2353 Final Project Life Exp

Team: A Plus
Yi Yang & Weiyi(David) Gong & Xiaolong Wang & Zeming Ren

2023-04-15

Contents

Data Preparation and cleaning

1.Data Cleaning and Descriptive

```
my_data <- read.csv("Life Expectancy Data.csv")
my_data1 <- my_data %>%
  na.omit() %>%
  mutate(Developing = as.integer(Status == "Developing")) # Change status to numeric
my_data1<-my_data1[,-c(1, 2, 3)] # remove country, year, status
pander(summary(my_data1),caption='Descriptive Statistics of The Data')
```

Table 1: Descriptive Statistics of The Data (continued below)

Life.expectancy	Adult.Mortality	infant.deaths	Alcohol
Min. :44.0	Min. : 1.0	Min. : 0.00	Min. : 0.010
1st Qu.:64.4	1st Qu.: 77.0	1st Qu.: 1.00	1st Qu.: 0.810
Median :71.7	Median :148.0	Median : 3.00	Median : 3.790
Mean :69.3	Mean :168.2	Mean : 32.55	Mean : 4.533
3rd Qu.:75.0	3rd Qu.:227.0	3rd Qu.: 22.00	3rd Qu.: 7.340
Max. :89.0	Max. :723.0	Max. :1600.00	Max. :17.870

Table 2: Table continues below

percentage.expenditure	Hepatitis.B	Measles	BMI
Min. : 0.00	Min. : 2.00	Min. : 0	Min. : 2.00
1st Qu.: 37.44	1st Qu.:74.00	1st Qu.: 0	1st Qu.:19.50
Median : 145.10	Median :89.00	Median : 15	Median :43.70
Mean : 698.97	Mean :79.22	Mean : 2224	Mean :38.13
3rd Qu.: 509.39	3rd Qu.:96.00	3rd Qu.: 373	3rd Qu.:55.80
Max. :18961.35	Max. :99.00	Max. :131441	Max. :77.10

Table 3: Table continues below

under.five.deaths	Polio	Total.expenditure	Diphtheria
Min. : 0.00	Min. : 3.00	Min. : 0.740	Min. : 2.00
1st Qu.: 1.00	1st Qu.:81.00	1st Qu.: 4.410	1st Qu.:82.00
Median : 4.00	Median :93.00	Median : 5.840	Median :92.00
Mean : 44.22	Mean :83.56	Mean : 5.956	Mean :84.16
3rd Qu.: 29.00	3rd Qu.:97.00	3rd Qu.: 7.470	3rd Qu.:97.00
Max. :2100.00	Max. :99.00	Max. :14.390	Max. :99.00

Table 4: Table continues below

HIV.AIDS	GDP	Population	thinness..1.19.years
Min. : 0.100	Min. : 1.68	Min. :3.400e+01	Min. : 0.100
1st Qu.: 0.100	1st Qu.: 462.15	1st Qu.:1.919e+05	1st Qu.: 1.600

HIV.AIDS	GDP	Population	thinness..1.19.years
Median : 0.100	Median : 1592.57	Median :1.420e+06	Median : 3.000
Mean : 1.984	Mean : 5566.03	Mean :1.465e+07	Mean : 4.851
3rd Qu.: 0.700	3rd Qu.: 4718.51	3rd Qu.:7.659e+06	3rd Qu.: 7.100
Max. :50.600	Max. :119172.74	Max. :1.294e+09	Max. :27.200

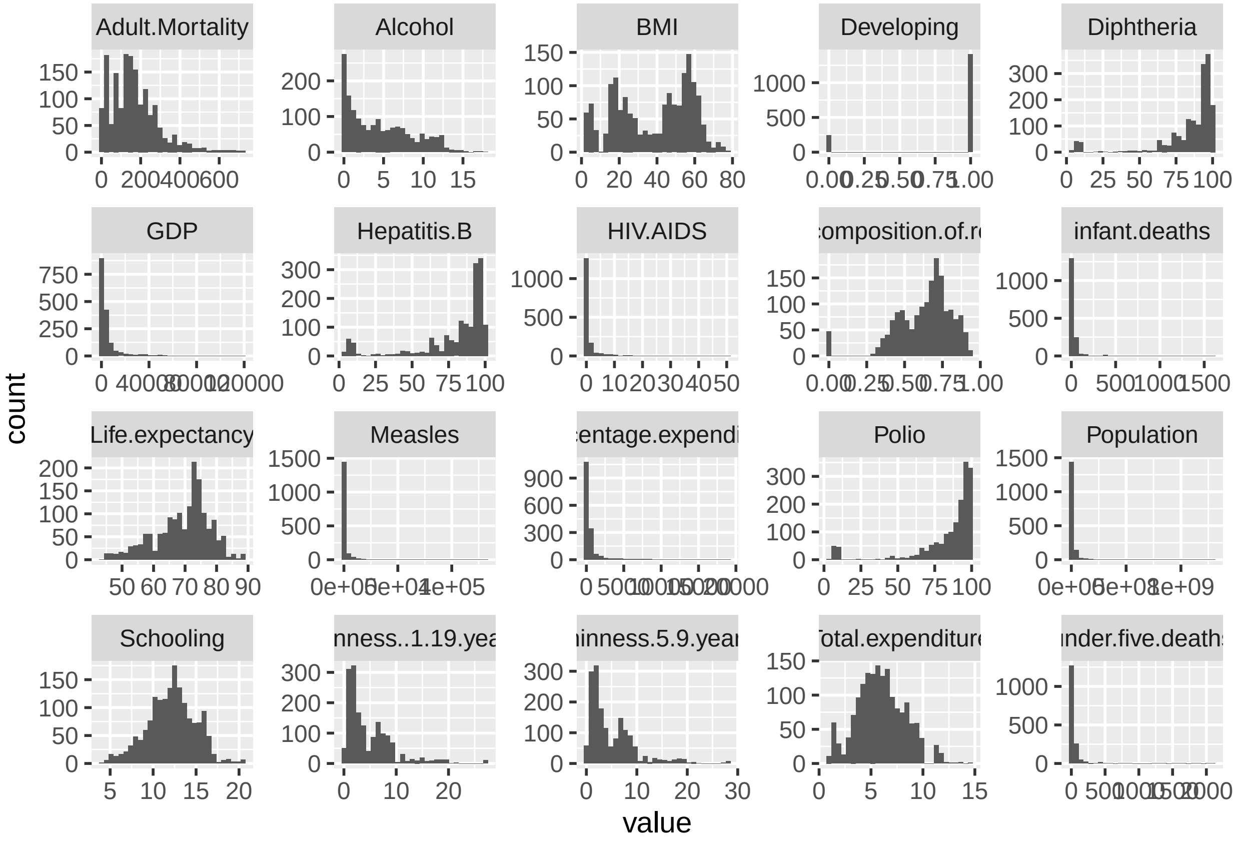
Table 5: Table continues below

thinness.5.9.years	Income.composition.of.resources	Schooling
Min. : 0.100	Min. :0.0000	Min. : 4.20
1st Qu.: 1.700	1st Qu.:0.5090	1st Qu.:10.30
Median : 3.200	Median :0.6730	Median :12.30
Mean : 4.908	Mean :0.6316	Mean :12.12
3rd Qu.: 7.100	3rd Qu.:0.7510	3rd Qu.:14.00
Max. :28.200	Max. :0.9360	Max. :20.70

Developing
Min. :0.0000
1st Qu.:1.0000
Median :1.0000
Mean :0.8532
3rd Qu.:1.0000
Max. :1.0000

```
my_data1 %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

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



```
pander(skewness(my_data1),caption='Skewness of numeric data')
```

Table 7: Table continues below

Life.expectancy	Adult.Mortality	infant.deaths	Alcohol
-0.6282	1.275	8.47	0.6619

Table 8: Table continues below

percentage.expenditure	Hepatitis.B	Measles	BMI	under.five.deaths
4.976	-1.792	7.951	-0.2334	8.333

Table 9: Table continues below

Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP	Population
-2.358	0.2132	-2.485	4.97	4.513	14.17

Table 10: Table continues below

thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources	Schooling	Developing
1.819	1.865	-1.154	-0.128	-1.997

```r
pander(head(my_data1),caption='First six rows of data')
```

Table 12: First six rows of data (continued below)

Life.expectancy	Adult.Mortality	infant.deaths	Alcohol
65	263	62	0.01
59.9	271	64	0.01
59.9	268	66	0.01
59.5	272	69	0.01
59.2	275	71	0.01
58.8	279	74	0.01

Table 13: Table continues below

percentage.expenditure	Hepatitis.B	Measles	BMI	under.five.deaths
71.28	65	1154	19.1	83
73.52	62	492	18.6	86
73.22	64	430	18.1	89
78.18	67	2787	17.6	93
7.097	68	3013	17.2	97
79.68	66	1989	16.7	102

Table 14: Table continues below

Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP	Population
6	8.16	65	0.1	584.3	33736494
58	8.18	62	0.1	612.7	327582
62	8.13	64	0.1	631.7	31731688
67	8.52	67	0.1	670	3696958
68	7.87	68	0.1	63.54	2978599
66	9.2	66	0.1	553.3	2883167

Table 15: Table continues below

thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources
17.2	17.3	0.479

thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources
17.5	17.5	0.476
17.7	17.7	0.47
17.9	18	0.463
18.2	18.2	0.454
18.4	18.4	0.448

Schooling	Developing
10.1	1
10	1
9.9	1
9.8	1
9.5	1
9.2	1

2.Define Training and test dataset

```
set.seed(0)
tr_size <- nrow(my_data1)*0.7 # training sample size
tr_ind <-sample(nrow(my_data1),tr_size)
data_tr <-my_data1[tr_ind, ] # training data
data_te <-my_data1[-tr_ind, ] # test data
ncol(my_data1)
```

