Final Project

In [112…

```
library(ggplot2)
library(ggpubr)
library(GGally)
library(dplyr)
library(tidyr)
library(moments)
library(car)
library(faraway)
library(faraway)
library(leaps)
library(pls)
library(MASS)
library(Metrics)
library(lars)
library(purrr)
library(leaps)
library(Metrics)
library(quantreg)
```

1 $Life.expectancy$: Life expectancy in years.\ 2 $Status$: Developing status for each country with 2-levels:\ – Developed\ – Developing\ 3 $infant.deaths$: Number of infant deaths per 1000 population; value should be less than or equal to 1000.\ 4 $Alcohol$: recorded per capita (15+) consumption (in litres of pure alcohol).\ 5 $Hepatitis.B$: Hepatitis B (HepB) immunization coverage among 1-year-olds (%).\ 6 BMI: Average Body Mass Index of entire population.\ 7 $under.five.deaths$: Number of under-five deaths per 1000 population; value should be less than or equal to 1000.\ 8 $Polio$: Polio (Pol3) immunization coverage among 1-year-olds (%).\ 9 $Diphtheria$: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).\ 10 GDP : Gross Domestic Product per capita (in USD).\ 11 $Schooling$: Number of years of Schooling (in years).

1 Lay Abstract (5 points)

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, Also, there maybe some inner relations between GDP and schooling since in developed countries people are more likey to receive education. Since this is a small dataset, each extreme point will somehow distract our prediction.

In [113…

```
# read data from RData file
life = get(load('FinalExam.RData'))

# change data type
life$Status = factor(life$Status)
```

In [114…

```
head(life)
dim(life)
```

A data.frame: 6 × 11

	Life.expectancy	Status	infant.deaths	Alcohol	Hepatitis.B	BMI	under.five.deaths
	<dbl>	<fct>	<int>	<dbl>	<int>	<dbl>	<int>
6	58.8	Developing	74	0.01	66	16.7	102
22	76.2	Developing	1	5.28	99	54.3	1
38	74.7	Developing	21	0.45	95	53.9	24
54	49.6	Developing	78	7.80	77	2.4	121
70	75.6	Developing	0	7.84	98	44.4	0
86	75.5	Developing	10	8.15	94	59.8	11

183 · 11

2 Introduction and Data Summary (10 points)

In this report, our scientific goal is to understand how these health and economical factors impact the life expectancy of countries. Firstly, check the missingness or entry errors of the data. Report notable issues. Report the missing data and potential data entry errors. If you find any missing data and potential data entry errors, please clean the dataset by deleting the countries with missing data and potential data entry in your analysis. Secondly, summarize the demographics in this sample by reporting summary statistics for each variable. Remember to code categorical data appropriately. Thirdly, comment the collinearity between predictors in the training dataset.

(1) Check the missingness or entry errors of the data

a. missingness

In [115…

```
# check missing values
life[!complete.cases(life),]
```

A data.frame: 41 × 11

	Life.expectancy	Status	infant.deaths	Alcohol	Hepatitis.B	BMI	under.five.deaths
	<dbl>	<fct>	<int>	<dbl>	<int>	<dbl>	<int>
166	75.0	Developing	0	9.19	98	61.3	0
310	68.7	Developing	9	3.95	91	49.3	12
438	51.5	Developing	60	3.15	85	25.0	84
614	62.0	Developing	7	3.53	74	24.5	10
695	77.5	Developed	0	12.69	99	63.6	0
711	69.0	Developing	8	3.12	93	3.3	10
727	57.4	Developing	239	1.81	6	19.1	321
743	79.2	Developed	0	10.28	NA	57.0	0
808	70.0	Developing	54	0.22	97	57.0	64
840	56.1	Developing	3	9.93	NA	22.1	4
920	79.9	Developing	0	9.72	NA	6.2	0
968	59.3	Developing	3	3.48	97	24.1	6
1128	36.3	Developing	23	5.76	NA	44.2	58
1160	74.5	Developed	0	10.78	NA	61.7	1
1176	81.8	Developed	0	8.25	NA	58.9	0

1224	74.1	Developing	22	0.03	99	53.6	26
1320	83.0	Developed	3	6.90	NA	26.9	4
1416	68.8	Developing	4	2.73	96	4.7	4
1432	63.6	Developing	10	5.95	74	18.0	13
1705	68.7	Developing	0	1.76	88	66.4	0
1835	88.0	Developed	1	9.33	NA	59.3	1
1916	81.0	Developed	0	6.59	NA	58.9	0
2093	87.0	Developing	2	9.23	94	29.5	2
2109	68.8	Developing	1	8.25	98	5.4	1
2174	74.2	Developing	0	10.87	97	43.8	0
2190	72.5	Developing	0	7.00	99	49.3	0
2335	75.1	Developed	0	10.13	99	55.9	0
2351	79.5	Developed	0	10.32	NA	57.0	0
2383	52.4	Developing	52	0.01	NA	22.0	83
2415	55.0	Developing	27	NA	NA	NA	41
2463	62.5	Developing	62	1.77	75	NA	92
2511	81.5	Developed	0	7.20	NA	57.3	0
2527	82.3	Developed	0	10.01	NA	55.4	0
2543	73.7	Developing	7	0.78	84	52.3	9
2591	74.7	Developing	0	1.47	9	57.4	0
2768	82.0	Developed	3	10.88	NA	63.6	4
2784	57.5	Developing	89	4.19	91	2.7	131
2800	78.7	Developed	25	8.55	92	66.9	30
2864	73.7	Developing	9	7.22	78	59.3	10
2880	75.2	Developing	29	3.93	88	14.0	35
2896	64.4	Developing	35	0.06	76	37.2	45

There are 41 rows in our dataframe that contains missing values. And we are supposed to remove these rows

In [116…

```
# remove rows containing na
life <- na.omit(life)
```

b.entry errors

In [117…

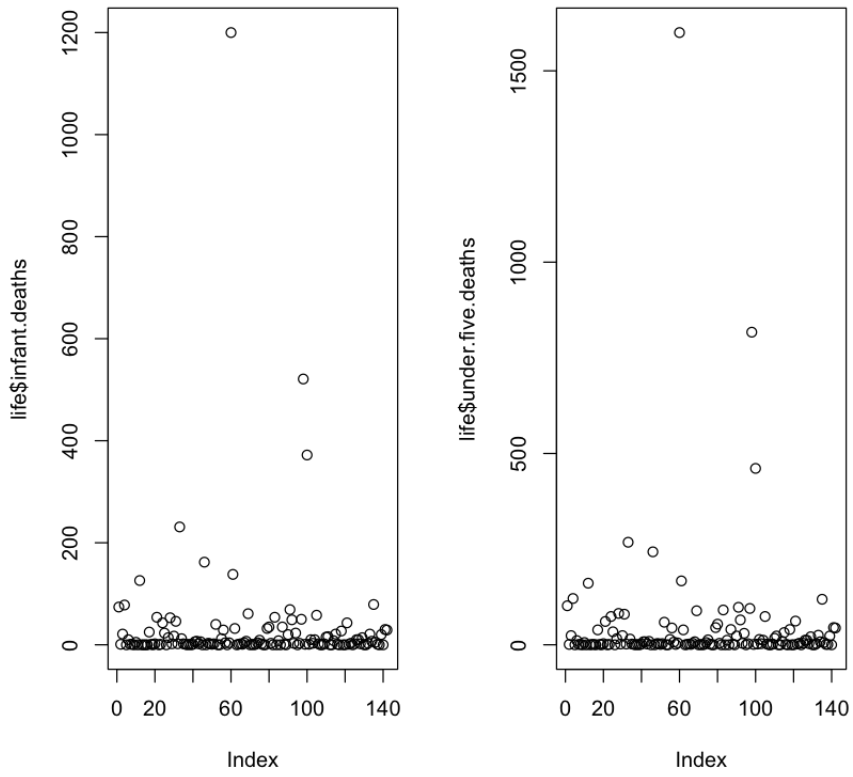
```
# summarize data
summary(life)
```

```
Life.expectancy        Status    infant.deaths        Alcohol
Min.   :48.10   Developed : 19   Min.   :   0.00   Min.   : 0.010
1st Qu.:63.55   Developing:123   1st Qu.:   0.00   1st Qu.: 1.173
Median :72.80                   Median :   2.50   Median : 3.895
Mean   :69.88                   Mean   :  30.45   Mean   : 4.639
3rd Qu.:75.60                   3rd Qu.:  20.75   3rd Qu.: 7.565
Max.   :89.00                   Max.   :1200.00   Max.   :14.970
 Hepatitis.B         BMI        under.five.deaths     Polio
Min.   : 7.00   Min.   : 2.20   Min.   :   0.00   Min.   : 7.00
1st Qu.:75.25   1st Qu.:19.12   1st Qu.:   1.00   1st Qu.:82.00
Median :92.00   Median :43.15   Median :   3.00   Median :93.50
Mean   :80.18   Mean   :37.87   Mean   :  41.92   Mean   :84.07
3rd Qu.:96.00   3rd Qu.:57.38   3rd Qu.:  24.00   3rd Qu.:97.00
Max.   :99.00   Max.   :75.20   Max.   :1600.00   Max.   :99.00
  Diphtheria         GDP              Schooling
Min.   : 7.00   Min.   :    8.38   Min.   : 4.50
1st Qu.:82.00   1st Qu.:  602.66   1st Qu.:10.53
Median :93.00   Median : 1930.80   Median :12.60
Mean   :83.75   Mean   : 6265.66   Mean   :12.39
3rd Qu.:97.00   3rd Qu.: 5441.72   3rd Qu.:14.30
Max.   :99.00   Max.   :51874.85   Max.   :20.30
```

According to the description, we know that $infant.deaths$ and $under.five.deaths$ should be equal or less than 1000. However the maximum of $i$nfant.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.

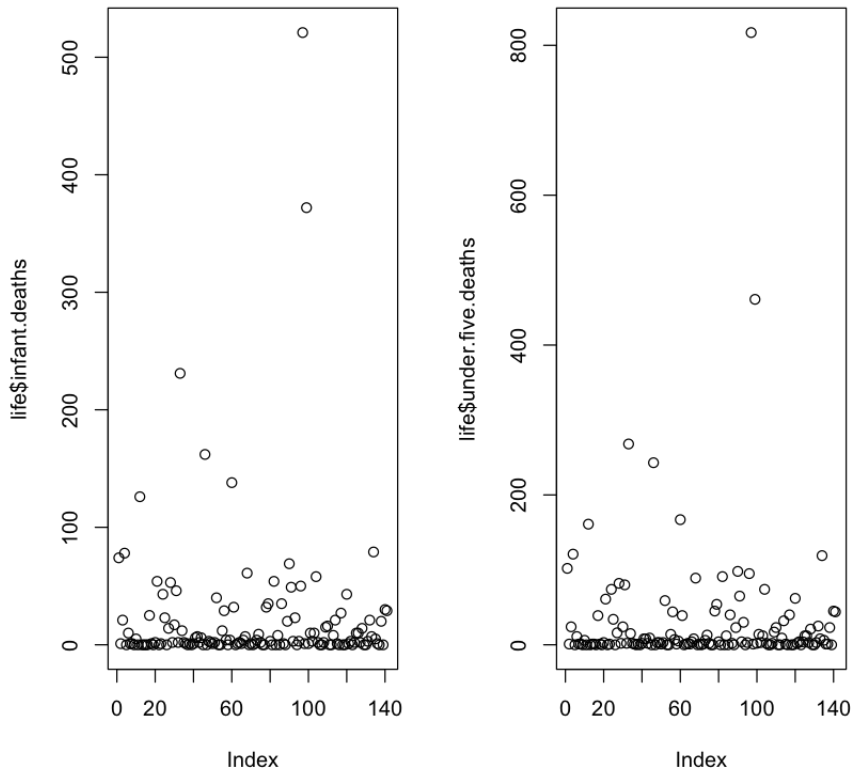
In [118…

```
par(mfrow=c(1,2))
plot(life$infant.deaths)
plot(life$under.five.deaths)
```



