Report for STATS 513 Final (2022 Winter)

Yifei Sun

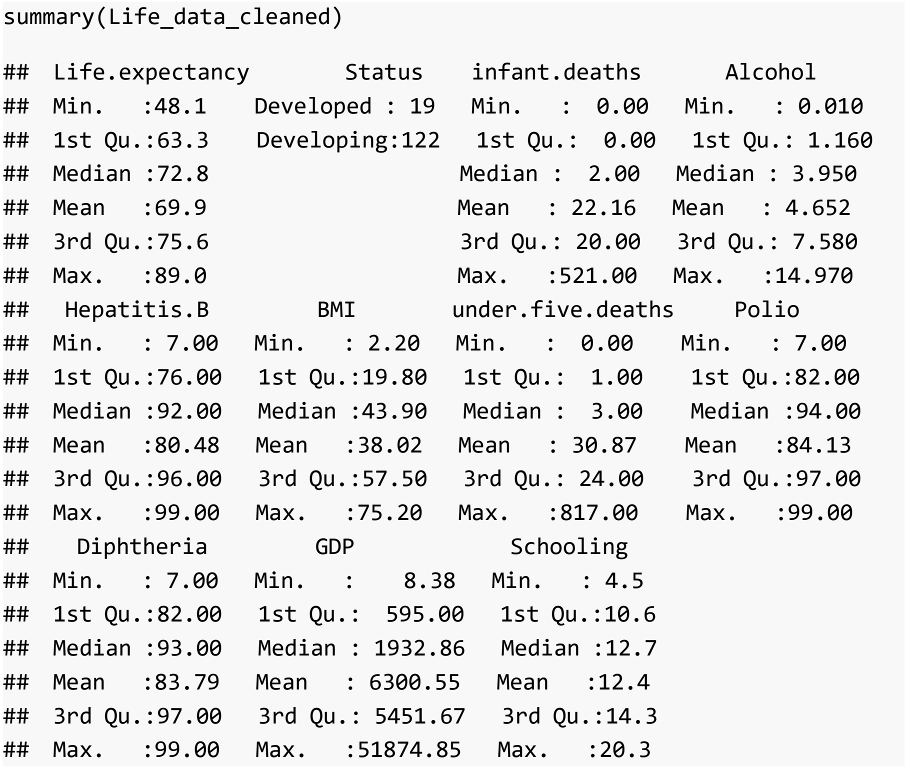
April 27, 2022

1 Lay Abstract

This work is conducted based on data from the Global Health Observatory. There are 183 observation (183 countries) and 11 variables, including Life.expectancy, status (developed or developing) and child or infant death rates, immunization rate, and other health and economic factors. From the result we know that holding all other factors constant, developing countries reduce the cubic of life expectancy by an average of 47520 relative to developed countries, a unit more alcohol consumption reduces the cubic of life expectancy by an average of 3405, a unit more BMI increases the cubic of life expectancy by an average of 756, a unit more GDP increases the cubic of life expectancy by an average of 1.169, a unit more Schooling increase the cubic of life expectancy by an average of 27380. Also, I find that the effects of status (developed, developing) are not different on BMI, but are more different on Schooling.

