Project

Xinrui Hu

2023-04-20

```
library(dplyr)
library(tidyverse)
library(knitr)
library(car)
options(scipen = 999)
df <- read.csv('newdf.csv')
df <- df %>% rename( "age" = "RIDAGEYR",
                     "gender" = "RIAGENDR",
                     "edu" = "DMDEDUC2",
                     "marry" = "DMDMARTL",
                     "race" = "RIDRETH3",
                     "height" = "WHD010",
                     "weight" = "WHD020",
                     "hi" = "HIQ011",
                     "ogtt" = "LBXGLT") %>% na.omit(df$ogtt)
```

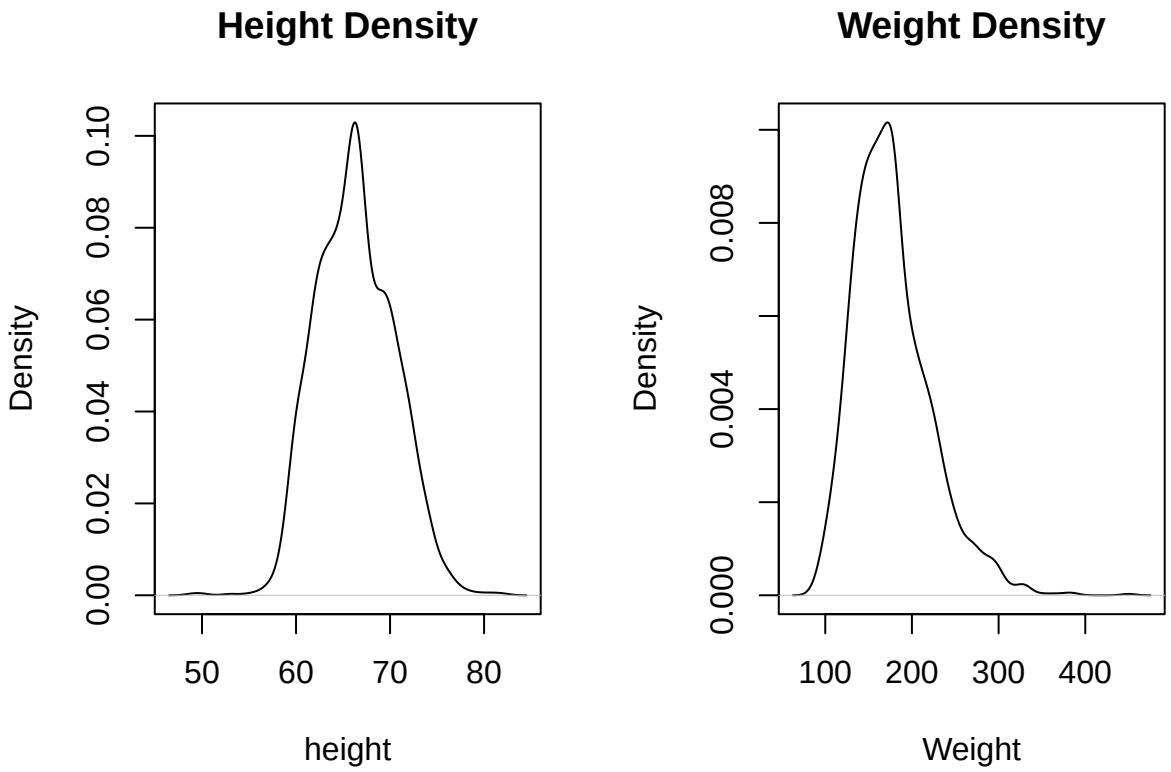
```
# Deal with the abnormal value
df$height[df$height > 90] <- NA
df$weight[df$weight > 700] <- NA
df$hi[df$hi != 1] <- 2
# put the missing value with the mean of the data
df$weight[is.na(df$weight)] <- mean(df$weight, na.rm = TRUE)
df$height[is.na(df$height)] <- mean(df$height, na.rm = TRUE)
df$bmi <- signif((df$weight / (df$height^2))*703,4)
```

```
summary(df)
```

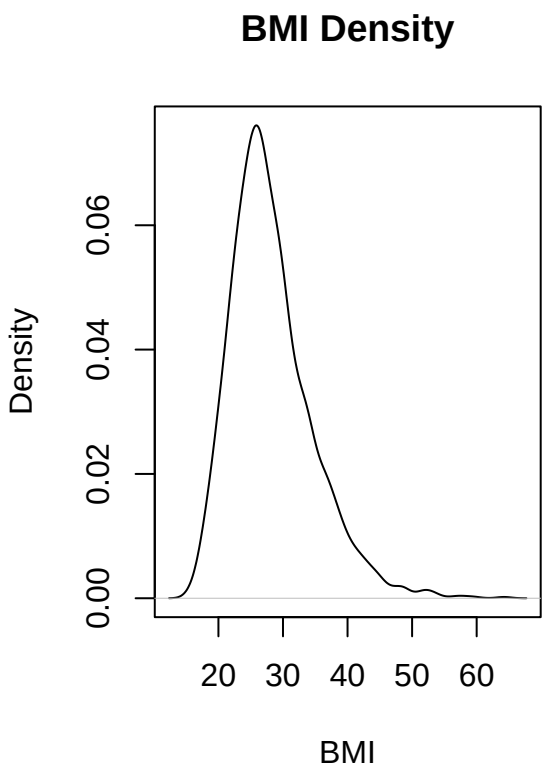
```
##      height          weight          hi             ogtt            age
##  Min.   :49.00   Min.   : 88   Min.   :1.000   Min.   : 35.0   Min.   :20.0
##  1st Qu.:63.00   1st Qu.:145   1st Qu.:1.000   1st Qu.: 92.0   1st Qu.:34.0
##  Median :66.00   Median :170   Median :1.000   Median :111.0   Median :47.5
##  Mean   :66.31   Mean   :177   Mean   :1.206   Mean   :123.6   Mean   :48.3
##  3rd Qu.:69.00   3rd Qu.:200   3rd Qu.:1.000   3rd Qu.:142.0   3rd Qu.:62.0
##  Max.   :82.00   Max.   :450   Max.   :2.000   Max.   :542.0   Max.   :80.0
##      gender           edu            marry            race
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000   Median :4.000   Median :1.000   Median :3.000
##  Mean   :1.509   Mean   :3.504   Mean   :2.634   Mean   :3.203
##  3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :2.000   Max.   :5.000   Max.   :6.000   Max.   :7.000
##       bmi
##  Min.   :15.83
```

