2353 Final Project Life Exp

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

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

Table 7: 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 8: 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 9: 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 10: Table continues below

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

```
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
```

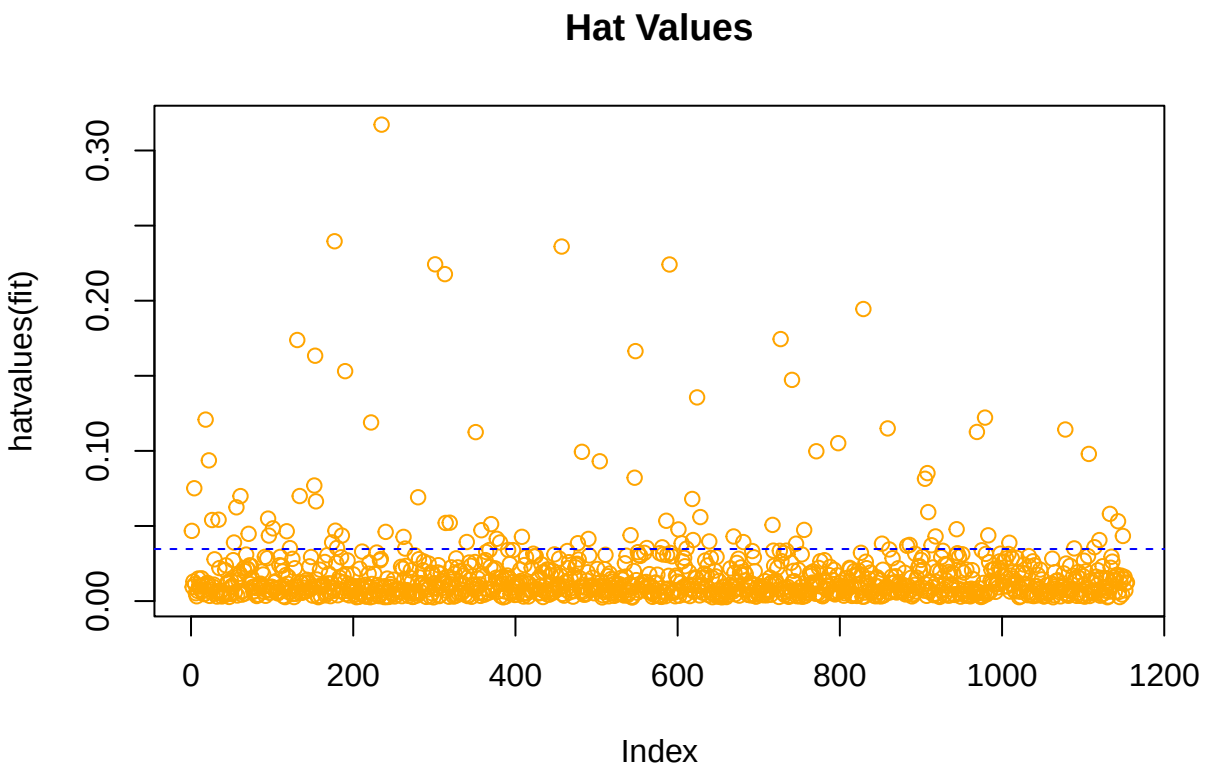
Base on the adjusted R-squared and the P-value of the full model,Using a linear model is appropriate.

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

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)
```



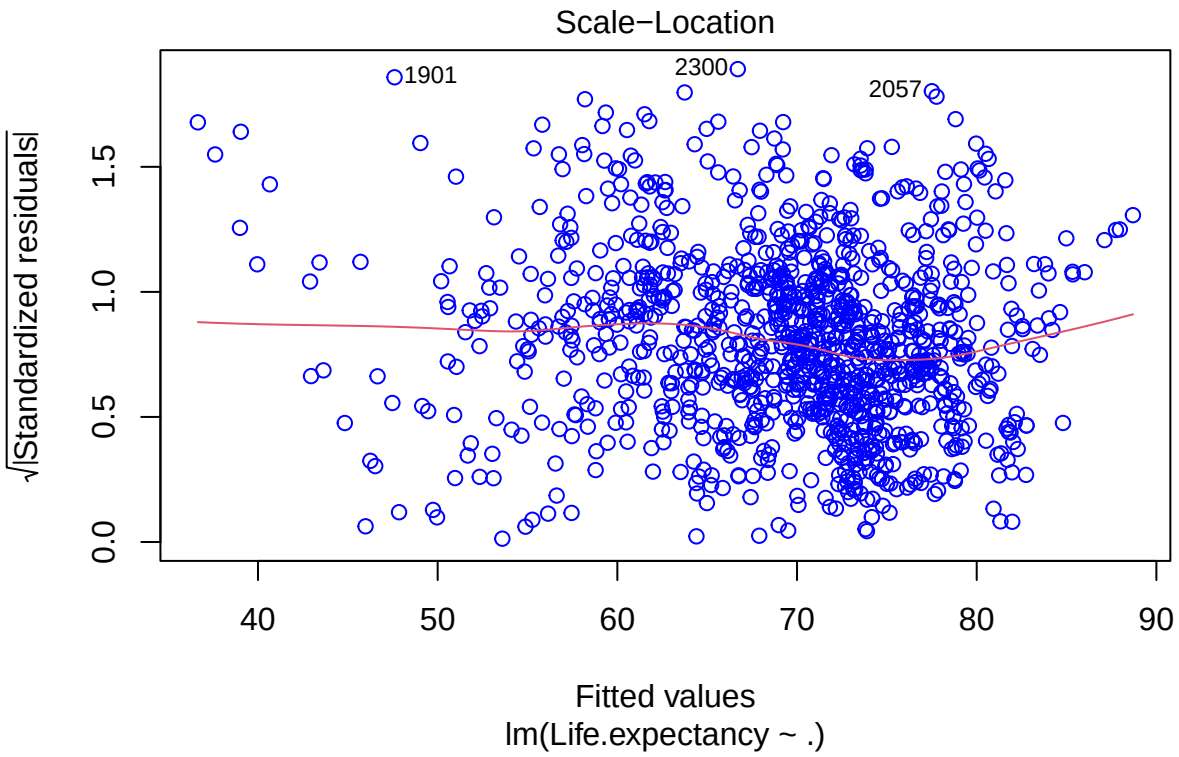
By combining the definition of high leverage points with the diagram above we can see that there are many high leverage points in the model.

Outliers

```
check_outliers(model)
```

```
## OK: No outliers detected.
## - Based on the following method and threshold: cook (1).
## - For variable: (Whole model)
```

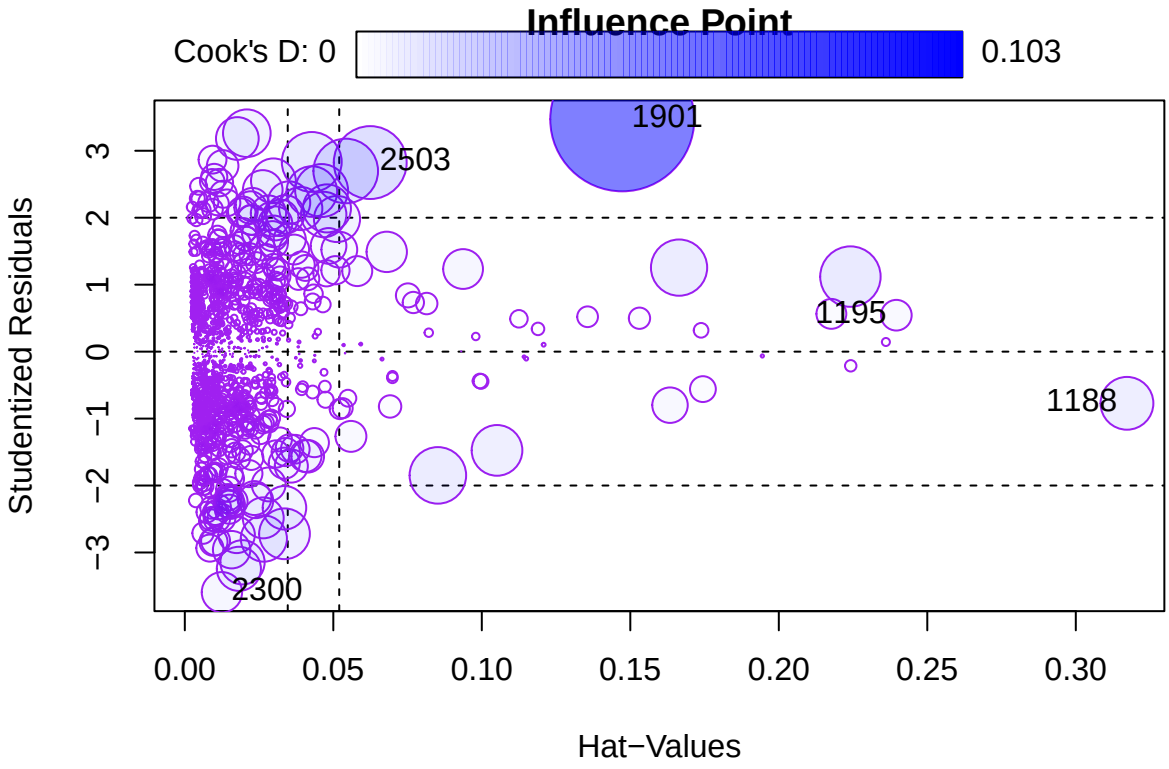
```
plot(model,which=3,col='blue')
```



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

Influential Point

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




