Data analysis

Thinh Hong

2023-03-16

The study involves 250 individuals and 19 variables, including the dependent variable. These variables are considered as biomarkers of cardiovascular health, which also include relevant demographic variables. The objective of the study is to evaluate the relationship between Triglyceride (dependent variable) and the other biomarkers of cardiovascular health. The data/variable dictionary is as follows.

1.Age in years 2. Weight 3. Height 4. Sex (male=1, female=2) 5. Body Mass Index (BMI) 6. Triglyceride (Dependent Variable) 7. Systolic Blood Pressure (SBP) 8. Diastolic Blood Pressure (DBP) 9. Hypertension Treatment (coded as HT_trt: 1=yes, 0=no) 10. Total Serum Cholesterol (coded as TSC) 11. Low-density Lipoprotein (LDL) Cholesterol (coded as LDL) 12. High-density Lipoprotein(HDL) Cholesterol (coded as HDL) 13. Cholesterol Treatment (coded as Cholesterol_Trt: yes=1, no=0) 14. Smoking (0=never, 1= used to, 2=active smoker) 15. Number of cigarettes per week (coded as Cigarettes) 16. Diabetes Status (yes=1, no=0) 17. Blood Sugar Level 18. Alcohol Intake (coded as Alcohol, and gives the number of drinks per week) 19. Incidence of recognized and unrecognized myocardial infarction, coronary insufficiency, and coronary heart disease (coded as Hard_CHD: yes=1, no=0)

```r
library(SenSrivastava)
library(car)
library(ggplot2)
library(ggpubr)
library(whitestrap)
library(lmtest)
library(olsrr)
library(corrplot)
library(heatmaply)
library(gplots)
library(GGally)
library(tidyverse)
library(olsrr)
library(MASS)
library(Hmisc)

setwd("~/")
```

R Markdown

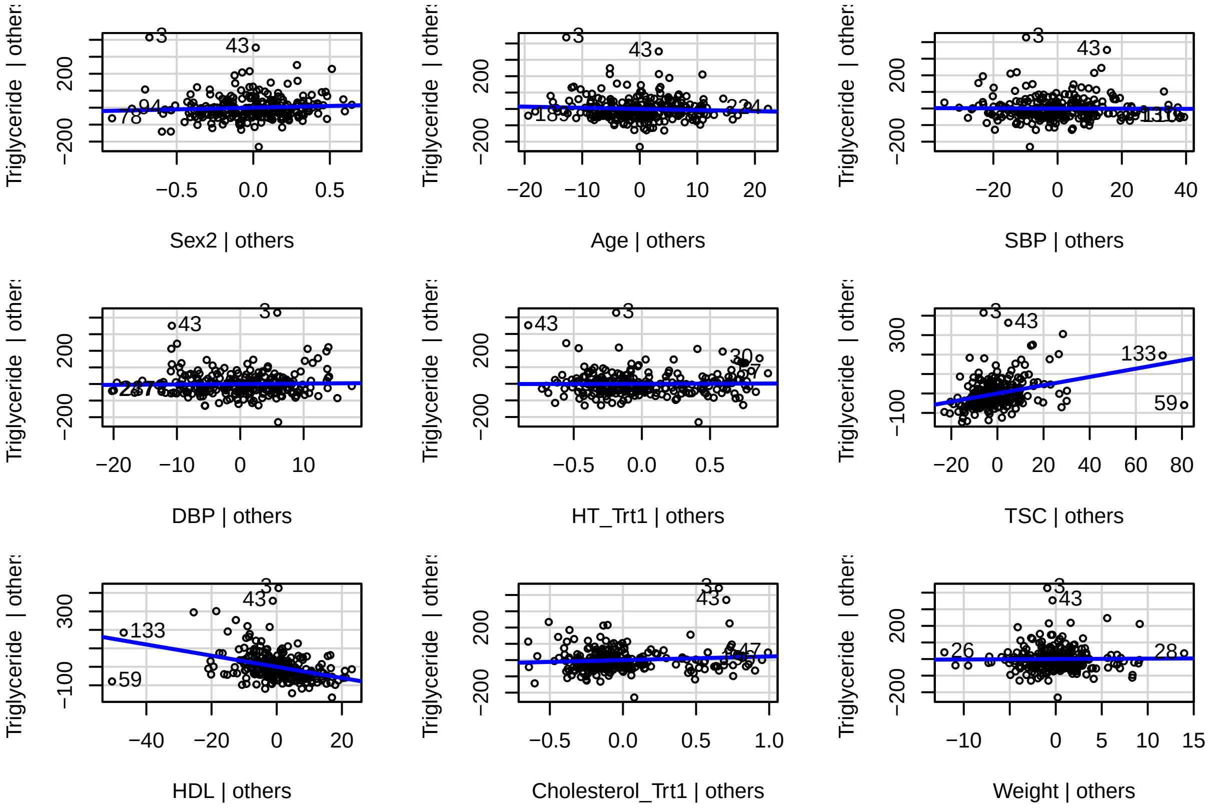
```r
#Read file in folder
df <- read.csv("Triglyceride.csv")
#Changes independent variables to character
df$Sex <- as.character(df$Sex)
df$HT_Trt <- as.character(df$HT_Trt)
```

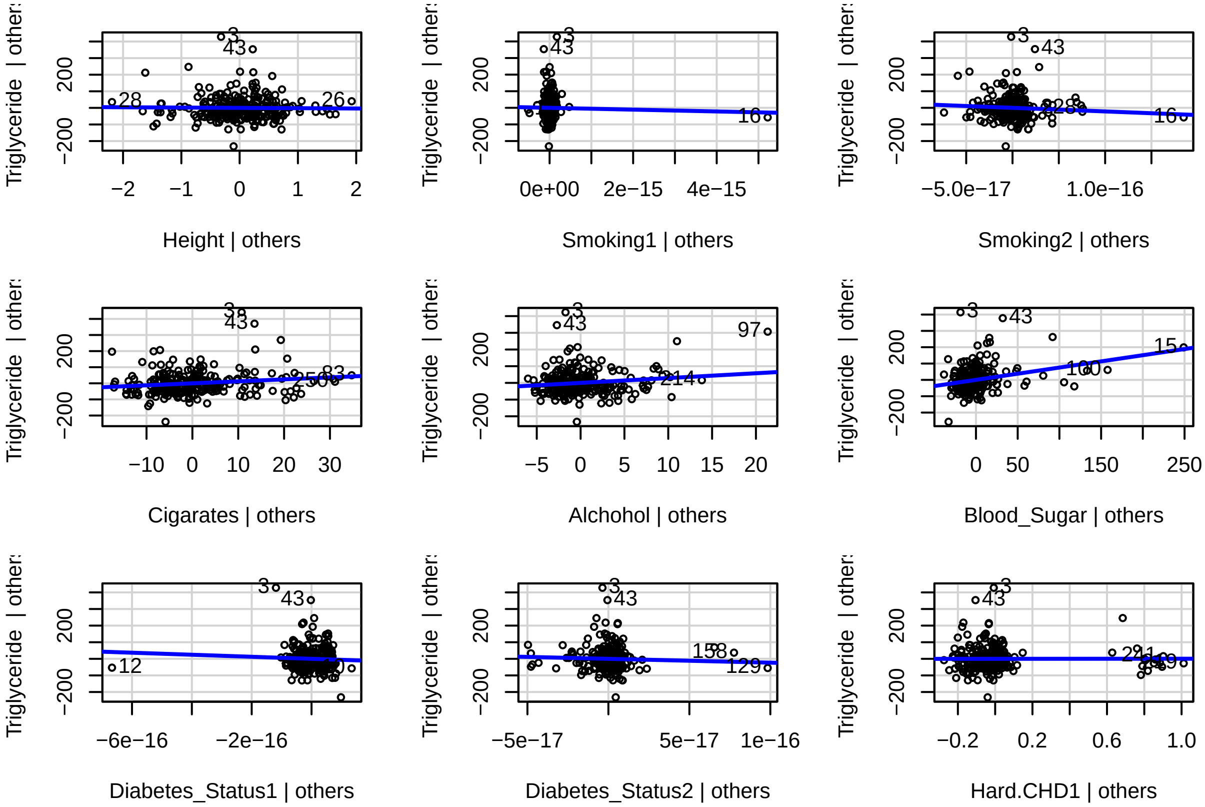
```
df$Cholesterol_Trt <- as.character(df$Cholesterol_Trt)
df$Smoking <- as.character(df$Smoking)
df$Diabetes_Status <- as.character(df$Smoking)
df$Hard.CHD <- as.character(df$Hard.CHD)

#Creating Linear Model
lm_full <- lm(Triglyceride ~ ., data= df)
summary(lm_full)
```

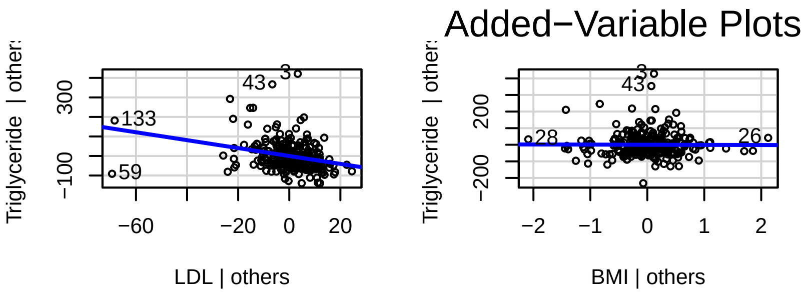
```
##
## Call:
## lm(formula = Triglyceride ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -231.76  -44.34   -9.10   30.80  427.86
##
## Coefficients: (2 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      164.23567  548.88433   0.299   0.7650
## Sex2              20.16817   18.07925   1.116   0.2658
## Age               -0.70011    0.68591  -1.021   0.3085
## SBP               -0.05583    0.35473  -0.157   0.8751
## DBP                0.25926    0.67103   0.386   0.6996
## HT_Trt1            1.61856   12.39117   0.131   0.8962
## TSC                2.12529    0.41713   5.095 7.24e-07 ***
## HDL               -3.01789    0.50769  -5.944 1.02e-08 ***
## Cholesterol_Trt1  22.36858   14.39193   1.554   0.1215
## Weight             0.25479    1.46129   0.174   0.8617
## Height            -1.85263    8.14548  -0.227   0.8203
## Smoking1         -11.59881   11.82797  -0.981   0.3278
## Smoking2         -41.50346   18.60217  -2.231   0.0266 *
## Cigarates          1.22992    0.52861   2.327   0.0208 *
## Alchohol           2.88538    1.36866   2.108   0.0361 *
## Blood_Sugar        0.75534    0.16927   4.462 1.27e-05 ***
## Diabetes_Status1        NA         NA      NA       NA
## Diabetes_Status2        NA         NA      NA       NA
## Hard.CHD1          0.53145   21.63147   0.025   0.9804
## LDL               -2.02526    0.45536  -4.448 1.35e-05 ***
## BMI               -0.84865    8.98371  -0.094   0.9248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.01 on 231 degrees of freedom
## Multiple R-squared:  0.3475, Adjusted R-squared:  0.2967
## F-statistic: 6.836 on 18 and 231 DF,  p-value: 9.873e-14
```

The amount of data points (250) indicates the data analysis test and variance are somewhat accurate There are 6 categorical variables in the data set, Sex, Hypertension Treatment, Cholesterol Treatment, Smoking(0,1,2), Diabetes Status and Hard_CHD. A simple linear regression model was created and analyzed. Based on the p-values in the linear model, independent variables that are significant are Cholesterol, HDL active smokers, cigarettes , Alcohol and Blood_sugar There appears to be much multicollinearity between the independent variables.