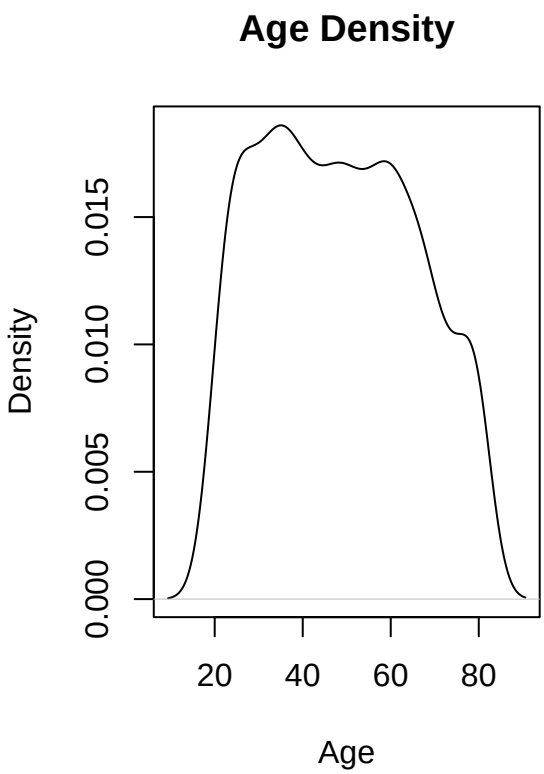
```
##  1st Qu.:23.80
##  Median :27.17
##  Mean   :28.22
##  3rd Qu.:31.41
##  Max.   :64.23
```

```r
# Set up the plotting window with two plots side-by-side
par(mfrow = c(1, 2))
plot(density(df$height), main = "Height Density", xlab = "height")
plot(density(df$weight), main = "Weight Density", xlab = "Weight")
```

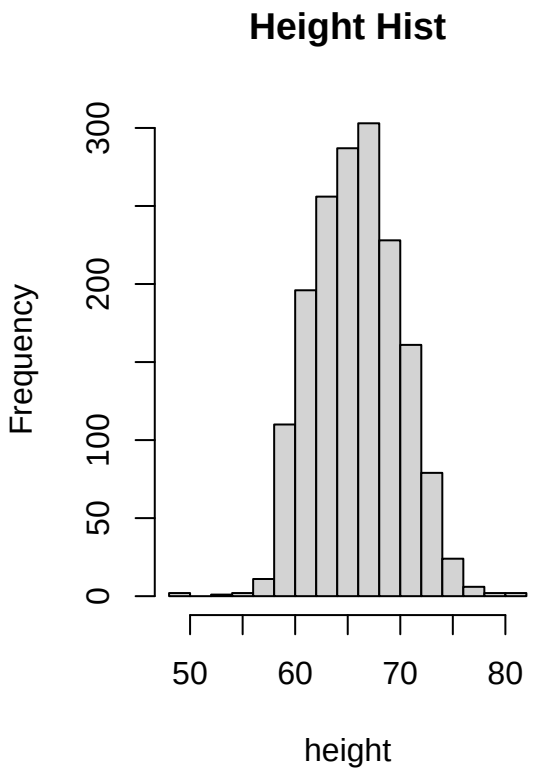


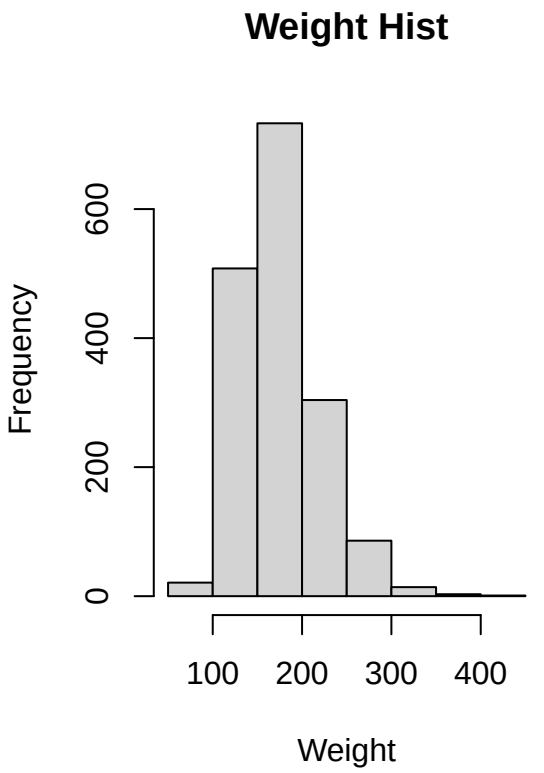

```r
par(mfrow = c(1, 2))
plot(density(df$bmi), main = "BMI Density", xlab = "BMI")
plot(density(df$age), main = "Age Density", xlab = "Age")
```



