STAT 420 Final project

## R Markdown

#part one  
train = read.csv(file="C:/Users/Hans/Desktop/training.csv", header=TRUE)  
#Log transformation  
fit9 = lm(log(health)~PCDH12+DLG5+BC038559+SHISA5+AF161342+CARKD+F2R+PHKG1+CDCP1+PLEKHM1+SMC2+PSMB6+BX440400+A\_24\_P936373+PPAN+BC007917+C14orf143+LOC440104+THC2578957+ANKIB1,data = train)  
summary(fit9)$r.squared

## [1] 0.650627

AIC(fit9)

## [1] 183.8983

#Box-cox transformation  
library(MASS)  
bc = boxcox(health~PCDH12+DLG5+BC038559+SHISA5+AF161342+CARKD+F2R+PHKG1+CDCP1+PLEKHM1+SMC2+PSMB6+BX440400+A\_24\_P936373+PPAN+BC007917+C14orf143+LOC440104+THC2578957+ANKIB1,data = train,plotit = FALSE)  
lambda = bc$x[which.max(bc$y)]  
health\_1 = ((train$health^lambda)-1)/lambda  
fit = lm((health\_1)~PCDH12+DLG5+BC038559+SHISA5+AF161342+CARKD+F2R+PHKG1+CDCP1+PLEKHM1+SMC2+PSMB6+BX440400+A\_24\_P936373+PPAN+BC007917+C14orf143+LOC440104+THC2578957+ANKIB1,data = train)  
summary(fit)$r.squared

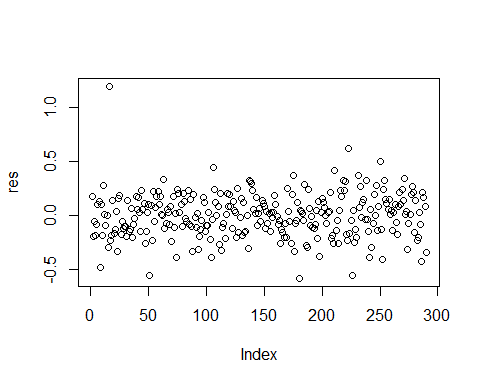
## [1] 0.6478009

AIC(fit)

## [1] -52.15716

# Based on the r-square for both y-transformation, even though box-cox has slight lower r-square than log transformation,   
# but box-cox's AIC and BIC is much smaller than log's, which means that box-cox has better goodness-of-fit, thus, we  
# choose box-cox transformation to do the diagnostics.

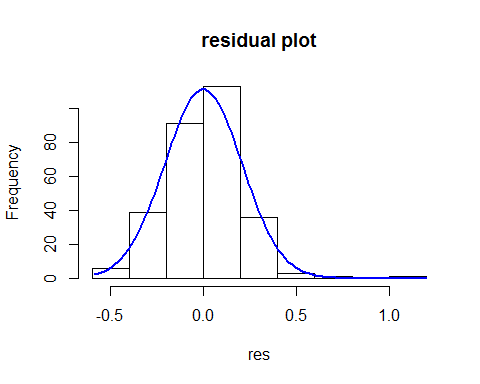
#Residuals and Normality check  
res = fit$residuals  
plot(res)



his = hist(res, main = "residual plot")  
shapiro.test(res)

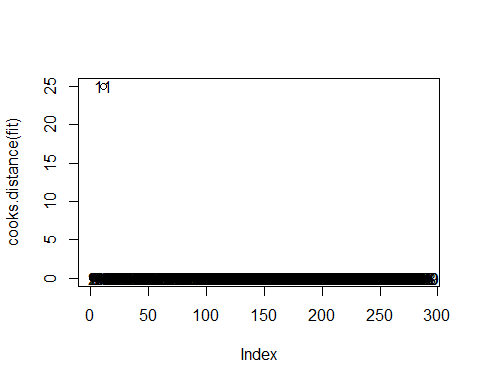
##   
## Shapiro-Wilk normality test  
##   
## data: res  
## W = 0.96694, p-value = 3.325e-06

xgrid =seq(min(res) ,max(res) ,length=100)  
yden =dnorm( xgrid ,mean=0,sd=0.2083)   
yden = yden\*length (res)\* diff (his$mids[1:2])  
lines(xgrid , yden, col="blue", lwd=2)



# Based on the graph shown below, the residuals plots does not have pattern but there are some points  
# far away from the main trend, Shapiro-Wilk normality test is less than 0.05, so it does not follow normality, so we should be careful about them. In addition, the historgram tell  
# us the similar result, which is the plot has slighly left-skewed, so it might need some adjustments.