In [119…

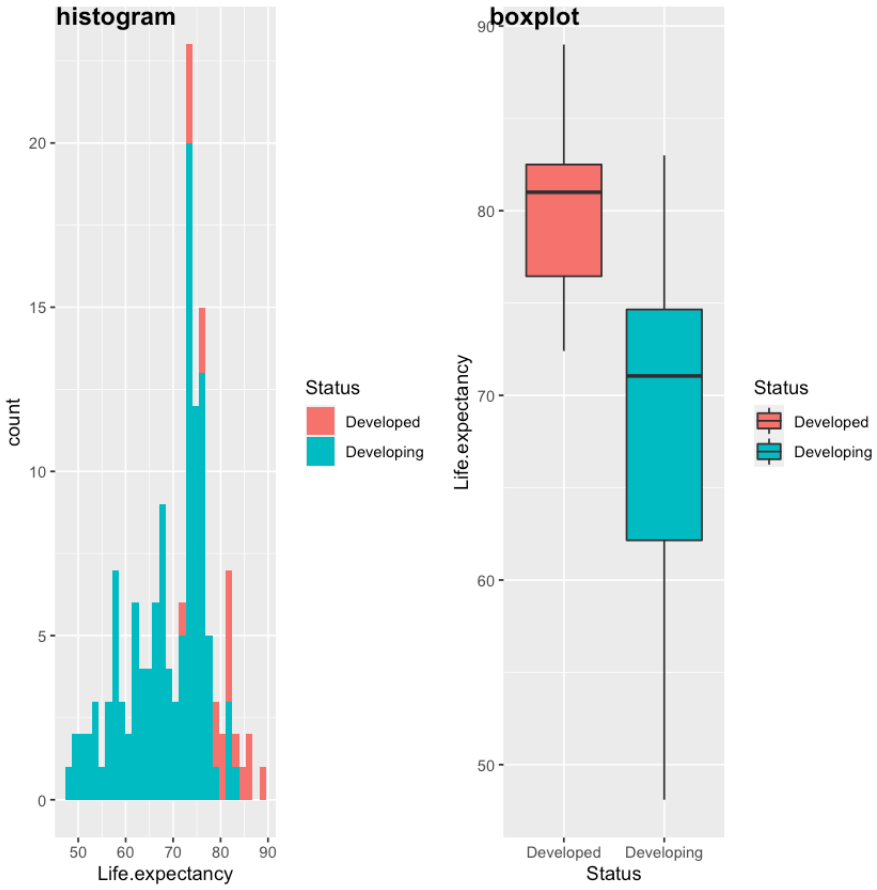
```
# filter the data
life = life %>% filter(infant.deaths <= 1000) %>% filter(under.five.deaths <=

par(mfrow=c(1,2))
plot(life$infant.deaths)
plot(life$under.five.deaths)
```



In [120…

```
# visulaize filtered data
hst = ggplot(data = life, aes(x=Life.expectancy,fill=Status))+geom_histogram(
box = ggplot(data=life,aes(x=Status,y=Life.expectancy,fill=Status))+geom_boxp
ggarrange(hst, box, labels = c("histogram", "boxplot"),ncol = 2, nrow = 1)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

(2) summarize the demographics

In [121…

```
summary(life)
```

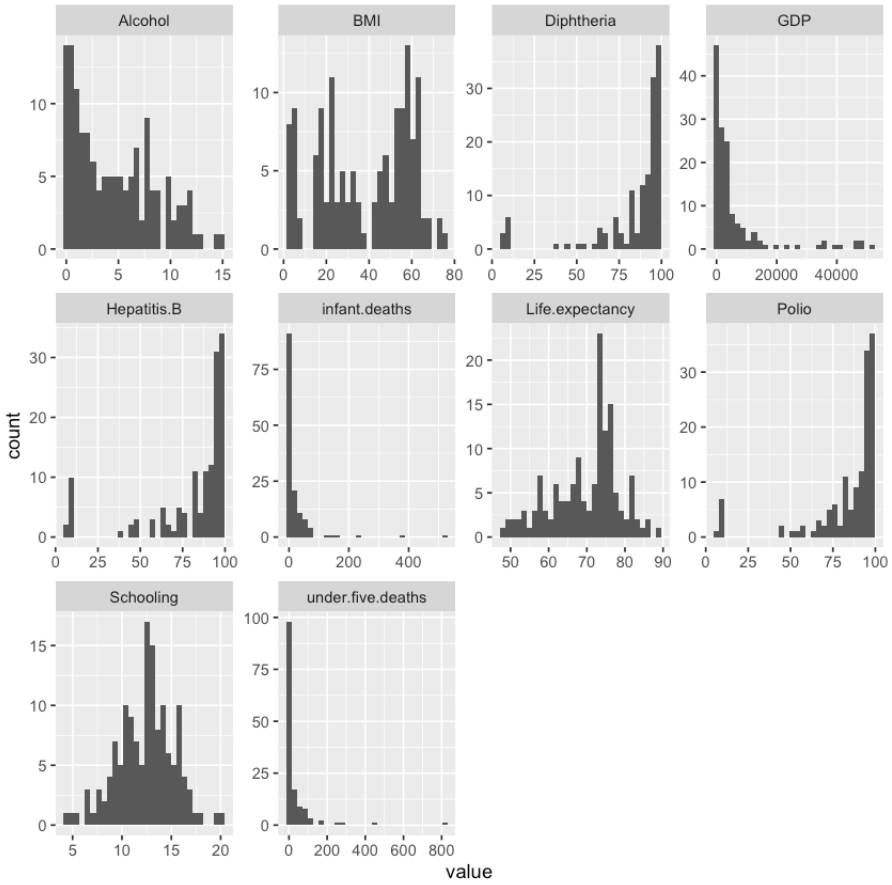
```
Life.expectancy        Status    infant.deaths       Alcohol
Min.   :48.1    Developed : 19   Min.   :  0.00   Min.   : 0.010
1st Qu.:63.3    Developing:122   1st Qu.:  0.00   1st Qu.: 1.160
Median :72.8                    Median :  2.00   Median : 3.950
Mean   :69.9                    Mean   : 22.16   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
```

a.distribution of the numerical variables

In [122…

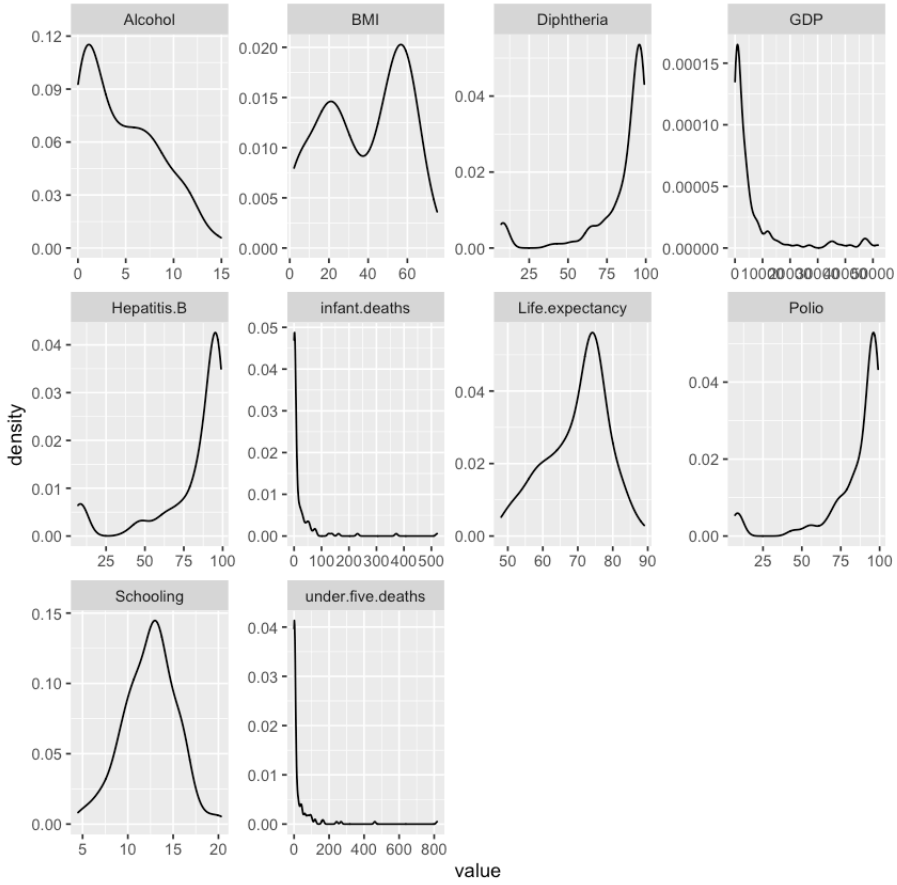
```
life %>% keep(is.numeric) %>% gather() %>% ggplot(aes(value)) + facet_wrap(~
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In [123…

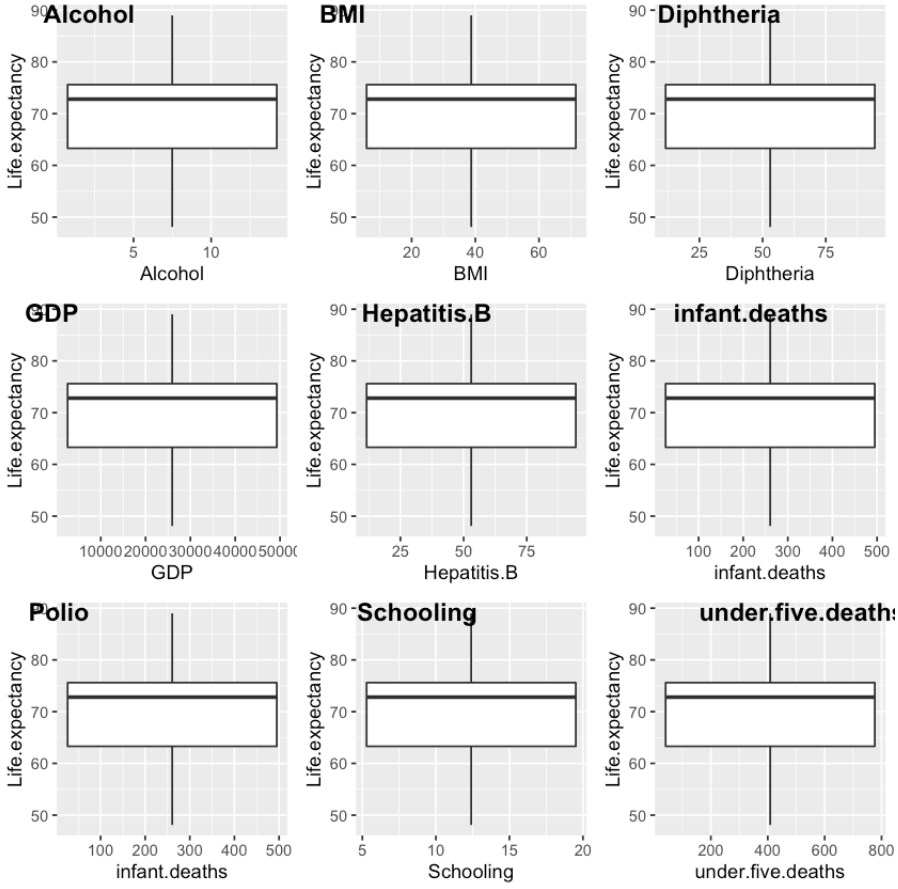
```
life %>% keep(is.numeric) %>% gather() %>% ggplot(aes(value)) + facet_wrap(~
```