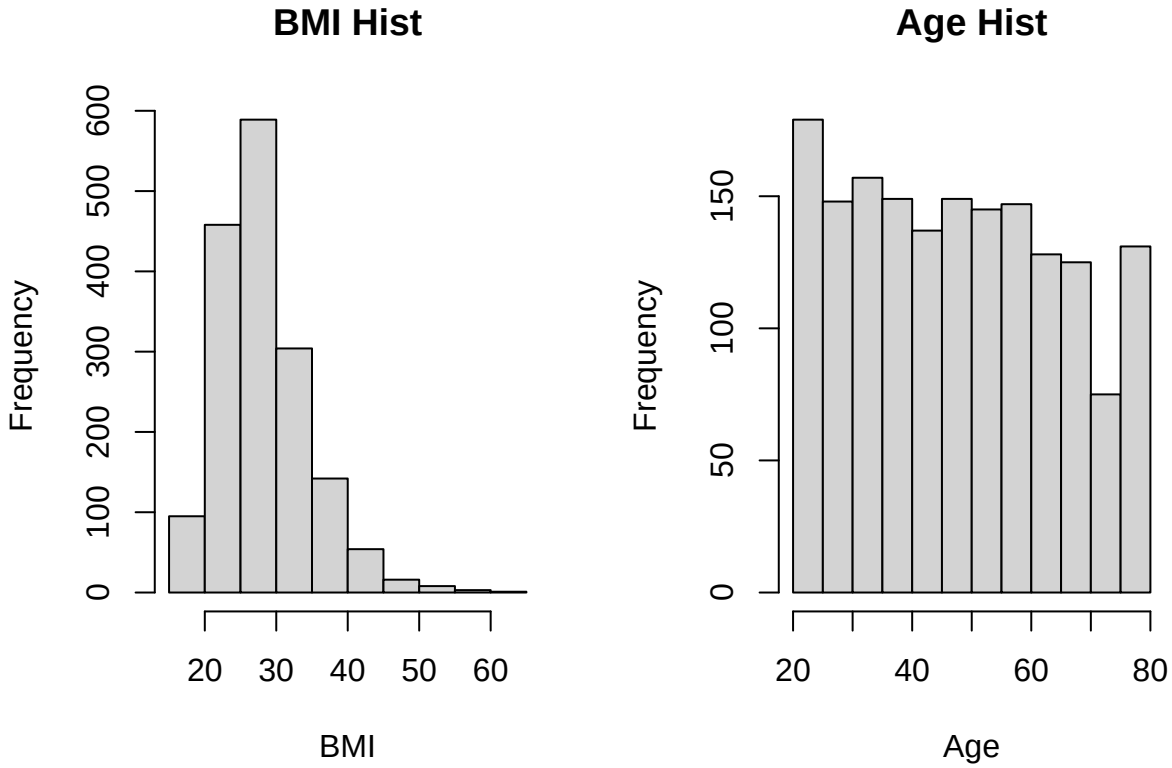



```r
par(mfrow = c(1, 2))
hist(df$height, main = "Height Hist", xlab = "height")
hist(df$weight, main = "Weight Hist", xlab = "Weight")
```

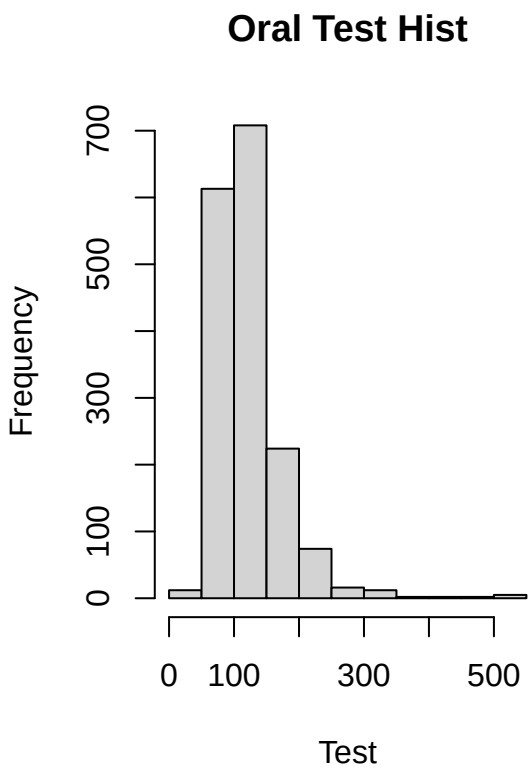


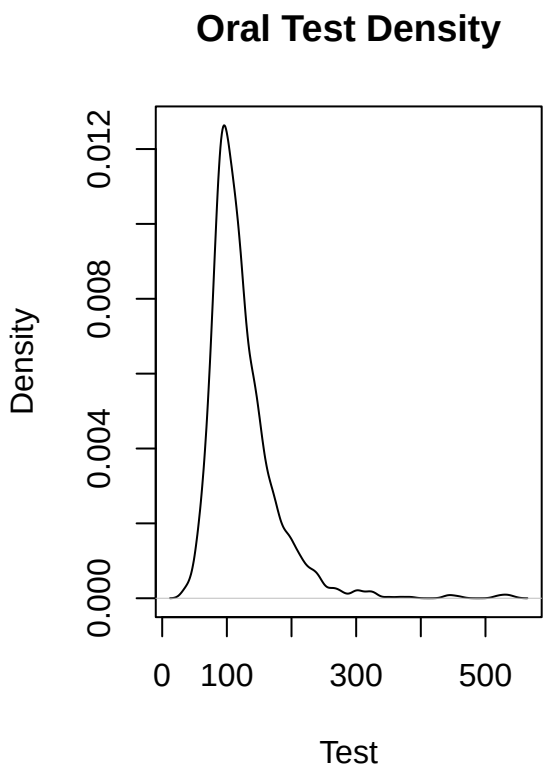


```r
par(mfrow = c(1, 2))
hist(df$bmi, main = "BMI Hist", xlab = "BMI")
hist(df$age, main = "Age Hist", xlab = "Age")
```




```r
par(mfrow = c(1, 2))
hist(df$ogtt, main = "Oral Test Hist", xlab = "Test")
plot(density(df$ogtt), main = "Oral Test Density", xlab = "Test")
```






```
library ("gridExtra")
```