# Cook's Distance & Outliers  
plot(cooks.distance(fit))  
text(cooks.distance(fit))

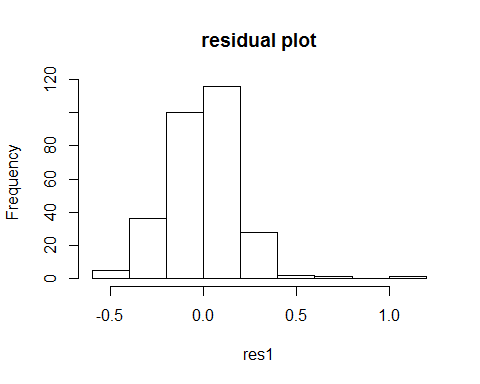


# After normality check, we use cook.distance() method to see which point is most far from other points  
# and we would consider removing it one at a time. The first point that will be toss out is 11.

#Remove point 11  
train1=train[c(-11),]  
health\_1 = ((train1$health^lambda)-1)/lambda  
fit1 = lm(health\_1~train1$PCDH12+DLG5+BC038559+SHISA5+AF161342+CARKD+F2R+PHKG1+CDCP1+PLEKHM1+SMC2+PSMB6+BX440400+A\_24\_P936373+PPAN+BC007917+C14orf143+LOC440104+THC2578957+ANKIB1,data = train1)  
summary(fit1)$r.squared

## [1] 0.682545

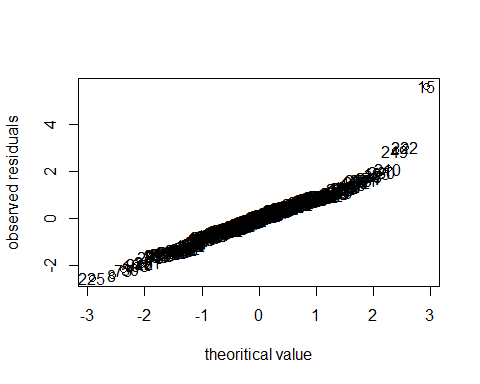
res1 = fit1$residuals  
his1 = hist(res1, main = "residual plot")



shapiro.test(res1)

##   
## Shapiro-Wilk normality test  
##   
## data: res1  
## W = 0.96535, p-value = 2.045e-06

sigma1 = sqrt(sum(fit1$residuals^2)/fit1$df.residual)  
perct1 = (rank(res1) - 0.5)/length(res1)  
z1 = qnorm(perct1)  
plot(z1, res1/sigma1, xlab = "theoritical value", ylab = "observed residuals")  
text(z1, res1/sigma1, c(1:length(res),pos=3))

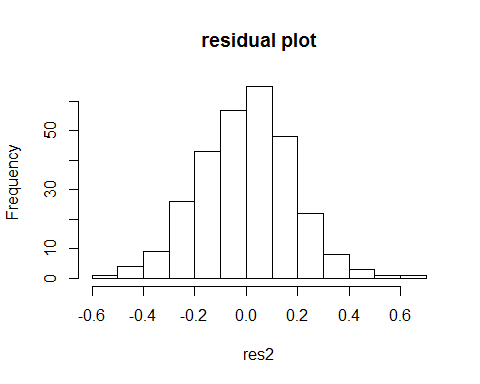


# After removing point 11, the r-squared has obviously increase from 0.6478009 to 0.6875, indicating that this model become better with removing   
# point 11. Also, the QQ plot and histogram look better than the previous one.

#Remove point 15  
train2=train1[c(-15),]  
health\_1 = ((train2$health^lambda)-1)/lambda  
fit2 = lm(health\_1~PCDH12+DLG5+BC038559+SHISA5+AF161342+CARKD+F2R+PHKG1+CDCP1+PLEKHM1+SMC2+PSMB6+BX440400+A\_24\_P936373+PPAN+BC007917+C14orf143+LOC440104+THC2578957+ANKIB1,data = train2)  
summary(fit2)$r.squared