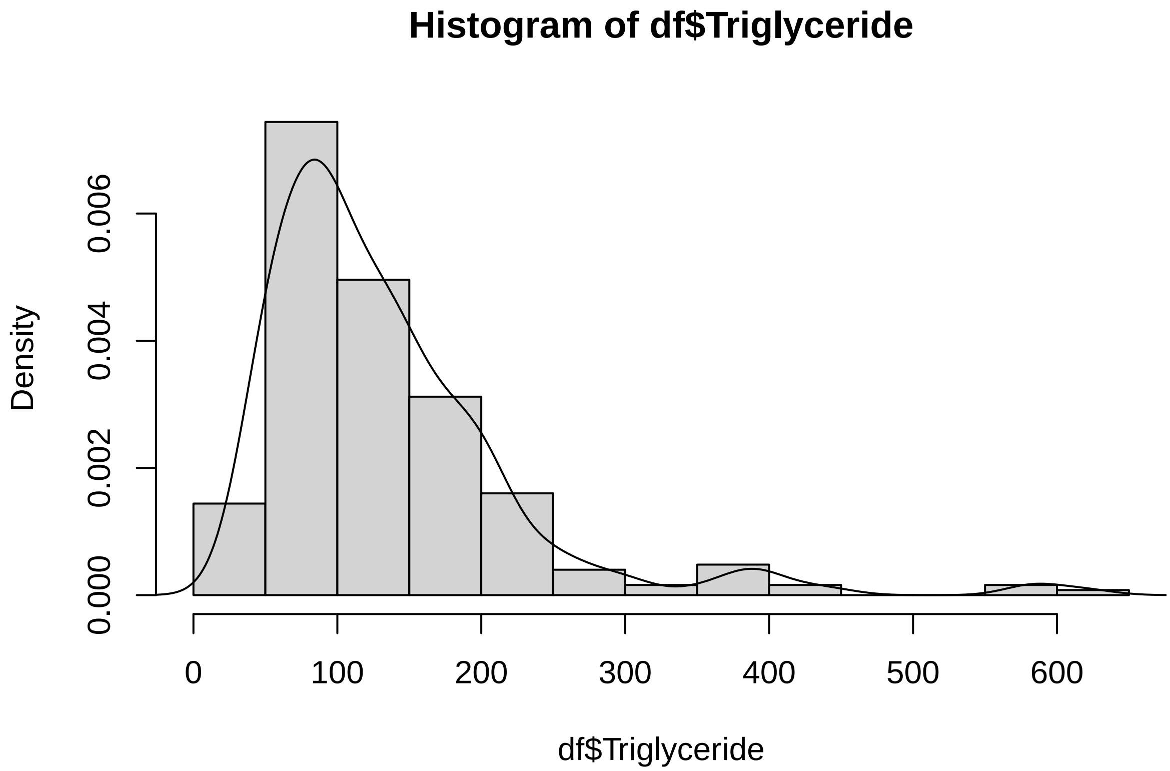
```
#creates plots of independent variables on Tricylecride
avPlots(lm_full)
```




Triglyceride | other:
200
−200
3
43
28
26
−2
−1
0
1
2
Height | others
0e+00
2e−15
4e−15
16
Smoking1 | others
−5.0e−17
1.0e−16
Smoking2 | others
−10
0
10
20
30
Cigarates | others
−5
0
5
10
15
20
97
214
Alchohol | others
0
50
150
250
15
100
Blood_Sugar | others
12
−6e−16
−2e−16
Diabetes_Status1 | others
−5e−17
5e−17
1e−16
158
129
Diabetes_Status2 | others
−0.2
0.2
0.6
1.0
Hard.CHD1 | others

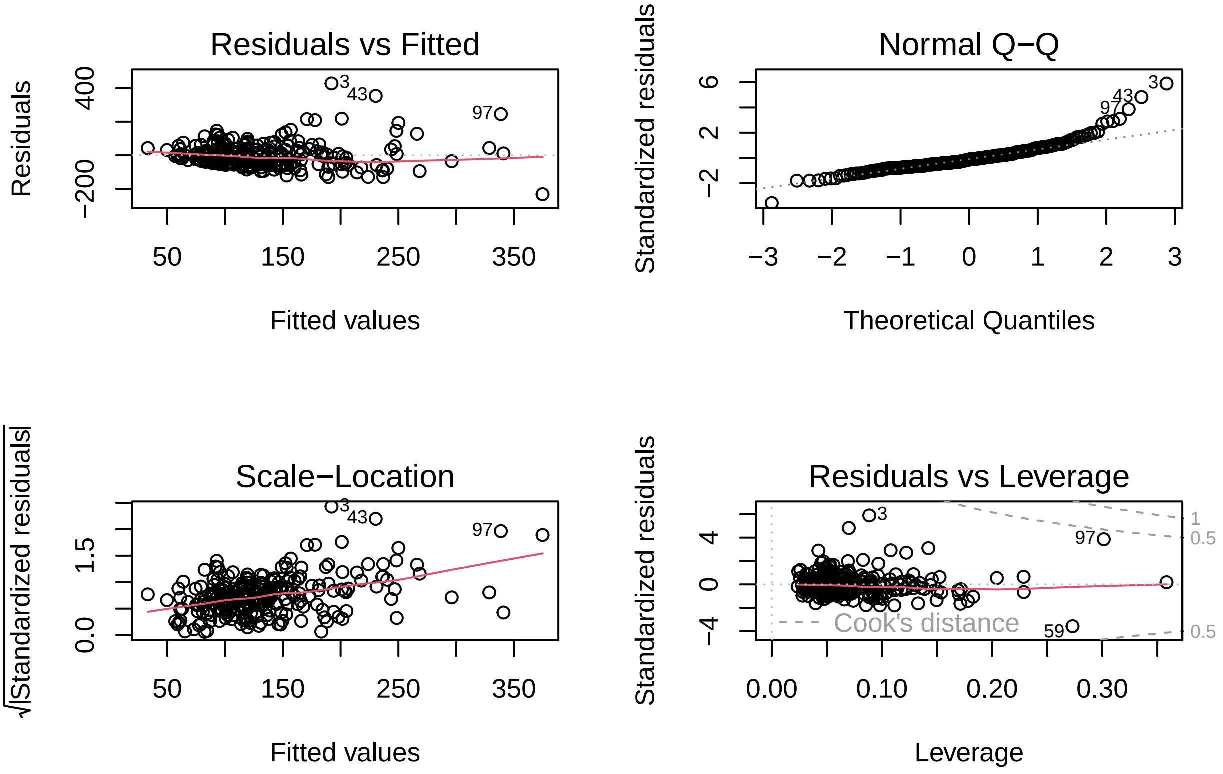


```r
# Histrogram plot on density
hist(df$Triglyceride,freq = F,prob = T)
lines(density(df$Triglyceride))
```



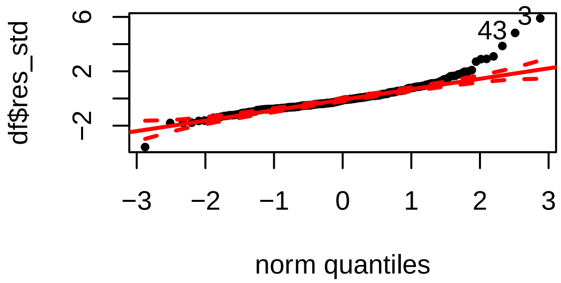
All independent variables are then plotted against Triglyceride to determine whether or not a correlation exist. Based on the plots, variables that appear to affect Triglyceride level are Cholesterol that is increasing, HDL decreasing, Smoking 2 decreasing, cigarettes increasing, Alcohol increasing, Blood_sugar increasing. By viewing the histogram, It appears the highest frequency of Triglyceride is around 50-100. Reaching its peak there and slowly lowering until 300. At which point it switches between up and down. Therefore the data is right skewed. Indicating the data is not normally distributed. There are also outliers when Triglyceride is around 600 .

```r
#Creates residual , normal QQ and standardized residuals plots
par(mfrow=c(2,2))
plot(lm_full)
```