In [124…

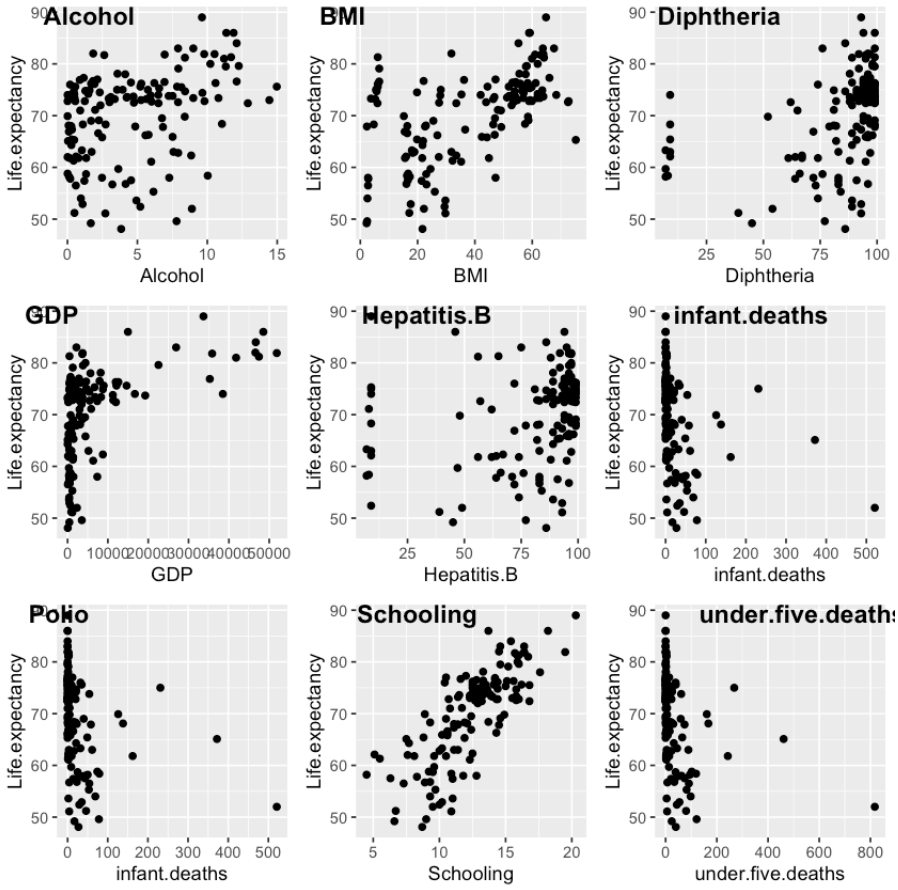
```
al_box = ggplot(data=life, aes(x = Alcohol, y=Life.expectancy)) + geom_boxplo
BM_box = ggplot(data=life, aes(x = BMI, y=Life.expectancy)) + geom_boxplot()
Di_box = ggplot(data=life, aes(x = Diphtheria, y=Life.expectancy)) + geom_box
GDP_box = ggplot(data=life, aes(x = GDP, y=Life.expectancy)) + geom_boxplot()
HB_box = ggplot(data=life, aes(x = Hepatitis.B, y=Life.expectancy)) + geom_bo
In_box = ggplot(data=life, aes(x = infant.deaths, y=Life.expectancy)) + geom_
Po_box = ggplot(data=life, aes(x = infant.deaths, y=Life.expectancy)) + geom_
Sc_box = ggplot(data=life, aes(x = Schooling, y=Life.expectancy)) + geom_boxp
Un_box = ggplot(data=life, aes(x = under.five.deaths, y=Life.expectancy)) + g
ggarrange(al_box,BM_box,Di_box,GDP_box,HB_box,In_box,Po_box, Sc_box,Un_box, l
"infant.deaths", "Polio","Schooling","under.five.deaths"),ncol = 3, nrow = 3)
```

```
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
Warning message:
“Continuous x aesthetic -- did you forget aes(group=...)?”
```



In [125…

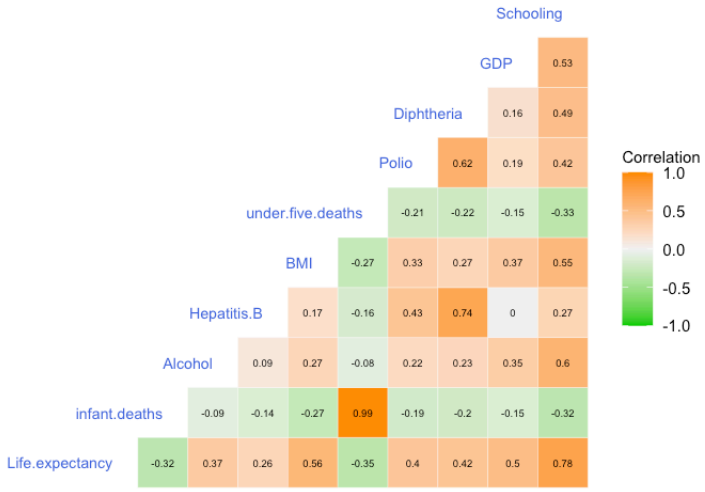
```
al = ggplot(data=life, aes(x = Alcohol, y=Life.expectancy)) + geom_point()
BM = ggplot(data=life, aes(x = BMI, y=Life.expectancy)) + geom_point()
Di = ggplot(data=life, aes(x = Diphtheria, y=Life.expectancy)) + geom_point()
GDP = ggplot(data=life, aes(x = GDP, y=Life.expectancy)) + geom_point()
HB = ggplot(data=life, aes(x = Hepatitis.B, y=Life.expectancy)) + geom_point(
In = ggplot(data=life, aes(x = infant.deaths, y=Life.expectancy)) + geom_poin
Po = ggplot(data=life, aes(x = infant.deaths, y=Life.expectancy)) + geom_poin
Sc = ggplot(data=life, aes(x = Schooling, y=Life.expectancy)) + geom_point()
Un = ggplot(data=life, aes(x = under.five.deaths, y=Life.expectancy)) + geom_
ggarrange(al,BM,Di,GDP,HB,In,Po, Sc, Un, labels = c("Alcohol", "BMI","Diphthe
"infant.deaths", "Polio","Schooling","under.five.deaths"),ncol = 3, nrow = 3)
```



b.correlation between numerical variables

In [126…

```
# numerical data
life_num = life %>% select_if(is.numeric)
ggcorr(life_num, label = T, label_size = 2, label_round = 2, hjust = 1, size
       color = "royalblue", layout.exp = 5, low = "green3", mid = "gray95", h
```



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

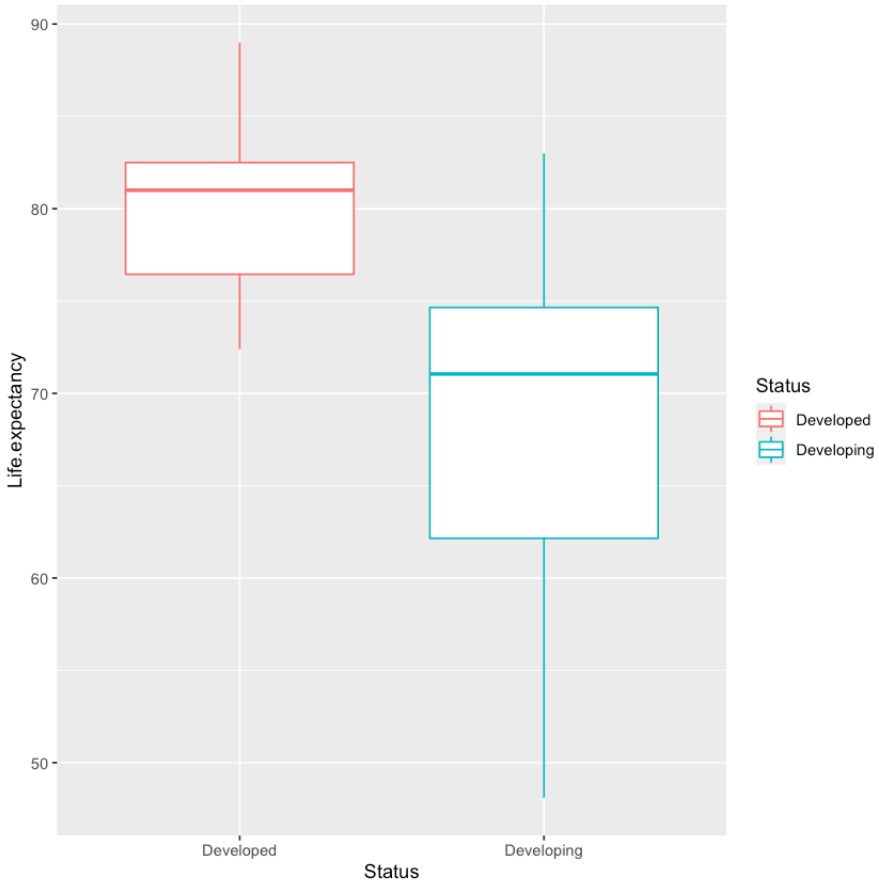
In [127…

```
# categorical data
life %>% group_by(Status) %>% summarise(count = n()) %>% mutate(percentage = 

ggplot(life, aes(Status, Life.expectancy, color = Status))+geom_boxplot(outli

summary(aov(Life.expectancy ~ Status, data = life))
```

A tibble: 2 × 3

Status	count	percentage
<fct>	<int>	<chr>
Developed	19	13.48%
Developing	122	86.52%

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Status        1   2252  2252.5   36.46 1.34e-08 ***
Residuals   139   8587    61.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



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) comment the colinearity between the data

In [128…

```
train_data = life[1:115,]
test_data = life[-c(1:115),]
life_model <- lm(Life.expectancy ~., data = train_data)
summary(life_model)
```

```
Call:
lm(formula = Life.expectancy ~ ., data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.0755  -2.8057   0.3676   3.0179  10.4312 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.779e+01  3.425e+00  13.953  < 2e-16 ***
StatusDeveloping -2.796e+00  1.774e+00  -1.576  0.11800    
infant.deaths     1.545e-01  5.626e-02   2.746  0.00711 ** 
Alcohol          -2.394e-01  1.692e-01  -1.415  0.16017    
Hepatitis.B      -2.170e-02  2.694e-02  -0.805  0.42249    
BMI               6.212e-02  2.675e-02   2.322  0.02217 *  
under.five.deaths -1.143e-01  3.960e-02  -2.885  0.00476 ** 
Polio            -9.302e-03  3.099e-02  -0.300  0.76468    
Diphtheria        5.114e-02  4.166e-02   1.227  0.22241    
GDP               4.416e-05  5.318e-05   0.830  0.40817    
Schooling         1.754e+00  2.717e-01   6.456 3.49e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.931 on 104 degrees of freedom
Multiple R-squared:  0.7084,    Adjusted R-squared:  0.6803 
F-statistic: 25.26 on 10 and 104 DF,  p-value: < 2.2e-16
```

One of the first things we should notice is that the F-test for the regression tells us that the regression is significant, however some individuals predictor is not. This happens as a result of the predictors being highly correlated.

In [129…

```
data.frame(vif(life_model))
```

A data.frame: 10 × 1

	vif.life_model.
	<dbl>
StatusDeveloping	1.874758
infant.deaths	66.175794
Alcohol	2.208582
Hepatitis.B	2.018049
BMI	1.554635
under.five.deaths	67.467059
Polio	2.295877
Diphtheria	3.781750
GDP	1.753947
Schooling	3.220494

In [130…

```
X = model.matrix(life_model)[, -1]
round(cor(X),2)
```

A matrix: 10 × 10 of type dbl

	StatusDeveloping	infant.deaths	Alcohol	Hepatitis.B	BMI	under.five.death
StatusDeveloping	1.00	0.15	-0.60	-0.06	-0.31	0.1
infant.deaths	0.15	1.00	-0.12	-0.15	-0.27	0.9
Alcohol	-0.60	-0.12	1.00	0.16	0.29	-0.1
Hepatitis.B	-0.06	-0.15	0.16	1.00	0.21	-0.1
BMI	-0.31	-0.27	0.29	0.21	1.00	-0.2
under.five.deaths	0.15	0.99	-0.11	-0.17	-0.27	1.0
Polio	-0.18	-0.19	0.25	0.50	0.33	-0.2
Diphtheria	-0.18	-0.22	0.33	0.69	0.34	-0.2
GDP	-0.52	-0.16	0.40	-0.01	0.42	-0.1
Schooling	-0.50	-0.33	0.63	0.27	0.56	-0.3

The VIF of $infant.deaths$ and $under.five.deaths$ are 66.175794 and 67.467059 respectively. And the correlation between the two variables is 0.99. Both tell us that the two variables are highly corelated. This strong correlation indicates multicollinearity among them. Therefore, we are going to deselect $under.5.deaths$, with consideration that other variables seems more related with conditions during infants period.

3 Data Analysis (20 points)

Hint: Note that the training dataset is used to create the model, and the testing dataset is used to evaluate the model fitting. After selecting the final model, please remember to use the whole cleaned dataset to refit the selected model and interpret your results.

3.1 Data Analysis A.1 (10 points)

The investigator want to find the health and economical factors can significantly affect the life expectancy. Create a model to investigate the association between these health and economical factors and the life expectancy of countries. In this step, you do not need to consider interaction terms in your model. To build the final model to answer this question, you need to use model diagnostic tools to evaluate the model. If you find any problems in model diagnostic, you may consider the following tools to build final model\ •Identify and deal with unusual points (leverage points, outliers, influential points)\ •Variable transformation\ •Variable selection\ •Robust method\ Please state your final model with justification. Check the performance of your final linear regression model. Summarize your findings from the final model.