## [1] 0.7154372

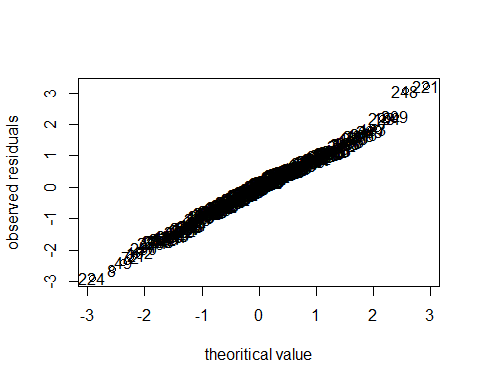
res2 = fit2$residuals  
his2 = hist(res2, main = "residual plot")



shapiro.test(res2)

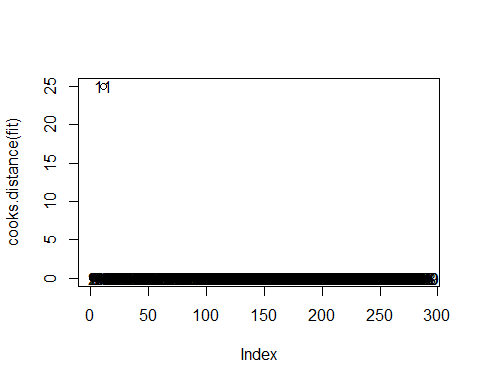
##   
## Shapiro-Wilk normality test  
##   
## data: res2  
## W = 0.99581, p-value = 0.6366

sigma2 = sqrt(sum(fit2$residuals^2)/fit2$df.residual)  
perct2 = (rank(res2) - 0.5)/length(res2)  
z2 = qnorm(perct2)  
plot(z2, res2/sigma2, xlab = "theoritical value", ylab = "observed residuals")  
text(z2, res2/sigma2, c(1:length(res2),pos=3))

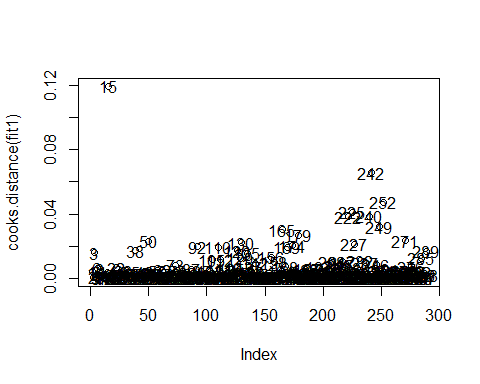


# With taking out point 15, the r-squared have a big increasement and the p-value for normality test is much   
# larger and it is greater than 0.05, which states that there are bigger variation the model could explain and this data is more normal. In addition to r-squared, histgarom and QQ plot look much better than the previous two data.

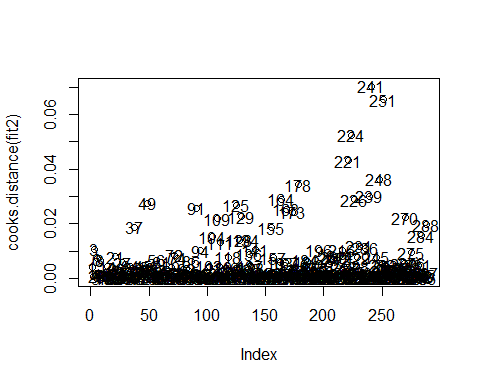
# Codes that located below is for verifying the outlier point that removed above.   
  
# Cook's Distance & Outliers  
plot(cooks.distance(fit))  
text(cooks.distance(fit))



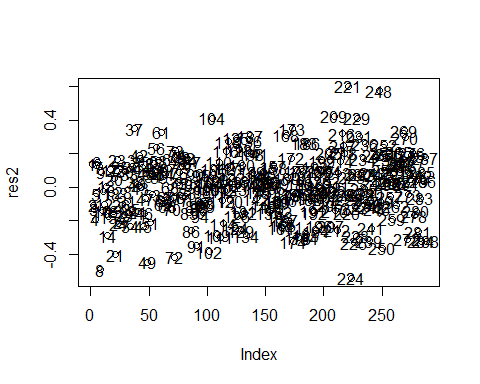
plot(cooks.distance(fit1))  
text(cooks.distance(fit1))



plot(cooks.distance(fit2))  
text(cooks.distance(fit2))



plot(res2)  
text(res2)



# Summary about Part one  
# # After removing outliers  
  
#Histogram plot  
hist(res2)  
  
#Normality test  
shapiro.test(res2)