```
##          StudRes        Hat       CookD
## 2503  2.8238475 0.06242648 0.026384768
## 2300 -3.5936519 0.01249462 0.008085122
## 1195  0.5439566 0.23957771 0.004664014
## 1188 -0.7719597 0.31722357 0.013848443
## 1901  3.4691472 0.14728525 0.102935519
```

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$residuals)
```

```
## [1] -9.834954e-17
```

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

Homoskedasticity Assumption

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 125.61, df = 19, p-value < 2.2e-16
```

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

```
##
##  Breusch-Pagan test
##
## data:  model
## BP = 162.73, df = 19, 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$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.99402, p-value = 0.0001406
```

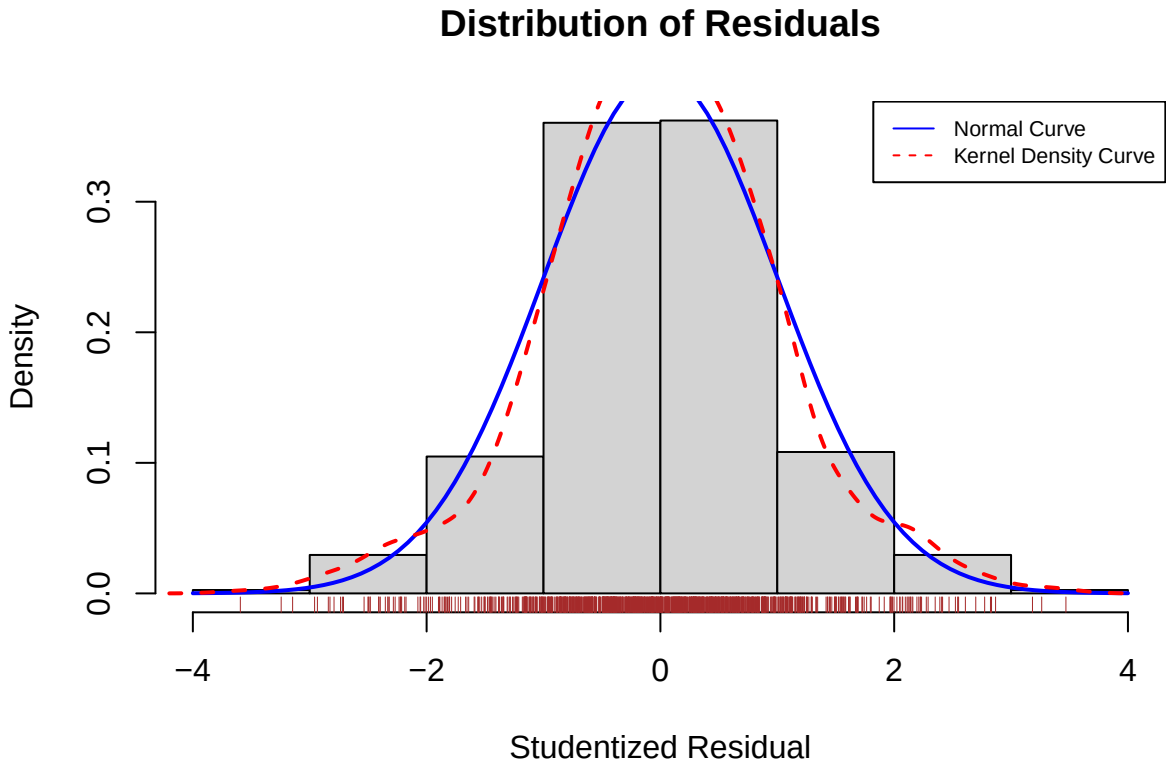
```
dev.new()
qqPlot(model,labels=row.names(df),id.method='identify',simulate=T,main='Q-Q Plot')
```

```
## 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,nbreaks=10){
z<-rstudent(model)
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)
```



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)
```

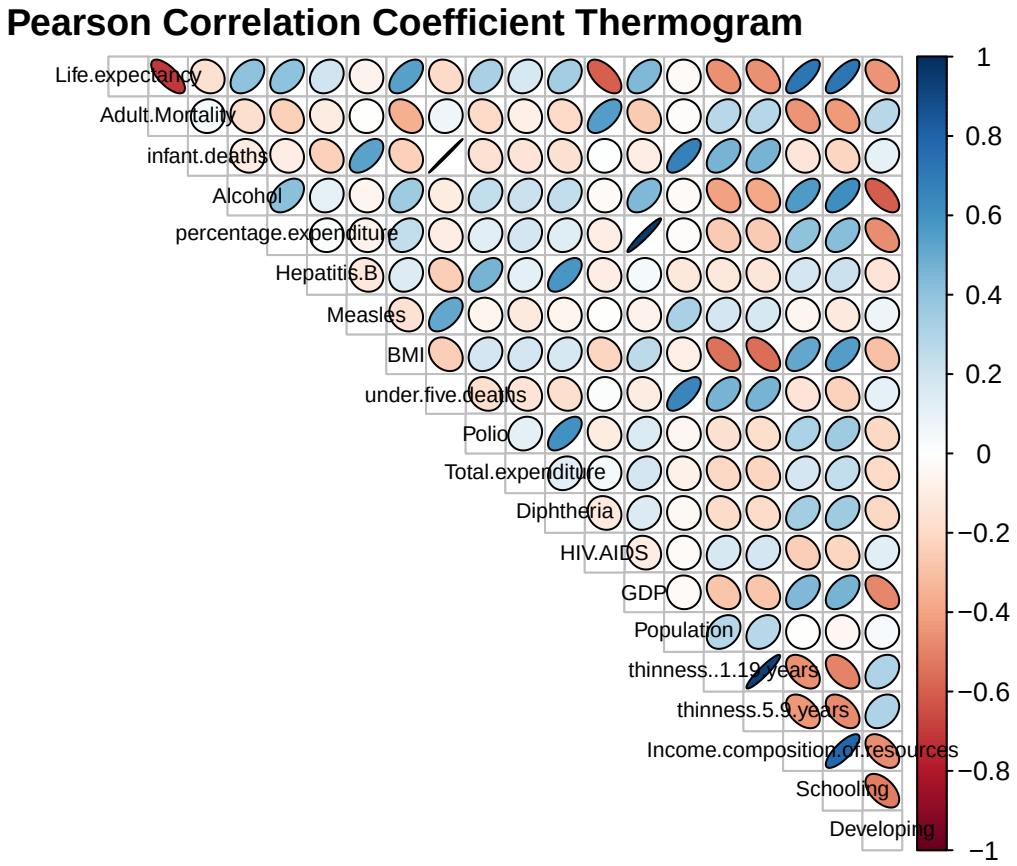
```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.0436, p-value = 0.7729
## alternative hypothesis: true autocorrelation is greater than 0
```