2 Introduction and Data Summary

The data we have is from the Global Health Observatory. There are 183 observation (183 countries) and 11 variables: Life.expectancy (years), Status (categorical variables: Developed or Developing), infant.deaths (Number of infant deaths per 1000 population), Alcohol (recorded per capita (15+) consumption in litres of pure alcohol), Hepatitis.B (Hepatitis B (HepB) immunization coverage among 1-year-olds (%)), BMI (Average Body Mass Index of entire population), under.five.deaths (Number of under-five deaths per 1000 population), Polio (Polio (Pol3) immunization coverage among 1-year-olds (%)), Diphtheria (Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)), GDP (Gross Domestic Product per capita (in USD)), Schooling (Number of years of Schooling (in years)). I will set Life.expectancy as response variable and the rest as predictors. First, I find there are rows with NA data. After eliminating these rows, there are 142 rows left. Also, I find there is one row with infant.deaths and under.five.deaths greater than 1000, which could be data entry errors, and I eliminate this row (141 rows left). Second, I create summary statistics for each variable.



```
summary(Life_data_cleaned)

##  Life.expectancy       Status    infant.deaths      Alcohol
##  Min.   :48.1    Developed : 19   Min.   :  0.00   Min.   : 0.010
##  1st Qu.:63.3    Developing:122   1st Qu.:  0.00   1st Qu.: 1.160
##  Median :72.8                     Median :  2.00   Median : 3.950
##  Mean   :69.9                     Mean   : 22.16   Mean   : 4.652
##  3rd Qu.:75.6                     3rd Qu.: 20.00   3rd Qu.: 7.580
##  Max.   :89.0                     Max.   :521.00   Max.   :14.970
##   Hepatitis.B        BMI       under.five.deaths     Polio
##  Min.   : 7.00   Min.   : 2.20   Min.   :  0.00    Min.   : 7.00
##  1st Qu.:76.00   1st Qu.:19.80   1st Qu.:  1.00    1st Qu.:82.00
##  Median :92.00   Median :43.90   Median :  3.00    Median :94.00
##  Mean   :80.48   Mean   :38.02   Mean   : 30.87    Mean   :84.13
##  3rd Qu.:96.00   3rd Qu.:57.50   3rd Qu.: 24.00    3rd Qu.:97.00
##  Max.   :99.00   Max.   :75.20   Max.   :817.00    Max.   :99.00
##    Diphtheria         GDP            Schooling
##  Min.   : 7.00   Min.   :    8.38   Min.   : 4.5
##  1st Qu.:82.00   1st Qu.:  595.00   1st Qu.:10.6
##  Median :93.00   Median : 1932.86   Median :12.7
##  Mean   :83.79   Mean   : 6300.55   Mean   :12.4
##  3rd Qu.:97.00   3rd Qu.: 5451.67   3rd Qu.:14.3
##  Max.   :99.00   Max.   :51874.85   Max.   :20.3
```

Figure 1 Summary Statistics of Each Variable

After separating the cleaned dataset into training dataset and testing dataset using the codes provided, we have 115 observations in the training dataset and 26 observations in the testing dataset. Third, we check the collinearity of the training dataset. I find that the conditional number is 6349.850, which is very large. The VIFs for under.five.deaths and infant.deaths are 67.299 and 66.049, which are far larger than 30. After checking the correlation matrix, I find there are some predictors that are highly correlated. under.five.deaths and infant.deaths has a correlation of 0.99. Diphtheria and Polio have a high correlation value of 0.74. Diphtheria and Hepatitis B have a high correlation value of 0.68. Schooling and Life.expectancy has a correlation value of 0.79. This is partly explained by the fact that many of these predictors are of the same categories, such as, infant or child death rate, immunization rate. We may delete some of those highly correlated predictors to avoid imprecise estimation of parameters and unstable results.

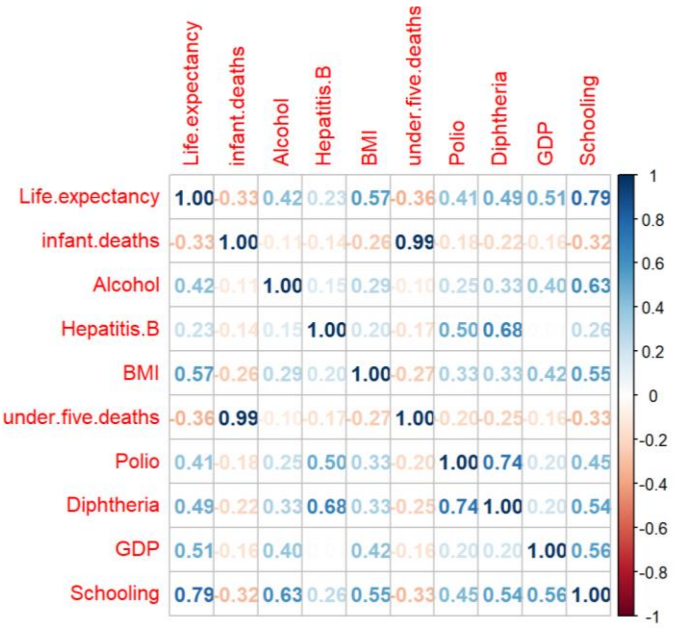


Figure 2 Correlation Matrix

3 Data Analysis

3.1 Data Analysis A.1

First, try to identify unusual points. Observations with indices 99 and 97 are regarded as leverage points, and observation 99 is regarded as influential points.