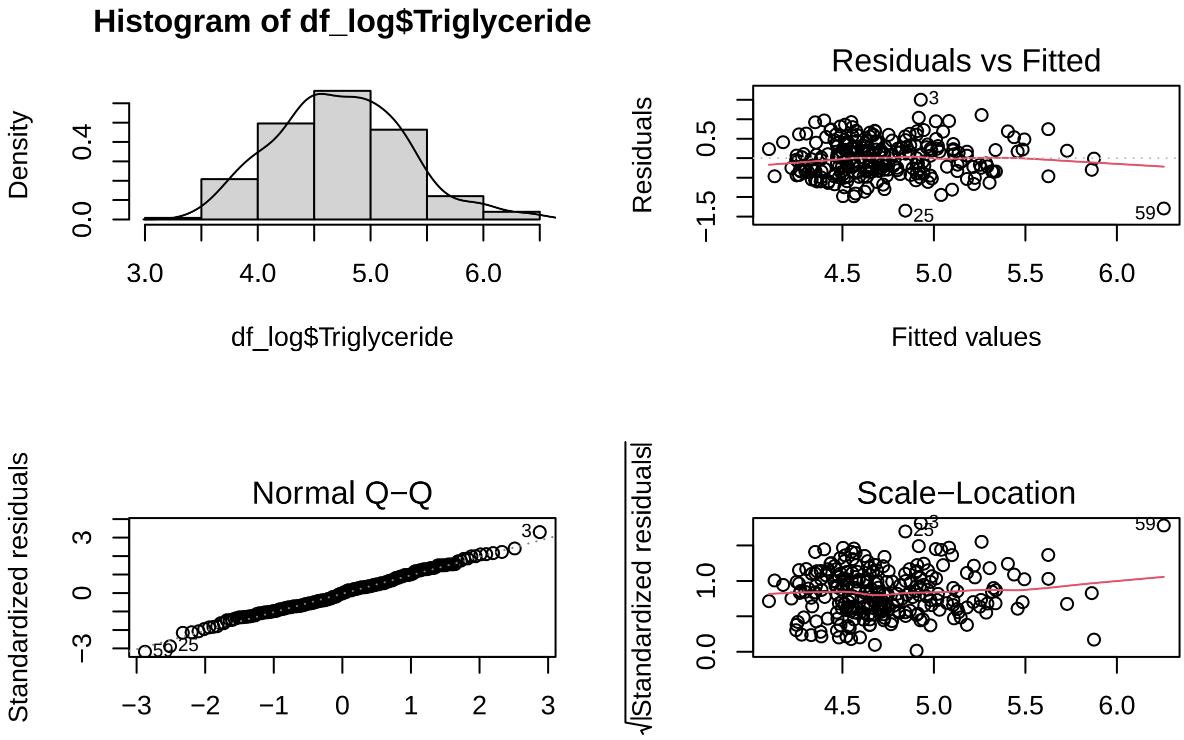
```
df$res<-residuals(lm_full)
df$res_std<-rstandard(lm_full)
qqPlot(df$res_std, envelope=list(style="lines"),col.lines="red",
id=T,pch=20,grid=F,lwd=2)
```

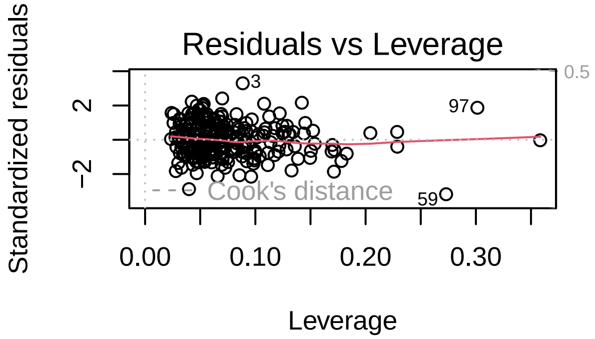
```
## [1]  3 43
```



The first gauss-markov condition appears to be violated because of many outlines (3,43,97) in the residual plot. The assumption of constant variance is also unsatisfied since there appears to be more variance towards the middle values of Triglyceride . There is a semi straight line going through the middle of the residuals. Which verifies the linearity assumption. Based on the qq plot, we can see that the data is mostly normally distributed except for the high tail end (3,43,96). At which the residuals quickly increase. Due to the highly skewed distribution at the tail end increasing, a log based transformation is considered. The log transformation can stabilize the variance since it appears to be proportional to the square of the mean. Error terms appear to be uncorrelated as there is no discernible pattern. Based on the square root of the standardized residuals, we can see the residuals increase with Triglyceride which indicates non-constant variance. The shape of the residuals appears to be uphill in the middle.

```
df_log <- read.csv("Triglyceride.csv")
df_log$Triglyceride <- log(df_log$Triglyceride)
df_log$Sex <- as.character(df$Sex)
df_log$HT_Trt <- as.character(df$HT_Trt)
df_log$Cholesterol <- as.character(df$Cholesterol)
df_log$Smoking <- as.character(df$Smoking)
df_log$Diabetes_Status <- as.character(df$Smoking)
df_log$Hard.CHD <- as.character(df$Hard.CHD)
lm_log <- lm(Triglyceride ~ ., data = df_log)
par(mfrow=c(2,2))
hist(df_log$Triglyceride,freq = F,prob = T)
lines(density(df_log$Triglyceride))
plot(lm_log)
```


Histogram of df_log$Triglyceride
Density
0.0
0.4
3.0
4.0
5.0
6.0
df_log$Triglyceride
Residuals vs Fitted
Residuals
−1.5
0.5
4.5
5.0
5.5
6.0
Fitted values
3
25
59
Normal Q−Q
Standardized residuals
−3
0
3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
3
59
25
Scale−Location
√|Standardized residuals|
0.0
1.0
4.5
5.0
5.5
6.0
Fitted values
25
3
59



After viewing the Density after the log transformation, the data appears to be more normal as the qqplot shows a straight line and the histogram shows more of normal curve when compared to the orginal. The same outliers appear indicating the error is not distributed around 0. The residual line is somewhat straight around 0 and verifies the linearity assumption. Based on the residuals against the fitted values, the plot lines do not follow any pattern therefore, the data appears to have no serial correlation. When viewing the square root of the standardized residuals against the fitted values, the line slighty increases towards the end, indication heteroskedasticity of the residuals. A problem is that their appears to be co linearity. This maybe due to the independent variables representing similar things. For example, smoking and cigarettes are very similar. As if a person does not smoke, their cigarettes count will also be zero.

```
white_test(lm_log)
```

```
## White's test results
##
## Null hypothesis: Homoskedasticity of the residuals
## Alternative hypothesis: Heteroskedasticity of the residuals
## Test Statistic: 7.36
## P-value: 0.025239
```

White's test p-value = to $0.025239 < 0.05$ indicates there is heteroskedasticity of the residuals and varies with expected value of a data point. Therefore the variance of the residuals increases with the fitted values.

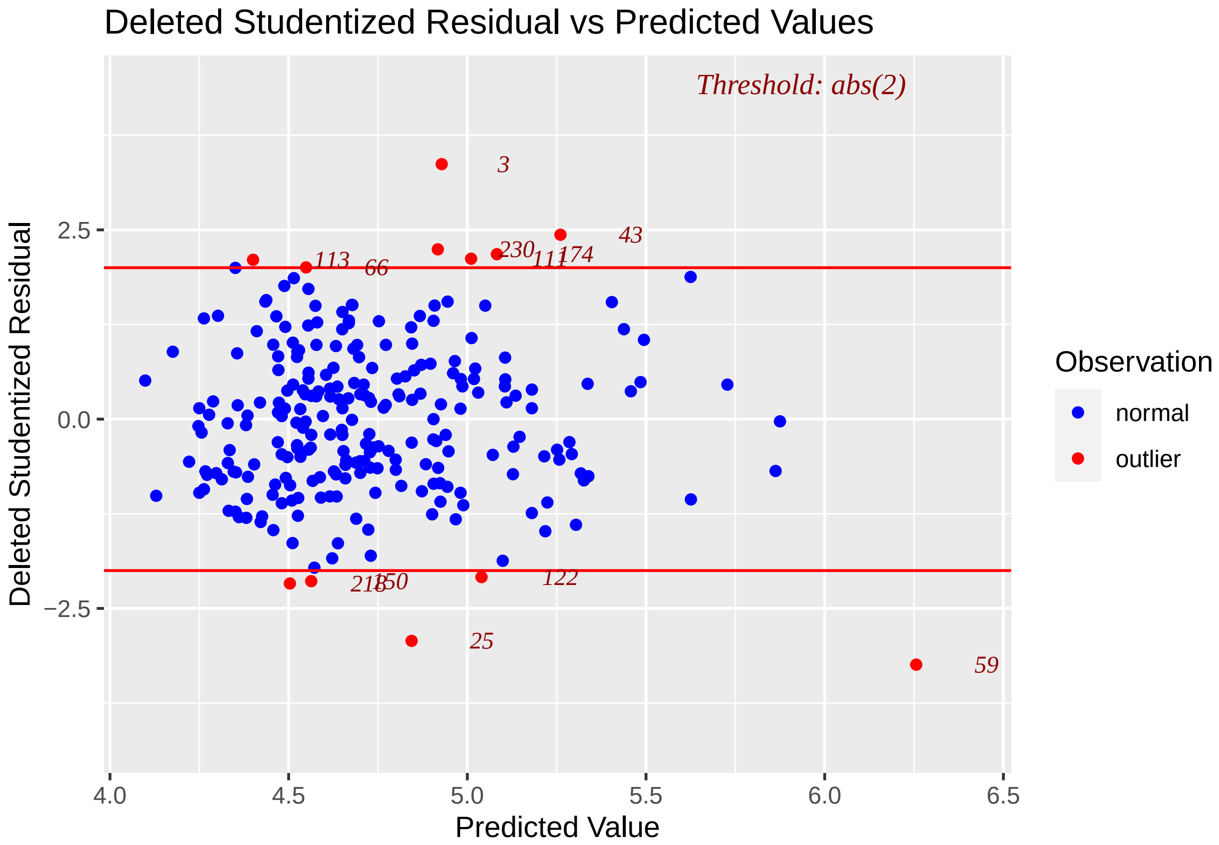
```
dwtest(lm_log, alternative = "two.sided")
```

```
##
```