In [131…

```
## define plotting functions
plot_residuals <- function(model, model_name){
  par(mfrow=c(1,2))
  ## Residuals vs Fitted
  par(col.lab="white")
  plot(model, which=1)
  par(col.lab="black")
  title(xlab=paste("Fitted Values (", model_name, ")", sep=""), ylab="Residua
  ## Q-Q Plot
  par(col.lab="white")
  plot(model, which=2)
  par(col.lab="black")
  title(xlab=paste("Theoretical Quantiles (", model_name, ")", sep=""), ylab=
}
```

a.regression-remove under.five.deaths

I.simple linear regression - drop $under.5.deaths$

In [132…

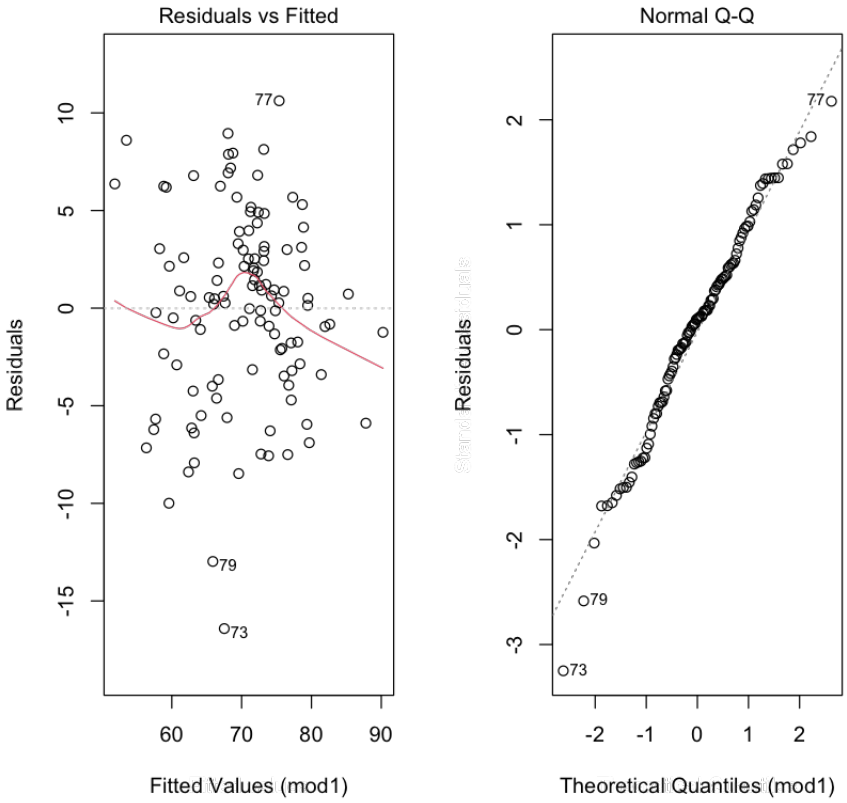
```
mod1 <- lm(Life.expectancy ~ Status + infant.deaths + Alcohol+ Hepatitis.B
           + BMI + Polio+ Diphtheria + GDP + Schooling, data = train_data)
summary(mod1)

(rmse(train_data$Life.expectancy, mod1$fit))

test_predict1 = predict.lm(mod1, test_data)
(rmse(test_data$Life.expectancy, test_predict1))

plot_residuals(mod1, 'mod1')
ggplot (data = test_data, aes (x=Life.expectancy,y= test_predict1)) +
  geom_smooth(se=F,method = "lm",colour = "gray35") + geom_point() +
  ggtitle('Linear Regression: Ground Truth vs Predicted') +
  xlab('Ground Truth') +
  ylab('Predicted')
```

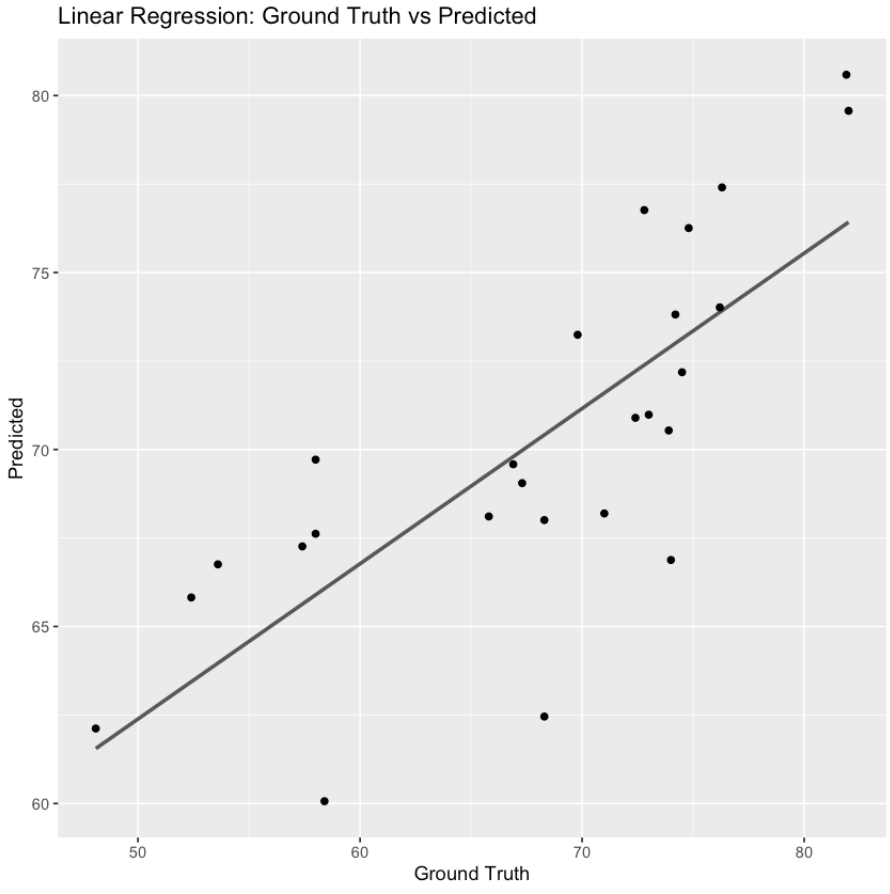
```
Call:
lm(formula = Life.expectancy ~ Status + infant.deaths + Alcohol +
    Hepatitis.B + BMI + Polio + Diphtheria + GDP + Schooling,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.4151  -3.1769   0.5027   3.0220  10.6207

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.603e+01  3.486e+00  13.205  < 2e-16 ***
StatusDeveloping -3.051e+00  1.832e+00  -1.665   0.0989 .
infant.deaths    -6.393e-03  7.686e-03  -0.832   0.4074
Alcohol          -3.474e-01  1.707e-01  -2.036   0.0443 *
Hepatitis.B      -1.495e-02  2.776e-02  -0.539   0.5913
BMI               5.924e-02  2.765e-02   2.142   0.0345 *
Polio            -6.419e-03  3.204e-02  -0.200   0.8416
Diphtheria        5.843e-02  4.301e-02   1.359   0.1772
GDP               3.875e-05  5.497e-05   0.705   0.4823
Schooling         1.857e+00  2.785e-01   6.669 1.24e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.1 on 105 degrees of freedom
Multiple R-squared:  0.685,     Adjusted R-squared:  0.658
F-statistic: 25.37 on 9 and 105 DF,  p-value: < 2.2e-16
```

4.87284324471395

6.36824927502708

```
`geom_smooth()` using formula 'y ~ x'
```


Residuals vs Fitted
Residuals
Fitted Values (mod1)
77
79
73
Normal Q-Q
Standardized Residuals
Theoretical Quantiles (mod1)
77
79
73

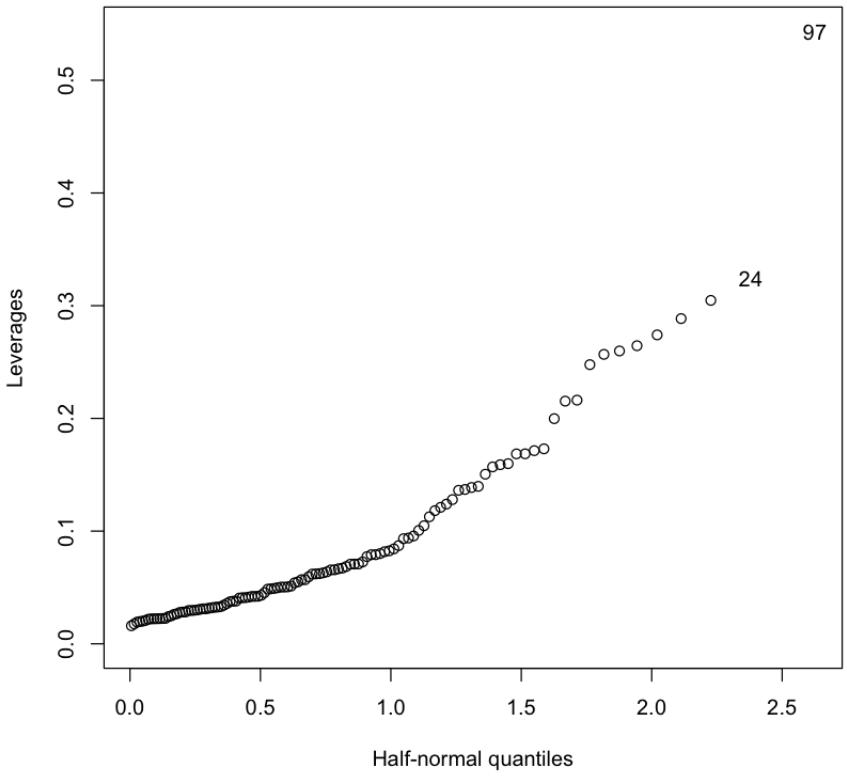

Linear Regression: Ground Truth vs Predicted
Predicted
Ground Truth

II. Unusual Points

Leverage Point:

In [133…

```
halfnorm(lm.influence(mod1)$hat, nlab = 2, ylab="Leverages")
```

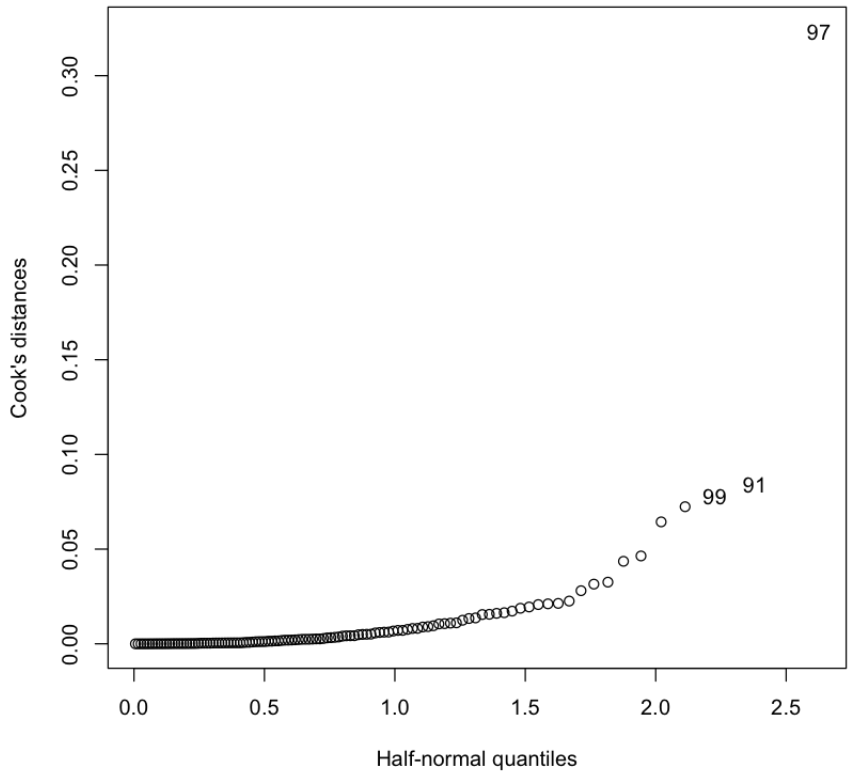


Influential Point:

In [134…

```
n =dim(life)[1]
p = dim(life)[2]
cook=cooks.distance(mod1)
cook[which(cook > 10/(n-p-1))]
halfnorm(cook, 3, ylab="Cook's distances")
```

91: 0.0839989838624467 **97:** 0.323296686573423 **99:** 0.0780290817945851



In [135…

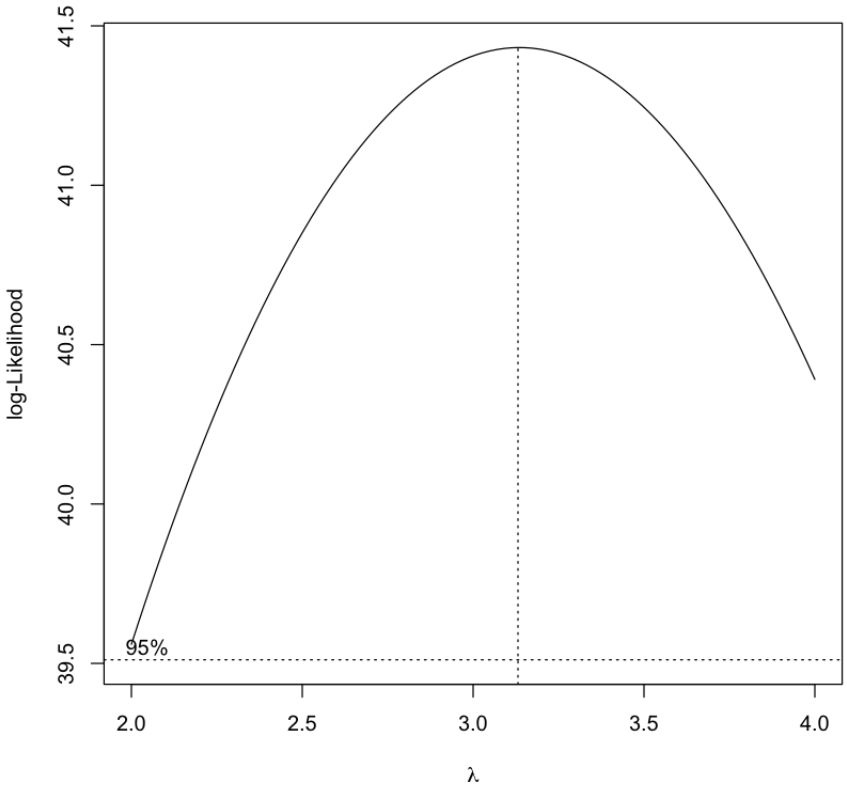
```
train_data = train_data[-c(97),]
```

From above, we know that the row 97 is an influential point, also a leverage point. To avoid the influence caused by some individual points, we decide to delete this row.