```
## [1] 20
```

```
nrow(my_data1)
```

```
## [1] 1649
```

```
nrow(data_tr)
```

```
## [1] 1154
```

```
nrow(data_te)
```

```
## [1] 495
```

Linear model building and statistical diagnosis

Model 1

```
set.seed(0)

model<-lm(Life.expectancy~.,data_tr)
summary(model)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = data_tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.704  -2.164   0.010   2.225  11.494
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.461e+01  1.021e+00  53.507  < 2e-16 ***
## Adult.Mortality                 -1.594e-02  1.137e-03 -14.021  < 2e-16 ***
## infant.deaths                    1.022e-01  1.487e-02   6.875 1.02e-11 ***
## Alcohol                         -1.207e-01  3.887e-02  -3.105  0.00195 **
## percentage.expenditure           3.241e-04  2.169e-04   1.494  0.13533
## Hepatitis.B                     -1.068e-02  5.205e-03  -2.052  0.04038 *
## Measles                         -1.038e-05  1.317e-05  -0.788  0.43066
## BMI                              3.401e-02  7.083e-03   4.802 1.78e-06 ***
## under.five.deaths               -7.607e-02  1.071e-02  -7.102 2.16e-12 ***
## Polio                            1.222e-02  6.156e-03   1.986  0.04731 *
## Total.expenditure                3.363e-02  4.801e-02   0.700  0.48377
## Diphtheria                       1.590e-02  7.184e-03   2.214  0.02705 *
## HIV.AIDS                        -4.386e-01  2.158e-02 -20.325  < 2e-16 ***
## GDP                              9.858e-06  3.440e-05   0.287  0.77450
## Population                      -1.724e-09  2.116e-09  -0.815  0.41542
## thinness..1.19.years            -1.341e-02  5.635e-02  -0.238  0.81193
## thinness.5.9.years              -5.432e-02  5.579e-02  -0.974  0.33043
## Income.composition.of.resources  1.045e+01  1.015e+00  10.293  < 2e-16 ***
## Schooling                        8.510e-01  6.999e-02  12.159  < 2e-16 ***
## Developing                      -1.144e+00  4.059e-01  -2.818  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.576 on 1134 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.8323
## F-statistic: 302.1 on 19 and 1134 DF,  p-value: < 2.2e-16
```

Model 2

```
model_select<-lm(Life.expectancy~ BMI + Adult.Mortality + infant.deaths + under.five.deaths + HIV.AIDS
           + Income.composition.of.resources + Schooling,data_tr)
summary(model_select)
```

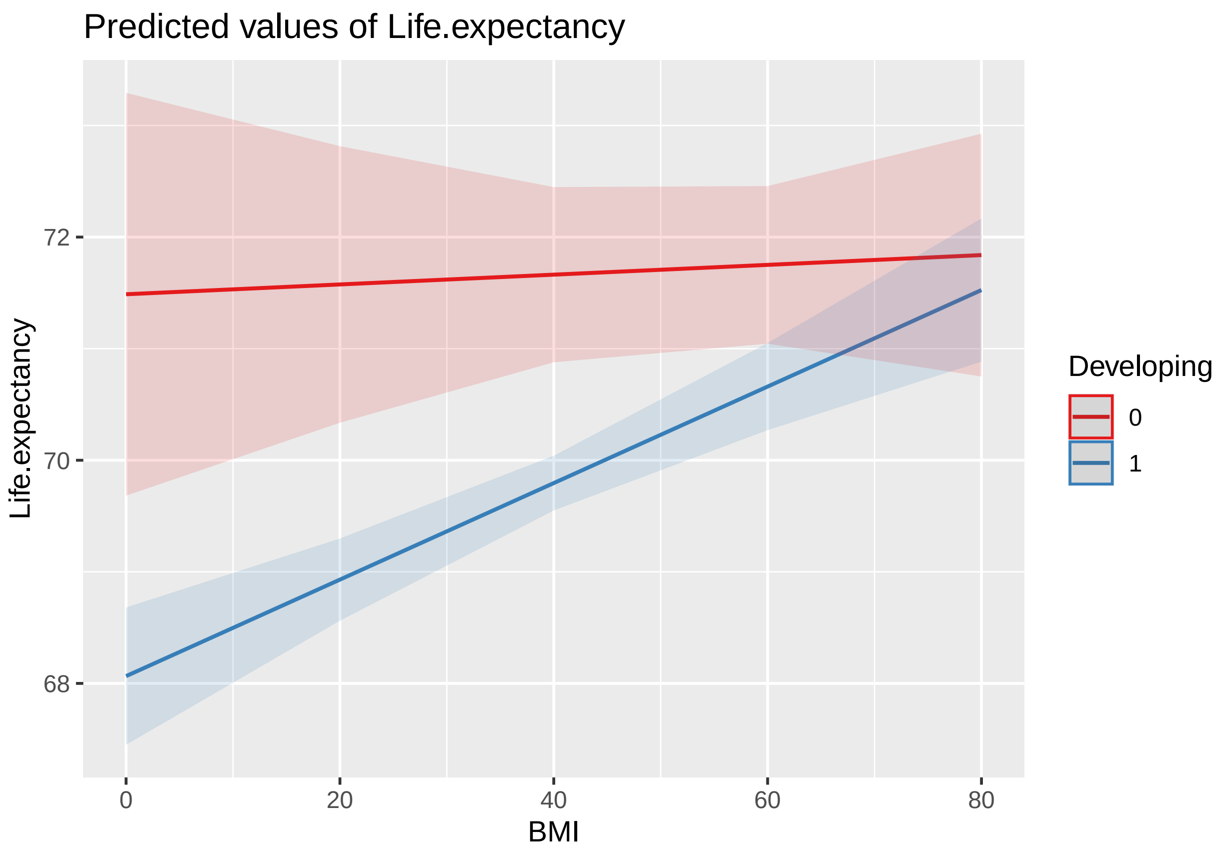
```
##
## Call:
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +
##     under.five.deaths + HIV.AIDS + Income.composition.of.resources +
##     Schooling, data = data_tr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1498  -2.1221  -0.0092   2.1907  11.5776
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     53.225637   0.629231  84.588  < 2e-16 ***
## BMI                              0.036609   0.006752   5.422 7.17e-08 ***
## Adult.Mortality                 -0.017282   0.001148 -15.060  < 2e-16 ***
## infant.deaths                    0.098287   0.013043   7.535 9.85e-14 ***
## under.five.deaths               -0.074952   0.009718  -7.713 2.67e-14 ***
## HIV.AIDS                        -0.439909   0.021793 -20.186  < 2e-16 ***
## Income.composition.of.resources 11.467158   1.002001  11.444  < 2e-16 ***
## Schooling                        0.936796   0.065821  14.233  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.672 on 1146 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8231
## F-statistic: 767.4 on 7 and 1146 DF,  p-value: < 2.2e-16
```

```
model_int<-lm(Life.expectancy~ BMI + Adult.Mortality+ infant.deaths + under.five.deaths + HIV.AIDS
           + Income.composition.of.resources + Schooling + Developing +  BMI : Developing, data_tr)
summary(model_int)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +
##     under.five.deaths + HIV.AIDS + Income.composition.of.resources +
##     Schooling + Developing + BMI:Developing, data = data_tr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1472  -2.1220  -0.0016   2.1714  10.9449
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     57.472249   1.267867  45.330  < 2e-16 ***
## BMI                              0.004386   0.016141   0.272 0.785878
## Adult.Mortality                 -0.016899   0.001142 -14.793  < 2e-16 ***
## infant.deaths                    0.103435   0.013006   7.953 4.36e-15 ***
## under.five.deaths               -0.078702   0.009690  -8.122 1.18e-15 ***
## HIV.AIDS                        -0.440082   0.021637 -20.339  < 2e-16 ***
## Income.composition.of.resources 11.020416   0.999842  11.022  < 2e-16 ***
## Schooling                        0.849761   0.068301  12.441  < 2e-16 ***
## Developing                      -3.422544   1.006038  -3.402 0.000692 ***
## BMI:Developing                   0.038866   0.017764   2.188 0.028877 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.645 on 1144 degrees of freedom
## Multiple R-squared:  0.8271, Adjusted R-squared:  0.8257
## F-statistic:   608 on 9 and 1144 DF,  p-value: < 2.2e-16
```

```
plot_model(model_int, type = "pred", terms = c("BMI", "Developing"))
```



```
anova(model_select, model_int)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +
##     HIV.AIDS + Income.composition.of.resources + Schooling
## Model 2: Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +
##     HIV.AIDS + Income.composition.of.resources + Schooling +
##     Developing + BMI:Developing
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   1146 15456
## 2   1144 15201  2     255.3 9.6068 7.285e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Base on the adjusted R-squared and the P-value of the full model,Using a linear model is appropriate. A multiple regression was used to study whether the effect of the BMI number on Country's Developing levels. Results indicated that both BMI and Country's Developing levels are both associated with the academic performance of the school. The interaction between BMI and Country's Developing levels is significant. Base on the Anova test, We reject the null hypothesis that the interaction is 0.

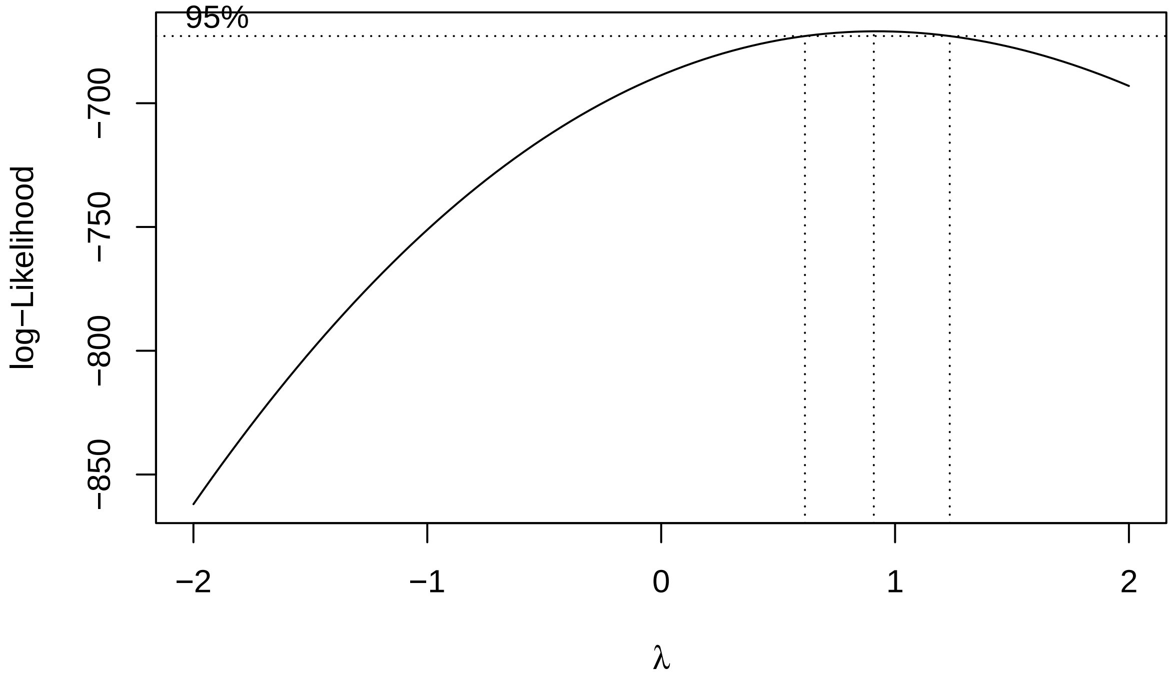
First, we need to do a thorough analysis of the interaction model.