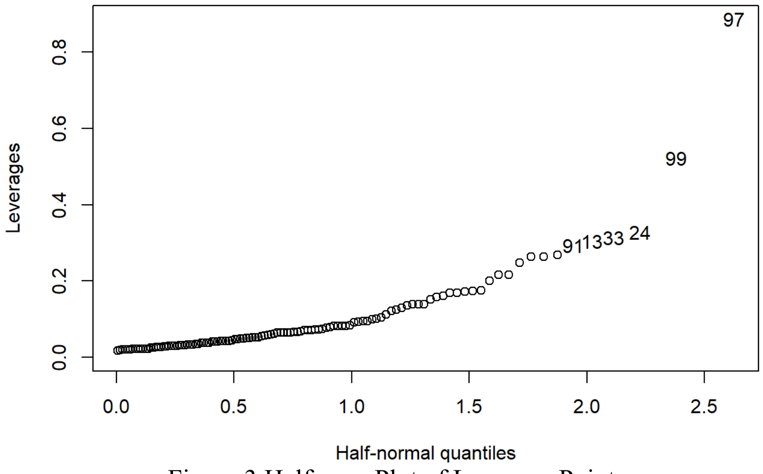


Figure 3 Halfnorm Plot of Leverage Points

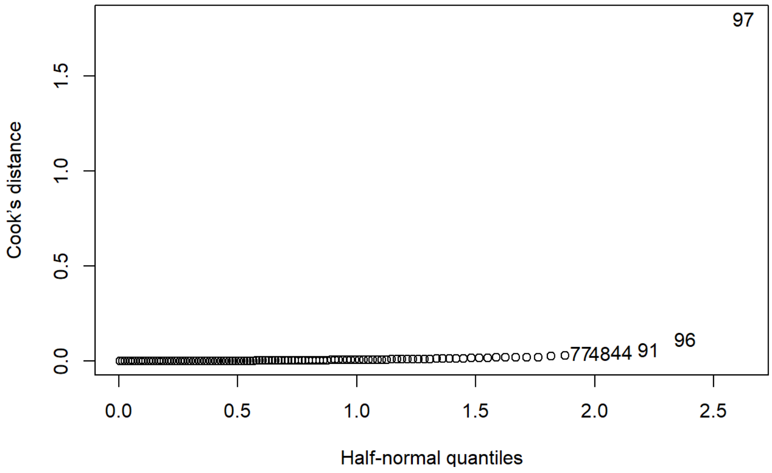


Figure 4 Halfnorm Plot of Influential Points

After calculating studentized residuals and compare them with thresholds decided by Bonferroni correction, I find there is no outliers. So, I decide to eliminate observation 99 and 97. (139 rows left)

First try linear model with all predictors. The parameter of infant.deaths is 0.2842, which is not reasonable as it indicates more infant.deaths lead to longer life expectancy. This is probably the result of collinearity. Using diagnostics tools, I think there is heteroscedasticity in the residuals.

Also, the residuals seem to follow long-tailed distribution. Will try robust methods and transformation.