III. Variable Transformation:

In [136…

```
lmod_bc = boxcox(lm(Life.expectancy ~ Status + infant.deaths + Alcohol+ Hepat
           + BMI + Polio+ Diphtheria + GDP + Schooling, data=train_data), la
```



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.

IIII.Fit a new linear model:

In [137…

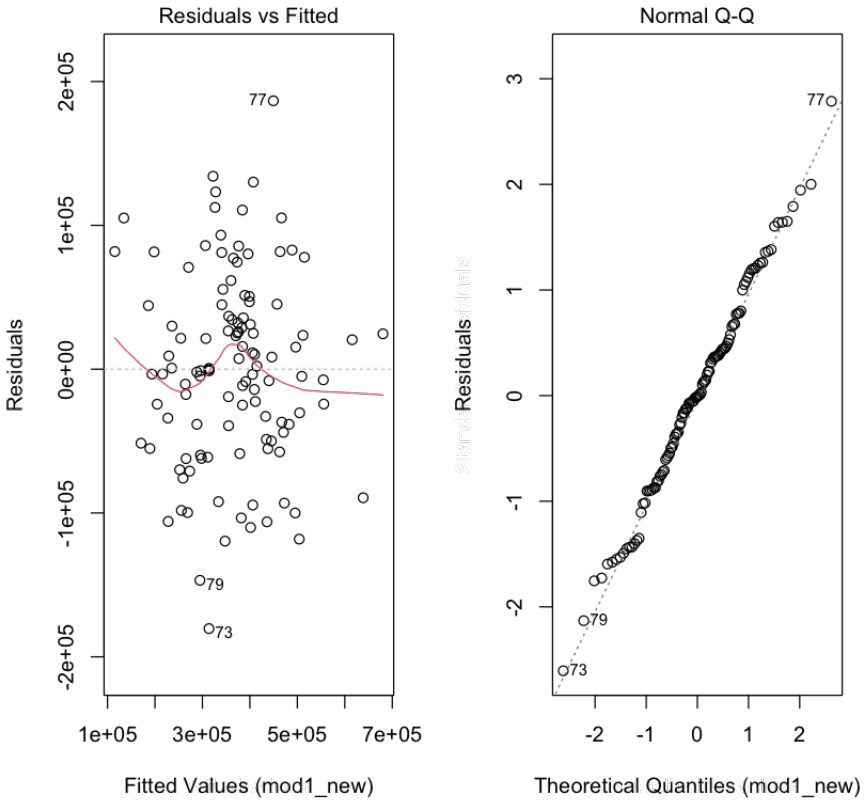
```
mod1_new=lm(Life.expectancy^3 ~ Status + infant.deaths + Alcohol+ Hepatitis.B
            + BMI + Polio+ Diphtheria + GDP + Schooling, data = train_data)
summary(mod1_new)

(rmse(train_data$Life.expectancy,mod1_new$fit^(1/3)))

test_predict1_new = (predict.lm(mod1_new, test_data))
(rmse(test_data$Life.expectancy,test_predict1_new^(1/3)))


plot_residuals(mod1_new, 'mod1_new')
ggplot (data = test_data, aes (x=Life.expectancy,y= test_predict1_new)) +
  geom_smooth(se=F,method = "lm",colour = "gray35") + geom_point() +
  ggtitle('Linear Regression: Ground Truth vs Predicted') +
  xlab('Ground Truth') +
  ylab('Predicted')
```

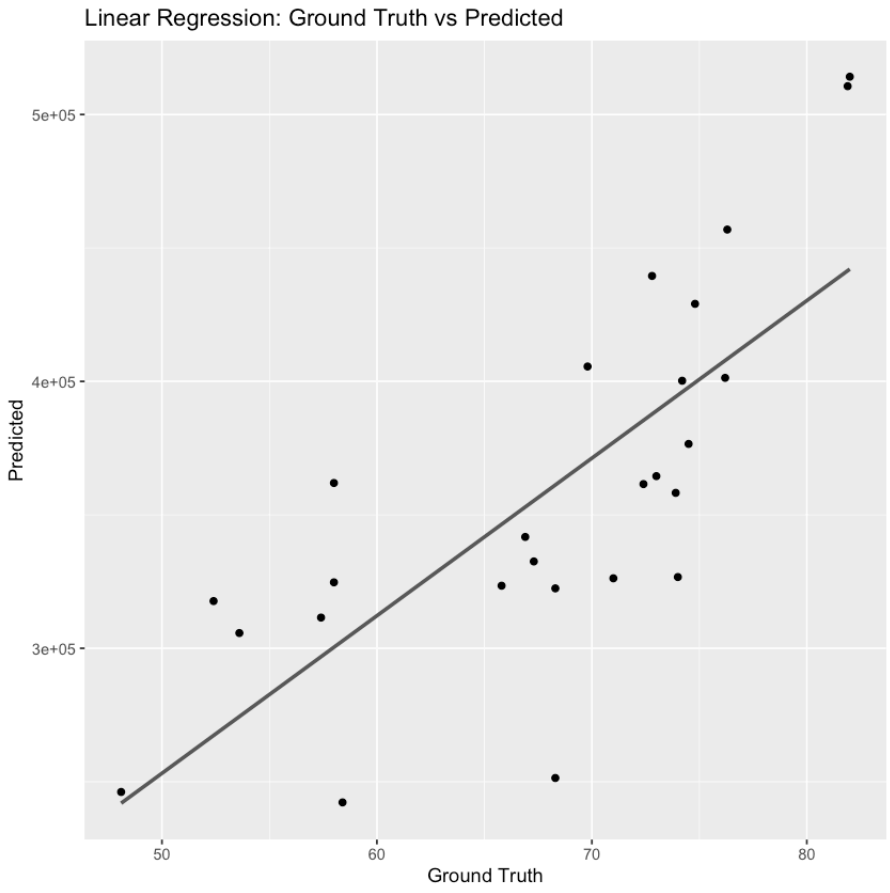
```
Call:
lm(formula = Life.expectancy^3 ~ Status + infant.deaths + Alcohol +
    Hepatitis.B + BMI + Polio + Diphtheria + GDP + Schooling,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-180376  -47606    -666   42338  186600

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.451e+04  4.828e+04   0.922   0.3588
StatusDeveloping -5.314e+04  2.526e+04  -2.104   0.0378 *
infant.deaths     7.614e+01  1.499e+02   0.508   0.6127
Alcohol          -3.615e+03  2.387e+03  -1.514   0.1330
Hepatitis.B      -4.494e+02  3.813e+02  -1.179   0.2412
BMI               7.949e+02  3.818e+02   2.082   0.0398 *
Polio            -1.356e+02  4.399e+02  -0.308   0.7585
Diphtheria        9.348e+02  5.903e+02   1.584   0.1163
GDP               1.127e+00  7.543e-01   1.494   0.1381
Schooling         2.517e+04  3.829e+03   6.574    2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69970 on 104 degrees of freedom
Multiple R-squared:  0.7024,    Adjusted R-squared:  0.6767
F-statistic: 27.28 on 9 and 104 DF,  p-value: < 2.2e-16
```

4.88508755036417

6.81139090761671

```
`geom_smooth()` using formula 'y ~ x'
```


Residuals vs Fitted
Residuals
Fitted Values (mod1_new)
77
79
73
Normal Q-Q
Standardized Residuals
Theoretical Quantiles (mod1_new)


Linear Regression: Ground Truth vs Predicted
Predicted
Ground Truth

VI. Variable Selection

1) AIC stepwise search

In [138…

```
step(mod1_new)
```

```
Start:  AIC=2553.07
Life.expectancy^3 ~ Status + infant.deaths + Alcohol + Hepatitis.B +
    BMI + Polio + Diphtheria + GDP + Schooling

                Df  Sum of Sq        RSS    AIC
- Polio          1 4.6514e+08 5.0965e+11 2551.2
- infant.deaths  1 1.2625e+09 5.1045e+11 2551.3
- Hepatitis.B    1 6.8011e+09 5.1599e+11 2552.6
<none>                       5.0919e+11 2553.1
- GDP            1 1.0931e+10 5.2012e+11 2553.5
- Alcohol        1 1.1227e+10 5.2042e+11 2553.6
- Diphtheria     1 1.2280e+10 5.2147e+11 2553.8
- BMI            1 2.1219e+10 5.3041e+11 2555.7
- Status         1 2.1673e+10 5.3086e+11 2555.8
- Schooling      1 2.1157e+11 7.2076e+11 2590.7

Step:  AIC=2551.17
Life.expectancy^3 ~ Status + infant.deaths + Alcohol + Hepatitis.B +
    BMI + Diphtheria + GDP + Schooling

                Df  Sum of Sq        RSS    AIC
- infant.deaths  1 1.2169e+09 5.1087e+11 2549.4
- Hepatitis.B    1 6.7924e+09 5.1645e+11 2550.7
<none>                       5.0965e+11 2551.2
- GDP            1 1.0882e+10 5.2054e+11 2551.6
- Alcohol        1 1.1021e+10 5.2067e+11 2551.6
- Diphtheria     1 1.4106e+10 5.2376e+11 2552.3
- BMI            1 2.0819e+10 5.3047e+11 2553.7
- Status         1 2.1433e+10 5.3109e+11 2553.9
- Schooling      1 2.1116e+11 7.2081e+11 2588.7

Step:  AIC=2549.44
Life.expectancy^3 ~ Status + Alcohol + Hepatitis.B + BMI + Diphtheria +
    GDP + Schooling

              Df  Sum of Sq        RSS    AIC
- Hepatitis.B  1 6.7202e+09 5.1759e+11 2548.9
<none>                      5.1087e+11 2549.4
- GDP          1 1.1187e+10 5.2206e+11 2549.9
- Alcohol      1 1.1517e+10 5.2239e+11 2550.0
- Diphtheria   1 1.4311e+10 5.2518e+11 2550.6
- BMI          1 1.9804e+10 5.3067e+11 2551.8
- Status       1 2.1875e+10 5.3275e+11 2552.2
- Schooling    1 2.1261e+11 7.2348e+11 2587.1

Step:  AIC=2548.93
```

```
Life.expectancy^3 ~ Status + Alcohol + BMI + Diphtheria + GDP +
    Schooling

             Df  Sum of Sq        RSS    AIC
- Diphtheria  1 7.6156e+09 5.2521e+11 2548.6
<none>                     5.1759e+11 2548.9
- Alcohol     1 1.1643e+10 5.2923e+11 2549.5
- GDP         1 1.4491e+10 5.3208e+11 2550.1
- BMI         1 1.8176e+10 5.3577e+11 2550.9
- Status      1 2.0829e+10 5.3842e+11 2551.4
- Schooling   1 2.2267e+11 7.4026e+11 2587.7

Step:  AIC=2548.6
Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooling

            Df  Sum of Sq        RSS    AIC
<none>                    5.2521e+11 2548.6
- Alcohol    1 1.0465e+10 5.3567e+11 2548.8
- GDP        1 1.2462e+10 5.3767e+11 2549.3
- Status     1 1.8327e+10 5.4353e+11 2550.5
- BMI        1 2.0160e+10 5.4537e+11 2550.9
- Schooling  1 3.0564e+11 8.3085e+11 2598.9
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling, data = train_data)

Coefficients:
     (Intercept)  StatusDeveloping           Alcohol               BMI
        49842.45          -48420.51          -3471.35            758.83
             GDP          Schooling
            1.18           27088.57
```