##   
## Shapiro-Wilk normality test  
##   
## data: res2  
## W = 0.99581, p-value = 0.6366

summary(fit2)

##   
## Call:  
## lm(formula = health\_1 ~ PCDH12 + DLG5 + BC038559 + SHISA5 + AF161342 +   
## CARKD + F2R + PHKG1 + CDCP1 + PLEKHM1 + SMC2 + PSMB6 + BX440400 +   
## A\_24\_P936373 + PPAN + BC007917 + C14orf143 + LOC440104 +   
## THC2578957 + ANKIB1, data = train2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.54242 -0.11932 0.00945 0.11711 0.60291   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.625281 0.119471 -5.234 3.37e-07 \*\*\*  
## PCDH12 -0.286096 0.042841 -6.678 1.40e-10 \*\*\*  
## DLG5 -0.302519 0.030917 -9.785 < 2e-16 \*\*\*  
## BC038559 -0.020688 0.050351 -0.411 0.681492   
## SHISA5 0.165758 0.044592 3.717 0.000245 \*\*\*  
## AF161342 0.136363 0.020579 6.626 1.89e-10 \*\*\*  
## CARKD 0.048259 0.013451 3.588 0.000397 \*\*\*  
## F2R 0.422555 0.063591 6.645 1.70e-10 \*\*\*  
## PHKG1 0.147251 0.032938 4.471 1.15e-05 \*\*\*  
## CDCP1 0.013843 0.025230 0.549 0.583678   
## PLEKHM1 0.732921 0.044501 16.470 < 2e-16 \*\*\*  
## SMC2 0.005896 0.005618 1.050 0.294875   
## PSMB6 -0.013602 0.050328 -0.270 0.787157   
## BX440400 0.006481 0.035993 0.180 0.857239   
## A\_24\_P936373 0.044472 0.027972 1.590 0.113040   
## PPAN 0.092661 0.035900 2.581 0.010383 \*   
## BC007917 -0.038175 0.041770 -0.914 0.361576   
## C14orf143 0.041637 0.015147 2.749 0.006387 \*\*   
## LOC440104 -0.021037 0.043384 -0.485 0.628140   
## THC2578957 0.064086 0.043693 1.467 0.143625   
## ANKIB1 -0.006277 0.046394 -0.135 0.892482   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1878 on 267 degrees of freedom  
## Multiple R-squared: 0.7154, Adjusted R-squared: 0.6941   
## F-statistic: 33.56 on 20 and 267 DF, p-value: < 2.2e-16

# In accordance with the result that summary function generated, the estimated parameter is (Intercept) -0.625281 PCDH12 -0.286096 DLG5 -0.302519 BC038559 -0.020688 SHISA5 0.165758 AF161342 0.136363CARKD 0.048259F2R 0.422555PHKG1 0.147251 CDCP1 0.013843 PLEKHM1 0.732921 SMC2 0.005896 PSMB6 -0.013602 BX440400 0.006481 A\_24\_P936373 0.044472 PPAN 0.092661 BC007917 -0.038175 C14orf143 0.041637 LOC440104 -0.021037 THC2578957 0.064086 ANKIB1 -0.006277 The residual standard error is 0.1878 on 267 degrees of freedom Multiple R-squared: 0.7154, Adjusted R-squared: 0.6941 F-statistic: 33.56 on 20 and 267 DF, p-value: < 2.2e-16 Among those variables, intercept, PCDH12,DLG5,SHISA5, AF161342, CARKD, F2R, PHKG1,PLEKHM1,PPAN,C14orf143   
  
  
# AIC  
AIC(fit2)

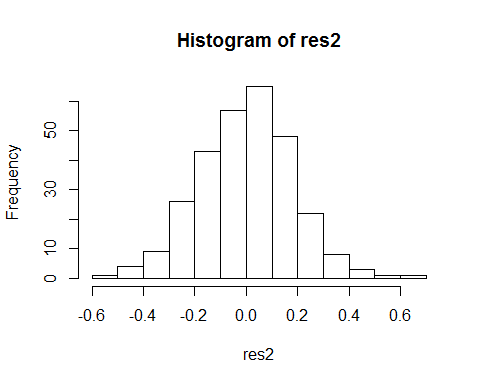
## [1] -123.7967

#BP test  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric



bptest(fit2)

##   
## studentized Breusch-Pagan test  
##   
## data: fit2  
## BP = 21.415, df = 20, p-value = 0.3731

# The histgram looks much better. BP test tells us that the constant variance of the residuals is accepted, indicating that this is a good model fitting.