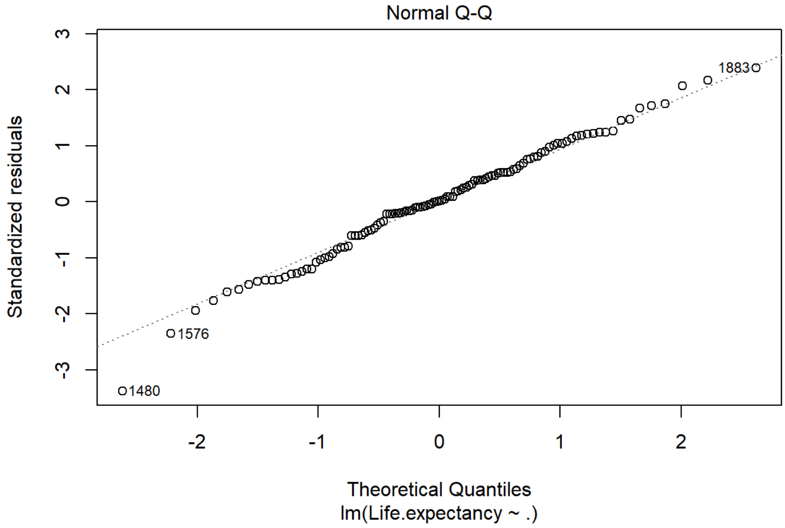


Figure 5 Norm QQ Plot of Standardized Residuals

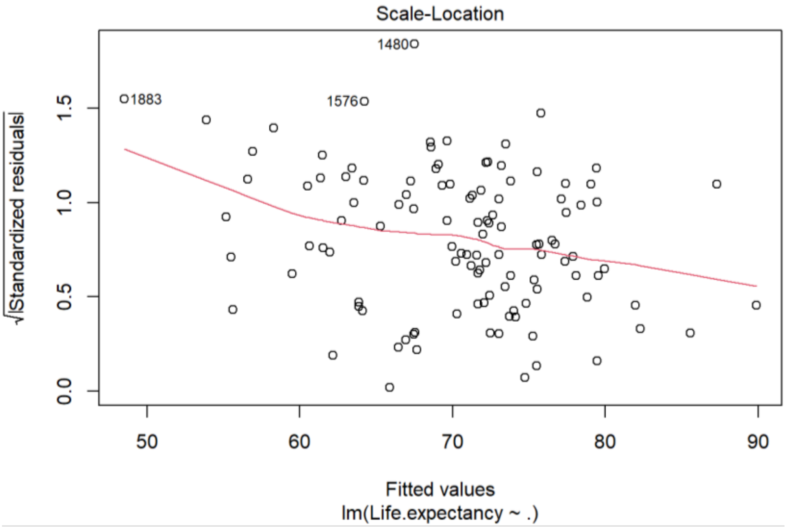


Figure 6 Fitted Values vs. $\sqrt{Standardized\ Residuals}$

First drop under.five.deaths because it is highly correlated with infant.deaths. Because there are many predictors, I will only try to transform response but not predictors. I find $\lambda = 3$ roughly maximize the log-likelihood function. After the Box Cox transformation, heteroscedasticity disappears, and the residuals seem to follow short-tailed distribution.