Model Transformation And Adjustment

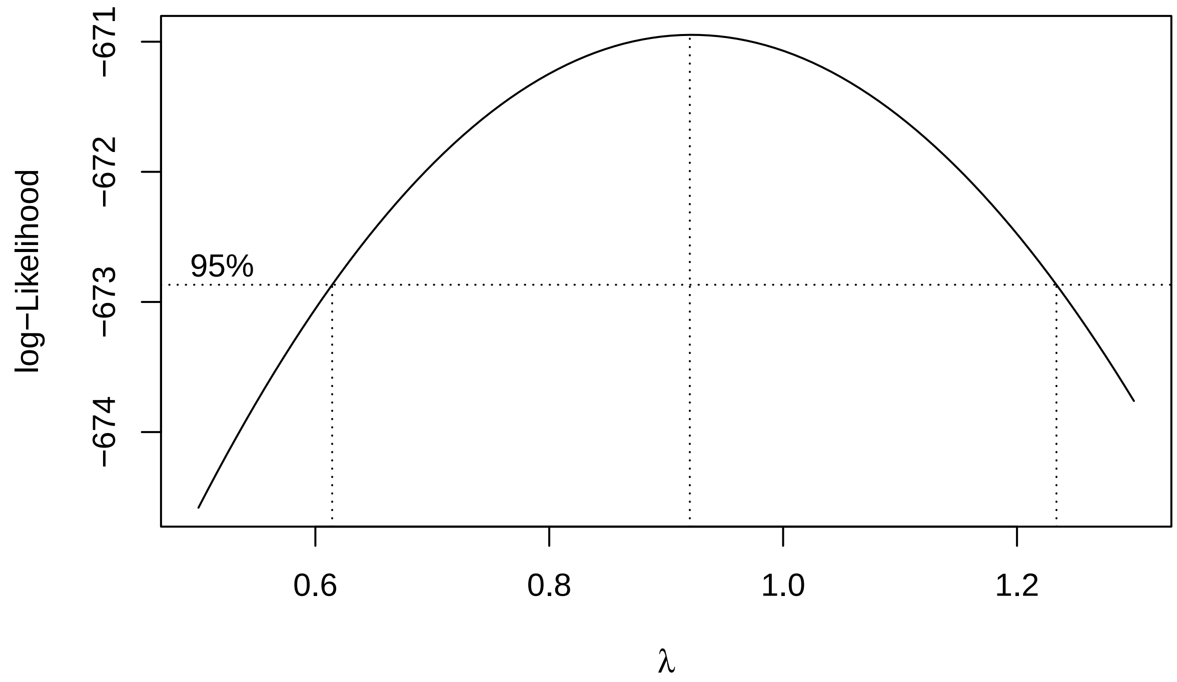
1.Box-Cox Transformation

In the previous section we found that there was a problem with the normality of the residuals of the full model, so we tried to solve it using the BOX-COX transform.

```
boxcox(model_int, plotit=T)
```



```
b<-boxcox(model_int, plotit=T,lambda=seq(0.5,1.3,by=0.01))
```



```
I=which(b$y==max(b$y))
b$x[I]
```

```
## [1] 0.920202
```

```
lmod_trans<-lm(Life.expectancy ^(0.920202) ~ BMI + Adult.Mortality+ infant.deaths + under.five.deaths +
           + Income.composition.of.resources + Schooling + Developing +  BMI : Developing, data = data_
summary(lmod_trans)
```

```
##
## Call:
## lm(formula = Life.expectancy^(0.920202) ~ BMI + Adult.Mortality +
##     infant.deaths + under.five.deaths + HIV.AIDS + Income.composition.of.resources +
##     Schooling + Developing + BMI:Developing, data = data_tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0468 -1.3810 -0.0042  1.4278  7.2789
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     41.6101751  0.8321685  50.002  < 2e-16 ***
## BMI                              0.0028999  0.0105939   0.274 0.784337
## Adult.Mortality                 -0.0111238  0.0007498 -14.836  < 2e-16 ***
## infant.deaths                    0.0687620  0.0085368   8.055 1.99e-15 ***
## under.five.deaths               -0.0523101  0.0063603  -8.224 5.28e-16 ***
## HIV.AIDS                        -0.2935748  0.0142014 -20.672  < 2e-16 ***
## Income.composition.of.resources  7.2360338  0.6562495  11.026  < 2e-16 ***
## Schooling                        0.5576471  0.0448294  12.439  < 2e-16 ***
## Developing                      -2.2208643  0.6603160  -3.363 0.000796 ***
## BMI:Developing                   0.0255797  0.0116594   2.194 0.028442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 1144 degrees of freedom
## Multiple R-squared:  0.8282, Adjusted R-squared:  0.8268
## F-statistic: 612.7 on 9 and 1144 DF,  p-value: < 2.2e-16
```

```
dwtest(lmod_trans)
```

```
##
##  Durbin-Watson test
##
## data:  lmod_trans
## DW = 2.0344, p-value = 0.7216
## alternative hypothesis: true autocorrelation is greater than 0
```

```
shapiro.test(lmod_trans$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmod_trans$residuals
## W = 0.99123, p-value = 2.28e-06
```

```
bptest(lmod_trans)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmod_trans
## BP = 105.14, df = 9, p-value < 2.2e-16
```

Based on the above graph we find that the 95% confidence interval for A contains 1, so we do not see the need to use the BOX-COX transformation.In fact, our model still fails the S-W test after the transformation using the best lambda values, which we believe may be due to problems with the variance of the model residuals.

2.Newey-West Adjustments

The presence of heteroskedasticity affects the fit of the linear model, making t-tests and F-tests no longer valid, so in the presence of heteroskedasticity we use heteroskedasticity robust standard errors instead of standard errors. We use white consistent standard errors for hypothesis testing. We use vcovHC() from the sandwich package for this purpose. Also using the NeweyWest() function allows for heteroskedasticity and autocorrelation robustness Newey-West adjustments.

```
model_nw<-NeweyWest(model_int)
(neweywest<-coeftest(model_int,vcov=NeweyWest(model_int)))
```

```
## 
## t test of coefficients:
## 
##                                   Estimate Std. Error  t value  Pr(>|t|)    
## (Intercept)                     57.4722487  1.2677962  45.3324 < 2.2e-16 ***
## BMI                              0.0043859  0.0166468   0.2635 0.7922382    
## Adult.Mortality                 -0.0168995  0.0014538 -11.6245 < 2.2e-16 ***
## infant.deaths                    0.1034353  0.0156014   6.6299 5.166e-11 ***
## under.five.deaths               -0.0787020  0.0119056  -6.6105 5.861e-11 ***
## HIV.AIDS                        -0.4400817  0.0282664 -15.5691 < 2.2e-16 ***
## Income.composition.of.resources 11.0204162  1.4317357   7.6972 2.993e-14 ***
## Schooling                        0.8497611  0.0842747  10.0832 < 2.2e-16 ***
## Developing                      -3.4225440  1.0002865  -3.4216 0.0006446 ***
## BMI:Developing                   0.0388659  0.0180444   2.1539 0.0314559 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model_int)
```

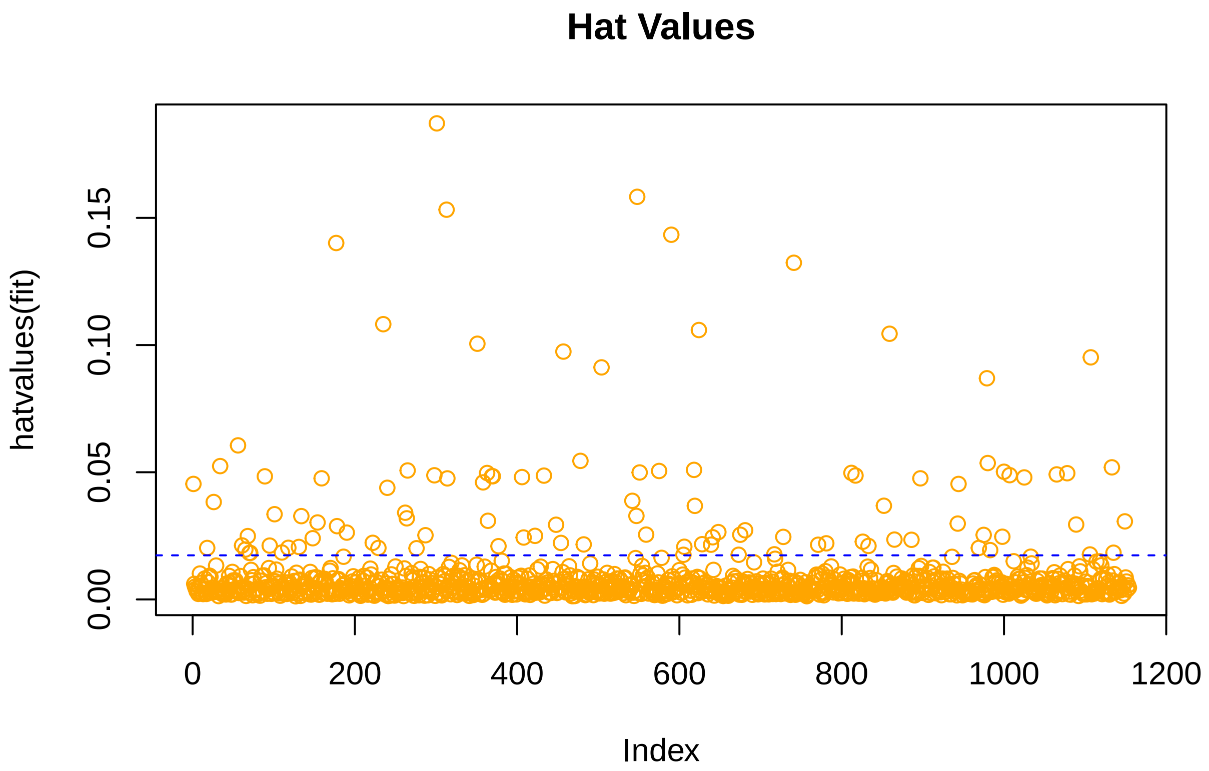
```
## 
## Call:
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + 
##     under.five.deaths + HIV.AIDS + Income.composition.of.resources + 
##     Schooling + Developing + BMI:Developing, data = data_tr)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.1472  -2.1220  -0.0016   2.1714  10.9449 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     57.472249   1.267867  45.330  < 2e-16 ***
## BMI                              0.004386   0.016141   0.272 0.785878    
## Adult.Mortality                 -0.016899   0.001142 -14.793  < 2e-16 ***
## infant.deaths                    0.103435   0.013006   7.953 4.36e-15 ***
## under.five.deaths               -0.078702   0.009690  -8.122 1.18e-15 ***
## HIV.AIDS                        -0.440082   0.021637 -20.339  < 2e-16 ***
## Income.composition.of.resources 11.020416   0.999842  11.022  < 2e-16 ***
## Schooling                        0.849761   0.068301  12.441  < 2e-16 ***
## Developing                      -3.422544   1.006038  -3.402 0.000692 ***
## BMI:Developing                   0.038866   0.017764   2.188 0.028877 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.645 on 1144 degrees of freedom
## Multiple R-squared:  0.8271, Adjusted R-squared:  0.8257 
## F-statistic:   608 on 9 and 1144 DF,  p-value: < 2.2e-16
```