# AIC & BIC before X transformation and model selection  
AIC(fit)

## [1] -52.15716

BIC(fit)

## [1] 28.58022

AIC(fit2)

## [1] -123.7967

BIC(fit2)

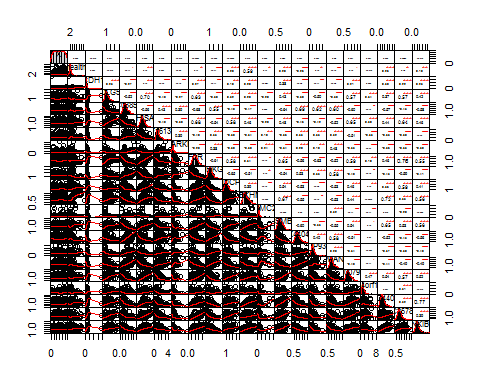
## [1] -43.21155

# Decreasing value means better goodness-of-fit

# PART TWO

library(PerformanceAnalytics)

suppressWarnings(chart.Correlation(train))

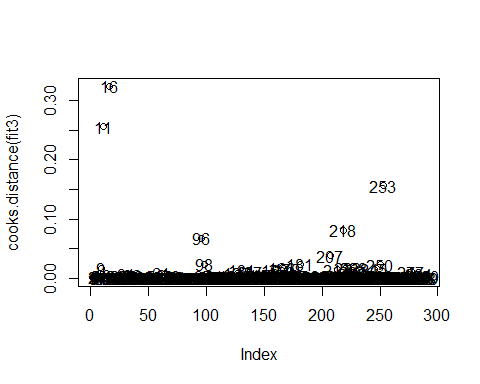


# After taking look at this graph, I decided to do six variables transformation.

# X Transformation  
#log  
PCDH12\_1 = log(train$PCDH12)  
AF161342\_1 = log(train$AF161342+1)  
CARKD\_1 = log(train$CARKD)  
SMC2\_1 = log(train$SMC2)  
A\_24\_P936373\_1 = log(train$A\_24\_P936373)  
C14orf143\_1 = log(train$C14orf143)  
  
fit3 = lm(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train)  
summary(fit3)$r.squared

## [1] 0.5782648

plot(cooks.distance(fit3))  
text(cooks.distance(fit3))



# In the plot of cook's distance that already transform six x variables, I see that both 11 and 16 point are far away from rest points, so removing them one at a time.

sort(cooks.distance(fit3),decreasing = TRUE)[1:3]

## 16 11 253   
## 0.3243052 0.2557437 0.1564330

# Removing point 11  
train4=train[c(-11),]  
PCDH12\_1 = log(train4$PCDH12)  
AF161342\_1 = log(train4$AF161342+1)  
CARKD\_1 = log(train4$CARKD)  
SMC2\_1 = log(train4$SMC2)  
A\_24\_P936373\_1 = log(train4$A\_24\_P936373)  
C14orf143\_1 = log(train4$C14orf143)  
  
fit4 = lm(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train4)  
summary(fit4)$r.squared

## [1] 0.5921738

# In the first, we have clue that 11 should be remove based on the sort function.

sort(cooks.distance(fit4),decreasing = TRUE)[1:3]

## 16 253 218   
## 0.33437916 0.18593035 0.08366595

# Removing point 15  
train5 = train4[c(-15),]  
PCDH12\_1 = log(train5$PCDH12)  
AF161342\_1 = log(train5$AF161342+1)  
CARKD\_1 = log(train5$CARKD)  
SMC2\_1 = log(train5$SMC2)  
A\_24\_P936373\_1 = log(train5$A\_24\_P936373)  
C14orf143\_1 = log(train5$C14orf143)  
  
fit5 = lm(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train5)  
summary(fit5)$r.squared

## [1] 0.6993818

# Aftering tossing out point 11, 16 and 253 have the biggest cook's distance and it is ten times of point 218. Thus, after removing 16(15 because 11 was removed in the previous steps), the r-squared have a big jump, and it turns to be 0.699

sort(cooks.distance(fit5),decreasing = TRUE)[1:3]