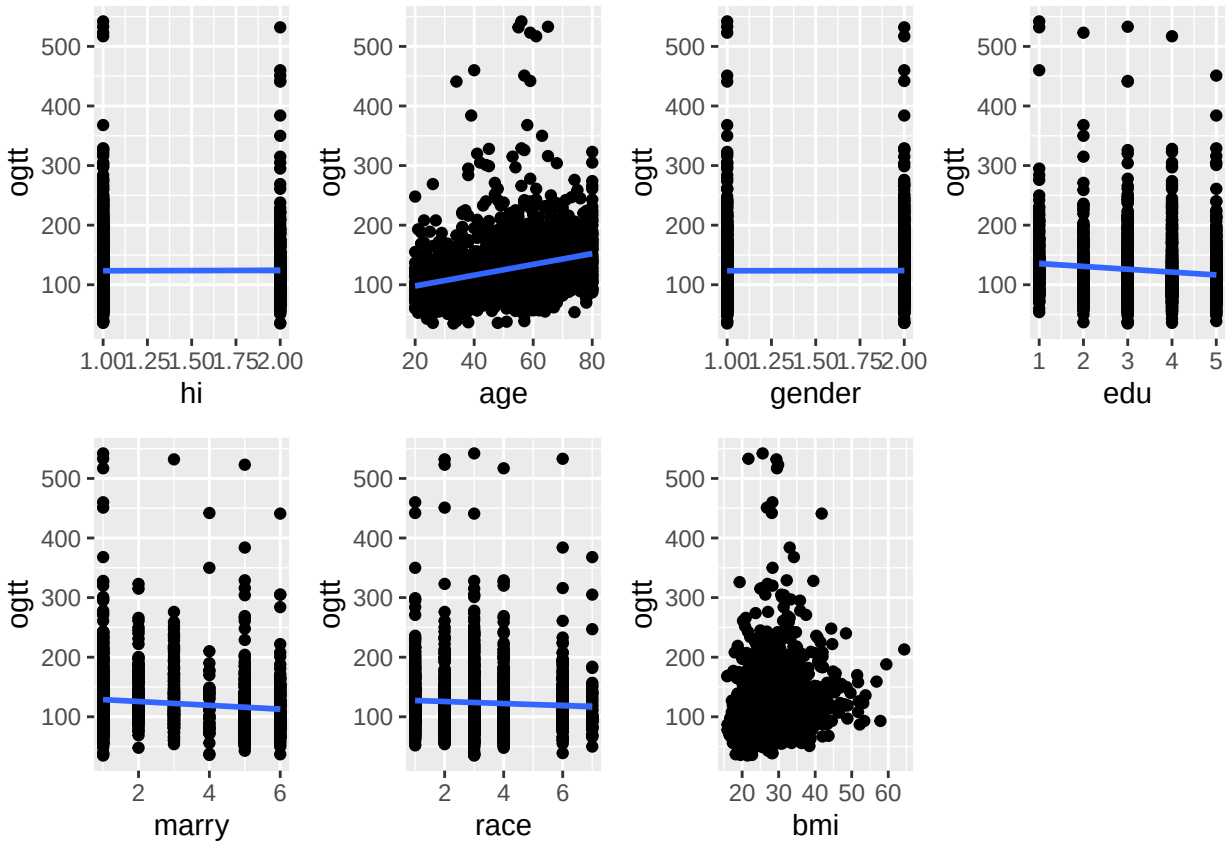
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
require(ggplot2)
p1 <- ggplot(df, aes(hi, ogtt)) + geom_point() + stat_smooth(method="lm")
p2 <- ggplot(df, aes(age, ogtt)) + geom_point() + stat_smooth(method="lm")
p3 <- ggplot(df, aes(gender, ogtt)) + geom_point() +stat_smooth(method="lm")
p4 <- ggplot(df, aes(edu, ogtt)) + geom_point() + stat_smooth( method="lm")
p5 <- ggplot(df, aes(marry, ogtt)) + geom_point() + stat_smooth(method="lm")
p6 <- ggplot(df, aes(race, ogtt)) + geom_point() + stat_smooth(method="lm")
p7 <- ggplot(df, aes(bmi, ogtt)) + geom_point(position = position_jitter(width = .2,height=0))
grid.arrange(p1, p2, p3, p4, p5, p6, p7, nrow = 2)
```

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

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

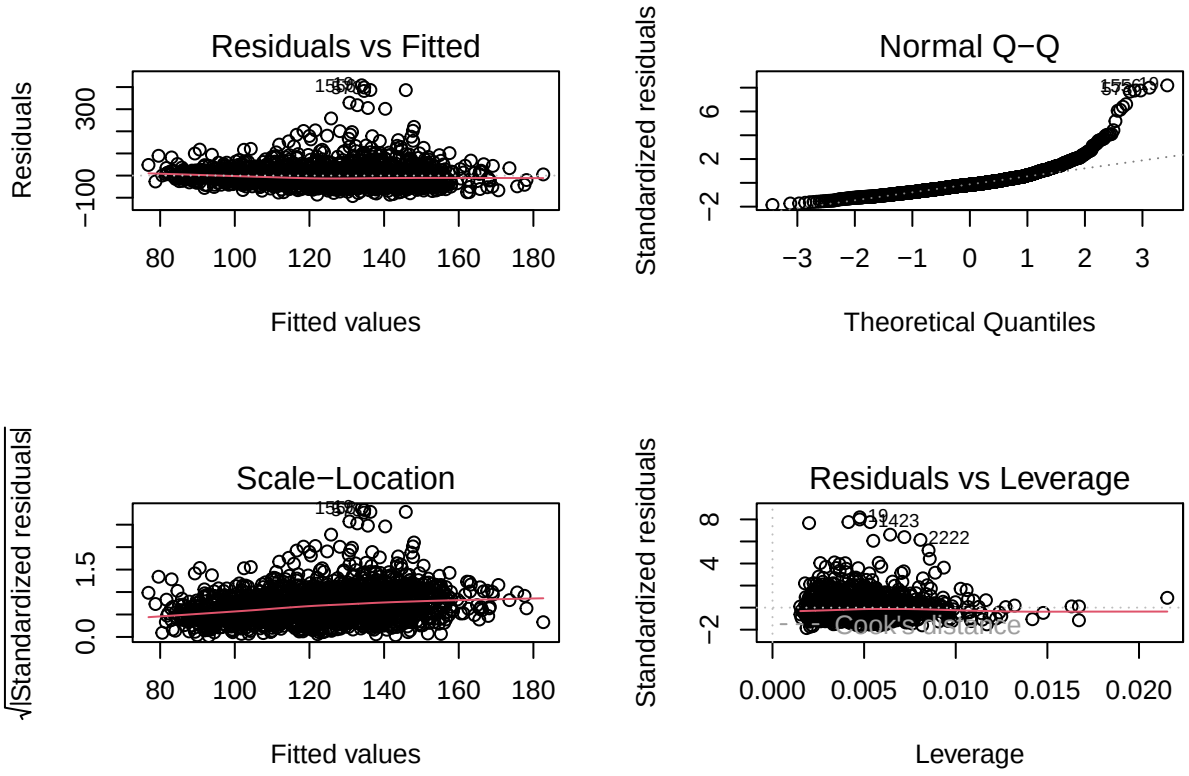