From the summary table we can see that the robustness estimates differ slightly from the initial estimates, with the variables ‘Polio’,‘Diphtheria’ in the initial estimates changing from significant to insignificant, which confirms the above statement. However, since this adjustment has little effect on either the fitted parameters of the model or the results of the y predictor x significance test j, we also do not intend to use

1.Anomaly Detection

leverage Points

```
hat_plot<-function(fit) {
p<-length(coefficients(fit))
n<-length(fitted(fit))
plot(hatvalues(fit),main='Hat Values',col='orange')
abline(h=2*p/n,col='blue',lty=2)
}
hat_plot(model_int)
```



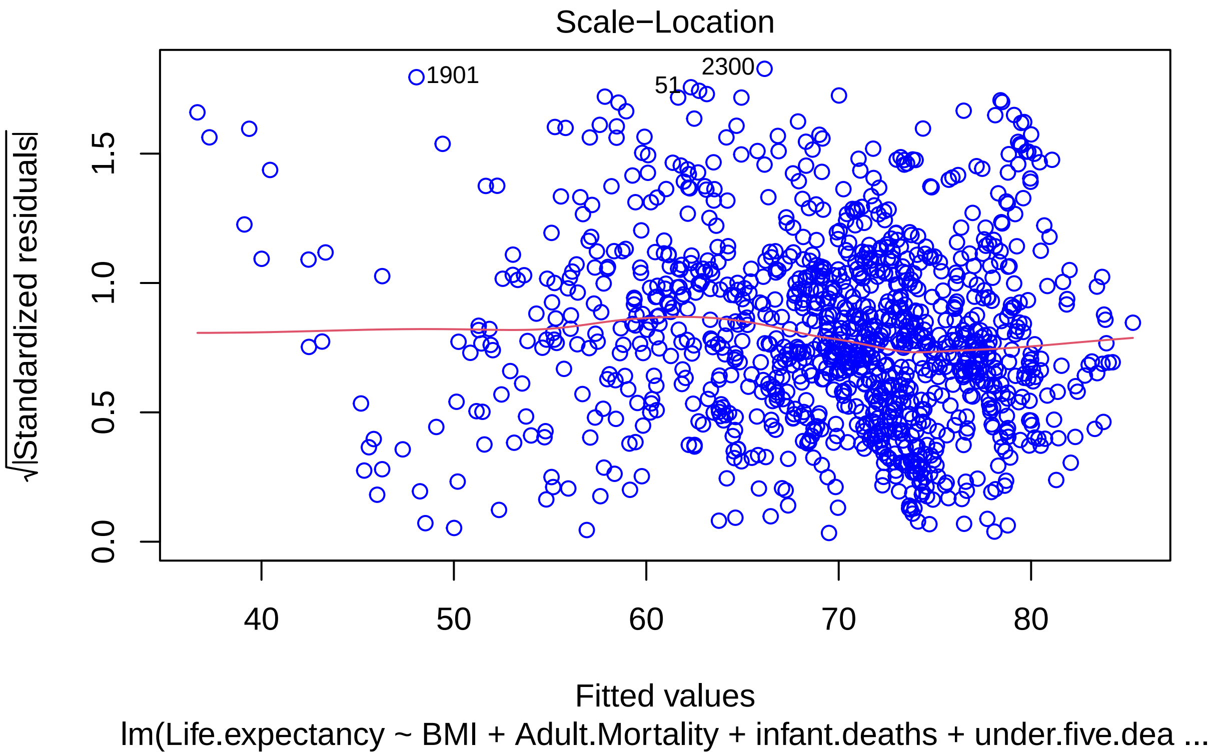
By combining the definition of high leverage points with the diagram above we can see that there are many high leverage points in the model.

Outliers

```
check_outliers(model_int)
```

```
## OK: No outliers detected.
## - Based on the following method and threshold: cook (0.928).
## - For variable: (Whole model)
```

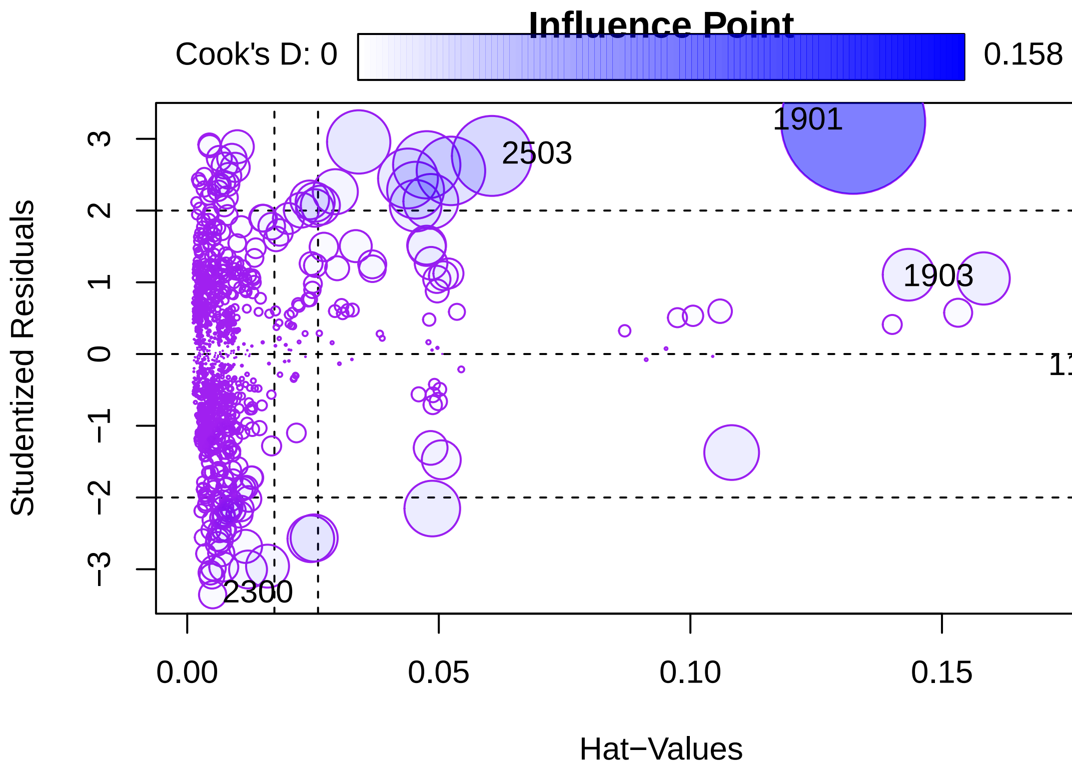
```
plot(model_int,which=3,col='blue')
```



Using the above graph and tests we can obtain that the initial model has no outliers.

Influential Point

```
influencePlot(model_int,id.method='identify',main='Influence Point',col='purple')
```



```
##          StudRes         Hat        CookD
## 2503  2.7618953 0.060551774 0.0488831847
## 2300 -3.3557666 0.005055401 0.0056710323
## 1198 -0.1862181 0.187161837 0.0007991414
## 1903  1.0537401 0.158277587 0.0208773630
## 1901  3.2367411 0.132349710 0.1584935050
```

Some anomalies are given in the above graph, but we found that the 1901st sample with the largest Cook distance has a Cook distance value of about 0.1029, which is less than 0.5, and this data sample is large, so we do not think there are strong influence points that need to be removed from this model.

Analysis of Gaussian-Markov Assumptions

Zero-mean Assumption

```
mean(model_int$residuals)
```

```
## [1] 2.365549e-16
```

Based on the above calculations, the model residuals are very close to 0.

Homoskedasticity Assumption

```
bptest(model_int)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_int
## BP = 102.01, df = 9, p-value < 2.2e-16
```

```
bptest(model_int,studentize=F)
```

```
##
##  Breusch-Pagan test
##
## data:  model_int
## BP = 133.23, df = 9, p-value < 2.2e-16
```

We found that although the p-values did not differ they were all less than 0.05, indicating that there was strong heteroskedasticity in the model. However, the BP values with studentisation removed increased, suggesting that studentisation played a role in correcting for heteroskedasticity, but not significantly in this case.