## 253 218 96   
## 0.29444572 0.11646603 0.06282447

# Removing point 252  
train6 = train4[c(-252),]  
PCDH12\_1 = log(train6$PCDH12)  
AF161342\_1 = log(train6$AF161342+1)  
CARKD\_1 = log(train6$CARKD)  
SMC2\_1 = log(train6$SMC2)  
A\_24\_P936373\_1 = log(train6$A\_24\_P936373)  
C14orf143\_1 = log(train6$C14orf143)  
  
fit6 = lm(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train6)  
summary(fit6)$r.squared

## [1] 0.6000812

# From this step, the points of 253 have more than two times of value than 218's, so deciding to take 253 out.

sort(cooks.distance(fit6),decreasing = TRUE)[1:3]

## 16 218 96   
## 0.34333613 0.08642587 0.05707395

# Removing point 15  
train7 = train6[c(-15),]  
PCDH12\_1 = log(train7$PCDH12)  
AF161342\_1 = log(train7$AF161342+1)  
CARKD\_1 = log(train7$CARKD)  
SMC2\_1 = log(train7$SMC2)  
A\_24\_P936373\_1 = log(train7$A\_24\_P936373)  
C14orf143\_1 = log(train7$C14orf143)  
  
fit7 = lm(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train7)  
summary(fit7)$r.squared

## [1] 0.7096539

# point 16 is another point should be removed. The r-squared become 0.7076539, which is a huge jump from the last model.

sort(cooks.distance(fit7),decreasing = TRUE)[1:3]

## 218 96 207   
## 0.12235599 0.05841315 0.05612518

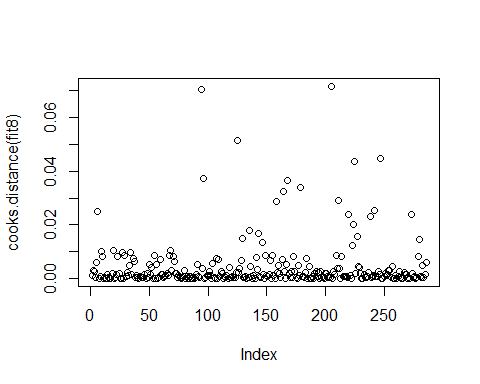
# Removing point 218  
train8 = train7[c(-216),]  
PCDH12\_1 = log(train8$PCDH12)  
AF161342\_1 = log(train8$AF161342+1)  
CARKD\_1 = log(train8$CARKD)  
SMC2\_1 = log(train8$SMC2)  
A\_24\_P936373\_1 = log(train8$A\_24\_P936373)  
C14orf143\_1 = log(train8$C14orf143)  
  
fit8 = lm(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train8)  
summary(fit8)$r.squared

## [1] 0.7183984

# After deciding tossing out point 218, because it have more than two times value of the second largest value point,   
sort(cooks.distance(fit8),decreasing = TRUE)[1:5]

## 207 96 127 250 228   
## 0.07167084 0.07036750 0.05161007 0.04464816 0.04358196

plot(cooks.distance(fit8))

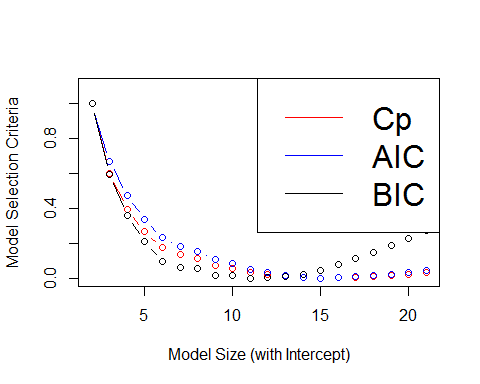


#upon here, we would like to conclude that all of the outliers are removed, since the   
# largest six cook's distance is very close and could seen in the graph.

# Best subset  
# install.packages("leaps")  
library("leaps")  
PCDH12\_1 = log(train8$PCDH12)  
AF161342\_1 = log(train8$AF161342+1)  
CARKD\_1 = log(train8$CARKD)  
SMC2\_1 = log(train8$SMC2)  
A\_24\_P936373\_1 = log(train8$A\_24\_P936373)  
C14orf143\_1 = log(train8$C14orf143)  
RSSleaps=regsubsets(health~PCDH12\_1+DLG5+BC038559+SHISA5+AF161342\_1+CARKD\_1+F2R+PHKG1+CDCP1+PLEKHM1+SMC2\_1+PSMB6+BX440400+A\_24\_P936373\_1+PPAN+BC007917+C14orf143\_1+LOC440104+THC2578957+ANKIB1,data = train8, nvmax = 20)  
RSSleaps=regsubsets(train8[,3:22],train8[,2],nvmax = 20)  
sumleaps=summary(RSSleaps,matrix=T)  
sumleaps$which