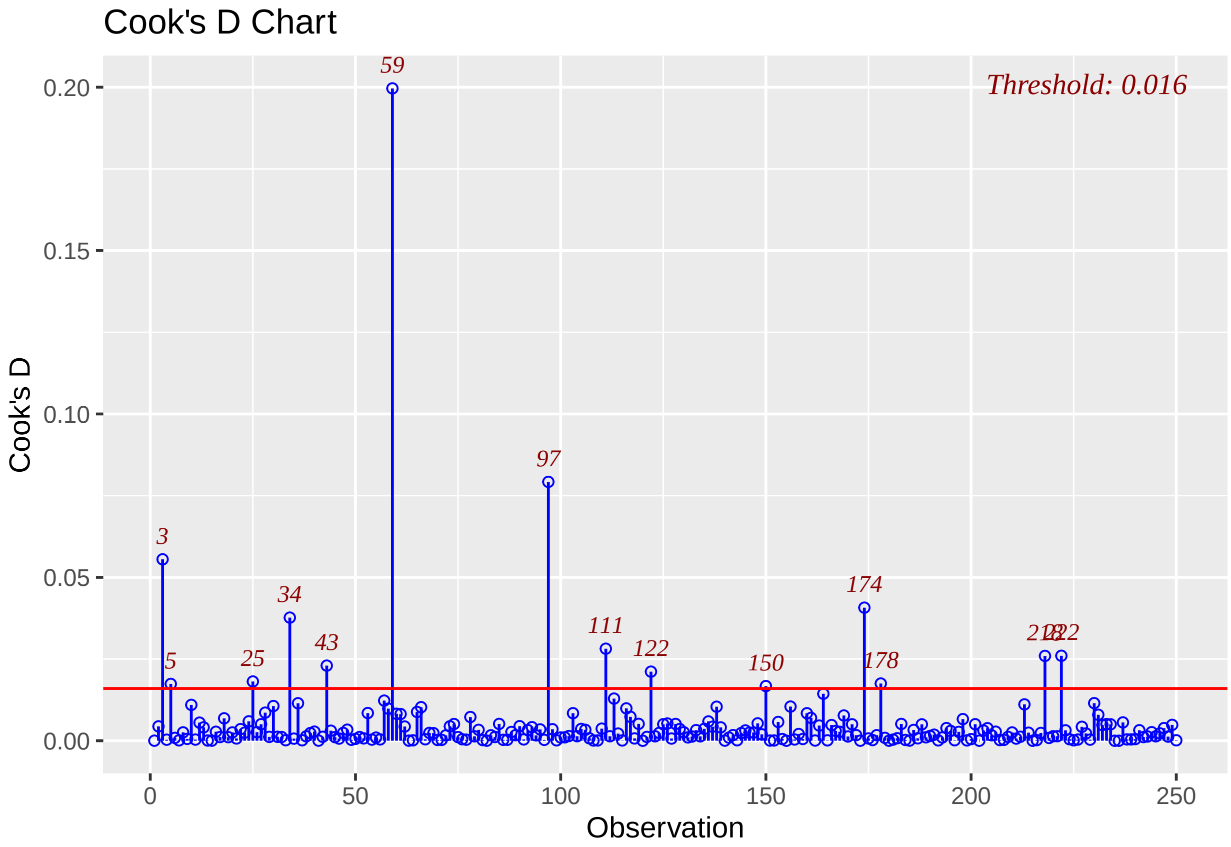
```
##  Durbin-Watson test
##
## data:  lm_log
## DW = 2.006, p-value = 0.3345
## alternative hypothesis: true autocorrelation is not 0
```

Since the Durbin- watson test has a value of 2.006 and pvalue = to 0.3345>0.05, there is no evidence that the true autocorrelation is not zero. Therefore generalized least squares is not necessary. It appears weighted regression is also not needed.

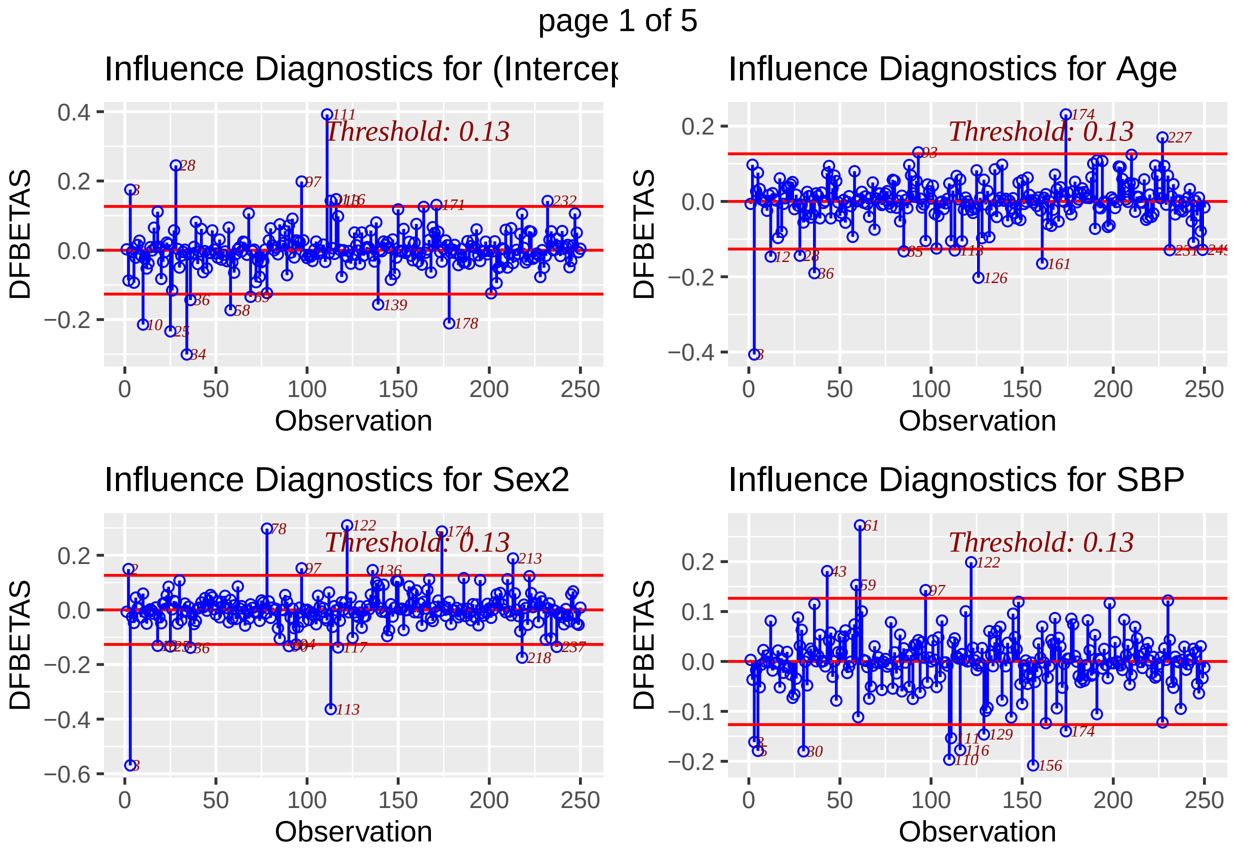
```
ols_plot_resid_stud_fit(lm_log)
```



```
ols_plot_cooksd_chart(lm_log)
```

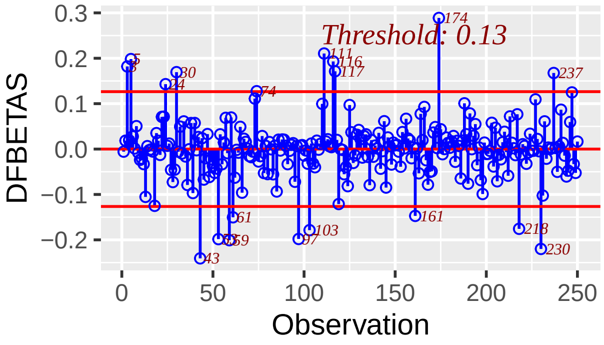


```
ols_plot_dfbetas(lm_log)
```

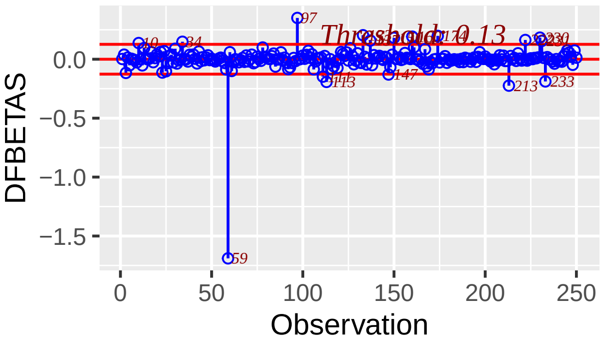

page 1 of 5
Influence Diagnostics for (Intercep
DFBETAS
Observation
Threshold: 0.13
0.4
0.2
0.0
−0.2
0
50
100
150
200
250
111
28
3
97
116
171
232
10
25
34
36
58
69
139
178
Influence Diagnostics for Age
DFBETAS
Observation
Threshold: 0.13
0.2
0.0
−0.2
−0.4
174
227
93
12
28
36
85
113
126
161
231
249
3
Influence Diagnostics for Sex2
DFBETAS
Observation
Threshold: 0.13
0.2
0.0
−0.2
−0.4
−0.6
78
122
174
97
136
213
36
104
117
218
237
113
3
Influence Diagnostics for SBP
DFBETAS
Observation
Threshold: 0.13
0.2
0.1
0.0
−0.1
−0.2
61
43
59
122
97
30
111
116
110
129
174
156
3
5