Normality Assumption

```
shapiro.test(model_int$residuals)
```

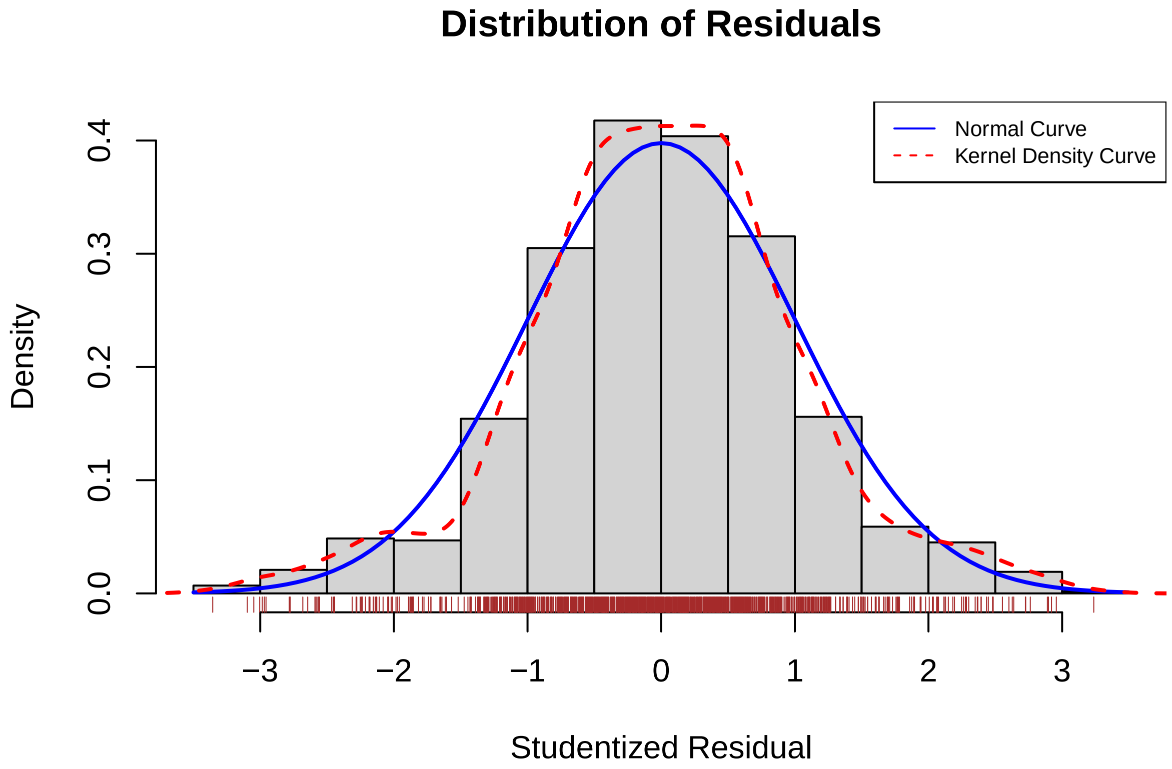
```
##
##  Shapiro-Wilk normality test
##
## data:  model_int$residuals
## W = 0.99143, p-value = 2.99e-06
```

```
dev.new()
qqPlot(model_int,labels=row.names(df),id.method='identify',simulate=T,main='Q-Q Plot')
```

```
## 2300 1901
##  135  741
```

In the Q-Q plot above, the blue shaded area is the 95% confidence interval and the two outlier sample points that were detected, for the 1901st and 2300th samples.

```
residplot<-function(model_int,nbreaks=10){
z<-rstudent(model_int)
hist(z,breaks=nbreaks,freq=F,
xlab='Studentized Residual',
main='Distribution of Residuals')
rug(jitter(z),col='brown')
curve(dnorm(x,mean=mean(z),sd=sd(z)),add=T,col='blue',lwd=2)
lines(density(z)$x,density(z)$y,col="red",lwd=2,lty=2)
legend('topright',legend=c('Normal Curve','Kernel Density Curve'),
lty=1:2,col=c('blue','red'),cex=.7)
}
residplot(model_int)
```



We can see from the residual distribution graph that the model residuals are almost completely unbiased. This is one of the reasons why subsequently when we used the BOX-COX variation to calculate the lambda we found that its confidence interval contained 1, i.e. the BOX-COX transformation was not necessary. In addition to this the problem of heteroskedasticity can also have an impact on the effectiveness of the BOX-COX transformation.

From the graphs above and the results of the tests we can conclude that the initial model residuals do not obey normality, but rather suffer from some heavy tails.

Linearity Assumption

We would have liked to use a deviation residual plot for this test, but the model has too many predictors and a large sample, and the RMD does not have enough computing power to give results. At the end of this section, we will use check_model() to find out about linearity.

Randomness Assumption

```
dwtest(model_int)
```

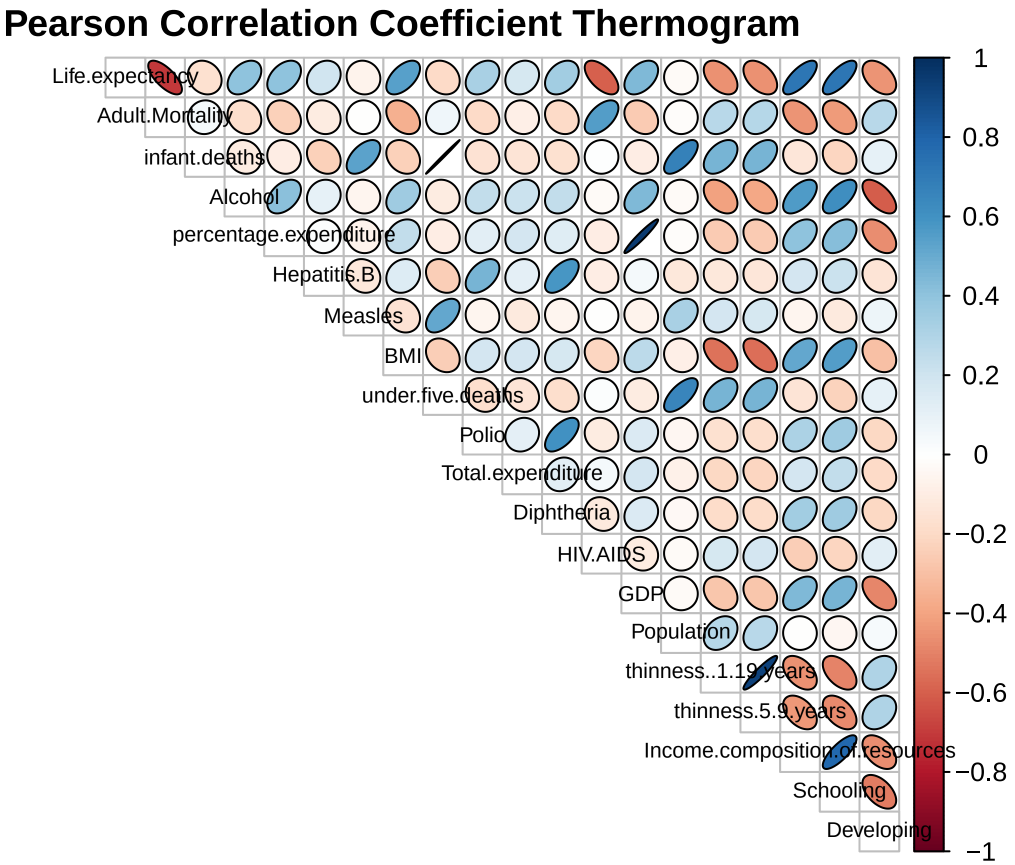
```
##
##  Durbin-Watson test
##
## data:  model_int
## DW = 2.033, p-value = 0.7136
## alternative hypothesis: true autocorrelation is greater than 0
```

From the p-values of the above results we can conclude that there is no first order autocorrelation problem with the model.

No Multicollinearity Assumption

First we can take a cursory look at the two-by-two correlation between the variables using a thermogram of the Pearson correlation coefficient matrix.

```
M=cor(my_data1)
corrplot(M,method='ellipse',type='upper',tl.col='black',tl.pos='d',tl.cex=0.7,show.legend=T,outline=T,t
```



However, Pearson's correlation coefficient can only show the correlation between two variables. In practical problems there may be problems with correlations between more than one variable, so for a further and clearer view we introduce the variance inflation factor.

```
alias(model_int)
```

```
## Model :
## Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +
##     HIV.AIDS + Income.composition.of.resources + Schooling +
##     Developing + BMI:Developing
```

The above checks revealed that none of the predictors in the data had a large number of identical data, leading to problems where parameters could not be fitted or vif could not be calculated.

```
pander(vif(model_int),caption='Vif of Full Model')
```

Table 17: Table continues below

BMI	Adult.Mortality	infant.deaths	under.five.deaths	HIV.AIDS
8.859	1.769	187.3	187.9	1.452

Income.composition.of.resources	Schooling	Developing	BMI:Developing
2.868	3.305	11.04	12.87

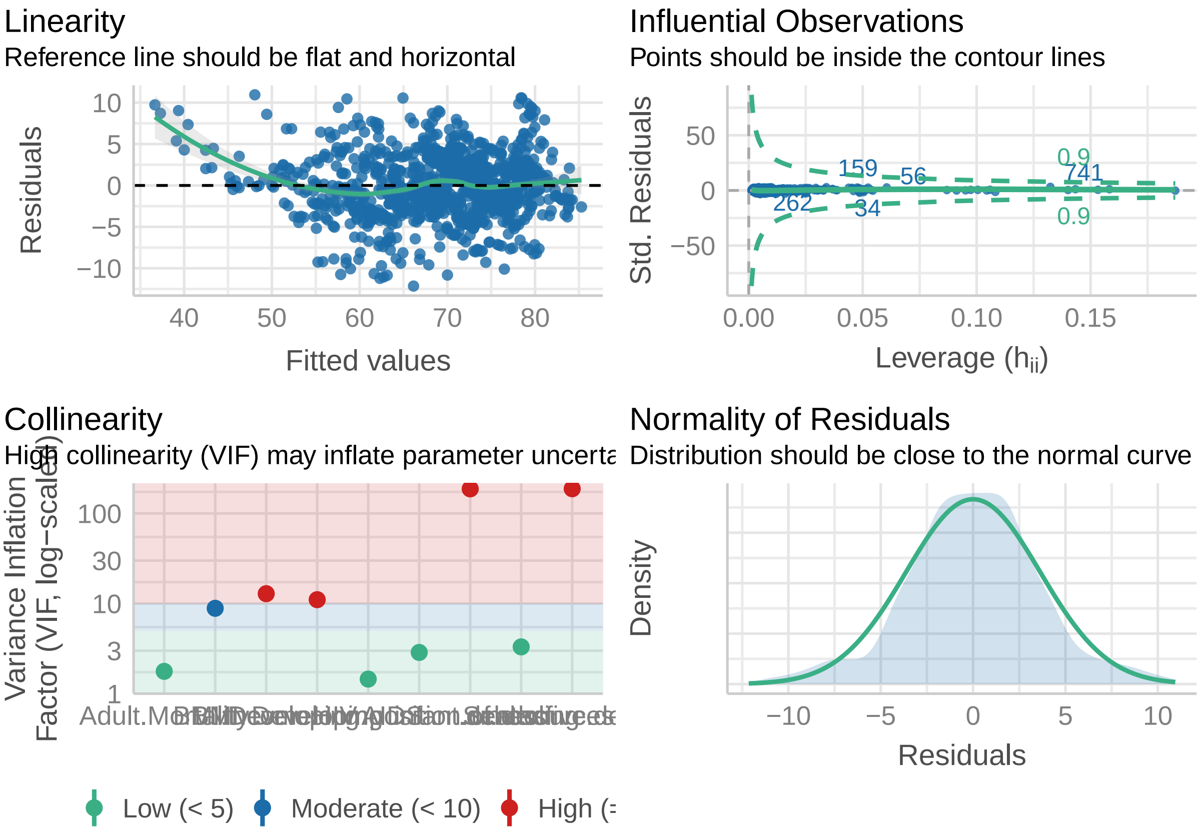
Using the above graphs we find that several predictors of 'infant.deaths', 'per centage.expenditure', 'under.five.deaths', 'GDP' have VIFs greater than 10 and their presence leads to serious multicollinearity problems.