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)
```

```
## Model :
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths +
##     Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP +
##     Population + thinness..1.19.years + thinness.5.9.years +
##     Income.composition.of.resources + Schooling + 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),caption='Vif of Full Model')
```

Table 12: Table continues below

Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure
1.819	254.4	2.258	14.42

Table 13: Table continues below

Hepatitis.B	Measles	BMI	under.five.deaths	Polio	Total.expenditure
1.66	1.543	1.773	238.5	1.712	1.115

Table 14: Table continues below

Diphtheria	HIV.AIDS	GDP	Population	thinness..1.19.years
2.05	1.5	15.12	2.265	5.839

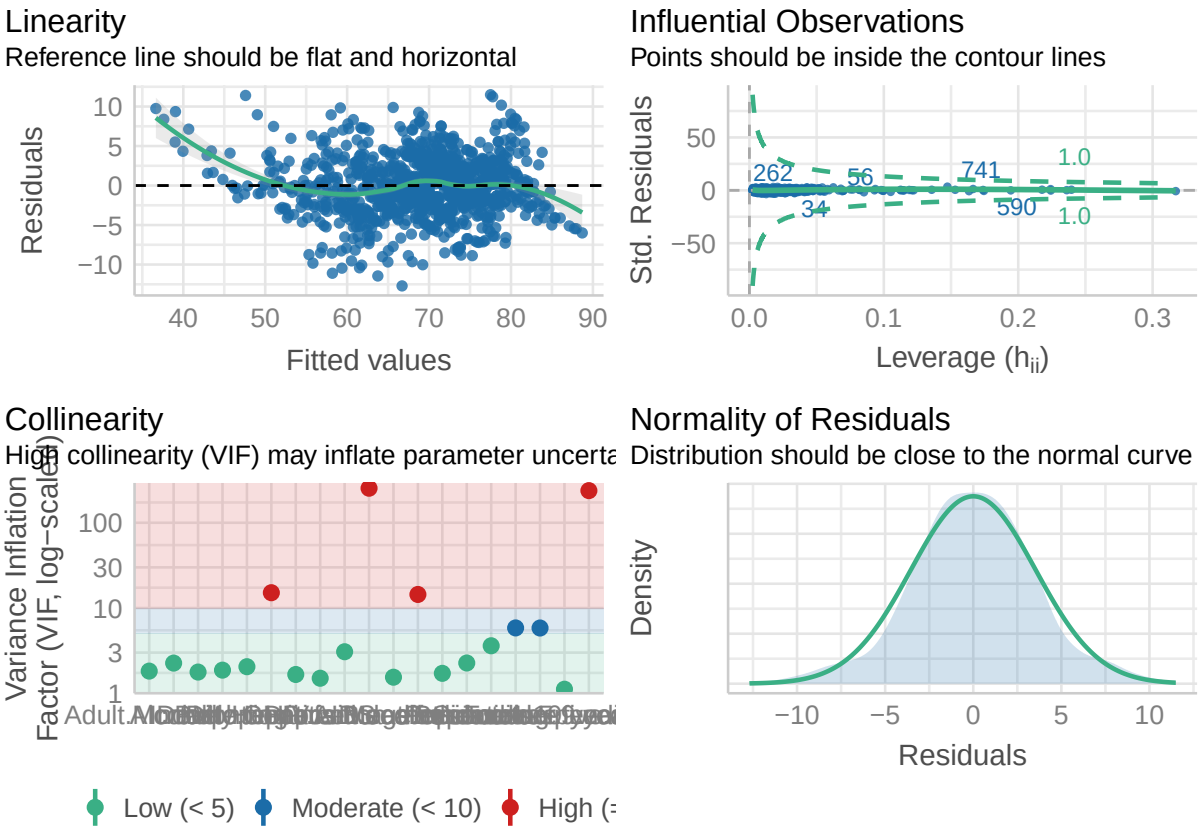
thinness.5.9.years	Income.composition.of.resources	Schooling	Developing
5.827	3.07	3.606	1.867

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



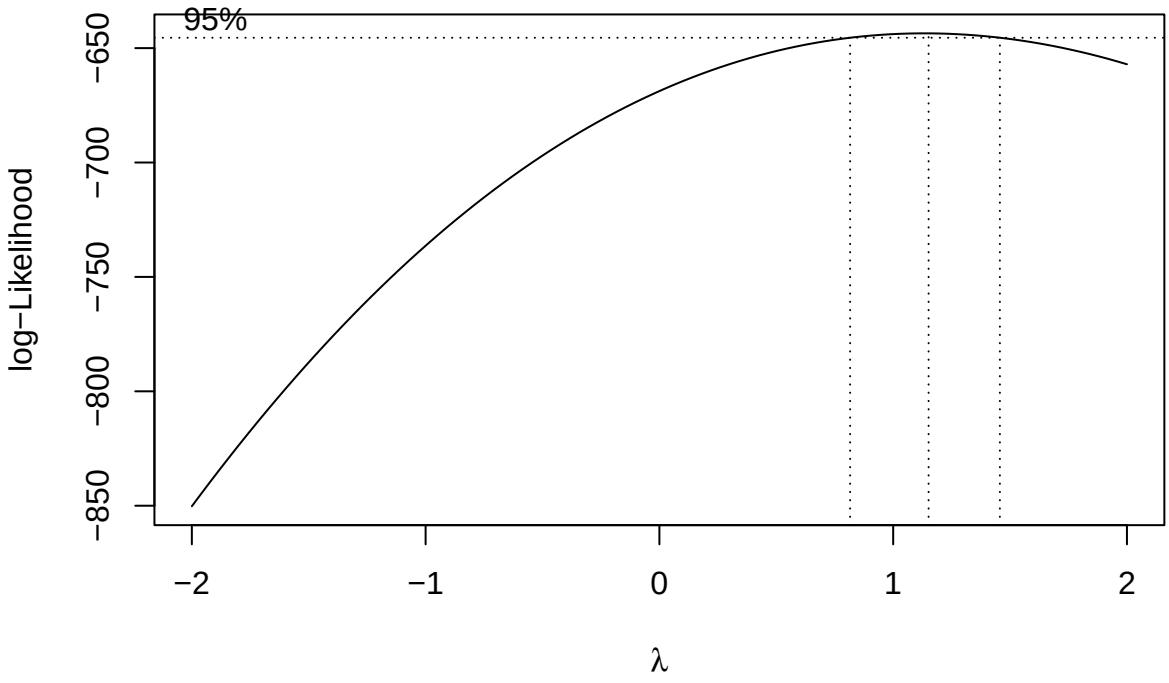
In this section we find that the pass test results for the full model, although fair overall, suffer mainly from multicollinearity, heteroskedasticity and non-normality. In the next section we will try to address these problems using the methods we have learned.

Model Transformation And Adjustment

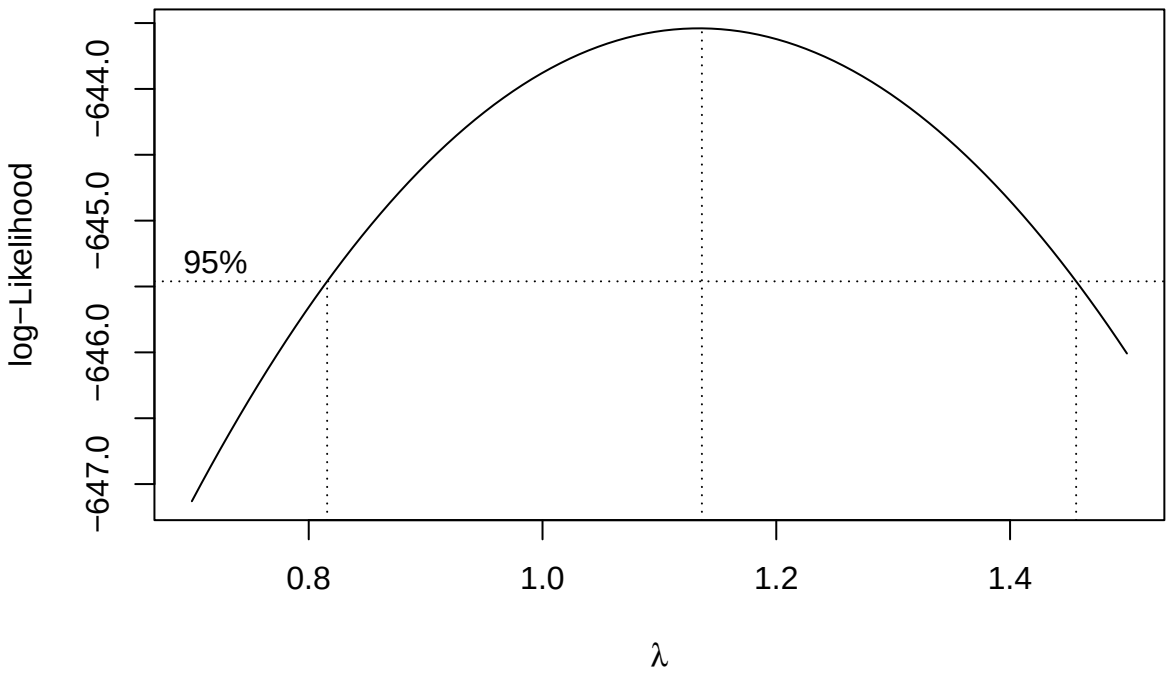
1.Box-Cox Transformation

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, plotit=T)
```




```
b<-boxcox(model, plotit=T,lambda=seq(0.7,1.5,by=0.01))
```