page 2 of 5

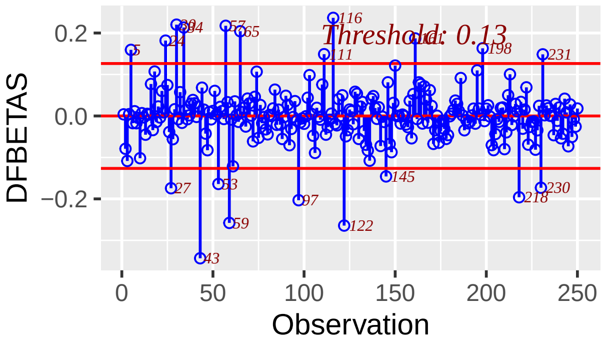
Influence Diagnostics for DBP



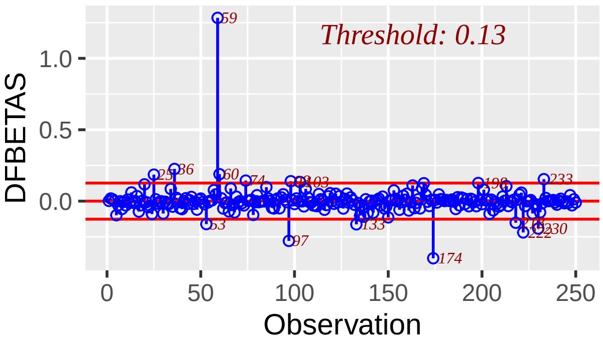
Influence Diagnostics for TSC



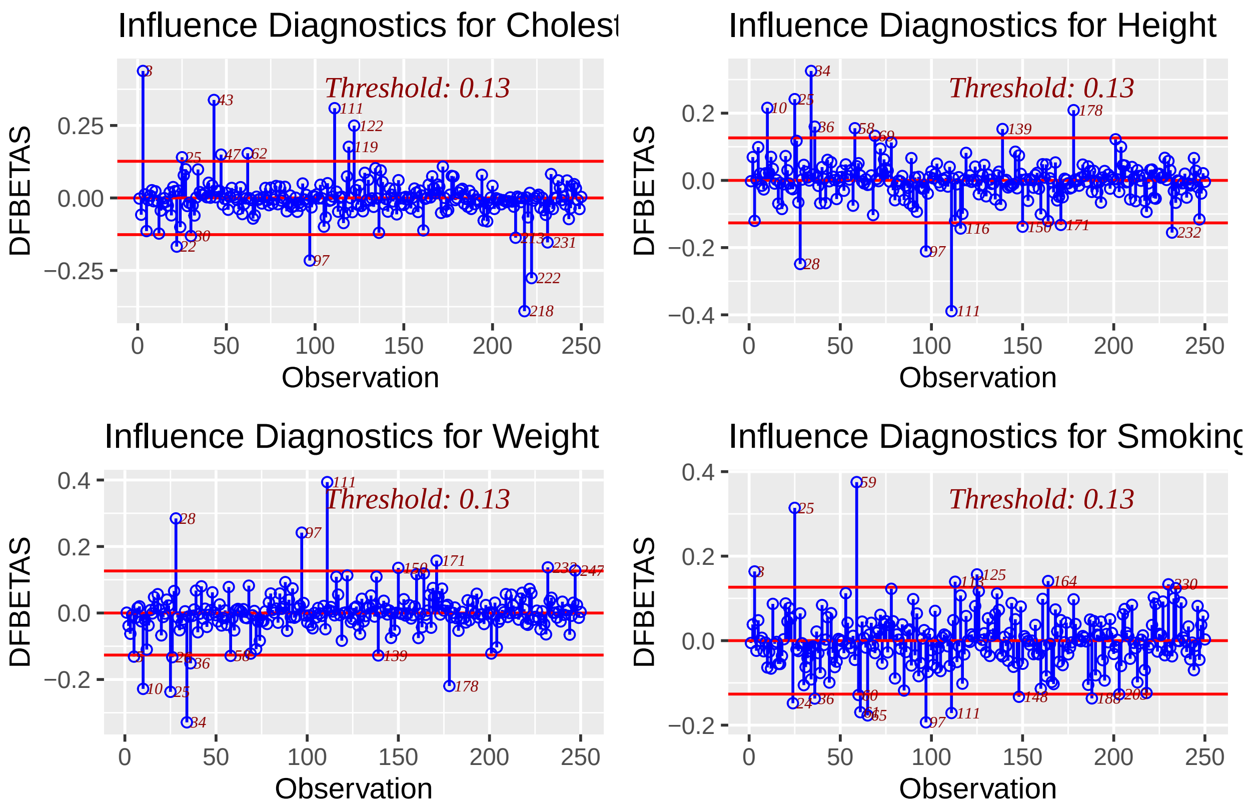
Influence Diagnostics for HT_Trt1



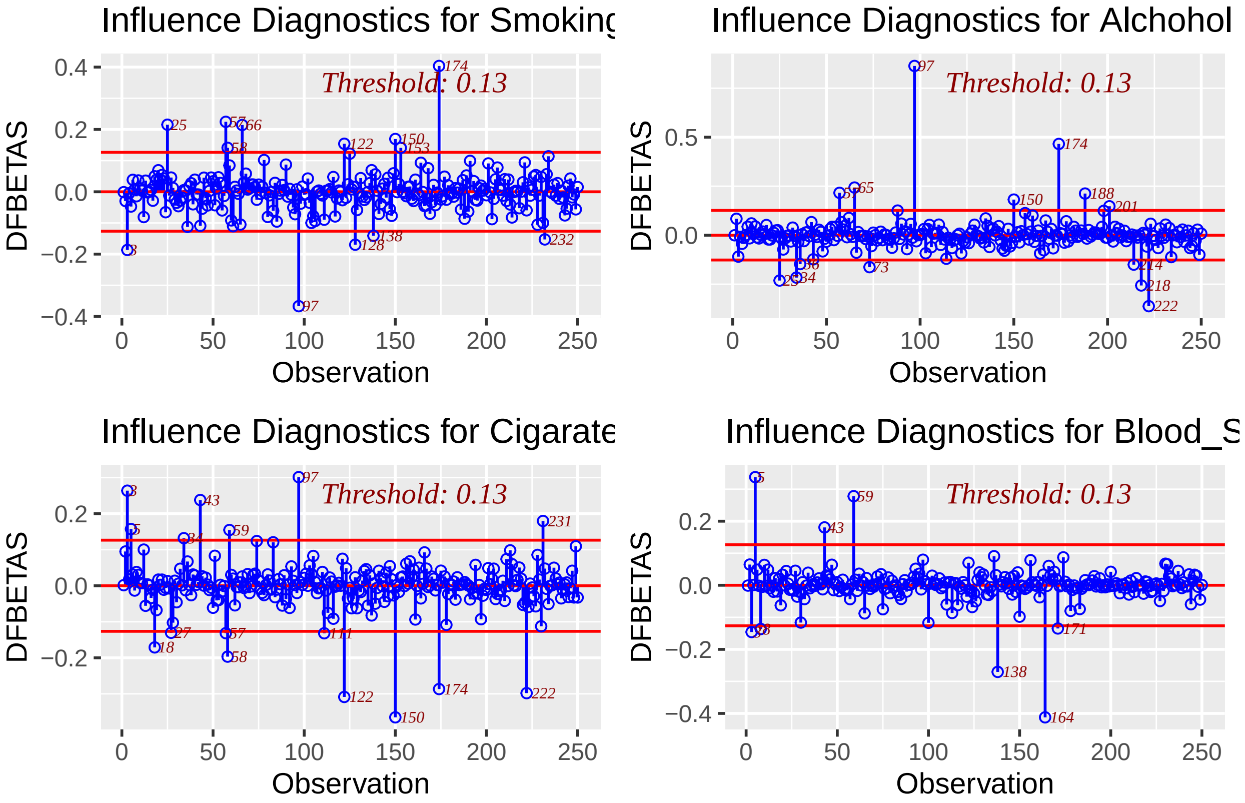
Influence Diagnostics for HDL

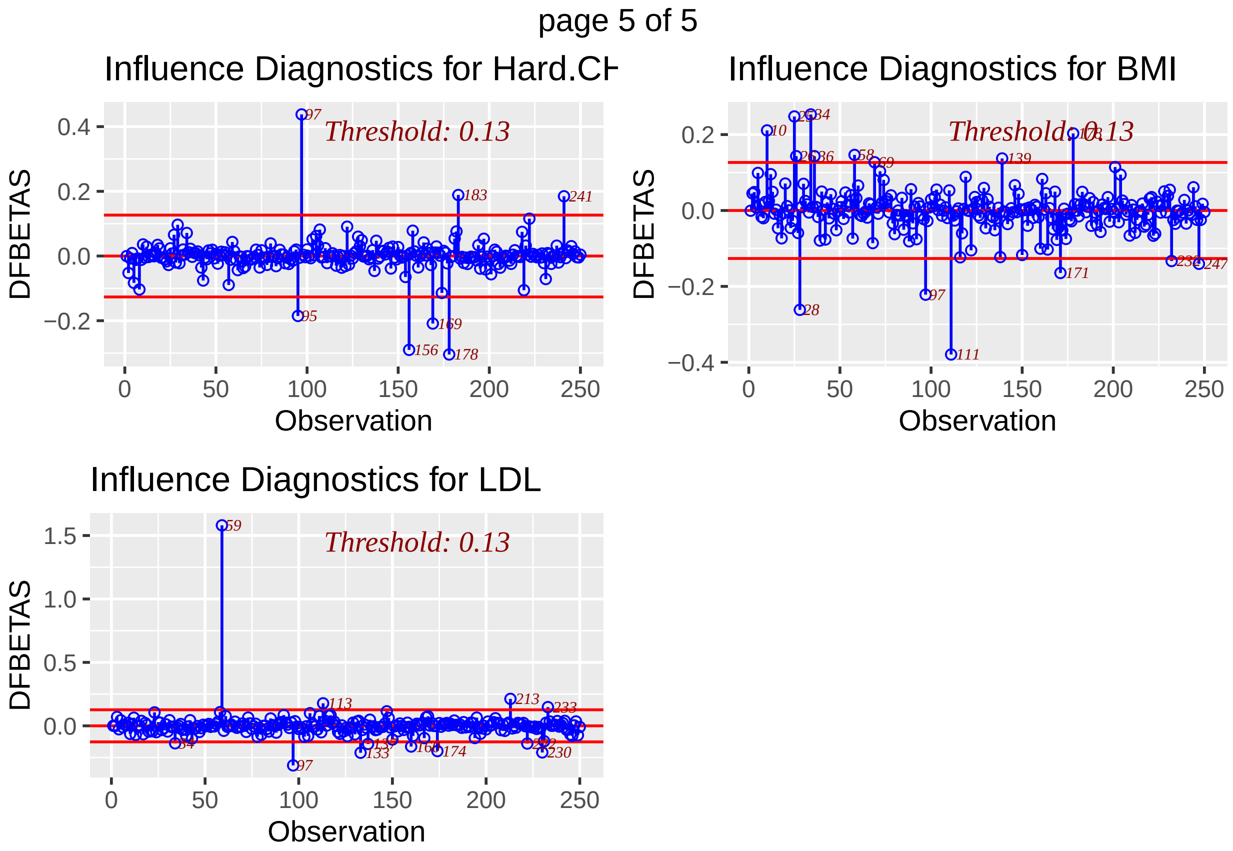


page 3 of 5

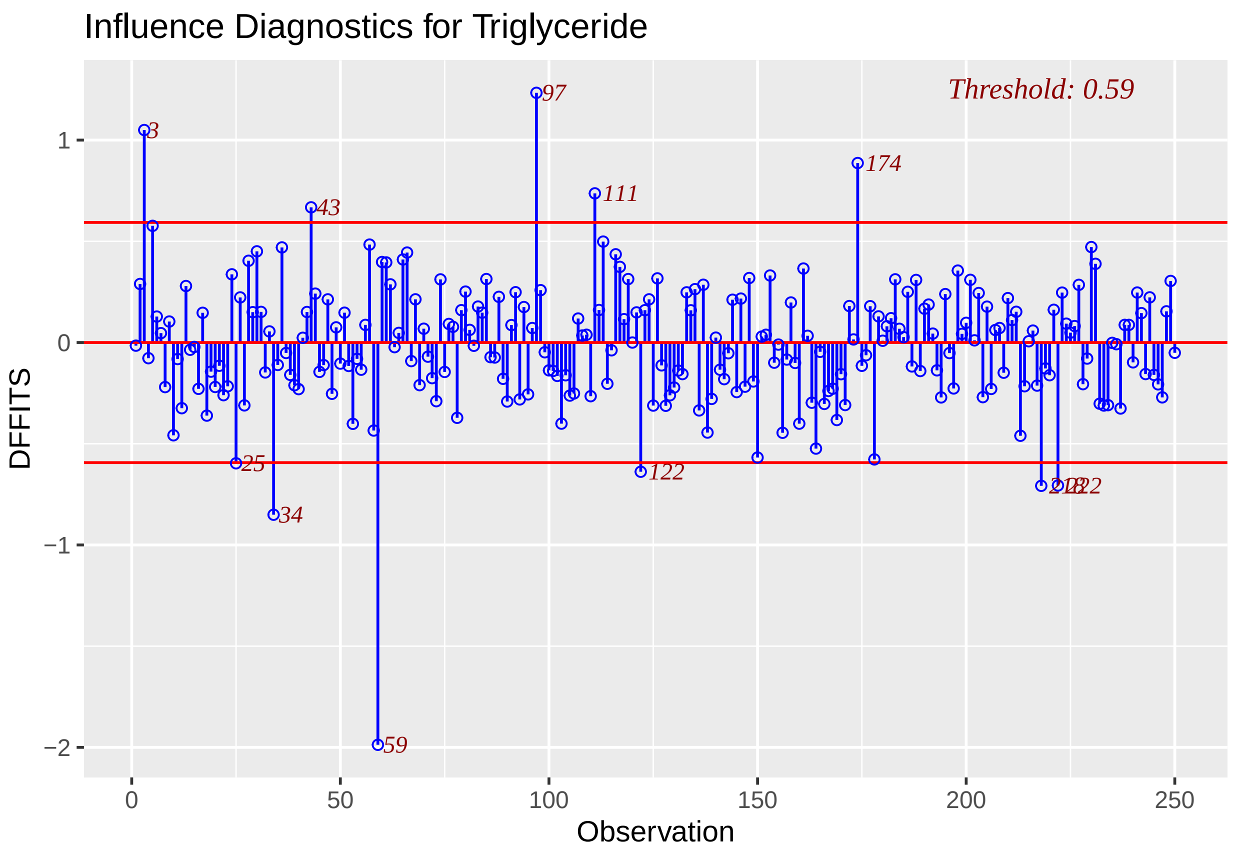

Influence Diagnostics for Cholest
Threshold: 0.13
DFBETAS
0.25
0.00
−0.25
Observation
0
50
100
150
200
250
3
43
111
122
119
25
47
62
22
30
97
213
231
222
218
Influence Diagnostics for Height
Threshold: 0.13
DFBETAS
0.2
0.0
−0.2
−0.4
Observation
34
10
25
36
58
59
139
178
97
28
116
111
156
171
232
Influence Diagnostics for Weight
Threshold: 0.13
DFBETAS
0.4
0.2
0.0
−0.2
Observation
111
28
97
150
171
230
247
10
25
34
36
56
139
178
Influence Diagnostics for Smoking
Threshold: 0.13
DFBETAS
0.4
0.2
0.0
−0.2
Observation
59
25
3
118
125
164
230
24
36
60
65
97
111
148
188
208

page 4 of 5

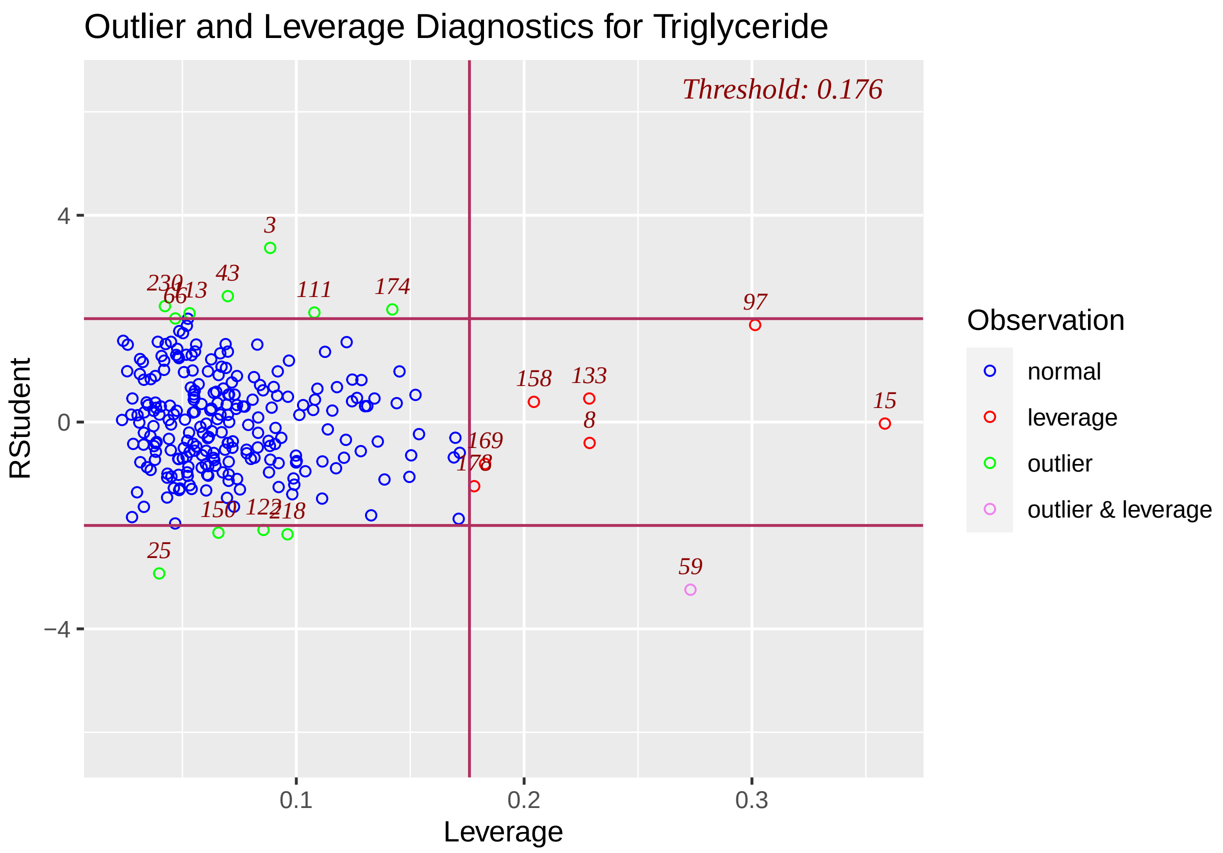

Influence Diagnostics for Smoking
Threshold: 0.13
DFBETAS
0.4
0.2
0.0
−0.2
−0.4
0
50
100
150
200
250
Observation
174
25
57
66
58
122
150
153
3
128
138
232
97
Influence Diagnostics for Alchohol
Threshold: 0.13
DFBETAS
0.5
0.0
97
174
65
150
188
201
23
34
38
73
214
218
222
Observation
Influence Diagnostics for Cigarate
Threshold: 0.13
DFBETAS
0.2
0.0
−0.2
3
43
97
5
59
34
231
27
18
57
58
111
122
150
174
222
Observation
Influence Diagnostics for Blood_S
Threshold: 0.13
DFBETAS
0.2
0.0
−0.2
−0.4
5
59
43
8
171
138
164
Observation



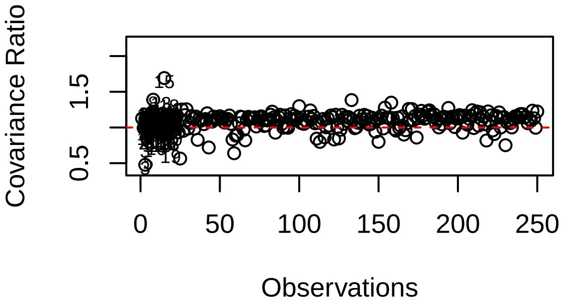
```
ols_plot_dffits(lm_log)
```



```
ols_plot_resid_lev(lm_log)