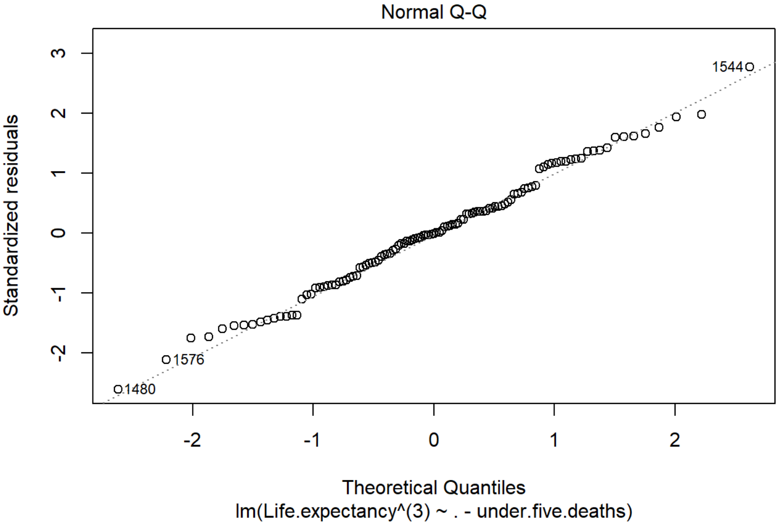


Figure 7 Norm QQ Plot of Standardized Residuals

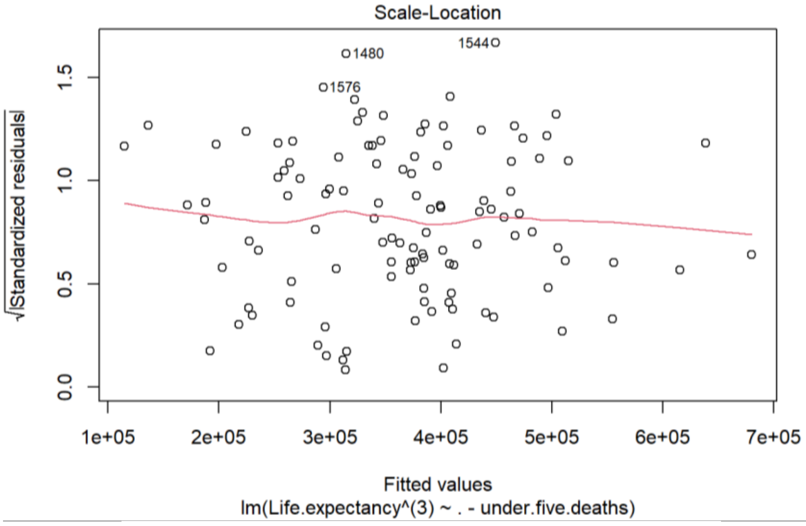


Figure 8 Fitted Values vs. $\sqrt{Standardized\ Residuals}$

Next will try criterion-based variable selection with under.five.deaths dropped. I choose five predictors: StatusDeveloping, Alcohol, BMI, GDP, Schooling, because they have the smallest AIC and Marlow's Cp. By computing RMSE, the model with these five predictors also has a pretty small training RMSE of 4.889 and testing RMSE of 6.485, compared with model selected using BIC and adjusted R2. I also tried robust methods such as Huber, LAD, and find that these methods give similar RMSE with OLS, so we will keep the simpler OLS model. I also tried GAM, it gives smaller RMSE than the previous model, but it is not conducive to interpretation so I abandoned it. So, the final model I choose is

$$\text{Life. expectancy} = (44780 - 47520 \times \text{StatusDeveloping} - 3405 \times \text{Alcohol} + 756.5 \times \text{BMI} + 1.169 \times \text{GDP} + 27380 \times \text{Schooling})^{1/3}$$

After transformation, the model is harder to interpret. From the result, it is concluded that The cubic of life expectancy is linearly related to BMI, Schooling, Alcohol, GDP, and Status. The coefficient for StatusDeveloping is -47520 (95% confidence interval is (-97099, 2057) not significant), demonstrating that holding all other factors constant, developing countries reduce the cubic of life expectancy by an average of 47520 relative to developed countries. The coefficient for Alcohol is -3405 (95% confidence interval is (-8106, 1296), not significant), demonstrating that holding all other factors constant, a unit more alcohol consumption reduces the cubic of life expectancy by an average of 3405. The coefficient for BMI is 756 (95% confidence interval is (163, 1496), significant), demonstrating that holding all other factors constant, a unit more BMI increases the cubic of life expectancy by an average of 756. The coefficient for GDP is 1.169 (95% confidence interval is (0.294, 2.633), not significant), demonstrating that holding all other factors constant, a unit more GDP increases the cubic of life expectancy by an average of 1.169. The coefficient for Schooling is 27380 (95% confidence interval is (20558, 34205), very significant), demonstrating that holding all other factors constant, a unit more Schooling increase the cubic of life expectancy by an average of 27380. We can see the performance of our final model is pretty good. The training data and testing data are closed to our predicted values. The adjusted R2 is close to 0.6791 which shows our model has pretty good explanatory ability. And it has has a pretty small training RMSE of 4.889 and testing RMSE of 6.485. Judging from the F test of all predictors, we find that they are significant when combined together. We found that schooling is very significant here. And BMI is also significant. The other predictors are not significant, but they help in the prediction.