In [139…

```
AICmod = lm(Life.expectancy^3~Status + Alcohol + BMI + GDP + Schooling, data =

summary(AICmod)

(rmse(train_data$Life.expectancy,AICmod$fit^(1/3)))

AICtest_predict = predict.lm(AICmod, test_data)
(rmse(test_data$Life.expectancy,AICtest_predict^(1/3)))

plot_residuals(AICmod, 'AICmod')
ggplot (data = test_data, aes (x=Life.expectancy,y= AICtest_predict)) +
  geom_smooth(se=F,method = "lm",colour = "gray35") + geom_point() +
  ggtitle('Linear Regression: Ground Truth vs Predicted') +
  xlab('Ground Truth') +
  ylab('Predicted')
```

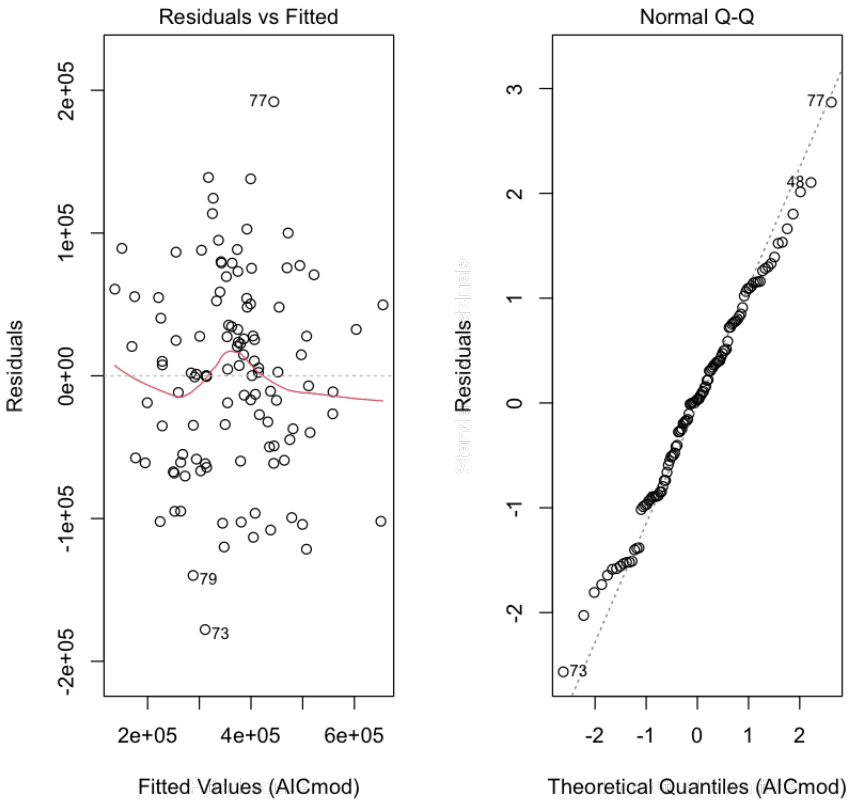
```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling, data = train_data)
```

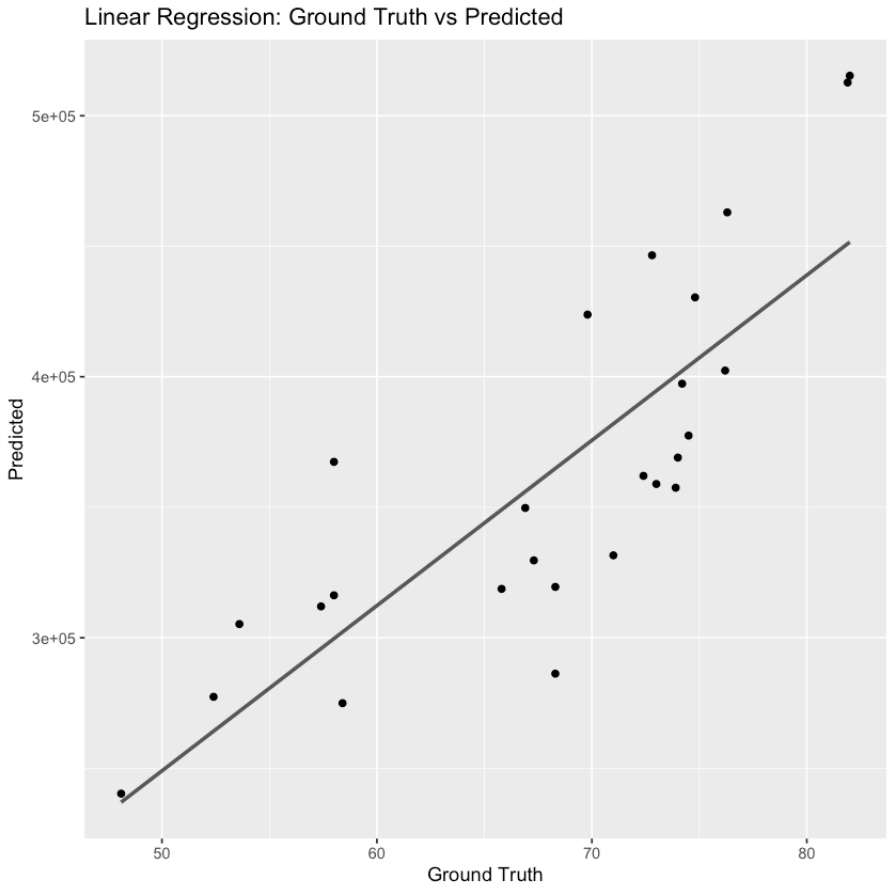
```
Residuals:
    Min      1Q  Median      3Q     Max
-177769  -53731    2206   49285  191950

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.984e+04  4.349e+04   1.146   0.2543
StatusDeveloping -4.842e+04  2.494e+04  -1.941   0.0548 .
Alcohol          -3.471e+03  2.366e+03  -1.467   0.1453
BMI               7.588e+02  3.727e+02   2.036   0.0442 *
GDP               1.180e+00  7.372e-01   1.601   0.1123
Schooling         2.709e+04  3.417e+03   7.928 2.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69740 on 108 degrees of freedom
Multiple R-squared:  0.6931,    Adjusted R-squared:  0.6789
F-statistic: 48.77 on 5 and 108 DF,  p-value: < 2.2e-16
```

4.88910104581347

6.51113351795897

```
`geom_smooth()` using formula 'y ~ x'
```





2) AIC exhaustive search

In [140…

```
AIC_e <- regsubsets(Life.expectancy^3 ~ Status + infant.deaths + Alcohol+ Hep
           + BMI + Polio+ Diphtheria + GDP + Schooling,train_data)
s = summary(AIC_e)
s
```

```
Subset selection object
Call: regsubsets.formula(Life.expectancy^3 ~ Status + infant.deaths +
    Alcohol + Hepatitis.B + BMI + Polio + Diphtheria + GDP +
    Schooling, train_data)
9 Variables  (and intercept)
                 Forced in Forced out
StatusDeveloping     FALSE      FALSE
infant.deaths        FALSE      FALSE
Alcohol              FALSE      FALSE
Hepatitis.B          FALSE      FALSE
BMI                  FALSE      FALSE
Polio                FALSE      FALSE
Diphtheria           FALSE      FALSE
GDP                  FALSE      FALSE
Schooling            FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         StatusDeveloping infant.deaths Alcohol Hepatitis.B BMI Polio
1  ( 1 ) " "              " "           " "     " "         " " " "
2  ( 1 ) " "              " "           " "     " "         " " " "
3  ( 1 ) " "              " "           " "     " "         "*" " "
4  ( 1 ) "*"              " "           " "     " "         "*" " "
5  ( 1 ) "*"              " "           "*"     " "         "*" " "
6  ( 1 ) "*"              " "           "*"     " "         "*" " "
7  ( 1 ) "*"              " "           "*"     "*"         "*" " "
8  ( 1 ) "*"              "*"           "*"     "*"         "*" " "
         Diphtheria GDP Schooling
1  ( 1 ) " "        " " "*"
2  ( 1 ) " "        "*" "*"
3  ( 1 ) " "        "*" "*"
4  ( 1 ) " "        "*" "*"
5  ( 1 ) " "        "*" "*"
6  ( 1 ) "*"        "*" "*"
7  ( 1 ) "*"        "*" "*"
8  ( 1 ) "*"        "*" "*"
```

In [141…

```
aic = 114*log(s$rss/114) + (2:10)*2
which.min(aic)
```

```
Warning message in 114 * log(s$rss/114) + (2:10) * 2:
“longer object length is not a multiple of shorter object length”
```

5

1. BIC

In [142…

```
BICSelect = step(mod1_new, k = log(9))
```

```
Start:  AIC=2555.04
Life.expectancy^3 ~ Status + infant.deaths + Alcohol + Hepatitis.B +
    BMI + Polio + Diphtheria + GDP + Schooling
```

```
                Df  Sum of Sq        RSS    AIC
- Polio          1 4.6514e+08 5.0965e+11 2552.9
- infant.deaths  1 1.2625e+09 5.1045e+11 2553.1
- Hepatitis.B    1 6.8011e+09 5.1599e+11 2554.3
<none>                       5.0919e+11 2555.0
- GDP            1 1.0931e+10 5.2012e+11 2555.3
- Alcohol        1 1.1227e+10 5.2042e+11 2555.3
- Diphtheria     1 1.2280e+10 5.2147e+11 2555.6
- BMI            1 2.1219e+10 5.3041e+11 2557.5
- Status         1 2.1673e+10 5.3086e+11 2557.6
- Schooling      1 2.1157e+11 7.2076e+11 2592.4

Step:  AIC=2552.95
Life.expectancy^3 ~ Status + infant.deaths + Alcohol + Hepatitis.B +
    BMI + Diphtheria + GDP + Schooling

                Df  Sum of Sq        RSS    AIC
- infant.deaths  1 1.2169e+09 5.1087e+11 2551.0
- Hepatitis.B    1 6.7924e+09 5.1645e+11 2552.3
<none>                       5.0965e+11 2552.9
- GDP            1 1.0882e+10 5.2054e+11 2553.2
- Alcohol        1 1.1021e+10 5.2067e+11 2553.2
- Diphtheria     1 1.4106e+10 5.2376e+11 2553.9
- BMI            1 2.0819e+10 5.3047e+11 2555.3
- Status         1 2.1433e+10 5.3109e+11 2555.4
- Schooling      1 2.1116e+11 7.2081e+11 2590.3

Step:  AIC=2551.02
Life.expectancy^3 ~ Status + Alcohol + Hepatitis.B + BMI + Diphtheria +
    GDP + Schooling

              Df  Sum of Sq        RSS    AIC
- Hepatitis.B  1 6.7202e+09 5.1759e+11 2550.3
<none>                      5.1087e+11 2551.0
- GDP          1 1.1187e+10 5.2206e+11 2551.3
- Alcohol      1 1.1517e+10 5.2239e+11 2551.4
- Diphtheria   1 1.4311e+10 5.2518e+11 2552.0
- BMI          1 1.9804e+10 5.3067e+11 2553.2
- Status       1 2.1875e+10 5.3275e+11 2553.6
- Schooling    1 2.1261e+11 7.2348e+11 2588.5

Step:  AIC=2550.31
Life.expectancy^3 ~ Status + Alcohol + BMI + Diphtheria + GDP +
    Schooling

             Df  Sum of Sq        RSS    AIC
- Diphtheria  1 7.6156e+09 5.2521e+11 2549.8
<none>                     5.1759e+11 2550.3
- Alcohol     1 1.1643e+10 5.2923e+11 2550.7
- GDP         1 1.4491e+10 5.3208e+11 2551.3
- BMI         1 1.8176e+10 5.3577e+11 2552.1
- Status      1 2.0829e+10 5.3842e+11 2552.6
```

```
- Schooling   1 2.2267e+11 7.4026e+11 2588.9

Step:  AIC=2549.78
Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooling

            Df  Sum of Sq        RSS    AIC
<none>                   5.2521e+11 2549.8
- Alcohol    1 1.0465e+10 5.3567e+11 2549.8
- GDP        1 1.2462e+10 5.3767e+11 2550.3
- Status     1 1.8327e+10 5.4353e+11 2551.5
- BMI        1 2.0160e+10 5.4537e+11 2551.9
- Schooling  1 3.0564e+11 8.3085e+11 2599.9
```