```



```r
df_log$DFBETAS<-dfbeta(lm_log)
df_log$DFFITS<-dffits(lm_log)
df_log$CookD<-cooks.distance(lm_log)
df_log$Lev<-hatvalues(lm_log)
df_log$CVR<-covratio(lm_log)
ind=1:20
par(mfrow=c(2,2))
plot(df_log$CVR,pch=1,xlab="Observations",ylab="Covariance Ratio",ylim=c(0.4,2.2))
abline(h=1,col="red",lty=2)
text(ind,df_log$CVR-0.05,labels=ind,cex=0.75)
```



There are many observations that require looking at since they are potential outliers (3 45,174,111,2,30,115,66,122,21,150,25,59) that require more attention and researched. Based on Cook's distance observations that have a significant influence on the least square estimates and predict/fitted values (59,3,5,25,34,45,97,111,122,150,174,178,212,222), 59 having the largest influence From the rstudent vs leverage graph, 59 is a value that is both high leverage and a large influencer on the least square estimates and predicted/fitted values. Based on the covariance plot, there are many observations that have a negative impact on the precision of the regression model, including the flag ones. These values require special attention as they may come from source of measuring error. Such as faulty measurement, analysis and others. Perhaps these errors are important information as some people are more affected by the independent variables and others. A few outliers cause the residual plot to not be completely random and have a variance sigma^2 around the 0. There should be further analysis on the outliers from experts in the field to support the deletion of outliers from data. From the covariance plot, it is showed that are many observations that have a negative impact on the regression model(3,5,19).