```
#Fitting a model
lmod <- lm(ogtt ~ .-weight-height, df)
summary(lmod)
```

```
##
## Call:
## lm(formula = ogtt ~ . - weight - height, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -92.58 -28.78  -7.93  16.32 408.00
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 40.47390   11.29572   3.583             0.000349 ***
## hi          10.36588    3.28110   3.159             0.001610 **
## age          0.93173    0.07835  11.892 < 0.0000000000000002 ***
## gender      -0.46345    2.45359  -0.189             0.850205
## edu         -3.21982    1.03195  -3.120             0.001839 **
## marry       -0.48940    0.68555  -0.714             0.475405
## race         1.04505    0.82420   1.268             0.204994
## bmi          1.25956    0.19532   6.449       0.000000000147 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.94 on 1662 degrees of freedom
## Multiple R-squared:  0.1227, Adjusted R-squared:  0.119
```

```
## F-statistic:  33.2 on 7 and 1662 DF,  p-value: < 0.00000000000000022
```

```
#Producing diagnistics plots:
par (mfrow = c(2,2))
plot (lmod)
```




```
par (mfrow = c(1,1))
```

```
cook <- cooks.distance(lmod)
cook[which(cook>0.5)]
```

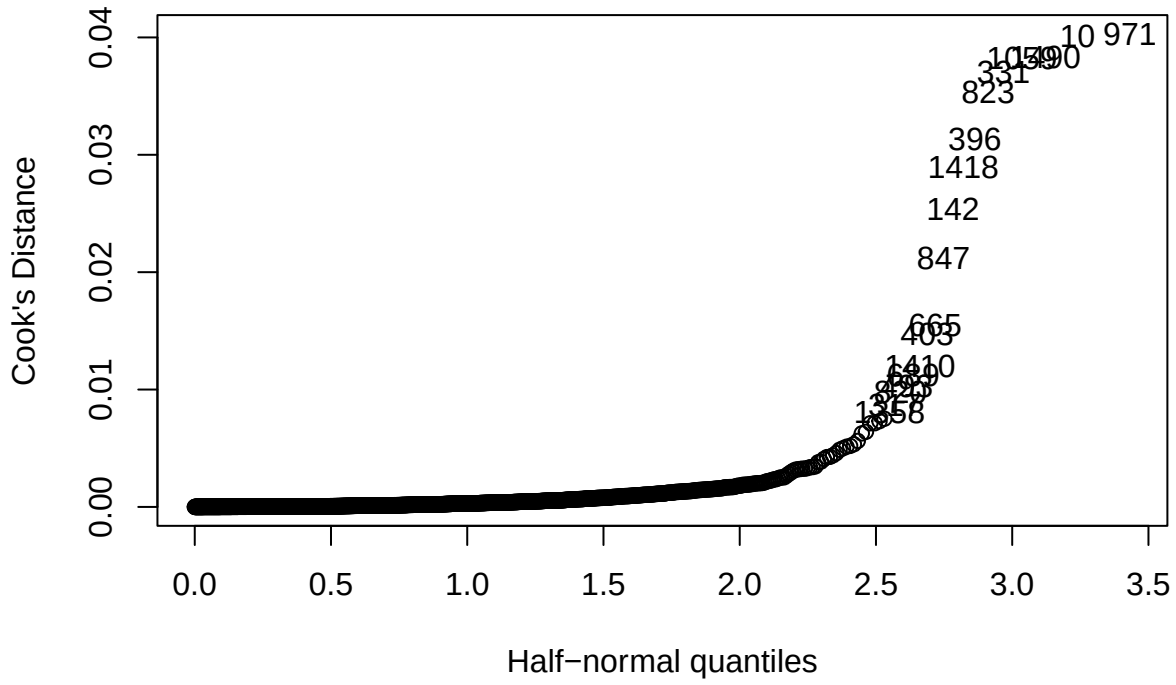
```
## named numeric(0)
```

```
library("faraway")
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
halfnorm(cook, 18, ylab = "Cook's Distance")
```




```
df[971,]
```

```
##      height weight hi ogtt age gender edu marry race   bmi
## 1423     60    150  2  532  55      2   1     3    2 29.29
```

#oberservation 10 and 971 has the high distance

```
subsetdf <- df[c(-10, -19, -221, -488, -573, -583, -1205, -1242, -1423, -1556, -142, -331, -396, -403, -
```