3.Model overview

Finally, let's look at the statistical diagnosis of the full model as a whole.

```
check_model(model_int,verbose=T,check=c('outliers','vif','normality','linearity'))
```

```
## Model has interaction terms. VIFs might be inflated.
##   You may check multicollinearity among predictors of a model without
##   interaction terms.
```



In this section we find that the pass test results for the full model, although fair overall, suffer mainly from multicollinearity, heteroskedasticity and non-normality. In the next section we will try to address these problems using the methods we have learned.

Model Variables Selection

1.AIC Selection

```
step(model_int)
```

```
## Start:  AIC=2995.16
## Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +
##     HIV.AIDS + Income.composition.of.resources + Schooling +
##     Developing + BMI:Developing
##
##                                   Df Sum of Sq   RSS    AIC
## <none>                                         15201 2995.2
## - BMI:Developing                   1      63.6 15265 2998.0
## - infant.deaths                    1     840.4 16041 3055.3
## - under.five.deaths                1     876.5 16078 3057.9
## - Income.composition.of.resources  1    1614.3 16815 3109.6
## - Schooling                        1    2056.8 17258 3139.6
## - Adult.Mortality                  1    2907.8 18109 3195.2
## - HIV.AIDS                         1    5497.0 20698 3349.4
```

```
## 
## Call:
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + 
##     under.five.deaths + HIV.AIDS + Income.composition.of.resources + 
##     Schooling + Developing + BMI:Developing, data = data_tr)
## 
## Coefficients:
##                     (Intercept)                              BMI  
##                       57.472249                         0.004386  
##                 Adult.Mortality                    infant.deaths  
##                       -0.016899                         0.103435  
##               under.five.deaths                         HIV.AIDS  
##                       -0.078702                        -0.440082  
## Income.composition.of.resources                        Schooling  
##                       11.020416                         0.849761  
##                      Developing                   BMI:Developing  
##                       -3.422544                         0.038866
```

```
# Find model with lowest AIC
lmod_AIC_B<-lm(Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +
    HIV.AIDS + Income.composition.of.resources + Schooling +
    Developing + BMI:Developing, data = data_tr) # AIC selected model
sum_AIC_B<-summary(lmod_AIC_B)
sum_AIC_B
```

```
## 
## Call:
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + 
##     under.five.deaths + HIV.AIDS + Income.composition.of.resources + 
##     Schooling + Developing + BMI:Developing, data = data_tr)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.1472  -2.1220  -0.0016   2.1714  10.9449 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     57.472249   1.267867  45.330  < 2e-16 ***
## BMI                              0.004386   0.016141   0.272 0.785878    
## Adult.Mortality                 -0.016899   0.001142 -14.793  < 2e-16 ***
## infant.deaths                    0.103435   0.013006   7.953 4.36e-15 ***
## under.five.deaths               -0.078702   0.009690  -8.122 1.18e-15 ***
## HIV.AIDS                        -0.440082   0.021637 -20.339  < 2e-16 ***
## Income.composition.of.resources 11.020416   0.999842  11.022  < 2e-16 ***
## Schooling                        0.849761   0.068301  12.441  < 2e-16 ***
## Developing                      -3.422544   1.006038  -3.402 0.000692 ***
## BMI:Developing                   0.038866   0.017764   2.188 0.028877 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.645 on 1144 degrees of freedom
## Multiple R-squared:  0.8271, Adjusted R-squared:  0.8257 
## F-statistic:   608 on 9 and 1144 DF,  p-value: < 2.2e-16
```

From the summary we can find that the model selected by the backward iterative AIC method, most certainly all predictors are statistically significant, but the adjusted R-squared does not change much compared to the full model, and we will subsequently judge whether this model should be used by the model's prediction error perspective

2.BIC Selection

```
set.seed(0)
fit_null<-lm(Life.expectancy~1,data_tr)
step(fit_null, scope = list(lower = fit_null, upper = model_int), direction = "both",
criterion = "BIC")
```

```
## Start:  AIC=5002.38
## Life.expectancy ~ 1
## 
##                                    Df Sum of Sq   RSS    AIC
## + Schooling                         1     46348 41563 4139.9
## + Income.composition.of.resources  1     46324 41586 4140.6
## + Adult.Mortality                   1     42969 44941 4230.1
## + HIV.AIDS                          1     30508 57403 4512.5
## + BMI                               1     24795 63116 4622.0
## + Developing                        1     17741 70169 4744.3
## + under.five.deaths                 1      3146 84765 4962.3
## + infant.deaths                     1      2460 85450 4971.6
## <none>                                          87911 5002.4
## 
## Step:  AIC=4139.91
## Life.expectancy ~ Schooling
## 
##                                    Df Sum of Sq   RSS    AIC
## + HIV.AIDS                          1     17966 23597 3488.7
## + Adult.Mortality                   1     16736 24827 3547.3
## + Income.composition.of.resources  1      5400 36164 3981.3
## + BMI                               1      2063 39500 4083.2
## + Developing                        1       555 41008 4126.4
## <none>                                          41563 4139.9
## + under.five.deaths                 1        33 41530 4141.0
## + infant.deaths                     1         2 41561 4141.9
## - Schooling                         1     46348 87911 5002.4
## 
## Step:  AIC=3488.65
## Life.expectancy ~ Schooling + HIV.AIDS
## 
##                                    Df Sum of Sq   RSS    AIC
## + Adult.Mortality                   1      4770 18827 3230.1
## + Income.composition.of.resources  1      3110 20487 3327.5
## + BMI                               1       984 22613 3441.5
## + Developing                        1       395 23202 3471.2
## + under.five.deaths                 1        98 23499 3485.8
## + infant.deaths                    1        47 23550 3488.3
## <none>                                         23597 3488.7
## - HIV.AIDS                         1     17966 41563 4139.9
## - Schooling                        1     33805 57403 4512.5
##
## Step:  AIC=3230.06
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality
##
##                                   Df Sum of Sq   RSS    AIC
## + Income.composition.of.resources  1    2088.5 16739 3096.4
## + BMI                              1     623.1 18204 3193.2
## + Developing                       1     231.8 18596 3217.8
## + under.five.deaths                1     105.0 18722 3225.6
## + infant.deaths                    1      63.3 18764 3228.2
## <none>                                         18827 3230.1
## - Adult.Mortality                  1    4769.9 23597 3488.7
## - HIV.AIDS                         1    5999.7 24827 3547.3
## - Schooling                        1   20937.3 39765 4090.9
##
## Step:  AIC=3096.37
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources
##
##                                   Df Sum of Sq   RSS    AIC
## + BMI                              1     425.0 16314 3068.7
## + under.five.deaths                1     157.3 16582 3087.5
## + Developing                       1     124.6 16614 3089.7
## + infant.deaths                    1     109.6 16629 3090.8
## <none>                                         16739 3096.4
## - Income.composition.of.resources  1    2088.5 18827 3230.1
## - Adult.Mortality                  1    3748.0 20487 3327.5
## - Schooling                        1    3916.8 20656 3337.0
## - HIV.AIDS                         1    5740.0 22479 3434.6
##
## Step:  AIC=3068.7
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     BMI
##
##                                   Df Sum of Sq   RSS    AIC
## + Developing                       1     122.2 16192 3062.0
## + under.five.deaths                1      91.7 16222 3064.2
## + infant.deaths                    1      55.3 16259 3066.8
## <none>                                         16314 3068.7
## - BMI                              1     425.0 16739 3096.4
## - Income.composition.of.resources  1    1890.3 18204 3193.2
## - Schooling                        1    2911.2 19225 3256.2
## - Adult.Mortality                  1    3528.9 19843 3292.7
## - HIV.AIDS                         1    5614.7 21929 3408.0
##
## Step:  AIC=3062.02
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     BMI + Developing
##
##                                   Df Sum of Sq   RSS    AIC
## + under.five.deaths                1      94.3 16097 3057.3
## + BMI:Developing                    1      65.9 16126 3059.3
## + infant.deaths                     1      56.1 16136 3060.0
## <none>                                         16192 3062.0
## - Developing                        1     122.2 16314 3068.7
## - BMI                               1     422.5 16614 3089.7
## - Income.composition.of.resources  1    1791.3 17983 3181.1
## - Schooling                         1    2408.6 18600 3220.1
## - Adult.Mortality                   1    3449.9 19642 3282.9
## - HIV.AIDS                          1    5639.6 21831 3404.9
##
## Step:  AIC=3057.28
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     BMI + Developing + under.five.deaths
##
##                                   Df Sum of Sq   RSS    AIC
## + infant.deaths                    1     832.8 15265 2998.0
## + BMI:Developing                   1      56.0 16041 3055.3
## <none>                                         16097 3057.3
## - under.five.deaths                1      94.3 16192 3062.0
## - Developing                       1     124.8 16222 3064.2
## - BMI                              1     356.4 16454 3080.6
## - Income.composition.of.resources  1    1836.5 17934 3180.0
## - Schooling                        1    2237.1 18335 3205.4
## - Adult.Mortality                  1    3459.3 19557 3279.9
## - HIV.AIDS                         1    5676.0 21773 3403.8
##
## Step:  AIC=2997.98
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     BMI + Developing + under.five.deaths + infant.deaths
##
##                                   Df Sum of Sq   RSS    AIC
## + BMI:Developing                   1      63.6 15201 2995.2
## <none>                                         15265 2998.0
## - Developing                       1     191.7 15456 3010.4
## - BMI                              1     394.4 15659 3025.4
## - infant.deaths                    1     832.8 16097 3057.3
## - under.five.deaths                1     871.0 16136 3060.0
## - Income.composition.of.resources  1    1642.9 16908 3113.9
## - Schooling                        1    2182.4 17447 3150.2
## - Adult.Mortality                  1    2947.2 18212 3199.7
## - HIV.AIDS                         1    5521.1 20786 3352.2
##
## Step:  AIC=2995.16
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     BMI + Developing + under.five.deaths + infant.deaths + BMI:Developing
##
##                                   Df Sum of Sq   RSS    AIC
## <none>                                         15201 2995.2
## - BMI:Developing                   1      63.6 15265 2998.0
## - infant.deaths                    1     840.4 16041 3055.3
## - under.five.deaths                1     876.5 16078 3057.9
## - Income.composition.of.resources  1    1614.3 16815 3109.6
## - Schooling                        1    2056.8 17258 3139.6
## - Adult.Mortality                  1    2907.8 18109 3195.2
## - HIV.AIDS                        1    5497.0 20698 3349.4
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##     Income.composition.of.resources + BMI + Developing + under.five.deaths +
##     infant.deaths + BMI:Developing, data = data_tr)
##
## Coefficients:
##                     (Intercept)                        Schooling
##                       57.472249                         0.849761
##                        HIV.AIDS                  Adult.Mortality
##                       -0.440082                        -0.016899
## Income.composition.of.resources                              BMI
##                       11.020416                         0.004386
##                      Developing                under.five.deaths
##                       -3.422544                        -0.078702
##                   infant.deaths                   BMI:Developing
##                        0.103435                         0.038866
```

```
lmod_BIC_BO<-lm(Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
                  Income.composition.of.resources +
                  percentage.expenditure + BMI + Diphtheria + Alcohol,data_tr) # BIC selected model
sum_BIC_BO<-summary(lmod_BIC_BO)
sum_BIC_BO
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##     Income.composition.of.resources + percentage.expenditure +
##     BMI + Diphtheria + Alcohol, data = data_tr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7365  -2.1572   0.0879   2.3716  12.3880
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     51.6253207  0.7516596  68.682  < 2e-16 ***
## Schooling                        0.9161955  0.0700758  13.074  < 2e-16 ***
## HIV.AIDS                        -0.4441776  0.0218750 -20.305  < 2e-16 ***
## Adult.Mortality                 -0.0174209  0.0011497 -15.153  < 2e-16 ***
## Income.composition.of.resources 11.1794749  1.0273773  10.882  < 2e-16 ***
## percentage.expenditure           0.0004363  0.0000669   6.521 1.05e-10 ***
## BMI                              0.0401836  0.0066937   6.003 2.59e-09 ***
## Diphtheria                       0.0240417  0.0055439   4.337 1.57e-05 ***
## Alcohol                         -0.1124538  0.0359347  -3.129   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.676 on 1145 degrees of freedom
## Multiple R-squared:  0.824,  Adjusted R-squared:  0.8228
## F-statistic: 670.1 on 8 and 1145 DF,  p-value: < 2.2e-16
```