## (Intercept) PCDH12 DLG5 BC038559 SHISA5 AF161342 CARKD F2R PHKG1  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## 4 TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE  
## 5 TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE  
## 6 TRUE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE  
## 7 TRUE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE  
## 8 TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE  
## 9 TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE FALSE  
## 10 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE FALSE  
## 11 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 12 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 13 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 14 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 15 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 16 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 17 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 18 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 19 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 20 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## CDCP1 PLEKHM1 SMC2 PSMB6 BX440400 A\_24\_P936373 PPAN BC007917  
## 1 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 4 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 5 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 6 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 7 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 8 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 9 FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  
## 10 FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  
## 11 FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  
## 12 FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  
## 13 FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE  
## 14 FALSE TRUE TRUE TRUE FALSE TRUE FALSE FALSE  
## 15 FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE  
## 16 FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE  
## 17 FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 18 FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 19 FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 20 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## C14orf143 LOC440104 THC2578957 ANKIB1  
## 1 FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE  
## 4 FALSE FALSE FALSE FALSE  
## 5 FALSE FALSE TRUE FALSE  
## 6 FALSE TRUE FALSE FALSE  
## 7 TRUE TRUE FALSE FALSE  
## 8 TRUE TRUE FALSE FALSE  
## 9 FALSE TRUE TRUE FALSE  
## 10 TRUE TRUE FALSE FALSE  
## 11 TRUE TRUE FALSE FALSE  
## 12 TRUE TRUE TRUE FALSE  
## 13 TRUE TRUE TRUE FALSE  
## 14 TRUE TRUE TRUE FALSE  
## 15 TRUE TRUE TRUE FALSE  
## 16 TRUE TRUE TRUE FALSE  
## 17 TRUE TRUE TRUE FALSE  
## 18 TRUE TRUE TRUE TRUE  
## 19 TRUE TRUE TRUE TRUE  
## 20 TRUE TRUE TRUE TRUE

inrange = function(x) { (x - min(x)) / (max(x) - min(x)) }  
  
sumleaps = summary(RSSleaps,matrix=T)  
msize = apply(sumleaps$which,1,sum)  
n=nrow(train8)  
p=nrow(train8)  
Cp = sumleaps$rss/(summary(fit8)$sigma^2) + 2\*msize - n  
AIC = n\*log(sumleaps$rss/n) + 2\*msize  
BIC = n\*log(sumleaps$rss/n) + msize\*log(n)  
Cp1=inrange(Cp)  
BIC1 = inrange(BIC)  
AIC1 = inrange(AIC)  
plot(range(msize), c(0, 1.1), type="n", xlab="Model Size (with Intercept)", ylab="Model Selection Criteria")  
points(msize, Cp1, col="red", type="b")  
points(msize, AIC1, col="blue", type="b")  
points(msize, BIC1, col="black", type="b")  
legend("topright", lty=rep(1,3), col=c("red", "blue", "black"), cex = 2, legend=c("Cp", "AIC", "BIC"))



AIC

## 1 2 3 4 5 6 7   
## 187.46692 131.31194 97.75627 74.42245 56.32625 48.29741 43.53173   
## 8 9 10 11 12 13 14   
## 35.04176 31.04388 25.48161 22.64939 19.70310 17.65634 16.80900   
## 15 16 17 18 19 20   
## 17.41099 18.40993 19.21760 20.44584 22.39072 24.32353

BIC

## 1 2 3 4 5 6 7   
## 194.77891 142.27991 112.38023 92.70240 78.26220 73.88935 72.77967   
## 8 9 10 11 12 13 14   
## 67.94568 67.60380 65.69752 66.52129 67.23100 68.84022 71.64888   
## 15 16 17 18 19 20   
## 75.90686 80.56179 85.02545 89.90968 95.51055 101.09936

Cp

## 1 2 3 4 5 6 7   
## 261.52261 163.51525 113.66687 82.46530 60.14180 50.62788 45.08529   
## 8 9 10 11 12 13 14   
## 35.73789 31.43435 25.67281 22.81964 19.93612 18.00305 17.26887   
## 15 16 17 18 19 20   
## 17.93635 18.98618 19.85880 21.13157 23.07971 25.01650