```
corMatrix <- cor(subsetdf)
corMatrix <- round (corMatrix, 2)
corMatrix
```

```
##        height weight    hi  ogtt   age gender   edu marry  race   bmi
## height   1.00   0.46 -0.03 -0.05 -0.05  -0.68  0.13  0.03  0.10 -0.04
## weight   0.46   1.00 -0.04  0.14 -0.01  -0.29  0.01 -0.04 -0.09  0.86
## hi      -0.03  -0.04  1.00 -0.02 -0.27  -0.04 -0.24  0.17 -0.13 -0.03
## ogtt    -0.05   0.14 -0.02  1.00  0.32   0.01 -0.10 -0.12 -0.05  0.18
## age     -0.05  -0.01 -0.27  0.32  1.00   0.02 -0.07 -0.35 -0.08  0.02
## gender  -0.68  -0.29 -0.04  0.01  0.02   1.00  0.02  0.00 -0.05  0.05
## edu      0.13   0.01 -0.24 -0.10 -0.07   0.02  1.00 -0.04  0.30 -0.06
## marry    0.03  -0.04  0.17 -0.12 -0.35   0.00 -0.04  1.00  0.02 -0.06
## race     0.10  -0.09 -0.13 -0.05 -0.08  -0.05  0.30  0.02  1.00 -0.16
## bmi     -0.04   0.86 -0.03  0.18  0.02   0.05 -0.06 -0.06 -0.16  1.00
```

```
vif(lmod)
```

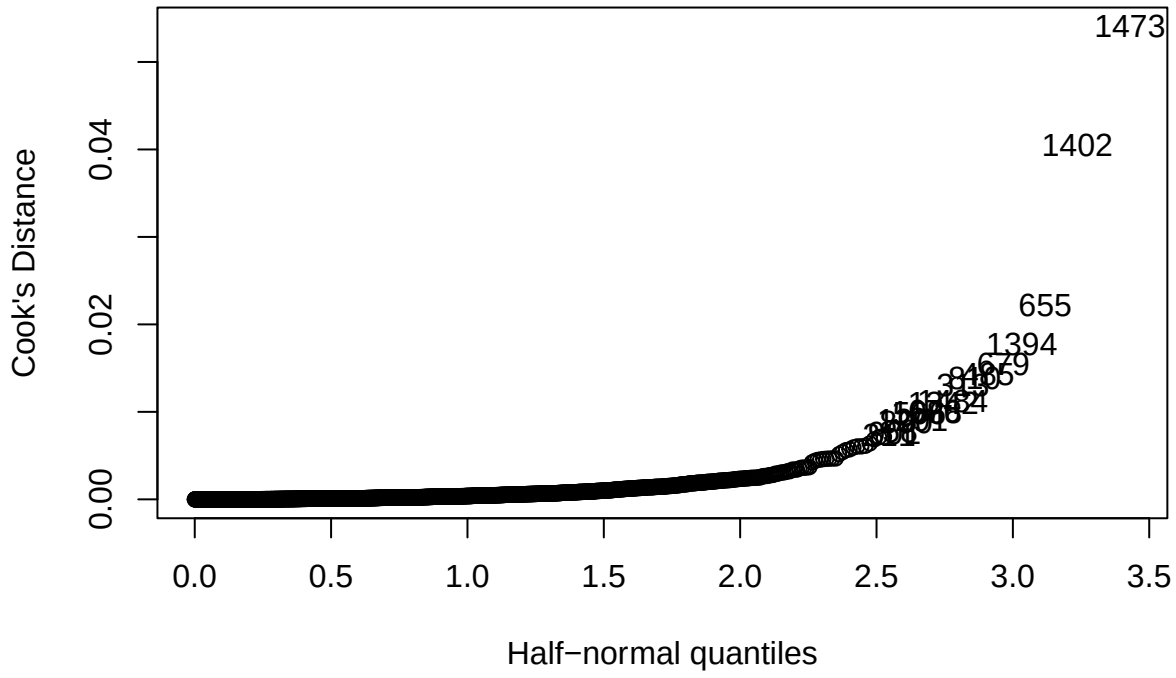
```
##       hi      age   gender      edu    marry     race      bmi
## 1.179021 1.231975 1.007441 1.170569 1.153592 1.144568 1.033909
```

```
#new model without outliers
model2 <- lm(ogtt ~ .-weight-height, subsetdf)
summary(model2)
```

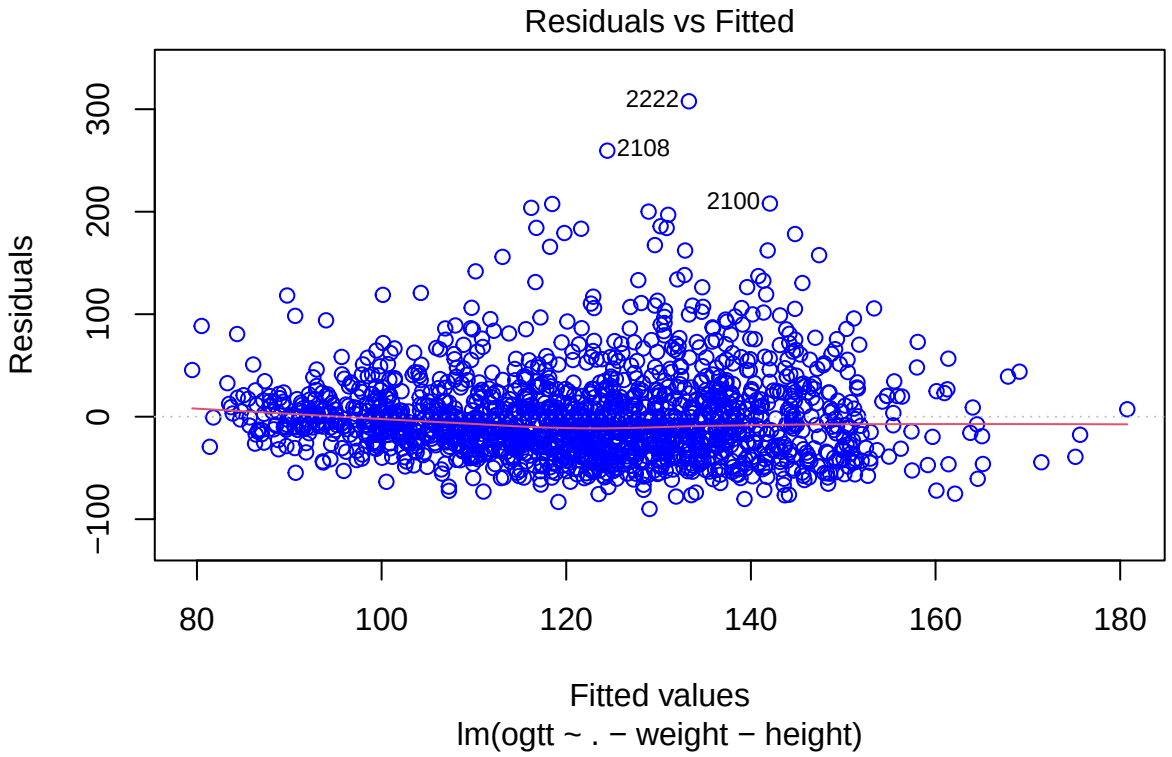
```
##
## Call:
```