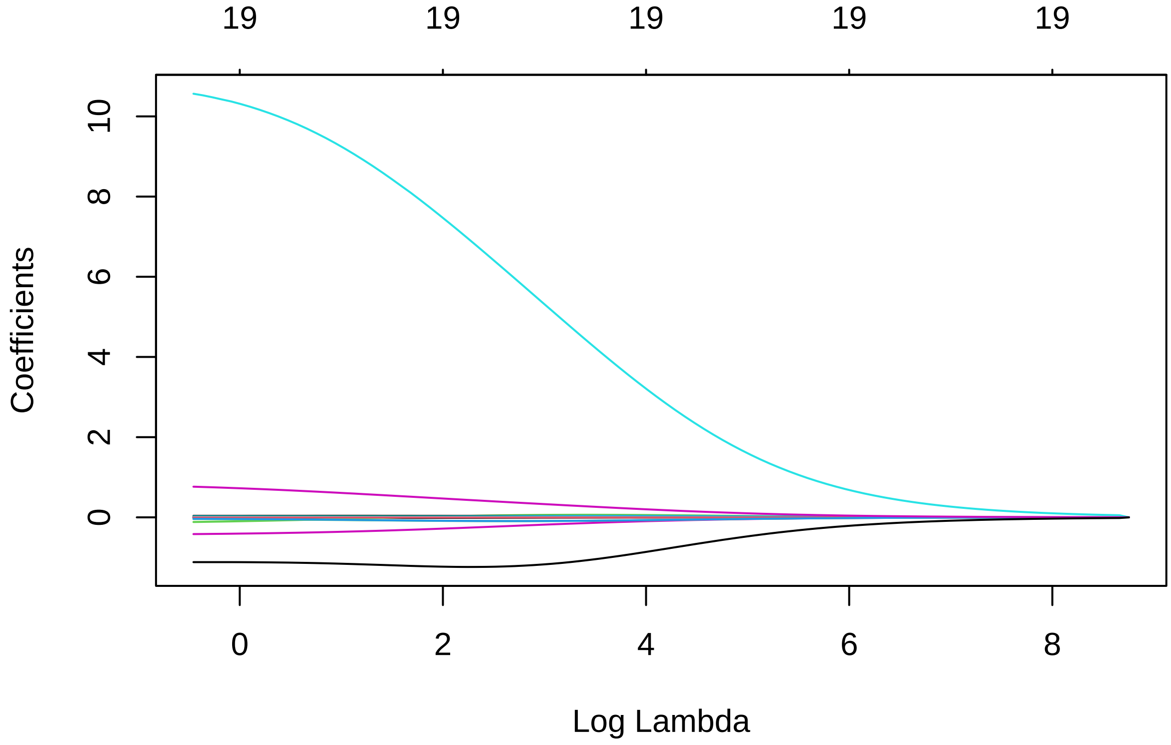
From the summary we can find that the model chosen by the backward iterative BIC method, most certainly all predictors are statistically significant, but the adjusted R-squared is even lower than the full model, and subsequently we will judge whether this model should be used by the model's prediction error perspective.

3.Selection ideas for other model selection methods

In the statistical diagnosis section of the model, we find that the full model suffers from multicollinearity and heteroskedasticity. GLS estimation is usually used when the m-model error term does not satisfy the "spherical perturbation assumption" (i.e. homoskedasticity assumption and no autocorrelation assumption in the G-M assumption). Ridge regression, lasso regression and adaptive lasso regression are all methods of constraining the fitted parameters by adding penalty factors. We will try each of these below.

4.Ridge Selection

```
set.seed(0)
x_tr<-as.matrix(data_tr[,c(2:ncol(data_tr))])
y_tr<-as.matrix(data_tr[,1])
x_te<-as.matrix(data_te[,c(2:ncol(data_te))])
y_te<-as.matrix(data_te[,1])
set.seed(0)
ridge<-glmnet(x=x_tr,y=y_tr,alpha=0)
plot(ridge,xvar='lambda')
```



```
ridge_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=10,alpha=0)
```

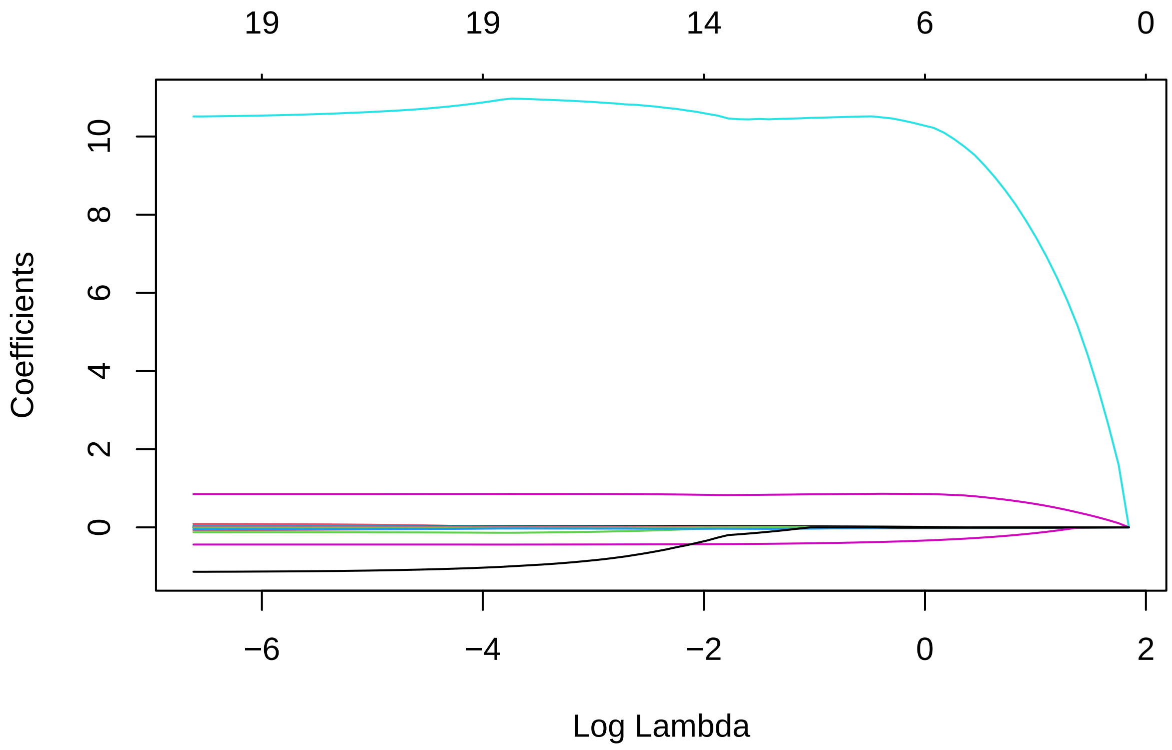
```
ridge_cv$lambda.min
```

```
## [1] 0.6337393
```

```
best_ridge<-coef(ridge_cv, s = ridge_cv$lambda.min)
```