#VIF

library(car)  
fit99 = lm(health~ PCDH12+DLG5+BC038559+SHISA5+AF161342+CARKD+F2R+PHKG1+CDCP1+PLEKHM1+SMC2+PSMB6+BX440400+A\_24\_P936373+PPAN+BC007917+C14orf143+LOC440104+THC2578957+ANKIB1,data = train)  
vif(fit99)

## PCDH12 DLG5 BC038559 SHISA5 AF161342   
## 1.218889 3.018192 4.453258 3.715247 1.689680   
## CARKD F2R PHKG1 CDCP1 PLEKHM1   
## 1.552932 5.687746 2.718229 2.627118 3.610882   
## SMC2 PSMB6 BX440400 A\_24\_P936373 PPAN   
## 1.446932 3.642281 2.852932 2.426480 2.521328   
## BC007917 C14orf143 LOC440104 THC2578957 ANKIB1   
## 3.092049 1.725649 4.481528 3.506056 3.521259

# F2R variable is greater than 5, so we might consider remove it and it is problematic.

# By implementing best subset method, generally consider AIC BIC and CP, we get the result that keep   
# PCDH12\_1+DLG5+SHISA5+AF161342\_1+CARKD\_1 +F2R+PHKG1+PLEKHM1+ SMC2\_1+PSMB6+A\_24\_P936373\_1+C14orf143\_1 +LOC440104+THC2578957  
PCDH12\_1 = log(train8$PCDH12)  
AF161342\_1 = log(train8$AF161342+1)  
CARKD\_1 = log(train8$CARKD)  
SMC2\_1 = log(train8$SMC2)  
A\_24\_P936373\_1 = log(train8$A\_24\_P936373)  
C14orf143\_1 = log(train8$C14orf143)  
fit11 = lm(health ~ PCDH12\_1+DLG5+SHISA5+AF161342\_1+CARKD\_1 +F2R+PHKG1+PLEKHM1+ SMC2\_1+PSMB6+A\_24\_P936373\_1+C14orf143\_1 +LOC440104+THC2578957, data = train8)  
summary(fit11)

##   
## Call:  
## lm(formula = health ~ PCDH12\_1 + DLG5 + SHISA5 + AF161342\_1 +   
## CARKD\_1 + F2R + PHKG1 + PLEKHM1 + SMC2\_1 + PSMB6 + A\_24\_P936373\_1 +   
## C14orf143\_1 + LOC440104 + THC2578957, data = train8)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5048 -0.6627 -0.0751 0.5239 4.2181   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.75342 0.59735 -6.283 1.32e-09 \*\*\*  
## PCDH12\_1 -1.81790 0.23682 -7.676 2.98e-13 \*\*\*  
## DLG5 -1.06695 0.16164 -6.601 2.14e-10 \*\*\*  
## SHISA5 0.61089 0.23662 2.582 0.010357 \*   
## AF161342\_1 0.78227 0.29263 2.673 0.007968 \*\*   
## CARKD\_1 0.61139 0.15958 3.831 0.000159 \*\*\*  
## F2R 1.10450 0.32894 3.358 0.000898 \*\*\*  
## PHKG1 0.51463 0.16881 3.049 0.002526 \*\*   
## PLEKHM1 4.65861 0.23410 19.900 < 2e-16 \*\*\*  
## SMC2\_1 0.17465 0.12103 1.443 0.150167   
## PSMB6 -0.14156 0.25924 -0.546 0.585482   
## A\_24\_P936373\_1 0.05389 0.16936 0.318 0.750595   
## C14orf143\_1 0.23883 0.12048 1.982 0.048460 \*   
## LOC440104 -0.54827 0.18654 -2.939 0.003574 \*\*   
## THC2578957 0.50786 0.22461 2.261 0.024548 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9972 on 271 degrees of freedom  
## Multiple R-squared: 0.7135, Adjusted R-squared: 0.6987   
## F-statistic: 48.2 on 14 and 271 DF, p-value: < 2.2e-16

write.csv(predict(fit11), file = "fitted value")  
# so the final model would be   
# health = -3.75342+ PCDH12\_1+DLG5+SHISA5+AF161342\_1+CARKD\_1 +F2R+PHKG1+PLEKHM1+ SMC2\_1+PSMB6+A\_24\_P936373\_1+C14orf143\_1 +LOC440104+THC2578957