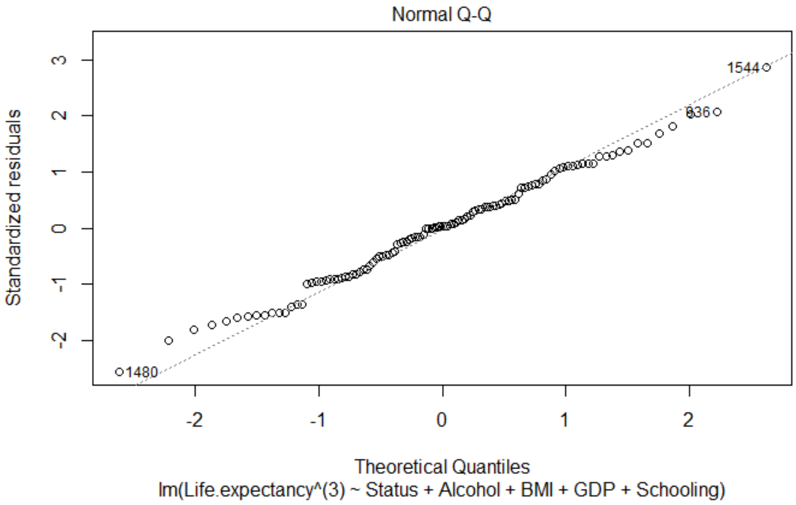


Figure 9 Norm QQ Plot of Standardized Residuals

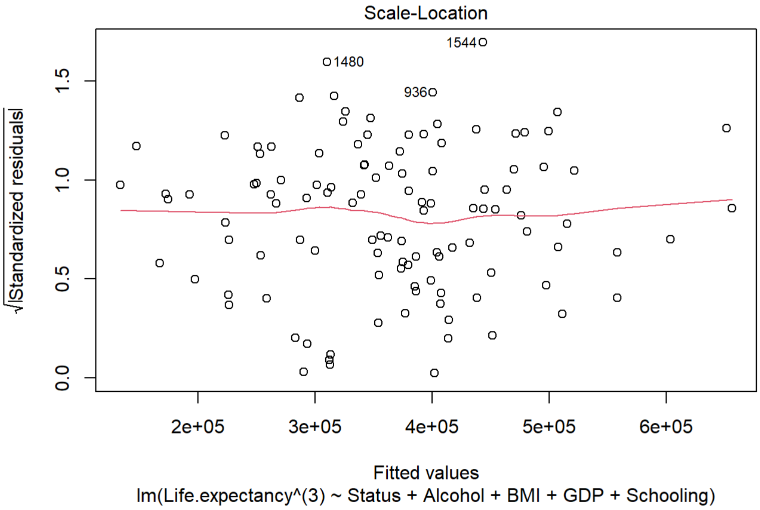


Figure 10 Fitted Values vs. $\sqrt{Standardized\ Residuals}$

3.2 Data Analysis A.2

To explore whether some of these significant health and economic factors (which I found is Schooling and BMI) have different effects between developed and developing countries, we change the model to the following two model, do two linear regressions respectively and compare the result with original model. (Add the interaction term for schooling and BMI respectively)

$$\text{Life. expectancy} = (\beta_0 + \beta_1 \times \text{I(Status} = \text{Developing)} + \beta_2 \times \text{Alcohol} + \beta_3 \times \text{BMI} + \beta_4 \times \text{GDP} + \beta_5 \times \text{Schooling} + \beta_8 \times \text{I(Status} = \text{Developing)} \times \text{BMI})^{1/3}$$

$$\begin{aligned}\text{Life.expectancy} = (&\beta_0 + \beta_1 \times \text{I(Status} \\ &= \text{Developing)} + \beta_2 \times \text{Alcohol} + \beta_3 \times \text{BMI} + \beta_4 \times \text{GDP} + \beta_5 \times \text{Schooling} \\ &+ \beta_8 \times \text{I(Status = Developing)} \times \text{Schooling})^{1/3}\end{aligned}$$

After calculation we find the result models are

$$\begin{aligned}\text{Life.expectancy} = (&7915 - 11070 \times \text{I(Status} \\ &= \text{Developing)} - 3266 \times \text{Alcohol} + 1358 \times \text{BMI} + 1.098 \times \text{GDP} \\ &+ 27660 \times \text{Schooling} - 688.3 \times \text{I(Status = Developing)} \times \text{BMI})^{1/3}\end{aligned}$$

After using ANOVA (F-Test), we get the P value is 0.5125, which is insignificant. It shows this model and the original model without interaction term do not have significantly different RSS. There is no evidence that the effects of BMI are difference among developing and developed countries.

$$\begin{aligned}\text{Life.expectancy} = (&349300 - 364500 \times \text{I(Status} \\ &= \text{Developing)} - 4172 \times \text{Alcohol} + 624.9 \times \text{BMI} + 1.806 \times \text{GDP} \\ &+ 8537 \times \text{Schooling} + 20300 \times \text{I(Status = Developing)} \times \text{Schooling})^{1/3}\end{aligned}$$

After using ANOVA (F-Test), we get the P value is 0.06578, which is insignificant at the level of 95%. But it is close to 0.05, indicating it is a relatively small value. It shows this model and the original model without schooling/status interaction term have more different RSS than the previous BMI/status pair. There is evidence that the effects of schooling are comparatively more difference among developing and developed countries.

4 Discussion

After the study, the final model I select is

$$\begin{aligned}\text{Life.expectancy} = (&44780 - 47520 \times \text{StatusDeveloping} - 3405 \times \text{Alcohol} + 756.5 \times \text{BMI} \\ &+ 1.169 \times \text{GDP} + 27380 \times \text{Schooling})^{1/3}\end{aligned}$$

This indicates that holding all other factors constant, developing countries reduce the cubic of life expectancy by an average of 47520 relative to developed countries, a unit more alcohol consumption reduces the cubic of life expectancy by an average of 3405, a unit more BMI increases the cubic of life expectancy by an average of 756, a unit more GDP increases the cubic of life expectancy by an average of 1.169, a unit more Schooling increase the cubic of life expectancy by an average of 27380. Also, I find that the effects of status (developed, developing) are not different on BMI, but are more different on Schooling.