```
## lm(formula = ogtt ~ . - weight - height, data = subsetdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.020 -27.379  -7.127  17.223 307.717
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 39.05794    9.72200   4.017   0.0000614781032709 ***
## hi           7.74127    2.83449   2.731              0.00638 **
## age          0.87713    0.06725  13.043 < 0.0000000000000002 ***
## gender      -0.16010    2.11031  -0.076              0.93953
## edu         -2.08812    0.88954  -2.347              0.01902 *
## marry       -0.21537    0.58997  -0.365              0.71512
## race         0.81716    0.70994   1.151              0.24989
## bmi          1.29507    0.16784   7.716   0.0000000000000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.73 on 1644 degrees of freedom
## Multiple R-squared:  0.1433, Adjusted R-squared:  0.1396
## F-statistic: 39.28 on 7 and 1644 DF,  p-value: < 0.00000000000000022
```

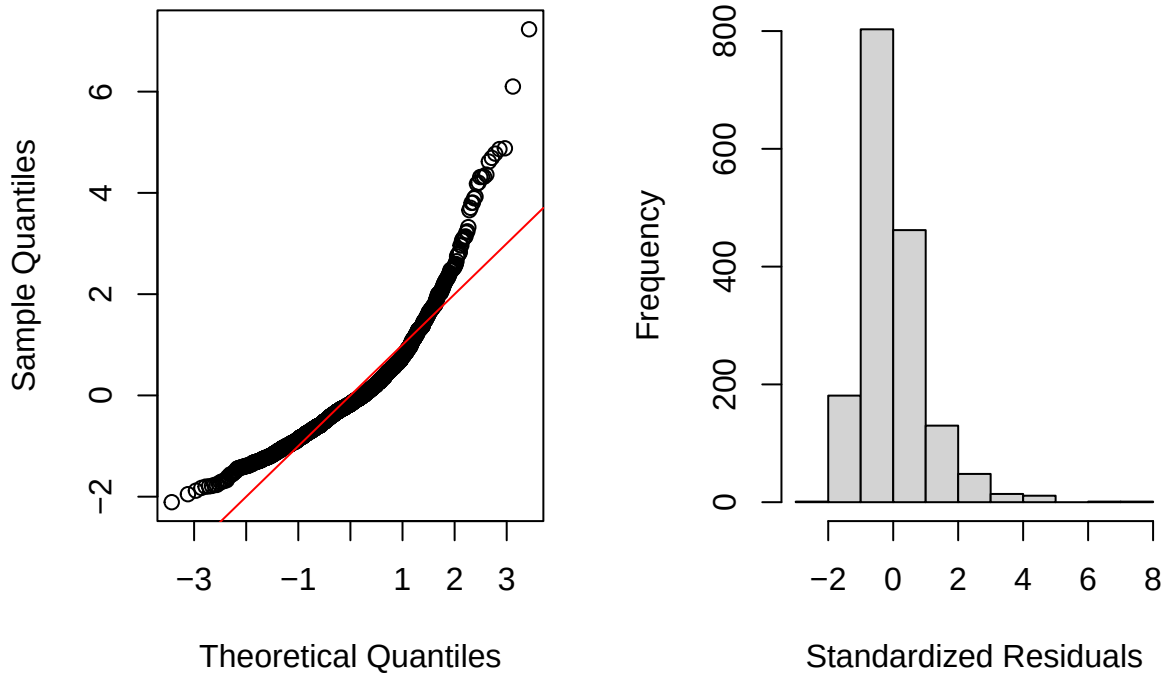
```
cook2 <- cooks.distance(model2)
halfnorm(cook2, 18, ylab = "Cook's Distance")
```




```
plot(model2, which = 1, col = "blue")
```




```
#Quantile-Quantile plot and a histogram based on the standardized residuals
par (mfrow = c (1,2))
qqnorm(rstandard(model2), main = "")
abline(0,1, col = "red")
hist (rstandard(model2), main = "", xlab = "Standardized Residuals")
```