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

```
## [1] 1.136364
```

```
lmod_trans<-lm(Life.expectancy ^(1.136) ~ Adult.Mortality + infant.deaths + Alcohol +
                 percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
                 Polio + Diphtheria + HIV.AIDS + thinness.5.9.years +
                 Income.composition.of.resources + Schooling + Developing,
               data = data_tr)
summary(lmod_trans)
```

```
##
## Call:
## lm(formula = Life.expectancy^(1.136) ~ Adult.Mortality + infant.deaths +
##     Alcohol + percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
##     Polio + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources +
##     Schooling + Developing, data = data_tr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8488  -4.4612   0.0275   4.5137  23.9967
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                     94.2283971  1.9688465  47.860  < 2e-16 ***
## Adult.Mortality                 -0.0321801  0.0022856 -14.080  < 2e-16 ***
## infant.deaths                    0.1870637  0.0265329   7.050 3.09e-12 ***
## Alcohol                         -0.2451812  0.0781263  -3.138  0.00174 **
## percentage.expenditure           0.0008019  0.0001372   5.845 6.59e-09 ***
## Hepatitis.B                     -0.0212295  0.0104670  -2.028  0.04277 *
## BMI                              0.0691134  0.0141619   4.880 1.21e-06 ***
## under.five.deaths               -0.1411592  0.0197774  -7.137 1.69e-12 ***
## Polio                            0.0242620  0.0123743   1.961  0.05016 .
## Diphtheria                       0.0320646  0.0144609   2.217  0.02680 *
## HIV.AIDS                        -0.8588569  0.0431957 -19.883  < 2e-16 ***
## thinness.5.9.years              -0.1363722  0.0641206  -2.127  0.03365 *
## Income.composition.of.resources 21.1613696  2.0368972  10.389  < 2e-16 ***
## Schooling                        1.7195401  0.1396837  12.310  < 2e-16 ***
## Developing                      -2.4605663  0.8170104  -3.012  0.00266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.212 on 1139 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8312
## F-statistic: 406.5 on 14 and 1139 DF,  p-value: < 2.2e-16
```

```
dwtest(lmod_trans)
```

```
##
##  Durbin-Watson test
##
## data:  lmod_trans
## DW = 2.0477, p-value = 0.7923
## alternative hypothesis: true autocorrelation is greater than 0
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  lmod_trans$residuals
## W = 0.9945, p-value = 0.0003119
```

```
bptest(lmod_trans)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmod_trans
## BP = 113.96, df = 14, 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)
(neweywest<-coeftest(model,vcov=NeweyWest(model)))
```

```
##
## t test of coefficients:
##
##                                   Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)                      5.4609e+01  1.0683e+00  51.1159 < 2.2e-16 ***
## Adult.Mortality                 -1.5936e-02  1.4231e-03 -11.1982 < 2.2e-16 ***
## infant.deaths                    1.0223e-01  1.6763e-02   6.0983 1.468e-09 ***
## Alcohol                         -1.2071e-01  3.6920e-02  -3.2695  0.001110 **
## percentage.expenditure           3.2409e-04  1.7148e-04   1.8900  0.059016 .
## Hepatitis.B                     -1.0681e-02  4.6906e-03  -2.2770  0.022971 *
## Measles                         -1.0380e-05  8.1187e-06  -1.2785  0.201331
## BMI                              3.4010e-02  7.3144e-03   4.6498 3.713e-06 ***
## under.five.deaths               -7.6070e-02  1.2743e-02  -5.9694 3.181e-09 ***
## Polio                            1.2224e-02  6.8749e-03   1.7780  0.075666 .
## Total.expenditure                3.3632e-02  5.8434e-02   0.5756  0.565024
## Diphtheria                       1.5903e-02  8.3198e-03   1.9114  0.056203 .
## HIV.AIDS                        -4.3859e-01  2.8818e-02 -15.2194 < 2.2e-16 ***
## GDP                              9.8581e-06  2.6362e-05   0.3739  0.708514
## Population                      -1.7242e-09  1.3060e-09  -1.3202  0.187020
## thinness..1.19.years            -1.3409e-02  4.1870e-02  -0.3203  0.748825
## thinness.5.9.years              -5.4317e-02  3.9691e-02  -1.3685  0.171430
## Income.composition.of.resources  1.0445e+01  1.4104e+00   7.4057 2.539e-13 ***
## Schooling                        8.5100e-01  8.4399e-02  10.0830 < 2.2e-16 ***
## Developing                      -1.1438e+00  3.7314e-01  -3.0654  0.002225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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
```