There are several limitations in this work. Firstly, in this work, I exclude the missing data and possible data entry errors when we analyze this dataset. This may create potential bias, as there may be some underlying patterns that I ignored since I simply deleted them. Besides, the errors could potentially not follow normal distribution and homoscedasticity, but I assumed they follow. Also, I assumed the underlying relationship is linear, and used linear regression, but the true relation is unknown. May try to use different models in the future. Because there are too many predictors, I did not do transformation on predictors. In the future, we could use more powerful computer and use different combinations and predictor transformation methods. And I only added interaction terms for BMI and Schooling, but other predictors may also have interaction terms, which can be explored in the future.

A Appendix: R codes

```
library(faraway)
library(corrplot)
library(GGally)
library(leaps)
library(MASS)

# 1 Introduction and Data Summary
load("FinalExam.RData")
head(Life_data)
dim(Life_data)
# check the missingness or entry errors of the data
# status is categorical, change it to categorical
Life_data$Status=factor(Life_data$Status)
# after using summary, there are many NA in data, remove these NA rows
Life_data_cleaned=Life_data[complete.cases(Life_data), ]
# after removing all NAs, the data dimensions we have is
dim(Life_data_cleaned)
# now all rows with one or more NAs are removed
# because we know for infant.deaths and under.five.deaths, the number
can not exceed 1000, and we found there are some data exceeding 1000,
which are clearly errors and need to be removed
Life_data_cleaned=subset(Life_data_cleaned,
infant.deaths<=1000|under.five.deaths<=1000)
# after removing data entry errors, the data dimensions we have is
dim(Life_data_cleaned)
# use histograms to see the distribution of data
units=c("(years)"," Developed/Developing","(Number per 1000)","
(litres)","(%)"," ","(Number per 1000)","(%)","(%)","(USD)","(years)")
xlimit=array(c(c(0,100),c(0,100),c(0,1000),c(0,20),c(0,100),c(0,100),c(0,1
000),c(0,100),c(0,100),c(0,60000),c(0,25)),dim=c(2,11,1))
summary(Life_data_cleaned)
train_data<-Life_data_cleaned[1:115,]
dim(train_data)
test_data<-Life_data_cleaned[-c(1:115),]
dim(test_data)
p_tr=dim(train_data)[2]-1
p_tr
n_tr=dim(train_data)[1]
n_tr
# do data summary for training dataset and etsting dataset separately
summary(train_data)
summary(test_data)
result <- lm(Life.expectancy ~ ., data=train_data)
# check correlation matrix
plot=cor(train_data[-1,-2])
```

```
corrplot(plot,method = 'number')

# check condition number
X=model.matrix(result)[,c(-1,-2)]
e <- eigen(t(X) %*% X)
e$val
round(sqrt(e$val[1]/e$val), 3)
# check variance inflation factor
round(sort(vif(X),decreasing=TRUE), 3)
par(mfcol=c(1,2))
# check for leverage points
halfnorm(lm.influence(result)$hat, nlab = 6,ylab="Leverages")
# check for influential points
cook <- cooks.distance(result)
halfnorm(cook, nlab = 6, ylab="Cook’s distance")
# check for outliers
## Compute studentized residuals
ti <- rstudent(result)
sorted_ti=sort(ti,decreasing=TRUE)
## Compute p-value
sorted_ti_p=2*(1-pt(abs(sorted_ti), df=n_tr-p_tr-1))
## compare to alpha/n
0.05/n_tr
sum(sorted_ti_p<(0.05/n_tr))
# remove 97th and 99th observations
dim(train_data)
train_data=train_data[c(-97,-99),]
dim(train_data)
# first try linear model
model_linear=lm(Life.expectancy ~., train_data)
summary(model_linear)