In [143…

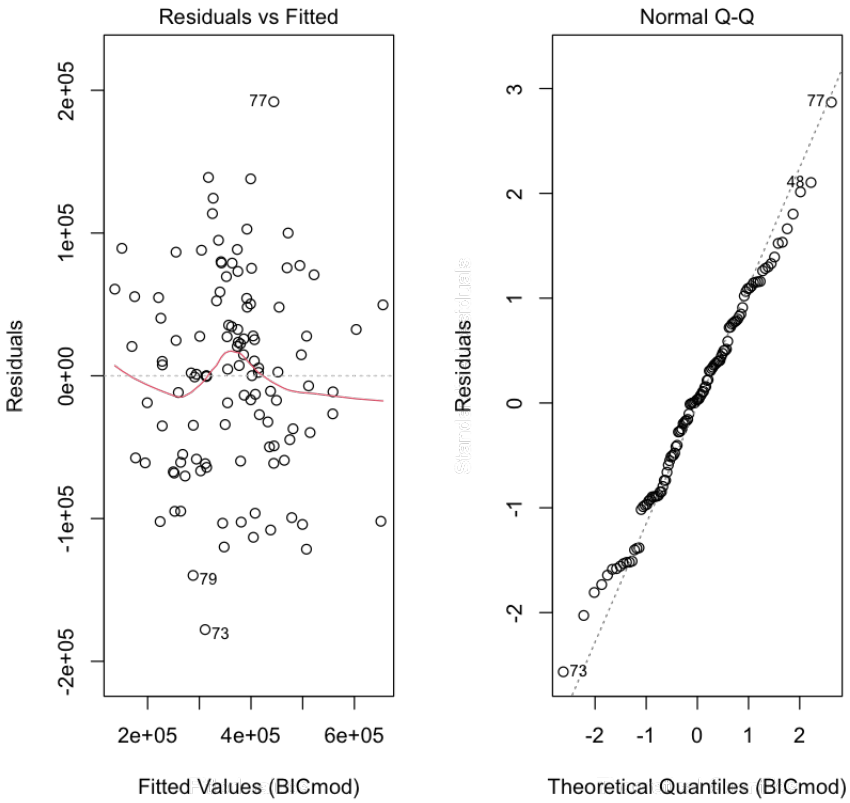
```
BICmod = lm(Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooling, dat

summary(BICmod)

(rmse(train_data$Life.expectancy,BICmod$fit^(1/3)))

BICtest_predict = predict.lm(BICmod, test_data)
(rmse(test_data$Life.expectancy,BICtest_predict^(1/3)))

plot_residuals(BICmod, 'BICmod')
ggplot (data = test_data, aes (x=Life.expectancy,y= BICtest_predict)) +
  geom_smooth(se=F,method = "lm",colour = "gray35") + geom_point() +
  ggtitle('Linear Regression: Ground Truth vs Predicted') +
  xlab('Ground Truth') +
  ylab('Predicted')
```

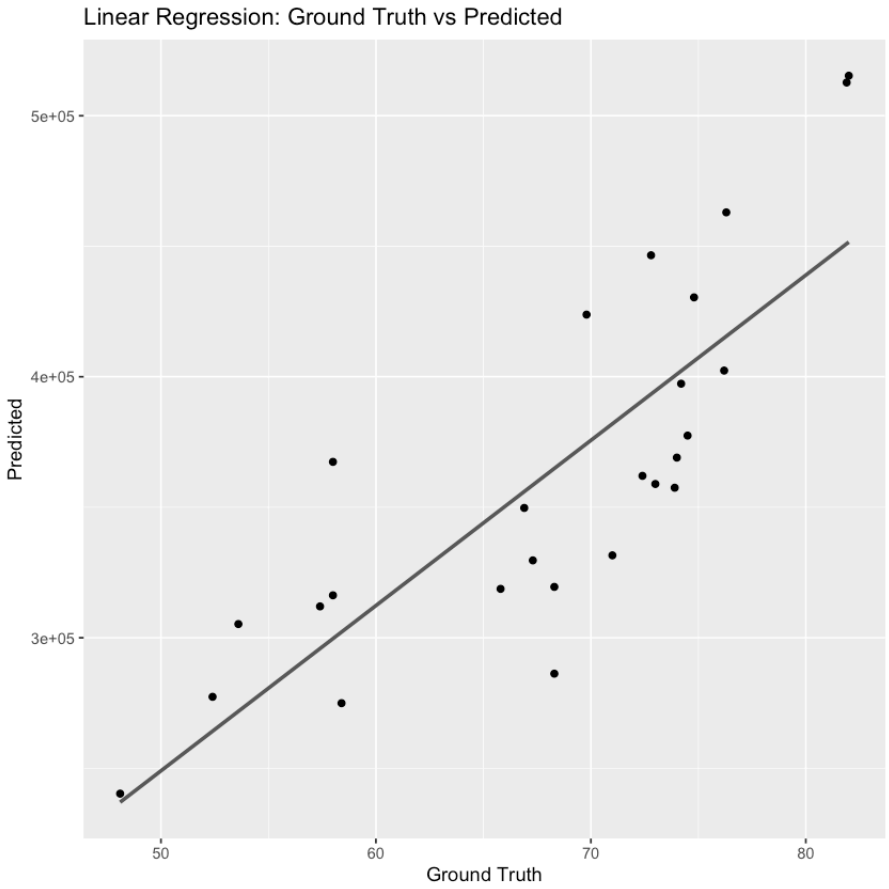
```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-177769  -53731    2206   49285  191950

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.984e+04  4.349e+04   1.146   0.2543
StatusDeveloping -4.842e+04  2.494e+04  -1.941   0.0548 .
Alcohol          -3.471e+03  2.366e+03  -1.467   0.1453
BMI               7.588e+02  3.727e+02   2.036   0.0442 *
GDP               1.180e+00  7.372e-01   1.601   0.1123
Schooling         2.709e+04  3.417e+03   7.928 2.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69740 on 108 degrees of freedom
Multiple R-squared:  0.6931,    Adjusted R-squared:  0.6789
F-statistic: 48.77 on 5 and 108 DF,  p-value: < 2.2e-16
```

4.88910104581347

6.51113351795897

```
`geom_smooth()` using formula 'y ~ x'
```



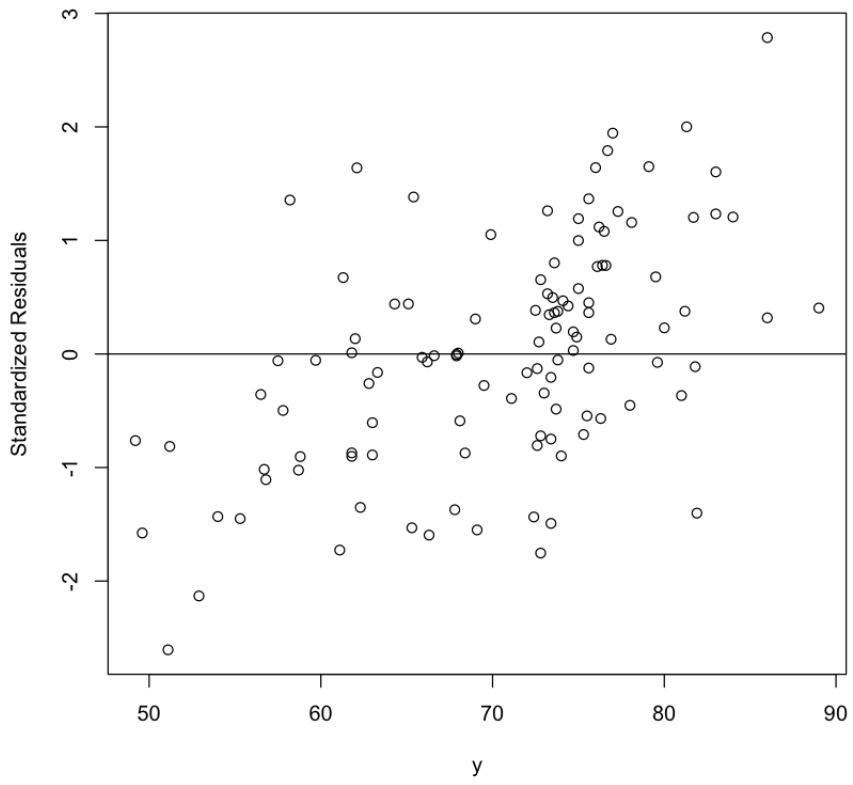


From results above, we decide to remain 5 variables regard of $RMSE$ and adjusted R^2. Though BIC selects 3 variables also has good prediction. BUt since we have a very small dataset, with low complexity pf our data, according to vias-variance tradeoff, maybe we will have high bias when using our model to predict larger unseen dataset.

V. Robust Method

In [144…

```
#create plot of y-values vs. standardized residuals
plot(train_data$Life.expectancy, rstandard(mod1_new), ylab='Standardized Resi
abline(h=0)
```



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 do huber regression.

In [151…

```
HRR = rlm(Life.expectancy^3~Status + Alcohol + BMI + GDP + Schooling,data = t
summary(HRR)

(rmse(train_data$Life.expectancy,HRR$fit^(1/3)))

HRR_predict = predict.lm(HRR, test_data)
(rmse(test_data$Life.expectancy,HRR_predict^(1/3)))
```

```
Call: rlm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling, data = train_data)
Residuals:
      Min         1Q    Median        3Q       Max
-178417.1  -54171.9     584.8   52108.9  198478.9

Coefficients:
                 Value       Std. Error  t value
(Intercept)       44198.8443  44270.2869      0.9984
StatusDeveloping -44589.1172  25392.0727     -1.7560
Alcohol           -3926.4430   2409.0575     -1.6299
BMI                 724.5189    379.4117      1.9096
GDP                   1.2093      0.7505      1.6113
Schooling         27517.7810   3478.5068      7.9108

Residual standard error: 78840 on 108 degrees of freedom
```

4.89201156964594

6.50992545794026

In [150…

```
# least absolute deviations
LAD = rq(Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooling, data =
summary(LAD)

(rmse(train_data$Life.expectancy, LAD$fit^(1/3)))

LAD_predict = predict(LAD, test_data)
(rmse(test_data$Life.expectancy,LAD_predict^(1/3)))
```

```
Call: rq(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling, data = train_data)

tau: [1] 0.5

Coefficients:
                 coefficients  lower bd      upper bd
(Intercept)        56540.71459 -127877.57113  172973.18172
StatusDeveloping  -43818.13316 -107275.92731    6555.91736
Alcohol            -3852.39267   -7066.79144    -362.27061
BMI                  856.80770     221.06389    2222.07512
GDP                    1.26217      -1.06745       3.45166
Schooling          26048.83556   18136.81106   38556.70232
```

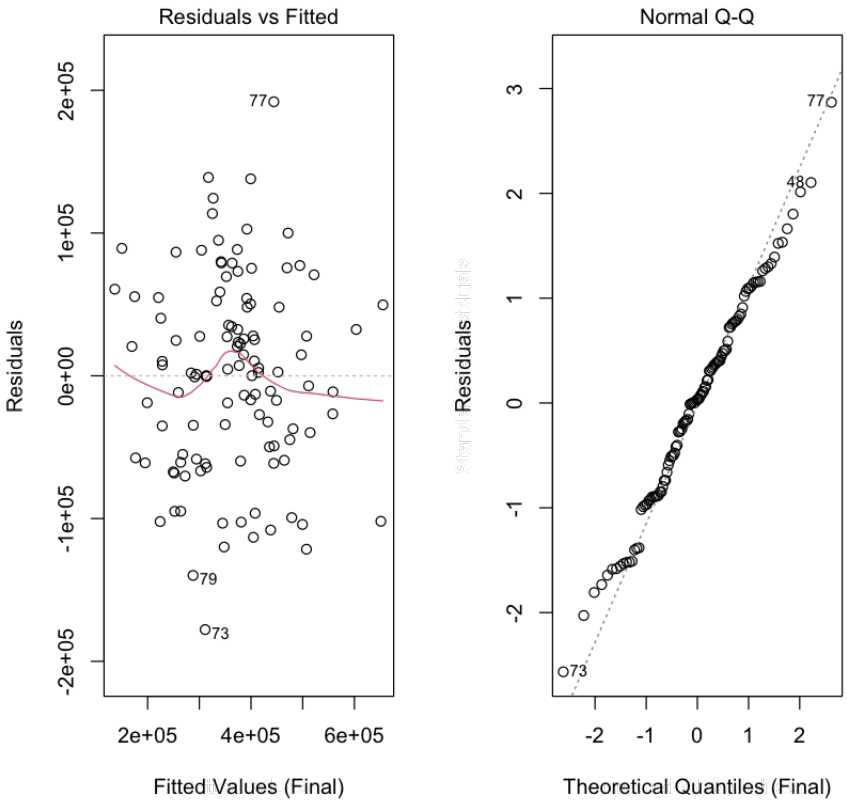
4.88917160674631

6.54918442354358

In [155…

```
Final = lm(Life.expectancy^3~Status + Alcohol + BMI + GDP + Schooling, train_
summary(Final)
plot_residuals(Final, 'Final')
```

```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-177769  -53731    2206   49285  191950

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.984e+04  4.349e+04   1.146   0.2543
StatusDeveloping -4.842e+04  2.494e+04  -1.941   0.0548 .
Alcohol          -3.471e+03  2.366e+03  -1.467   0.1453
BMI               7.588e+02  3.727e+02   2.036   0.0442 *
GDP               1.180e+00  7.372e-01   1.601   0.1123
Schooling         2.709e+04  3.417e+03   7.928 2.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69740 on 108 degrees of freedom
Multiple R-squared:  0.6931,    Adjusted R-squared:  0.6789
F-statistic: 48.77 on 5 and 108 DF,  p-value: < 2.2e-16
```



3.2 Data Analysis A.2 (10 points)

Based on the significant factors you found in Analysis A.1, the investigator want to find whether some of these significant health and economical factors (except for Status) have different effects between developed and developing countries. Please create a model with justification to answer this question.