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

Model Variables Selection

1.AIC Selection

```
step(model)
```

```
## Start:  AIC=2960.78
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths +
##     Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP +
##     Population + thinness..1.19.years + thinness.5.9.years +
##     Income.composition.of.resources + Schooling + Developing
##
##                                 Df Sum of Sq   RSS    AIC
## - thinness..1.19.years           1       0.7 14502 2958.8
## - GDP                            1       1.1 14502 2958.9
## - Total.expenditure              1       6.3 14508 2959.3
## - Measles                        1       7.9 14509 2959.4
## - Population                     1       8.5 14510 2959.5
## - thinness.5.9.years             1      12.1 14513 2959.7
## <none>                                       14501 2960.8
```

```
## - percentage.expenditure           1      28.6 14530 2961.0
## - Polio                            1      50.4 14552 2962.8
## - Hepatitis.B                      1      53.9 14555 2963.1
## - Diphtheria                       1      62.7 14564 2963.8
## - Developing                       1     101.6 14603 2966.8
## - Alcohol                          1     123.3 14625 2968.5
## - BMI                              1     294.8 14796 2982.0
## - infant.deaths                    1     604.4 15106 3005.9
## - under.five.deaths                1     645.1 15146 3009.0
## - Income.composition.of.resources  1    1354.7 15856 3061.8
## - Schooling                        1    1890.4 16392 3100.2
## - Adult.Mortality                  1    2514.0 17015 3143.3
## - HIV.AIDS                         1    5282.4 19784 3317.2
##
## Step:  AIC=2958.84
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths +
##     Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP +
##     Population + thinness.5.9.years + Income.composition.of.resources +
##     Schooling + Developing
##
##                                   Df Sum of Sq   RSS    AIC
## - GDP                              1       1.0 14503 2956.9
## - Total.expenditure                1       6.3 14508 2957.3
## - Measles                          1       8.0 14510 2957.5
## - Population                       1       8.8 14511 2957.5
## <none>                                         14502 2958.8
## - percentage.expenditure           1      28.6 14531 2959.1
## - Polio                            1      49.8 14552 2960.8
## - thinness.5.9.years               1      52.6 14555 2961.0
## - Hepatitis.B                      1      54.3 14556 2961.1
## - Diphtheria                       1      63.2 14565 2961.9
## - Developing                       1     101.2 14603 2964.9
## - Alcohol                          1     122.6 14625 2966.6
## - BMI                              1     297.2 14799 2980.3
## - infant.deaths                    1     606.4 15108 3004.1
## - under.five.deaths                1     647.7 15150 3007.3
## - Income.composition.of.resources  1    1361.3 15863 3060.4
## - Schooling                        1    1906.4 16408 3099.4
## - Adult.Mortality                  1    2514.6 17017 3141.4
## - HIV.AIDS                         1    5293.1 19795 3315.9
##
## Step:  AIC=2956.92
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths +
##     Polio + Total.expenditure + Diphtheria + HIV.AIDS + Population +
##     thinness.5.9.years + Income.composition.of.resources + Schooling +
##     Developing
##
##                                   Df Sum of Sq   RSS    AIC
## - Total.expenditure                1       6.1 14509 2955.4
## - Measles                          1       7.9 14511 2955.5
## - Population                       1       9.0 14512 2955.6
## <none>                                         14503 2956.9
```

```
## - Polio                             1      50.4 14554 2958.9
## - thinness.5.9.years                1      52.7 14556 2959.1
## - Hepatitis.B                       1      53.7 14557 2959.2
## - Diphtheria                        1      62.7 14566 2959.9
## - Developing                        1     102.2 14605 2963.0
## - Alcohol                           1     122.2 14625 2964.6
## - BMI                               1     297.7 14801 2978.4
## - percentage.expenditure            1     403.1 14906 2986.6
## - infant.deaths                     1     607.1 15110 3002.2
## - under.five.deaths                 1     648.2 15151 3005.4
## - Income.composition.of.resources   1    1374.4 15878 3059.4
## - Schooling                         1    1921.7 16425 3098.5
## - Adult.Mortality                   1    2514.1 17017 3139.4
## - HIV.AIDS                          1    5292.7 19796 3313.9
##
## Step:  AIC=2955.4
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths +
##     Polio + Diphtheria + HIV.AIDS + Population + thinness.5.9.years +
##     Income.composition.of.resources + Schooling + Developing
##
##                                  Df Sum of Sq   RSS    AIC
## - Measles                         1       8.4 14518 2954.1
## - Population                      1       8.9 14518 2954.1
## <none>                                        14509 2955.4
## - Polio                           1      51.1 14560 2957.5
## - Hepatitis.B                     1      52.1 14561 2957.5
## - thinness.5.9.years              1      55.8 14565 2957.8
## - Diphtheria                      1      63.6 14573 2958.5
## - Developing                      1     103.8 14613 2961.6
## - Alcohol                         1     121.7 14631 2963.0
## - BMI                             1     303.0 14812 2977.3
## - percentage.expenditure          1     411.2 14920 2985.7
## - infant.deaths                   1     603.6 15113 3000.4
## - under.five.deaths               1     644.7 15154 3003.6
## - Income.composition.of.resources 1    1371.1 15880 3057.6
## - Schooling                       1    1938.1 16447 3098.1
## - Adult.Mortality                 1    2521.7 17031 3138.3
## - HIV.AIDS                        1    5322.5 19832 3314.0
##
## Step:  AIC=2954.07
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
##     Polio + Diphtheria + HIV.AIDS + Population + thinness.5.9.years +
##     Income.composition.of.resources + Schooling + Developing
##
##                                  Df Sum of Sq   RSS    AIC
## - Population                      1       6.9 14524 2952.6
## <none>                                        14518 2954.1
## - Polio                           1      50.6 14568 2956.1
## - Hepatitis.B                     1      51.5 14569 2956.2
## - thinness.5.9.years              1      51.7 14569 2956.2
## - Diphtheria                      1      64.1 14582 2957.2
## - Developing                      1     104.8 14622 2960.4
```

```
## - Alcohol                          1     125.2 14643 2962.0
## - BMI                              1     314.6 14832 2976.8
## - percentage.expenditure           1     413.3 14931 2984.5
## - infant.deaths                    1     644.6 15162 3002.2
## - under.five.deaths                1     675.0 15192 3004.5
## - Income.composition.of.resources  1    1375.5 15893 3056.5
## - Schooling                        1    1938.8 16456 3096.7
## - Adult.Mortality                  1    2536.2 17054 3137.9
## - HIV.AIDS                         1    5328.9 19846 3312.9
##
## Step:  AIC=2952.62
## Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
##     percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
##     Polio + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources +
##     Schooling + Developing
##
##                                   Df Sum of Sq   RSS    AIC
## <none>                                         14524 2952.6
## - Polio                            1      50.0 14574 2954.6
## - Hepatitis.B                      1      50.9 14575 2954.7
## - thinness.5.9.years               1      53.7 14578 2954.9
## - Diphtheria                       1      62.7 14587 2955.6
## - Developing                       1     105.1 14630 2958.9
## - Alcohol                          1     125.3 14650 2960.5
## - BMI                              1     310.1 14834 2975.0
## - percentage.expenditure           1     411.4 14936 2982.8
## - infant.deaths                    1     660.9 15185 3002.0
## - under.five.deaths                1     677.4 15202 3003.2
## - Income.composition.of.resources  1    1382.0 15906 3055.5
## - Schooling                        1    1933.8 16458 3094.9
## - Adult.Mortality                  1    2554.6 17079 3137.6
## - HIV.AIDS                         1    5324.4 19849 3311.0


##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     Alcohol + percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
##     Polio + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources +
##     Schooling + Developing, data = data_tr)
##
## Coefficients:
##                      (Intercept)                   Adult.Mortality
##                        54.7468812                       -0.0160171
##                     infant.deaths                          Alcohol
##                         0.0945725                       -0.1212779
##          percentage.expenditure                      Hepatitis.B
##                         0.0003858                       -0.0103587
##                               BMI                 under.five.deaths
##                         0.0345796                       -0.0713724
##                             Polio                       Diphtheria
##                         0.0121297                        0.0158760
##                          HIV.AIDS                thinness.5.9.years
##                        -0.4370189                       -0.0651349
## Income.composition.of.resources                         Schooling
```

```
##                       10.4991049                         0.8516676
##                       Developing
##                       -1.1615165
```

```
# Find model with lowest AIC
lmod_AIC_B<-lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
                 percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
                 Polio + Diphtheria + HIV.AIDS + thinness.5.9.years +
                 Income.composition.of.resources + Schooling + Developing,
               data = data_tr) # AIC selected model
sum_AIC_B<-summary(lmod_AIC_B)
sum_AIC_B
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     Alcohol + percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
##     Polio + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources +
##     Schooling + Developing, data = data_tr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4930  -2.1727   0.0338   2.2373  11.6085
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.475e+01  9.748e-01  56.161  < 2e-16 ***
## Adult.Mortality                 -1.602e-02  1.132e-03 -14.154  < 2e-16 ***
## infant.deaths                    9.457e-02  1.314e-02   7.199 1.10e-12 ***
## Alcohol                         -1.213e-01  3.868e-02  -3.135  0.00176 **
## percentage.expenditure           3.858e-04  6.793e-05   5.680 1.71e-08 ***
## Hepatitis.B                     -1.036e-02  5.182e-03  -1.999  0.04587 *
## BMI                              3.458e-02  7.012e-03   4.932 9.37e-07 ***
## under.five.deaths               -7.137e-02  9.792e-03  -7.289 5.84e-13 ***
## Polio                            1.213e-02  6.127e-03   1.980  0.04797 *
## Diphtheria                       1.588e-02  7.160e-03   2.217  0.02680 *
## HIV.AIDS                        -4.370e-01  2.139e-02 -20.434  < 2e-16 ***
## thinness.5.9.years              -6.513e-02  3.175e-02  -2.052  0.04043 *
## Income.composition.of.resources  1.050e+01  1.009e+00  10.411  < 2e-16 ***
## Schooling                        8.517e-01  6.916e-02  12.314  < 2e-16 ***
## Developing                      -1.162e+00  4.045e-01  -2.871  0.00416 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.571 on 1139 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8328
## F-statistic: 411.1 on 14 and 1139 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), 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
## + thinness.5.9.years               1     20496 67415 4698.0
## + thinness..1.19.years             1     20454 67456 4698.8
## + Developing                       1     17741 70169 4744.3
## + GDP                              1     16988 70923 4756.6
## + percentage.expenditure           1     14667 73244 4793.7
## + Alcohol                          1     14322 73589 4799.2
## + Polio                            1      9706 78205 4869.4
## + Diphtheria                       1      9526 78385 4872.0
## + under.five.deaths                1      3146 84765 4962.3
## + Hepatitis.B                      1      2772 85139 4967.4
## + infant.deaths                    1      2460 85450 4971.6
## + Total.expenditure                1      1948 85962 4978.5
## + Measles                          1       276 87635 5000.8
## <none>                                         87911 5002.4
## + Population                       1        38 87873 5003.9
##
## 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
## + thinness.5.9.years               1      1844 39720 4089.6
## + thinness..1.19.years             1      1437 40126 4101.3
## + GDP                              1      1100 40463 4111.0
## + percentage.expenditure           1      1046 40517 4112.5
## + Polio                            1       776 40787 4120.2
## + Diphtheria                       1       644 40919 4123.9
## + Developing                       1       555 41008 4126.4
## + Alcohol                          1       355 41208 4132.0
## + Hepatitis.B                      1        75 41488 4139.8
## <none>                                         41563 4139.9
## + Measles                          1        38 41525 4140.9
## + under.five.deaths                1        33 41530 4141.0
## + Total.expenditure                1        25 41538 4141.2
```

```