5.Lasso Selection

```
set.seed(0)
lasso<-glmnet(x=x_tr,y=y_tr,alpha=1)
plot(lasso,xvar='lambda')
```

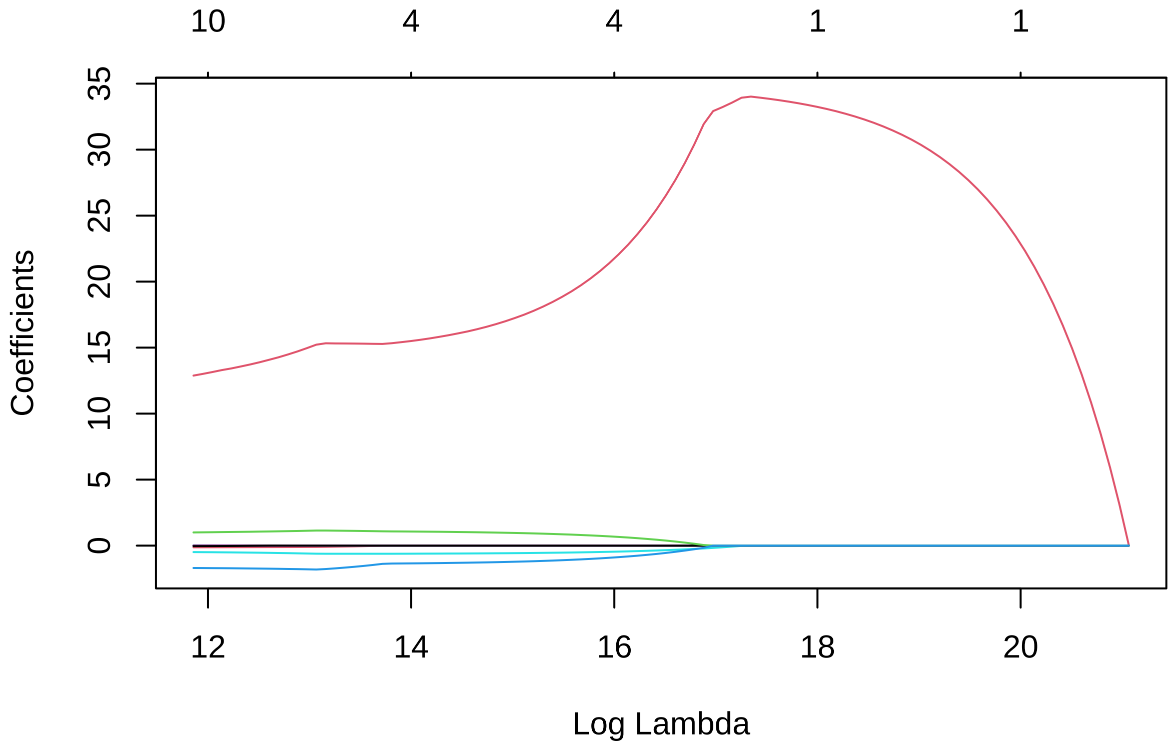


```
lasso_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=10,alpha=1,keep=T)
lasso_cv$lambda.min
```

```
## [1] 0.003081622
```

6.Adaptive Lasso Selection

```
set.seed(0)
alasso<-glmnet(x=x_tr,y=y_tr,alpha=1,penalty.factor=1/abs(best_ridge[-1]))
plot(alasso,xvar='lambda')
```



```
alasso_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=10,alpha=1,penalty.factor=1/abs(best_ridge[
```

```
alasso_cv$lambda.min
```

```
## [1] 140906.5
```

7.Error Comparison And Confirmation of Final Model

Next we will calculate the prediction error of each model in the training set:

```
result_full<-predict(model_int,newdata=data_te,interval='prediction')
(err_full<-mean((data_te$Life.expectancy-result_full)^2))
```

```
## [1] 48.02168
```

```
result_aic<-predict(lmod_AIC_B,newdata=data_te,interval='prediction')
(err_aic<-mean((data_te$Life.expectancy-result_aic)^2))
```

```
## [1] 48.02168
```

```
result_bic<-predict(lmod_BIC_B0,newdata = data_te,interval='prediction')
(err_bic<-mean((data_te$Life.expectancy-result_bic)^2))
```

```
## [1] 49.10929
```

```
result_ridge<-predict(ridge_cv,newx=x_te,interval='prediction')
(err_ridge<-mean((y_te-result_ridge)^2))
```

```
## [1] 14.77299
```

```
result_la<-predict(lasso_cv,newx=x_te,interval='prediction')
(err_la<-mean((y_te-result_la)^2))
```

```
## [1] 13.90835
```

```
result_adala<-predict(alasso_cv,newx=x_te,interval='prediction')
(err_adala<-mean((y_te-result_adala)^2))
```

```
## [1] 15.92209
```

```
which.min(c(err_full, err_aic, err_bic, err_ridge, err_la, err_adala))
```

```
## [1] 5
```

From the above results we can see that the model selected using the 10-fold lasso method has the smallest test error and a significant reduction compared to the original model, so we will finally choose this model.

```
(best_alasso_coef<-coef(alasso_cv,s=alasso_cv$lambda.min))
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                                           s1
## (Intercept)                     53.038475594
## Adult.Mortality                 -0.012413848
## infant.deaths                     .
## Alcohol                         -0.127001514
## percentage.expenditure            .
## Hepatitis.B                       .
## Measles                           .
## BMI                              0.024103630
## under.five.deaths                 .
## Polio                             .
## Total.expenditure                 .
## Diphtheria                       0.005072368
## HIV.AIDS                        -0.488547396
## GDP                               .
## Population                        .
## thinness..1.19.years            -0.003011737
## thinness.5.9.years              -0.038537048
## Income.composition.of.resources 12.882346015
## Schooling                        0.999554558
## Developing                      -1.695027771
```

So our final model will be:

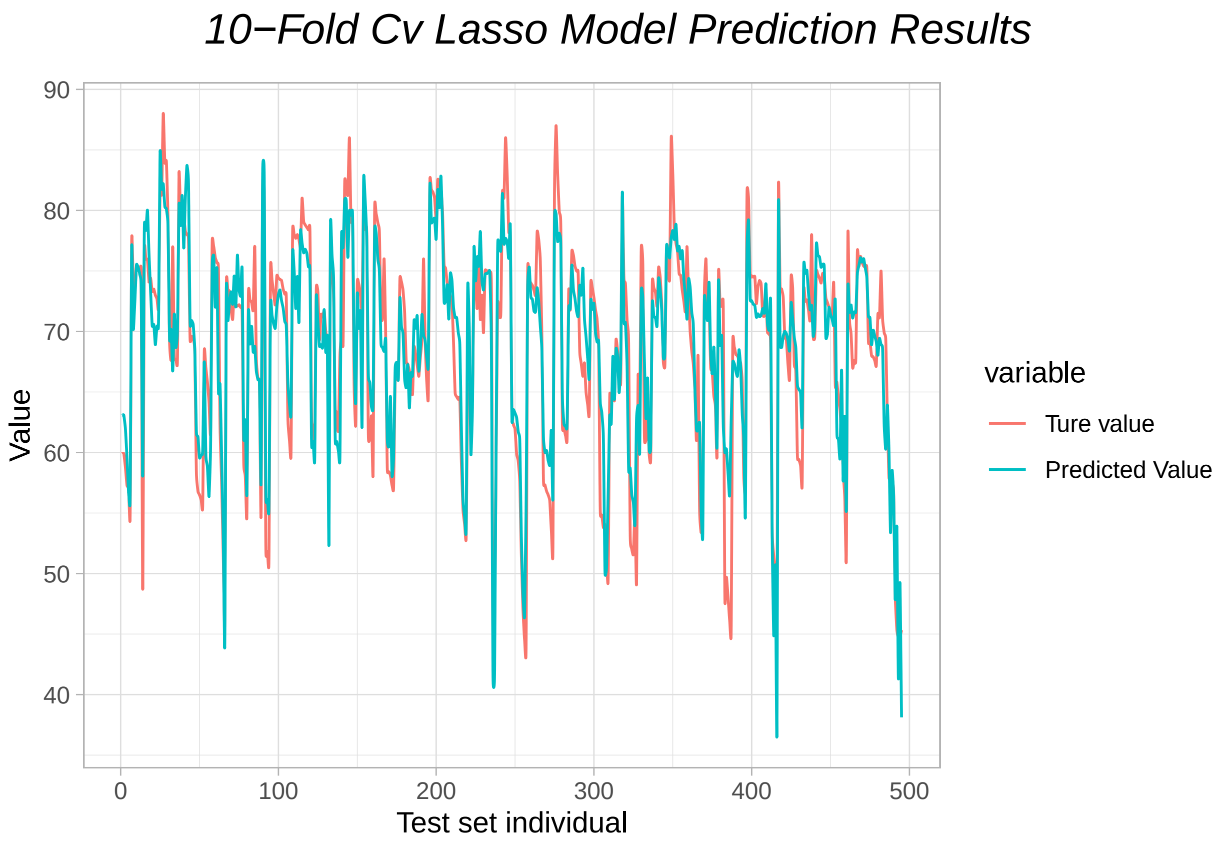
$$Life.expectancy = 53.038475594 - 0.012413848 * Adult.Mortality - 0.127001514 * Alcohol + 0.024103630 * BMI + 0.005072368 * L$$

Model prediction

In this section we will use our selected 10-fold lasso model to make predictions and compare them with the true values, by way of icons to see the predictions.

```
x_gr<-1:495
y_pred<-predict(lasso_cv,x_te)
df<-data.frame(x_gr,y_te,y_pred)
names(df)<-c('x_gr','Ture value','Predicted Value')
df_long<-melt(df,id.vars='x_gr')
P<-ggplot(df_long,aes(x_gr,value,col=variable))+
geom_xspline()+labs(x='Test set individual',y='Value')+theme_light()
grid.arrange(textGrob('10-Fold Cv Lasso Model Prediction Results',
             gp=gpar(fontsize =2*8, fontface ='italic')),
             P,
             heights=c(0.1,1))
```

```
## Warning: Using the `size` aesthetic in this geom was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` in the `default_aes` field and elsewhere instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



As we can see from the graph above, the predicted values are close to the true values, which means that the model is successful in its predictions.