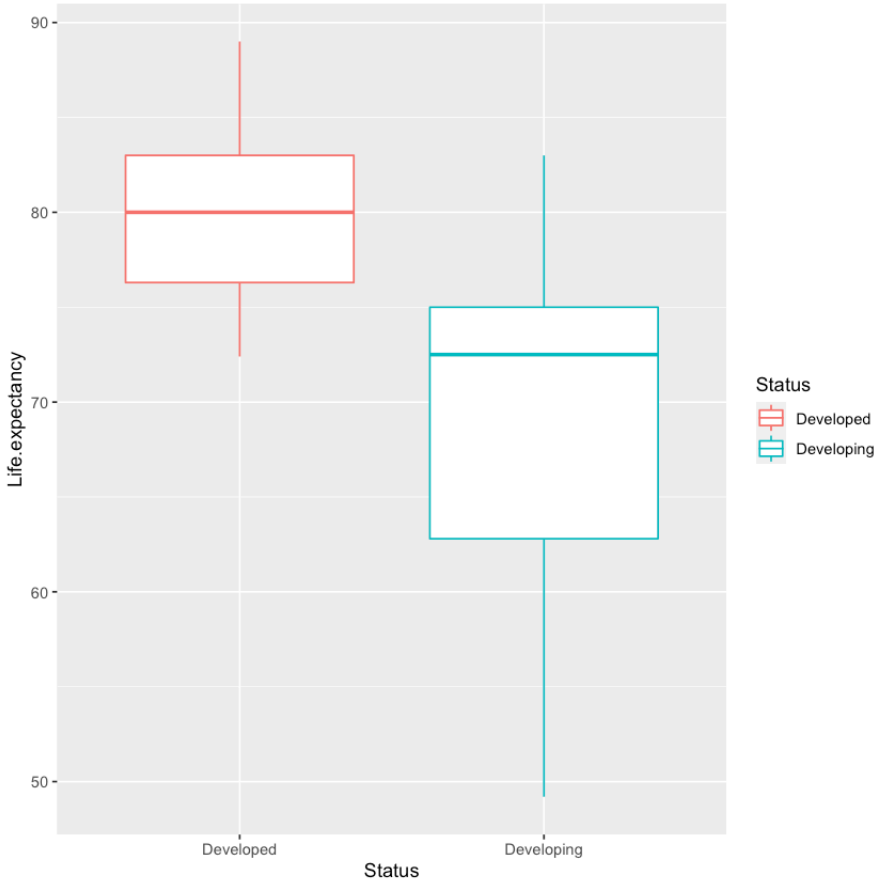
In [156…

```
# summary of the data
by(train_data, train_data$Status, summary)
```

```
train_data$Status: Developed
 Life.expectancy        Status   infant.deaths       Alcohol
 Min.   :72.40   Developed :17   Min.   :0.0000   Min.   : 6.95
 1st Qu.:76.30   Developing: 0   1st Qu.:0.0000   1st Qu.: 9.80
 Median :80.00                  Median :0.0000   Median :10.80
 Mean   :79.81                  Mean   :0.5882   Mean   :10.62
 3rd Qu.:83.00                  3rd Qu.:1.0000   3rd Qu.:11.88
 Max.   :89.00                  Max.   :2.0000   Max.   :12.90
  Hepatitis.B         BMI        under.five.deaths     Polio
 Min.   : 9.00   Min.   : 6.00   Min.   :0.0000    Min.   :76.00
 1st Qu.:88.00   1st Qu.:57.80   1st Qu.:0.0000    1st Qu.:93.00
 Median :94.00   Median :58.90   Median :0.0000    Median :94.00
 Mean   :85.24   Mean   :53.99   Mean   :0.7059    Mean   :93.47
 3rd Qu.:97.00   3rd Qu.:61.90   3rd Qu.:1.0000    3rd Qu.:96.00
 Max.   :98.00   Max.   :67.60   Max.   :3.0000    Max.   :99.00
   Diphtheria         GDP          Schooling
 Min.   :76.00   Min.   : 1356   Min.   :13.70
 1st Qu.:93.00   1st Qu.: 6843   1st Qu.:14.60
 Median :95.00   Median :12600   Median :15.90
 Mean   :93.94   Mean   :21103   Mean   :16.06
 3rd Qu.:98.00   3rd Qu.:35849   3rd Qu.:16.70
 Max.   :99.00   Max.   :51875   Max.   :20.30
------------------------------------------------------------
train_data$Status: Developing
 Life.expectancy        Status   infant.deaths       Alcohol
 Min.   :49.20   Developed : 0   Min.   :  0.00   Min.   : 0.010
 1st Qu.:62.80   Developing:97   1st Qu.:  1.00   1st Qu.: 0.600
 Median :72.50                  Median :  4.00   Median : 3.010
 Mean   :68.85                  Mean   : 23.57   Mean   : 3.768
 3rd Qu.:75.00                  3rd Qu.: 23.00   3rd Qu.: 6.020
 Max.   :83.00                  Max.   :372.00   Max.   :14.970
  Hepatitis.B         BMI        under.five.deaths     Polio
 Min.   : 7.00   Min.   : 2.20   Min.   :  0.00    Min.   : 7.00
 1st Qu.:77.00   1st Qu.:17.30   1st Qu.:  1.00    1st Qu.:78.00
 Median :92.00   Median :33.50   Median :  5.00    Median :92.00
 Mean   :81.37   Mean   :35.22   Mean   : 31.85    Mean   :82.38
 3rd Qu.:96.00   3rd Qu.:53.90   3rd Qu.: 34.00    3rd Qu.:97.00
 Max.   :99.00   Max.   :75.20   Max.   :461.00    Max.   :99.00
   Diphtheria         GDP             Schooling
 Min.   : 7.00   Min.   :    8.38   Min.   : 4.50
 1st Qu.:81.00   1st Qu.:  543.96   1st Qu.:10.20
 Median :93.00   Median : 1366.88   Median :12.30
 Mean   :83.57   Mean   : 4217.81   Mean   :11.78
 3rd Qu.:97.00   3rd Qu.: 4463.39   3rd Qu.:13.50
 Max.   :99.00   Max.   :47447.48   Max.   :17.60
```

In [158…

```
bxp = ggplot(train_data, aes(Status, Life.expectancy, color = Status))+geom_b
bxp
```



In [161…

```
# include consideration of interaction between Status and BMI
model_BMI = lm(Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooling +
summary(model_BMI)
```

```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + 
    Schooling + BMI:Status, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-178074  -54526    3127   44966  188938 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.323e+04  7.095e+04   0.187    0.852    
StatusDeveloping     -1.220e+04  6.076e+04  -0.201    0.841    
Alcohol              -3.333e+03  2.382e+03  -1.399    0.165    
BMI                   1.357e+03  9.875e+02   1.374    0.172    
GDP                   1.109e+00  7.471e-01   1.485    0.141    
Schooling             2.736e+04  3.451e+03   7.928 2.28e-12 ***
StatusDeveloping:BMI -6.839e+02  1.046e+03  -0.654    0.514    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69920 on 107 degrees of freedom
Multiple R-squared:  0.6943,    Adjusted R-squared:  0.6771 
F-statistic:  40.5 on 6 and 107 DF,  p-value: < 2.2e-16
```

In [163…

```
anova(model_BMI, Final)
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	107	523114206507	NA	NA	NA	NA
2	108	525206011990	-1	-2091805482	0.4278668	0.514441

In [164…

```
# include consideration of interaction between Status and Schooling
model_Schooling = lm(Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schoo
summary(model_Schooling)
```

```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP +
    Schooling + Schooling:Status, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-176128  -52140    1695   44173  153121

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.500e+05  1.693e+05   2.068   0.0411 *
StatusDeveloping           -3.605e+05  1.720e+05  -2.096   0.0385 *
Alcohol                    -4.231e+03  2.377e+03  -1.780   0.0779 .
BMI                         6.295e+02  3.754e+02   1.677   0.0965 .
GDP                         1.808e+00  8.057e-01   2.244   0.0269 *
Schooling                   8.515e+03  1.068e+04   0.797   0.4270
StatusDeveloping:Schooling  1.999e+04  1.090e+04   1.833   0.0695 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68990 on 107 degrees of freedom
Multiple R-squared:  0.7024,    Adjusted R-squared:  0.6857
F-statistic: 42.09 on 6 and 107 DF,  p-value: < 2.2e-16
```

In [165…

```
anova(model_Schooling, Final)
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	107	509210947299	NA	NA	NA	NA
2	108	525206011990	-1	-15995064691	3.361027	0.06953631

In [166…

```
# include consideration of interaction between Status and Alcohol
model_Alcohol = lm(Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooli
summary(model_Alcohol)
```

```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + 
    Schooling + Alcohol:Status, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-177836  -53876    1920   47631  189413 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.308e+04  1.301e+05   0.100    0.920    
StatusDeveloping          -1.253e+04  1.223e+05  -0.102    0.919    
Alcohol                   -1.720e+02  1.125e+04  -0.015    0.988    
BMI                        7.793e+02  3.805e+02   2.048    0.043 *  
GDP                        1.155e+00  7.452e-01   1.550    0.124    
Schooling                  2.716e+04  3.440e+03   7.896 2.69e-12 ***
StatusDeveloping:Alcohol -3.457e+03  1.153e+04  -0.300    0.765    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70030 on 107 degrees of freedom
Multiple R-squared:  0.6933,    Adjusted R-squared:  0.6761 
F-statistic: 40.32 on 6 and 107 DF,  p-value: < 2.2e-16
```

In [167…

```
anova(model_Alcohol, Final)
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	107	524764877333	NA	NA	NA	NA
2	108	525206011990	-1	-441134657	0.08994773	0.7648255

In [168…

```
# include consideration of interaction between Status and GDP
model_GDP = lm(Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + Schooling + 
summary(model_GDP)
```

```
Call:
lm(formula = Life.expectancy^3 ~ Status + Alcohol + BMI + GDP + 
    Schooling + GDP:Status, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-177222  -53881    2041   47191  190545 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           5.690e+04  4.821e+04   1.180   0.2405    
StatusDeveloping     -5.517e+04  3.177e+04  -1.737   0.0853 .  
Alcohol              -3.517e+03  2.380e+03  -1.478   0.1424    
BMI                   7.411e+02  3.777e+02   1.962   0.0523 .  
GDP                   9.480e-01  1.000e+00   0.948   0.3453    
Schooling             2.704e+04  3.433e+03   7.877 2.96e-12 ***
StatusDeveloping:GDP  4.725e-01  1.369e+00   0.345   0.7306    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70020 on 107 degrees of freedom
Multiple R-squared:  0.6934,    Adjusted R-squared:  0.6762 
F-statistic: 40.33 on 6 and 107 DF,  p-value: < 2.2e-16
```

In [169…

```
anova(model_GDP, Final)
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	107	524621482403	NA	NA	NA	NA
2	108	525206011990	-1	-584529587	0.1192187	0.7305613

4 Discussion (5 points)

At 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 MSE exploded even reached 1000+, so finally I gave up adding new features. With many features anad 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.