## + Population                         1         8 41555 4141.7
## + 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
## + GDP                              1      1017 22581 3439.8
## + percentage.expenditure           1       992 22605 3441.1
## + BMI                              1       984 22613 3441.5
## + thinness.5.9.years               1       641 22956 3458.9
## + thinness..1.19.years             1       500 23097 3465.9
## + Diphtheria                       1       486 23111 3466.6
## + Polio                            1       452 23146 3468.4
## + Developing                       1       395 23202 3471.2
## + under.five.deaths                1        98 23499 3485.8
## + Total.expenditure                1        90 23507 3486.2
## + infant.deaths                    1        47 23550 3488.3
## <none>                                         23597 3488.7
## + Hepatitis.B                      1        21 23577 3489.6
## + Measles                          1         8 23590 3490.3
## + Population                       1         5 23592 3490.4
## + Alcohol                          1         1 23596 3490.6
## - 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
## + GDP                              1     725.8 18102 3186.7
## + percentage.expenditure           1     704.4 18123 3188.1
## + BMI                              1     623.1 18204 3193.2
## + thinness.5.9.years               1     428.3 18399 3205.5
## + thinness..1.19.years             1     388.2 18439 3208.0
## + Diphtheria                       1     357.2 18470 3210.0
## + Polio                            1     285.0 18542 3214.5
## + Developing                       1     231.8 18596 3217.8
## + under.five.deaths                1     105.0 18722 3225.6
## + infant.deaths                    1      63.3 18764 3228.2
## + Total.expenditure                1      49.4 18778 3229.0
## <none>                                         18827 3230.1
## + Hepatitis.B                      1      17.6 18810 3231.0
## + Alcohol                          1      15.4 18812 3231.1
## + Population                       1       5.1 18822 3231.7
## + Measles                          1       2.9 18824 3231.9
## - 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
## + percentage.expenditure          1     459.1 16280 3066.3
## + GDP                             1     437.7 16301 3067.8
## + BMI                             1     425.0 16314 3068.7
## + thinness.5.9.years              1     249.5 16489 3081.0
## + thinness..1.19.years            1     206.5 16532 3084.0
## + Diphtheria                      1     194.5 16544 3084.9
## + Polio                           1     169.1 16570 3086.7
## + under.five.deaths               1     157.3 16582 3087.5
## + Developing                      1     124.6 16614 3089.7
## + infant.deaths                   1     109.6 16629 3090.8
## + Total.expenditure               1      44.7 16694 3095.3
## + Alcohol                         1      29.1 16710 3096.4
## <none>                                       16739 3096.4
## + Population                      1      14.7 16724 3097.4
## + Hepatitis.B                     1       7.9 16731 3097.8
## + Measles                         1       0.1 16739 3098.4
## - 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=3066.28
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure
##
##                                  Df Sum of Sq   RSS    AIC
## + BMI                             1     435.4 15844 3037.0
## + thinness.5.9.years              1     217.3 16062 3052.8
## + Diphtheria                      1     212.4 16067 3053.1
## + Polio                           1     193.5 16086 3054.5
## + thinness..1.19.years            1     185.3 16094 3055.1
## + under.five.deaths               1     158.8 16121 3057.0
## + infant.deaths                   1     108.4 16171 3060.6
## + Alcohol                         1      96.8 16183 3061.4
## <none>                                       16280 3066.3
## + Total.expenditure               1      24.8 16255 3066.5
## + Hepatitis.B                     1      22.7 16257 3066.7
## + Developing                      1      20.2 16260 3066.8
## + Population                      1      14.2 16266 3067.3
## + GDP                             1       2.4 16277 3068.1
## + Measles                         1       0.1 16280 3068.3
## - percentage.expenditure          1     459.1 16739 3096.4
## - Income.composition.of.resources  1    1843.1 18123 3188.1
## - Schooling                       1    3328.1 19608 3278.9
## - Adult.Mortality                 1    3594.7 19875 3294.5
## - HIV.AIDS                        1    5854.4 22134 3418.8
##
## Step:  AIC=3037
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI
```

```
##
##                                  Df Sum of Sq   RSS    AIC
## + Diphtheria                      1     239.9 15604 3021.4
## + Polio                           1     207.0 15637 3023.8
## + Alcohol                         1     118.1 15726 3030.4
## + under.five.deaths               1      92.2 15752 3032.3
## + thinness.5.9.years              1      56.2 15788 3034.9
## + infant.deaths                   1      53.9 15790 3035.1
## + thinness..1.19.years            1      46.1 15798 3035.6
## <none>                                       15844 3037.0
## + Hepatitis.B                     1      20.5 15824 3037.5
## + Developing                      1      18.5 15826 3037.7
## + Measles                         1       8.8 15836 3038.4
## + Total.expenditure               1       8.8 15836 3038.4
## + Population                      1       4.1 15840 3038.7
## + GDP                             1       1.7 15843 3038.9
## - BMI                             1     435.4 16280 3066.3
## - percentage.expenditure          1     469.6 16314 3068.7
## - Income.composition.of.resources 1    1655.0 17499 3149.6
## - Schooling                       1    2429.6 18274 3199.6
## - Adult.Mortality                 1    3376.7 19221 3257.9
## - HIV.AIDS                        1    5727.9 21572 3391.1
##
## Step:  AIC=3021.39
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI + Diphtheria
##
##                                  Df Sum of Sq   RSS    AIC
## + Alcohol                         1     132.3 15472 3013.6
## + under.five.deaths               1      56.1 15548 3019.2
## + Polio                           1      50.5 15554 3019.7
## + thinness.5.9.years              1      46.3 15558 3020.0
## + thinness..1.19.years            1      38.8 15566 3020.5
## + infant.deaths                   1      29.2 15575 3021.2
## <none>                                       15604 3021.4
## + Hepatitis.B                     1      24.9 15580 3021.5
## + Developing                      1      15.6 15589 3022.2
## + Measles                         1      11.4 15593 3022.5
## + Total.expenditure               1       3.2 15601 3023.2
## + Population                      1       2.3 15602 3023.2
## + GDP                             1       1.3 15603 3023.3
## - Diphtheria                      1     239.9 15844 3037.0
## - BMI                             1     462.9 16067 3053.1
## - percentage.expenditure          1     489.1 16094 3055.0
## - Income.composition.of.resources 1    1490.0 17095 3124.6
## - Schooling                       1    2186.1 17791 3170.7
## - Adult.Mortality                 1    3312.8 18917 3241.6
## - HIV.AIDS                        1    5743.2 21348 3381.0
##
## Step:  AIC=3013.56
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI + Diphtheria + Alcohol
##
##                                  Df Sum of Sq   RSS    AIC
```

```
## + Developing                         1      83.5 15389 3009.3
## + thinness.5.9.years                 1      66.5 15406 3010.6
## + thinness..1.19.years               1      58.4 15414 3011.2
## + Polio                              1      52.8 15419 3011.6
## + under.five.deaths                  1      49.6 15423 3011.9
## + Hepatitis.B                        1      27.6 15445 3013.5
## <none>                                         15472 3013.6
## + infant.deaths                      1      26.0 15446 3013.6
## + Measles                            1      12.8 15459 3014.6
## + Total.expenditure                  1       4.9 15467 3015.2
## + Population                         1       2.7 15470 3015.4
## + GDP                                1       2.4 15470 3015.4
## - Alcohol                            1     132.3 15604 3021.4
## - Diphtheria                         1     254.1 15726 3030.4
## - BMI                                1     487.0 15959 3047.3
## - percentage.expenditure             1     574.6 16047 3053.6
## - Income.composition.of.resources    1    1600.0 17072 3125.1
## - Schooling                          1    2309.9 17782 3172.1
## - Adult.Mortality                    1    3102.7 18575 3222.5
## - HIV.AIDS                           1    5571.4 21044 3366.5
##
## Step:  AIC=3009.32
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI + Diphtheria + Alcohol + Developing
##
##                                   Df Sum of Sq   RSS    AIC
## + thinness.5.9.years               1      67.3 15321 3006.3
## + thinness..1.19.years             1      61.5 15327 3006.7
## + Polio                            1      52.1 15337 3007.4
## + under.five.deaths                1      49.6 15339 3007.6
## + Hepatitis.B                      1      37.9 15351 3008.5
## <none>                                       15389 3009.3
## + infant.deaths                    1      25.9 15363 3009.4
## + Measles                          1      14.1 15375 3010.3
## + Total.expenditure                1       3.6 15385 3011.1
## + Population                       1       2.4 15386 3011.1
## + GDP                              1       1.3 15387 3011.2
## - Developing                       1      83.5 15472 3013.6
## - Alcohol                          1     200.2 15589 3022.2
## - Diphtheria                       1     252.0 15641 3026.1
## - percentage.expenditure           1     426.0 15815 3038.8
## - BMI                              1     490.9 15880 3043.6
## - Income.composition.of.resources  1    1612.3 17001 3122.3
## - Schooling                        1    2195.5 17584 3161.2
## - Adult.Mortality                  1    2986.0 18375 3212.0
## - HIV.AIDS                         1    5527.3 20916 3361.5
##
## Step:  AIC=3006.26
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI + Diphtheria + Alcohol + Developing +
##     thinness.5.9.years
##
##                                   Df Sum of Sq   RSS    AIC
## + Polio                            1      51.5 15270 3004.4
```

```
## + Hepatitis.B                      1      37.7 15284 3005.4
## <none>                                      15321 3006.3
## + Measles                          1      21.0 15300 3006.7
## + under.five.deaths                1      18.3 15303 3006.9
## + infant.deaths                    1       4.6 15317 3007.9
## + thinness..1.19.years             1       2.8 15318 3008.0
## + GDP                              1       1.3 15320 3008.2
## + Total.expenditure                1       1.3 15320 3008.2
## + Population                       1       1.0 15320 3008.2
## - thinness.5.9.years               1      67.3 15389 3009.3
## - Developing                       1      84.3 15406 3010.6
## - Alcohol                          1     223.3 15545 3021.0
## - Diphtheria                       1     241.0 15562 3022.3
## - BMI                              1     306.8 15628 3027.1
## - percentage.expenditure           1     413.2 15734 3035.0
## - Income.composition.of.resources  1    1578.9 16900 3117.4
## - Schooling                        1    2134.9 17456 3154.8
## - Adult.Mortality                  1    2954.2 18276 3207.7
## - HIV.AIDS                         1    5400.3 20722 3352.7
##
## Step:  AIC=3004.37
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI + Diphtheria + Alcohol + Developing +
##     thinness.5.9.years + Polio
##
##                                   Df Sum of Sq   RSS    AIC
## + Hepatitis.B                      1      59.7 15210 3001.9
## <none>                                      15270 3004.4
## + Measles                          1      21.4 15248 3004.8
## + under.five.deaths                1      14.8 15255 3005.3
## + thinness..1.19.years             1       4.7 15265 3006.0
## + infant.deaths                    1       3.0 15267 3006.1
## - Polio                            1      51.5 15321 3006.3
## + Population                       1       1.4 15268 3006.3
## + Total.expenditure                1       0.9 15269 3006.3
## + GDP                              1       0.7 15269 3006.3
## - thinness.5.9.years               1      66.7 15337 3007.4
## - Diphtheria                       1      83.5 15353 3008.7
## - Developing                       1      83.6 15354 3008.7
## - Alcohol                          1     225.5 15495 3019.3
## - BMI                              1     307.3 15577 3025.4
## - percentage.expenditure           1     422.6 15692 3033.9
## - Income.composition.of.resources  1    1560.0 16830 3114.6
## - Schooling                        1    2082.2 17352 3149.9
## - Adult.Mortality                  1    2915.1 18185 3204.0
## - HIV.AIDS                         1    5388.8 20659 3351.2
##
## Step:  AIC=3001.85
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
##     percentage.expenditure + BMI + Diphtheria + Alcohol + Developing +
##     thinness.5.9.years + Polio + Hepatitis.B
##
##                                   Df Sum of Sq   RSS    AIC
## <none>                                      15210 3001.9
```

```
## + under.five.deaths                1      24.9 15185 3002.0
## + Measles                          1      15.7 15194 3002.7
## + infant.deaths                    1       8.3 15202 3003.2
## + thinness..1.19.years             1       4.3 15206 3003.5
## + Total.expenditure                1       1.9 15208 3003.7
## + GDP                              1       1.3 15209 3003.8
## + Population                       1       0.1 15210 3003.8
## - Hepatitis.B                      1      59.7 15270 3004.4
## - thinness.5.9.years               1      66.4 15276 3004.9
## - Polio                            1      73.5 15284 3005.4
## - Developing                       1      97.1 15307 3007.2
## - Diphtheria                       1     133.6 15344 3009.9
## - Alcohol                          1     239.0 15449 3017.8
## - BMI                              1     318.6 15529 3023.8
## - percentage.expenditure           1     386.9 15597 3028.8
## - Income.composition.of.resources  1    1538.8 16749 3111.1
## - Schooling                        1    2094.6 17305 3148.7
## - Adult.Mortality                  1    2879.3 18089 3199.9
## - HIV.AIDS                         1    5410.3 20620 3351.0
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##     Income.composition.of.resources + percentage.expenditure +
##     BMI + Diphtheria + Alcohol + Developing + thinness.5.9.years +
##     Polio + Hepatitis.B, data = data_tr)
##
## Coefficients:
##                       (Intercept)                          Schooling
##                         53.576043                           0.881199
##                           HIV.AIDS                    Adult.Mortality
##                         -0.439333                          -0.016888
## Income.composition.of.resources             percentage.expenditure
##                         10.987582                           0.000374
##                               BMI                          Diphtheria
##                          0.035047                           0.022959
##                           Alcohol                          Developing
##                         -0.165250                          -1.115874
##                thinness.5.9.years                              Polio
##                         -0.066872                           0.014678
##                       Hepatitis.B
##                         -0.011087
```

```
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
```