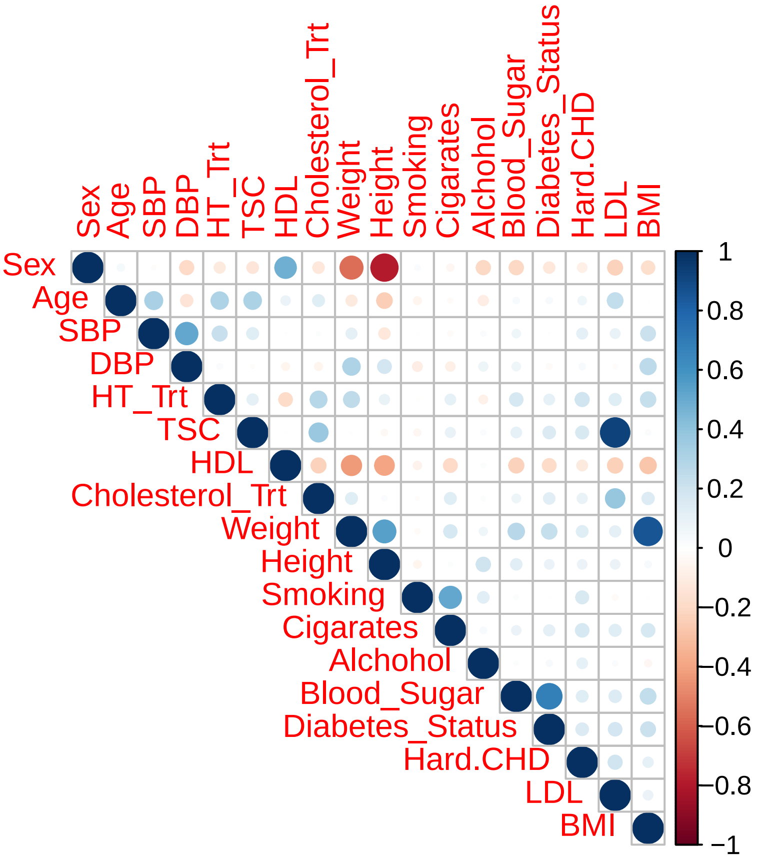
```
df.numeric <- read.csv("Triglyceride.csv")
df.numeric$Triglyceride <- log(df.numeric$Triglyceride)
X <- df.numeric [,2:19]
X <- as.matrix(X)

round(eigen(t(X)%*%X)$values,4)
```

```
##  [1] 3.239269e+07 4.925228e+05 2.071889e+05 1.892740e+05 5.839984e+04
##  [6] 3.069137e+04 2.129336e+04 1.626214e+04 1.287782e+04 6.584698e+03
## [11] 3.320247e+03 1.219486e+03 7.561570e+01 4.631840e+01 2.587920e+01
## [16] 1.912530e+01 1.334670e+01 1.165160e+01
```

```r
rho <- cor(X)
round(rho,2)
```

```
##                   Sex   Age   SBP   DBP HT_Trt   TSC   HDL Cholesterol_Trt
## Sex              1.00  0.05 -0.01 -0.20  -0.12 -0.13  0.48           -0.13
## Age              0.05  1.00  0.33 -0.15   0.31  0.31  0.08            0.14
## SBP             -0.01  0.33  1.00  0.52   0.23  0.14  0.00            0.01
## DBP             -0.20 -0.15  0.52  1.00   0.03 -0.01 -0.06           -0.06
## HT_Trt          -0.12  0.31  0.23  0.03   1.00  0.12 -0.18            0.29
## TSC             -0.13  0.31  0.14 -0.01   0.12  1.00  0.00            0.37
## HDL              0.48  0.08  0.00 -0.06  -0.18  0.00  1.00           -0.22
## Cholesterol_Trt -0.13  0.14  0.01 -0.06   0.29  0.37 -0.22            1.00
## Weight          -0.55 -0.12  0.11  0.31   0.26  0.01 -0.42            0.14
## Height          -0.79 -0.24 -0.12  0.19   0.10 -0.04 -0.40            0.02
## Smoking          0.03 -0.06  0.00 -0.09  -0.01 -0.04 -0.06           -0.01
## Cigarates       -0.05 -0.02 -0.02 -0.08   0.11  0.09 -0.19            0.14
## Alchohol        -0.20 -0.10  0.02  0.08  -0.07  0.02  0.02            0.02
## Blood_Sugar     -0.20  0.03  0.07  0.07   0.18  0.11 -0.23            0.07
## Diabetes_Status -0.12  0.04  0.00 -0.03   0.11  0.15 -0.19            0.13
## Hard.CHD        -0.09  0.07  0.11  0.04   0.20  0.16 -0.11            0.10
## LDL             -0.22  0.25  0.09 -0.02   0.14  0.92 -0.24            0.39
## BMI             -0.18  0.00  0.22  0.27   0.24  0.02 -0.27            0.15
##                 Weight Height Smoking Cigarates Alchohol Blood_Sugar
## Sex              -0.55  -0.79    0.03     -0.05    -0.20       -0.20
## Age              -0.12  -0.24   -0.06     -0.02    -0.10        0.03
## SBP               0.11  -0.12    0.00     -0.02     0.02        0.07
## DBP               0.31   0.19   -0.09     -0.08     0.08        0.07
## HT_Trt            0.26   0.10   -0.01      0.11    -0.07        0.18
## TSC               0.01  -0.04   -0.04      0.09     0.02        0.11
## HDL              -0.42  -0.40   -0.06     -0.19     0.02       -0.23
## Cholesterol_Trt   0.14   0.02   -0.01      0.14     0.02        0.07
## Weight            1.00   0.55   -0.02      0.18     0.06        0.27
## Height            0.55   1.00   -0.05      0.02     0.20        0.12
## Smoking          -0.02  -0.05    1.00      0.51     0.13        0.02
## Cigarates         0.18   0.02    0.51      1.00     0.04        0.08
## Alchohol          0.06   0.20    0.13      0.04     1.00        0.02
## Blood_Sugar       0.27   0.12    0.02      0.08     0.02        1.00
## Diabetes_Status   0.24   0.09    0.00      0.12     0.03        0.69
## Hard.CHD          0.13   0.09    0.16      0.18     0.11        0.13
## LDL               0.12   0.08   -0.03      0.14     0.02        0.15
## BMI               0.85   0.04   -0.01      0.18    -0.05        0.25
##                 Diabetes_Status Hard.CHD   LDL   BMI
## Sex                       -0.12    -0.09 -0.22 -0.18
## Age                        0.04     0.07  0.25  0.00
## SBP                        0.00     0.11  0.09  0.22
## DBP                       -0.03     0.04 -0.02  0.27
## HT_Trt                     0.11     0.20  0.14  0.24
## TSC                        0.15     0.16  0.92  0.02
## HDL                       -0.19    -0.11 -0.24 -0.27
## Cholesterol_Trt            0.13     0.10  0.39  0.15
## Weight                     0.24     0.13  0.12  0.85
## Height                     0.09     0.09  0.08  0.04
## Smoking                    0.00     0.16 -0.03 -0.01
## Cigarates                    0.12      0.18  0.14  0.18
## Alchohol                     0.03      0.11  0.02 -0.05
## Blood_Sugar                  0.69      0.13  0.15  0.25
## Diabetes_Status              1.00      0.16  0.18  0.22
## Hard.CHD                     0.16      1.00  0.20  0.10
## LDL                          0.18      0.20  1.00  0.09
## BMI                          0.22      0.10  0.09  1.00
```

```
corrplot(rho,type = "upper")
```