plot(model_linear)
summary(lm(sqrt(abs(model_linear$residuals))~model_linear$fitted.value))
g = lm(Life.expectancy ~.-under.five.deaths, train_data)
boxcox(g, plotit=T, lambda=seq(1, 5, by=0.1))
model_boxcox=lm(Life.expectancy^(3) ~ .-under.five.deaths, data=train_data)
summary(model_boxcox)
summary(lm(sqrt(abs(model_boxcox$residuals))~model_boxcox$fitted.value))
plot(model_boxcox)
b = regsubsets(Life.expectancy^(3) ~ .-under.five.deaths, data=train_data)
summary(b)
rs <- summary(b)
aic <- n_tr*log(rs$rss/n_tr) + (2:9)*2
which.min(aic)
which.min(rs$bic)
which.max(rs$adjr2)
which.min(rs$cp)
```

```
model_1=lm(Life.expectancy^(3) ~ Status+Alcohol+BMI+GDP+Schooling,
data=train_data)
summary(model_1)
plot(model_1)
gls <- lm(Life.expectancy ~ ., data=train_data)
summary(gls)
gls_2 <- lm(Life.expectancy ~ .-under.five.deaths, data=train_data)
summary(gls_2)
## Least absolute deviations
library(quantreg)
glad <- rq(Life.expectancy ~ .-under.five.deaths, data=train_data)
summary(glad)
## Huber’s method
ghuber <- rlm(Life.expectancy ~ .-under.five.deaths, data=train_data)
plot(ghuber)
summary(ghuber)
2*(1-pt(abs(coef(summary(ghuber)))[,"t value"],df=n_tr-p_tr-1-1))
# GAM
require(mgcv)
gamod=gam(Life.expectancy
~factor(Status)+s(infant.deaths)+s(Alcohol)+s(Hepatitis.B)+s(BMI)+s(under.five.d
eaths)+s(Polio)+s(Diphtheria)+s(GDP)+s(Schooling),data=train_data)
plot(gamod)
gamod
summary(gamod)
rmse <- function(x, y) { sqrt(mean( (x - y)^2 ))}
# boxcox transformation, lambda=3, Life.expectancy^(3) ~
Status+Alcohol+BMI+GDP+Schooling
rmse((model_1$fitted.values)^(1/3),train_data$Life.expectancy)
rmse((predict(model_1,newdata=test_data))^(1/3),test_data$Life.expectancy)
# linear model, no transformation, keep under.five.deaths
rmse((model_linear$fitted.values),train_data$Life.expectancy)
rmse((predict(model_linear,newdata=test_data)),test_data$Life.expectancy)
# linear model, no transformation, Life.expectancy ~.-under.five.deaths
rmse((g$fitted.values),train_data$Life.expectancy)
rmse((predict(g,newdata=test_data)),test_data$Life.expectancy)
# Life.expectancy^(3) ~ BMI+GDP+Schooling
rmse((model_2$fitted.values)^(1/3),train_data$Life.expectancy)
rmse((predict(model_2,newdata=test_data))^(1/3),test_data$Life.expectancy)
# Life.expectancy^(3) ~ .-infant.deaths-Polio-under.five.deaths
rmse((model_3$fitted.values)^(1/3),train_data$Life.expectancy)
rmse((predict(model_3,newdata=test_data))^(1/3),test_data$Life.expectancy)
# Huber Life.expectancy ~ .-under.five.deaths
rmse((ghuber$fitted.values),train_data$Life.expectancy)
rmse((predict(ghuber,newdata=test_data)),test_data$Life.expectancy)
# GAM Life.expectancy
~factor(Status)+s(infant.deaths)+s(Alcohol)+s(Hepatitis.B)+s(BMI)+s(under.five.d
eaths)+s(Polio)+s(Diphtheria)+s(GDP)+s(Schooling)
```

```
rmse((gamod$fitted.values),train_data$Life.expectancy)
rmse((predict(gamod,newdata=test_data)),test_data$Life.expectancy)
# Least absolute deviations
rmse((glad$fitted.values),train_data$Life.expectancy)
rmse((predict(glad,newdata=test_data)),test_data$Life.expectancy)
anova(model_1,model_2)
summary(model_1)
confint(model_1)
plot(model_1)
model_interact_BMI=lm(Life.expectancy^(3) ~
Status+Alcohol+BMI+GDP+Schooling+Status:BMI, data=train_data)
summary(model_interact_BMI)
anova(model_interact_BMI, model_1)
model_interact_Schooling=lm(Life.expectancy^(3) ~
Status+Alcohol+BMI+GDP+Schooling+Status:Schooling, data=train_data)
summary(model_interact_Schooling)
anova(model_interact_Schooling, model_1)
```