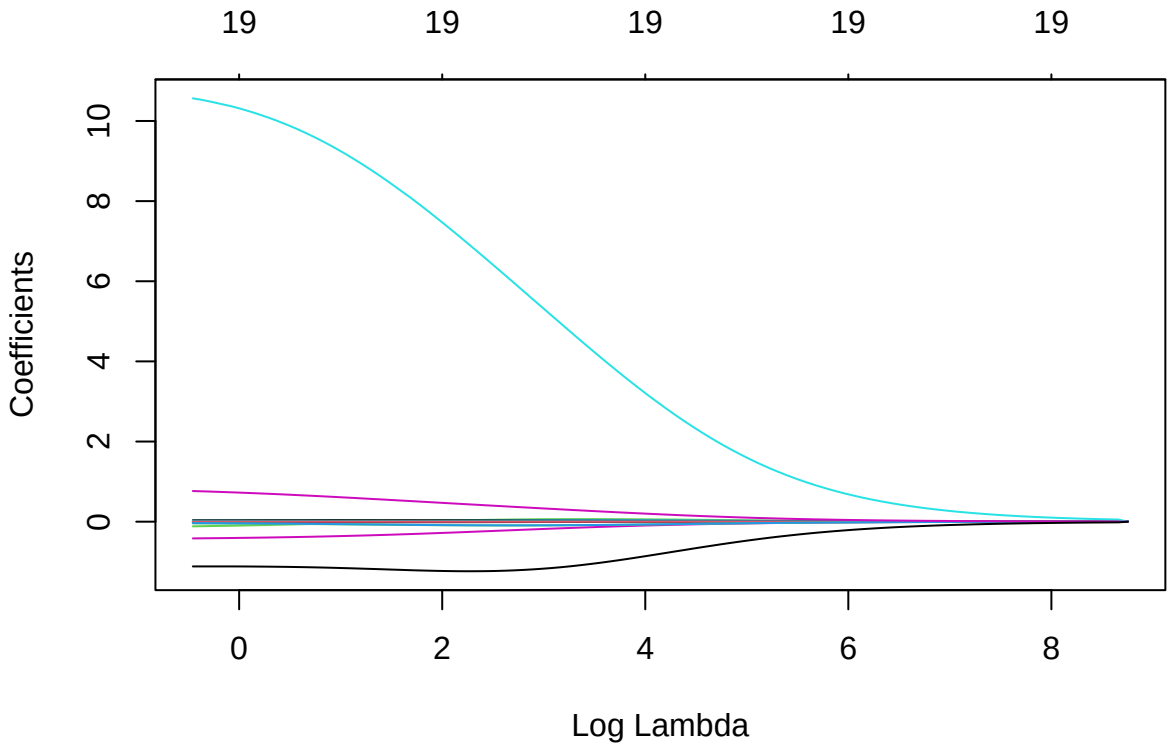
From the summary we can find that the model chosen by the backward iterative BIC method, most certainly all predictors are statistically significant, but the adjusted R-squared is even lower than the full model, and subsequently we will judge whether this model should be used by the model's prediction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