Based on the half heat map, we can see there are many variables that are correlated. For example, it appears sex is correlated with height and weight. This shows that the 2 value in sex indicates a lower mean in weight and height. Hdl also has a negative correlation with height and weight. It appears the following are positively correlated is slightly positively correlated with DBP, weight and height, weight and bmi, diabeties with blood sugar, smoking and cigaretts, Hdl and sex, Cholesterol with cholesterol.

```
model_in=lm(Triglyceride ~ 1,data= df_log)#initial model, with only intercept
model_s3 <- stepAIC(object=model_in,direction="both",scope = formula(lm_log))
```

```
## Start:  AIC=-280.22
## Triglyceride ~ 1
##
##                   Df Sum of Sq    RSS     AIC
## + HDL              1   10.4227 70.426 -312.72
## + Blood_Sugar      1    9.4602 71.389 -309.33
## + LDL              1    6.5491 74.300 -299.34
## + TSC              1    6.3786 74.471 -298.76
## + Cholesterol_Trt  1    5.8667 74.982 -297.05
## + Cholesterol      1    5.8667 74.982 -297.05
## + BMI              1    5.6065 75.243 -296.19
## + Weight           1    4.7973 76.052 -293.51
## + HT_Trt           1    3.4107 77.438 -288.99
## + Cigarates        1    2.2395 78.610 -285.24
## + Sex              1    1.3392 79.510 -282.39
## + Hard.CHD         1    0.9119 79.937 -281.06
## <none>                          80.849 -280.22
## + SBP              1    0.5462 80.303 -279.91
## + Age              1    0.3675 80.482 -279.36
## + DBP              1    0.2187 80.630 -278.90
## + Height           1    0.2158 80.633 -278.89
## + Alchohol         1    0.0000 80.849 -278.22
## + Smoking          2    0.1094 80.740 -276.56
## + Diabetes_Status  2    0.1094 80.740 -276.56
##
## Step:  AIC=-312.72
## Triglyceride ~ HDL
##
##                   Df Sum of Sq    RSS     AIC
## + TSC              1    6.3930 64.033 -334.51
## + Blood_Sugar      1    5.7566 64.670 -332.04
## + LDL              1    3.4243 67.002 -323.18
## + Cholesterol_Trt  1    3.0350 67.391 -321.74
## + Cholesterol      1    3.0350 67.391 -321.74
## + BMI              1    2.4021 68.024 -319.40
## + HT_Trt           1    1.6212 68.805 -316.55
## + Weight           1    0.8391 69.587 -313.72
## + Cigarates        1    0.8083 69.618 -313.61
## + Height           1    0.8065 69.620 -313.60
## + Age              1    0.7679 69.659 -313.46
## <none>                          70.426 -312.72
## + SBP              1    0.5498 69.877 -312.68
## + Hard.CHD         1    0.3592 70.067 -312.00
## + Sex              1    0.2028 70.224 -311.44
## + DBP              1    0.0835 70.343 -311.02
## + Alchohol         1    0.0041 70.422 -310.74
## + Smoking          2    0.1883 70.238 -309.39
## + Diabetes_Status  2    0.1883 70.238 -309.39
## - HDL              1   10.4227 80.849 -280.22
##
## Step:  AIC=-334.51
## Triglyceride ~ HDL + TSC
##
##                   Df Sum of Sq    RSS     AIC
## + Blood_Sugar      1    4.5525 59.481 -350.95
## + LDL              1    2.8963 61.137 -344.09
## + BMI              1    2.2084 61.825 -341.29
## + HT_Trt           1    0.9510 63.082 -336.25
## + Weight           1    0.8000 63.233 -335.66
## + Cholesterol_Trt  1    0.7118 63.322 -335.31
## + Cholesterol      1    0.7118 63.322 -335.31
## + Sex              1    0.7017 63.332 -335.27
## + Height           1    0.6427 63.391 -335.04
## <none>                         64.033 -334.51
## + Cigarates        1    0.4321 63.601 -334.21
## + SBP              1    0.1599 63.874 -333.14
## + DBP              1    0.1014 63.932 -332.91
## + Hard.CHD         1    0.0349 63.999 -332.65
## + Age              1    0.0070 64.026 -332.54
## + Alchohol         1    0.0002 64.033 -332.51
## + Smoking          2    0.1845 63.849 -331.24
## + Diabetes_Status  2    0.1845 63.849 -331.24
## - TSC              1    6.3930 70.426 -312.72
## - HDL              1   10.4371 74.471 -298.76
##
## Step:  AIC=-350.95
## Triglyceride ~ HDL + TSC + Blood_Sugar
##
##                   Df Sum of Sq    RSS     AIC
## + LDL              1    2.8254 56.656 -361.12
## + BMI              1    1.1814 58.300 -353.97
## + Sex              1    1.0909 58.390 -353.58
## + Cholesterol_Trt  1    0.7999 58.681 -352.34
## + Cholesterol      1    0.7999 58.681 -352.34
## + Height           1    0.7887 58.692 -352.29
## + HT_Trt           1    0.4950 58.986 -351.04
## <none>                         59.481 -350.95
## + Cigarates        1    0.3524 59.129 -350.44
## + Weight           1    0.2300 59.251 -349.92
## + SBP              1    0.0812 59.400 -349.29
## + DBP              1    0.0367 59.444 -349.11
## + Age              1    0.0014 59.480 -348.96
## + Alchohol         1    0.0010 59.480 -348.96
## + Hard.CHD         1    0.0002 59.481 -348.95
## + Smoking          2    0.1753 59.306 -347.69
## + Diabetes_Status  2    0.1753 59.306 -347.69
## - Blood_Sugar      1    4.5525 64.033 -334.51
## - TSC              1    5.1888 64.670 -332.04
## - HDL              1    7.0283 66.509 -325.03
##
## Step:  AIC=-361.12
## Triglyceride ~ HDL + TSC + Blood_Sugar + LDL
##
##                   Df Sum of Sq    RSS     AIC
## + Sex              1    1.2220 55.434 -364.57
## + BMI              1    1.2104 55.445 -364.52
## + Cholesterol_Trt  1    0.7232 55.932 -362.33
## + Cholesterol      1    0.7232 55.932 -362.33
## + Height           1    0.5708 56.085 -361.65
## <none>                         56.656 -361.12
## + HT_Trt           1    0.4033 56.252 -360.90
## + Cigarates        1    0.3919 56.264 -360.85
## + Weight           1    0.3111 56.345 -360.49
## + Hard.CHD         1    0.0093 56.646 -359.16
## + SBP              1    0.0085 56.647 -359.16
## + DBP              1    0.0081 56.648 -359.15
## + Age              1    0.0070 56.649 -359.15
## + Alchohol         1    0.0002 56.655 -359.12
## + Smoking          2    0.1917 56.464 -357.96
## + Diabetes_Status  2    0.1917 56.464 -357.96
## - LDL              1    2.8254 59.481 -350.95
## - Blood_Sugar      1    4.4815 61.137 -344.09
## - TSC              1    5.4032 62.059 -340.34
## - HDL              1    9.8011 66.457 -323.23
##
## Step:  AIC=-364.57
## Triglyceride ~ HDL + TSC + Blood_Sugar + LDL + Sex
##
##                   Df Sum of Sq    RSS     AIC
## + Weight           1    1.3490 54.085 -368.73
## + BMI              1    1.2968 54.137 -368.49
## + Cholesterol_Trt  1    0.6584 54.775 -365.56
## + Cholesterol      1    0.6584 54.775 -365.56
## <none>                         55.434 -364.57
## + HT_Trt           1    0.4069 55.027 -364.41
## + Cigarates        1    0.3020 55.132 -363.93
## + DBP              1    0.0966 55.337 -363.00
## + Alchohol         1    0.0820 55.352 -362.94
## + Age              1    0.0242 55.409 -362.68
## + Height           1    0.0174 55.416 -362.65
## + Hard.CHD         1    0.0109 55.423 -362.62
## + SBP              1    0.0059 55.428 -362.60
## + Smoking          2    0.2198 55.214 -361.56
## + Diabetes_Status  2    0.2198 55.214 -361.56
## - Sex              1    1.2220 56.656 -361.12
## - LDL              1    2.9565 58.390 -353.58
## - Blood_Sugar      1    4.8925 60.326 -345.42
## - TSC              1    5.7817 61.215 -341.77
## - HDL              1   10.9619 66.396 -321.46
##
## Step:  AIC=-368.73
## Triglyceride ~ HDL + TSC + Blood_Sugar + LDL + Sex + Weight
##
##                   Df Sum of Sq    RSS     AIC
## + Cholesterol_Trt  1    0.5233 53.561 -369.16
## + Cholesterol      1    0.5233 53.561 -369.16
## <none>                         54.085 -368.73
## + HT_Trt           1    0.1710 53.914 -367.52
## + Cigarates        1    0.1457 53.939 -367.40
## + Alchohol         1    0.1093 53.975 -367.23
## + BMI              1    0.0318 54.053 -366.87
## + Height           1    0.0124 54.072 -366.79
## + SBP              1    0.0085 54.076 -366.77
## + Age              1    0.0036 54.081 -366.74
## + Hard.CHD         1    0.0001 54.084 -366.73
## + DBP              1    0.0001 54.084 -366.73
## + Smoking          2    0.2092 53.875 -365.70
## + Diabetes_Status  2    0.2092 53.875 -365.70
## - Weight           1    1.3490 55.434 -364.57
## - Sex              1    2.2600 56.345 -360.49
## - LDL              1    3.2229 57.308 -356.26
## - Blood_Sugar      1    3.8882 57.973 -353.37
## - TSC              1    6.2543 60.339 -343.37
## - HDL              1    9.9680 64.053 -328.44
##
## Step:  AIC=-369.16
## Triglyceride ~ HDL + TSC + Blood_Sugar + LDL + Sex + Weight +
##     Cholesterol_Trt
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        53.561 -369.16
## - Cholesterol_Trt  1    0.5233 54.085 -368.73
## + Cigarates        1    0.1177 53.444 -367.71
## + Alchohol         1    0.0957 53.466 -367.61
## + HT_Trt           1    0.0668 53.494 -367.47
## + BMI              1    0.0172 53.544 -367.24
## + Age              1    0.0091 53.552 -367.20
## + DBP              1    0.0061 53.555 -367.19
## + Height           1    0.0035 53.558 -367.18
## + SBP              1    0.0027 53.559 -367.17
## + Hard.CHD         1    0.0000 53.561 -367.16
## + Smoking          2    0.2046 53.357 -366.12
## + Diabetes_Status  2    0.2046 53.357 -366.12
## - Weight           1    1.2140 54.775 -365.56
## - Sex              1    2.1084 55.670 -361.51
## - LDL              1    3.1344 56.696 -356.94
## - Blood_Sugar      1    3.9811 57.542 -353.23
## - TSC              1    5.5120 59.073 -346.67
## - HDL              1    8.7126 62.274 -333.48
```

```
model_out=lm(Triglyceride ~ .,data= df.numeric)#initial model with numerical values
ols_step_both_p(model_out)
```

```
##
##                                Stepwise Selection Summary
## ---------------------------------------------------------------------------------------
##                          Added/                   Adj.
## Step      Variable      Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ---------------------------------------------------------------------------------------
##    1         HDL        addition       0.129       0.125    65.2340    398.7462    0.5329
##    2         TSC        addition       0.208       0.202    38.9820    376.9555    0.5092
##    3     Blood_Sugar    addition       0.264       0.255    20.8630    360.5183    0.4917
##    4         LDL        addition       0.299       0.288    10.3770    350.3519    0.4809
##    5         Sex        addition       0.314       0.300     6.9760    346.9006    0.4766
##    6       Weight       addition       0.331       0.315     3.0150    342.7413    0.4718
## ---------------------------------------------------------------------------------------
```

Both addition and deletion variable selection is implemented.By comparing the regression coefficients between the full model and reduced model. Only 5 addition steps were done and no deletions occured. there is only a 0.3492 0.2542 = 0.095 less Rˆ2 . The akialkie's information Criterion values start at 2929.4338 and lowers slighty until the 5th variable Cholesterol_Trt is added. The high AIC value indicates the model is not good. variables Only 5/19 variables are seen to be significant.

```
lm_reduced <- lm(Triglyceride ~ Diabetes_Status + HDL + TSC + LDL + Cholesterol_Trt, data = df_log)
summary(lm_reduced)
```

```
##
## Call:
## lm(formula = Triglyceride ~ Diabetes_Status + HDL + TSC + LDL +
##     Cholesterol_Trt, data = df_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45855 -0.34334  0.00105  0.31957  1.36433
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.647196   0.206999  22.450  < 2e-16 ***
## Diabetes_Status1 -0.063352   0.071496  -0.886 0.376447
## Diabetes_Status2 -0.047867   0.100087  -0.478 0.632897
## HDL              -0.019483   0.002874  -6.780 9.06e-11 ***
## TSC               0.011982   0.002618   4.577 7.51e-06 ***
## LDL              -0.009854   0.002913  -3.383 0.000836 ***
## Cholesterol_Trt   0.144415   0.090254   1.600 0.110878
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4981 on 243 degrees of freedom
## Multiple R-squared:  0.2542, Adjusted R-squared:  0.2357
## F-statistic:  13.8 on 6 and 243 DF,  p-value: 1.724e-13
```