of under five deaths increases to 10-fold, there would be a 1.95-year (95% CI: [0.48, 3.42]) decrease of life expectancy; if the GDP increases to 10-fold, there would be a 1.43-year (95% CI: [0.10, 2.96]) decrease of life expectancy.

Compare to the fitting results between linear model and WLS, although the estimation on coefficients did not have large difference, WLS gave narrower 95% confidence intervals for all predictors and higher adjusted R^2. We chose the WLS as our final model.

3.2 Data Analysis A.2

According to WLS, all predictors including `Status`, `Alcohol`, `Schooling`, `log.under.five.death` and `log.GDP` are significant at the level of 0.05. To find whether some of these significant health and economical factors (except for Status) have different effects between developed and developing countries, we need to compare the following two linear models:

$$\begin{aligned}
\texttt{Life.expectancy} =& \beta_0 + \beta_1 \texttt{I(Status=Developing)} + \beta_2 \texttt{Alcohol} + \beta_3 \texttt{Schooling} \\
&+ \beta_4 \texttt{log.under.five.deaths} + \beta_5 \texttt{log.GDP}, \\
\texttt{Life.expectancy} =& \beta_0 + \beta_1 \texttt{I(Status=Developing)} + \beta_2 \texttt{Alcohol} + \beta_3 \texttt{Schooling} \\
&+ \beta_4 \texttt{log.under.five.deaths} + \beta_5 \texttt{log.GDP} \\
&+ \beta_6 \texttt{I(Status=Developing)} \times \texttt{Alcohol} \\
&+ \beta_7 \texttt{I(Status=Developing)} \times \texttt{Schooling} \\
&+ \beta_8 \texttt{I(Status=Developing)} \times \texttt{log.under.five.deaths} \\
&+ \beta_9 \texttt{I(Status=Developing)} \times \texttt{log.GDP}.
\end{aligned}$$

After adding these interacting terms, the adjusted R^2 almost did not change (from 0.745 to 0.740), but the predictors `Status`, `log.under.five.death` and `log.GDP` became not significant at the level of 0.05. Moreover, based on the anova results, the p-value is 0.8127, which indicates that these interacting terms are not significant at all. Taking the significance of predictors and anova results into consideration together, it can be concluded that there is no evidence that the effects of these significant factors are different among developing and developed countries.

4 Discussion

From this work, the weighted least square model performs better than the linear model in terms of homoskedasticity and adjusted R^2, which can explain approximately 74% of variation of the life expectancy. We applied the WLS to fit the whole cleaned dataset. The results from the WLS indicates that it can be conclude that developed countries with less alcohol consumption, longer schooling years, lower under-five deaths and higher GDP tend to larger life expectancy.

Keeping the other variables not changing, we found from the fitting results from the WLS that there would be 2.59-year (95% CI: [0.28, 4.90]) decrease of life expectancy for developing countries compared with developed countries; 1 litres increase of alcohol consumption per capita (15+) would lead to 0.39-year ((95% CI: [0.14, 0.64])) decrease of life expectancy; 1-year increase of schooling would leads to a 1.92-year (95% CI: [1.33, 2.50]) increase of life expectancy; if the number of under five deaths increases to 10-fold, there would be a 1.95-year (95% CI: [0.48, 3.42]) decrease of life expectancy; if the GDP increases to 10-fold, there would be a 1.43-year (95% CI: [0.10, 2.96]) decrease of life expectancy. Furthermore, the effects of these significant factors (except status) are the same among developing and developed countries.

There are several limitations in this work. Firstly, in this work, we removed the missing data directly when analyzing this dataset, which made our analysis probably biased. Secondly, the residuals of our fitted WLS model was right-skewed, violating the normal distribution assumption. After trying the square-transformation (suggested by box-cox plot) of the response, the normality was still violated. In the future, we may consider the more transformation on the response and more various transformations on the predictors to obtain better estimates.

Appendix

All codes with comments can be found at STATS513_Final_Exam_Code.ipynb (needs to be downloaded to view) or STATS513_Final_Exam_Code.pdf (supports preview).