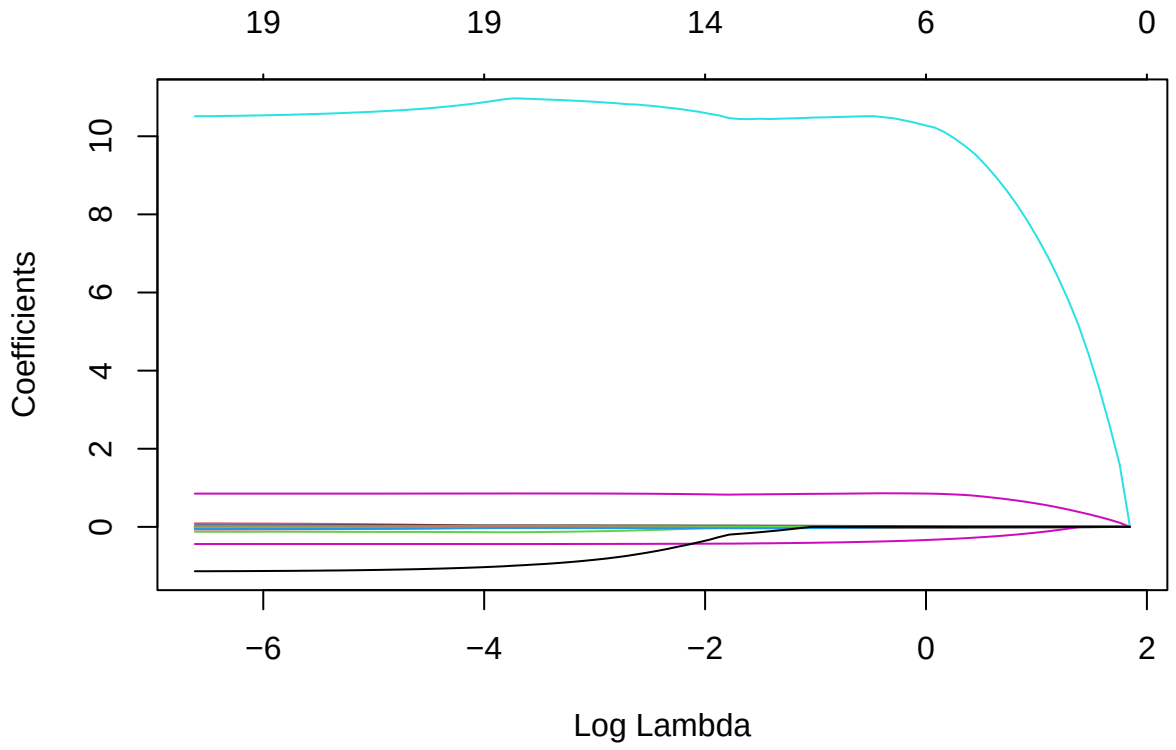
```
ridge_cv$lambda.min
```

```
## [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')
```

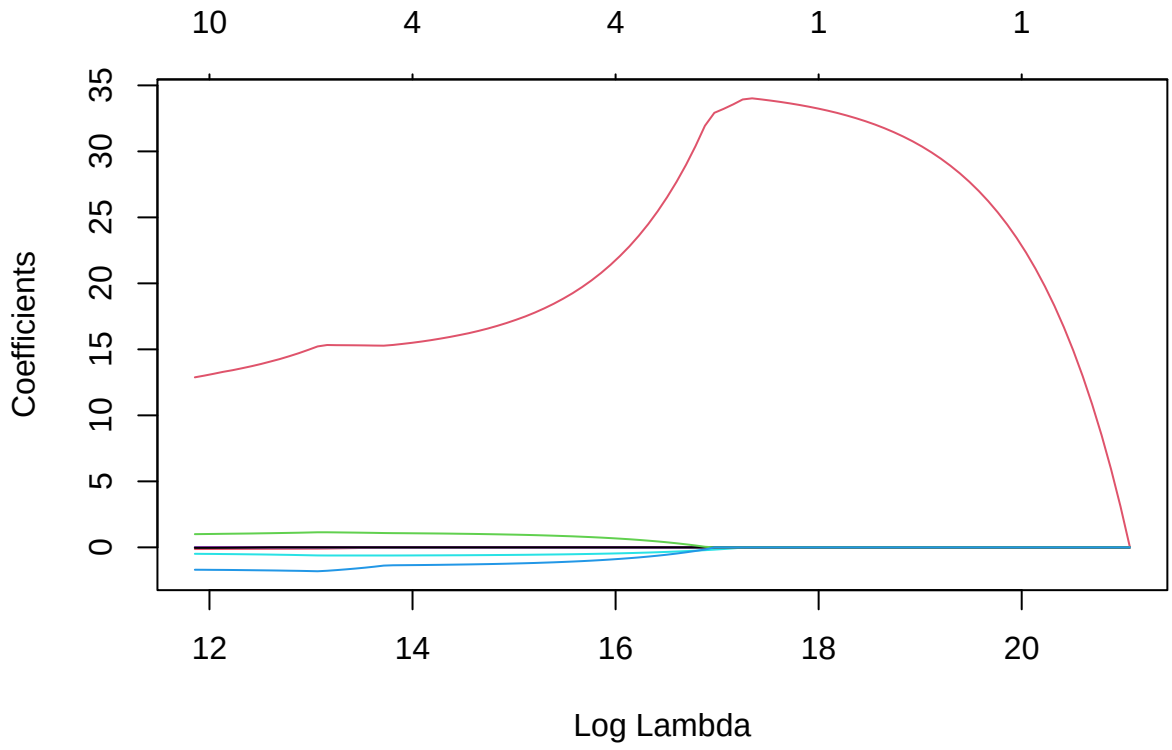



```
lasso_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=10,alpha=1,keep=T)
#plot(lasso_cv)
lasso_cv$lambda.min
```

```
## [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[
#plot(alasso_cv)
```

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

```
## [1] 46.65954
```

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

```
## [1] 46.44081
```

```
result_bic<-predict(lmod_BIC_BO,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:

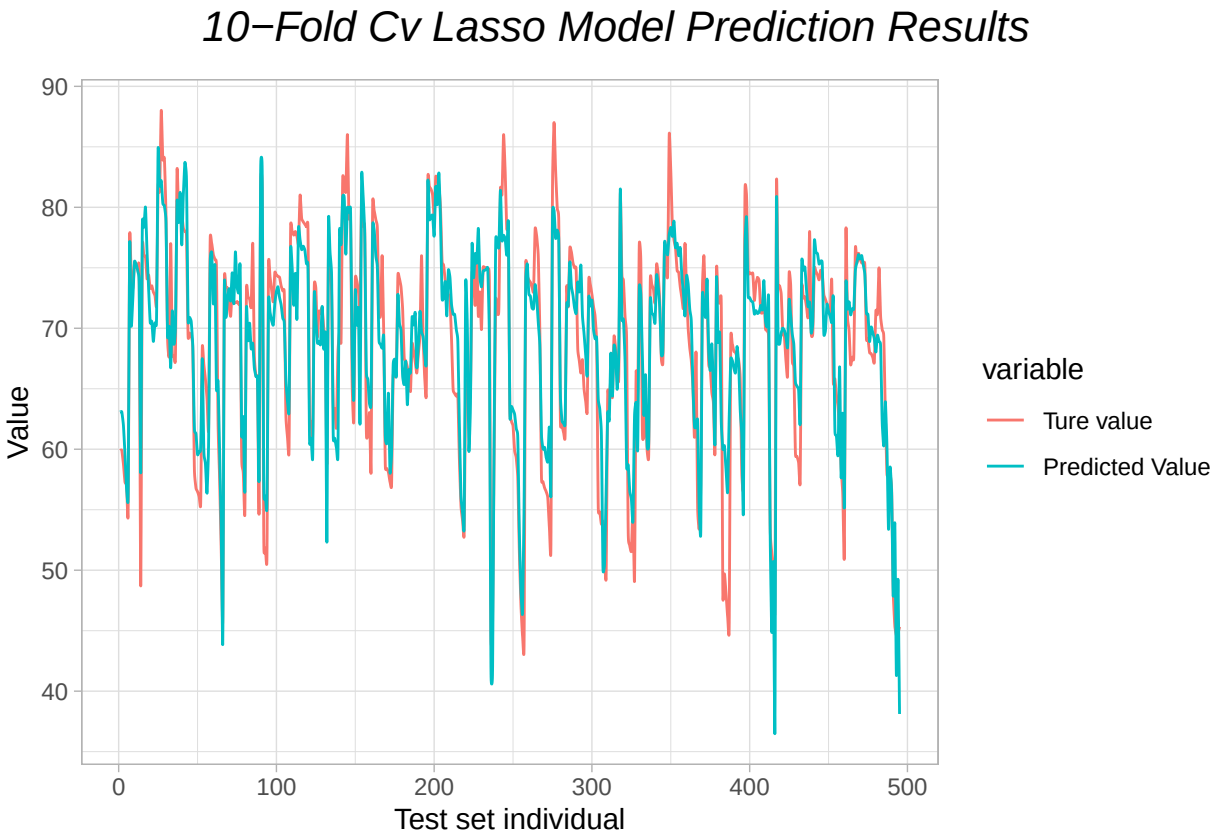
$Life.expectancy = 53.038475594 - 0.012413848 * Adult.Mortality - 0.127001514 * Alcohol + 0.024103630 * BMI + 0.005072368 * L$

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.