```
# AIC
require(leaps)
```

```
## Loading required package: leaps
```

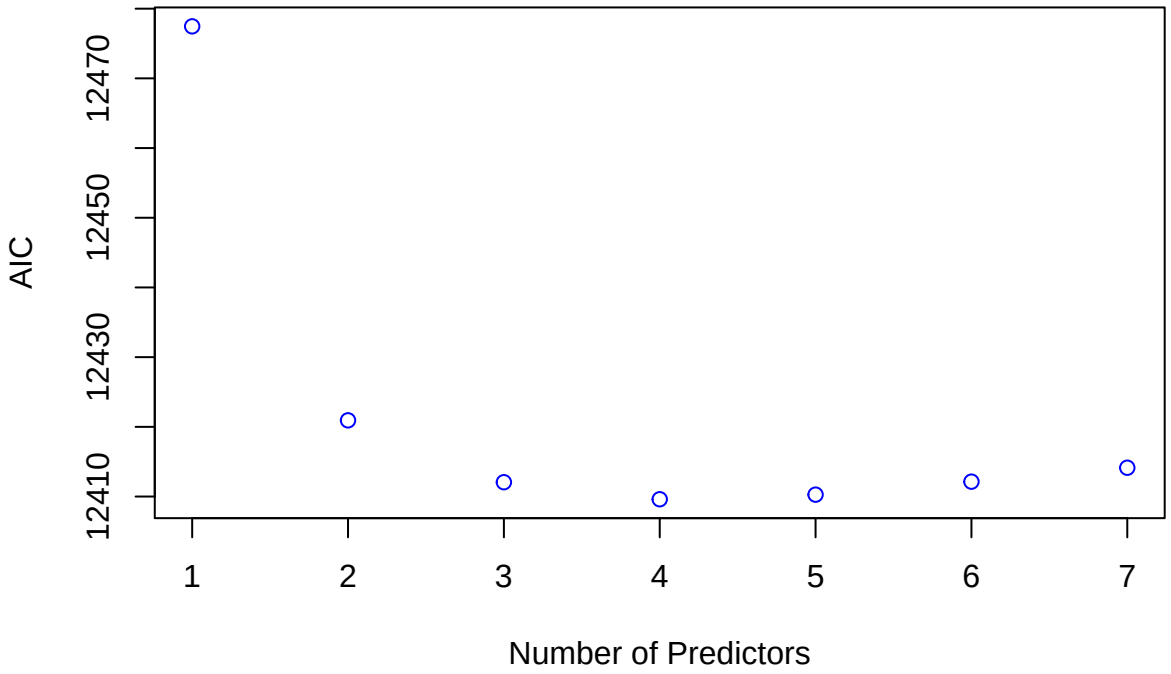
```
amod <- regsubsets(ogtt ~ .-weight-height, subsetdf)
rs <- summary(amod)
rs$which
```

```
##   (Intercept)    hi  age gender   edu marry  race   bmi
## 1        TRUE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
## 2        TRUE FALSE TRUE  FALSE FALSE FALSE FALSE  TRUE
## 3        TRUE  TRUE TRUE  FALSE FALSE FALSE FALSE  TRUE
## 4        TRUE  TRUE TRUE  FALSE  TRUE FALSE FALSE  TRUE
## 5        TRUE  TRUE TRUE  FALSE  TRUE FALSE  TRUE  TRUE
## 6        TRUE  TRUE TRUE  FALSE  TRUE  TRUE  TRUE  TRUE
## 7        TRUE  TRUE TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
```

```
rs$rss
```

```
## [1] 3141665 3032310 3012408 3004317 3001882 3001637 3001626
```

```
n <- nrow(subsetdf)
p <- 2:8
AIC <- n*log(rs$rss / n) + 2 * p
plot(AIC ~ I(p - 1), ylab = "AIC", xlab = "Number of Predictors", col = "blue")
```




```
# Fourth has the loest AIC
model1 <- lm(ogtt ~  hi+age + edu +bmi, subsetdf)
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = ogtt ~ hi + age + edu + bmi, data = subsetdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -89.854 -27.492  -7.108  17.607 307.738
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  41.0001     8.3493   4.911   0.0000009980082390 ***
## hi            7.3757     2.8113   2.624              0.00878 **
## age           0.8787     0.0635  13.838 < 0.0000000000000002 ***
## edu          -1.8033     0.8562  -2.106              0.03535 *
## bmi           1.2678     0.1653   7.667   0.0000000000000299 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.71 on 1647 degrees of freedom
## Multiple R-squared:  0.1425, Adjusted R-squared:  0.1404
## F-statistic: 68.43 on 4 and 1647 DF,  p-value: < 0.00000000000000022
```

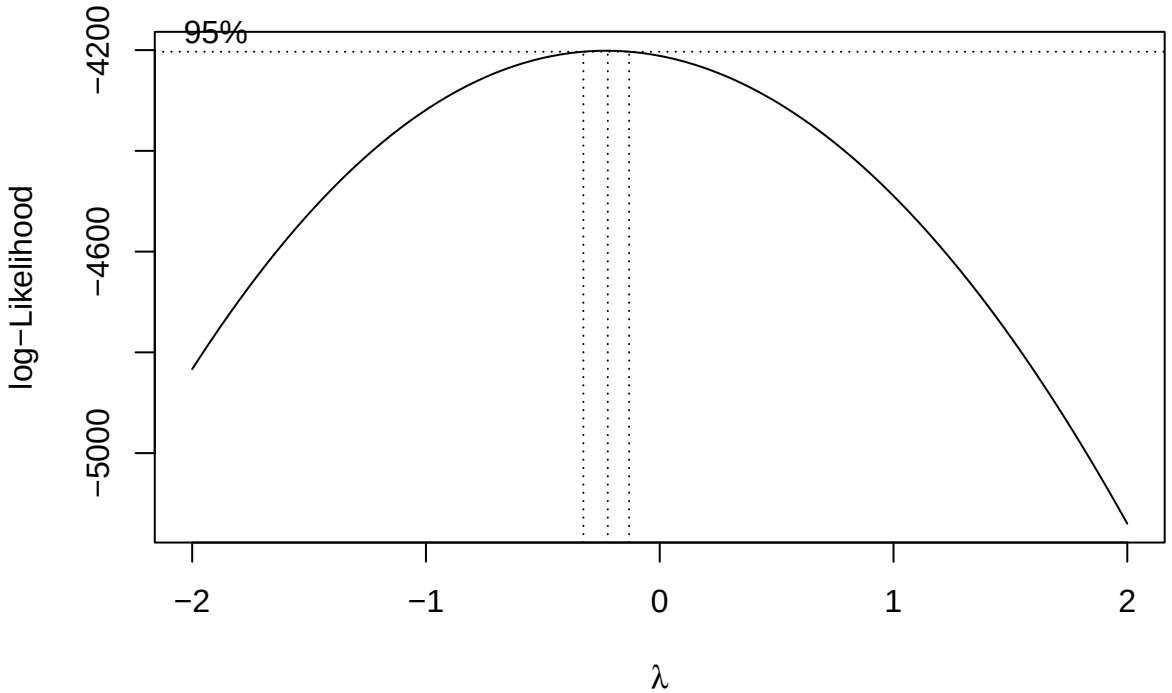
```
require(MASS)
```

```
## Loading required package: MASS
```

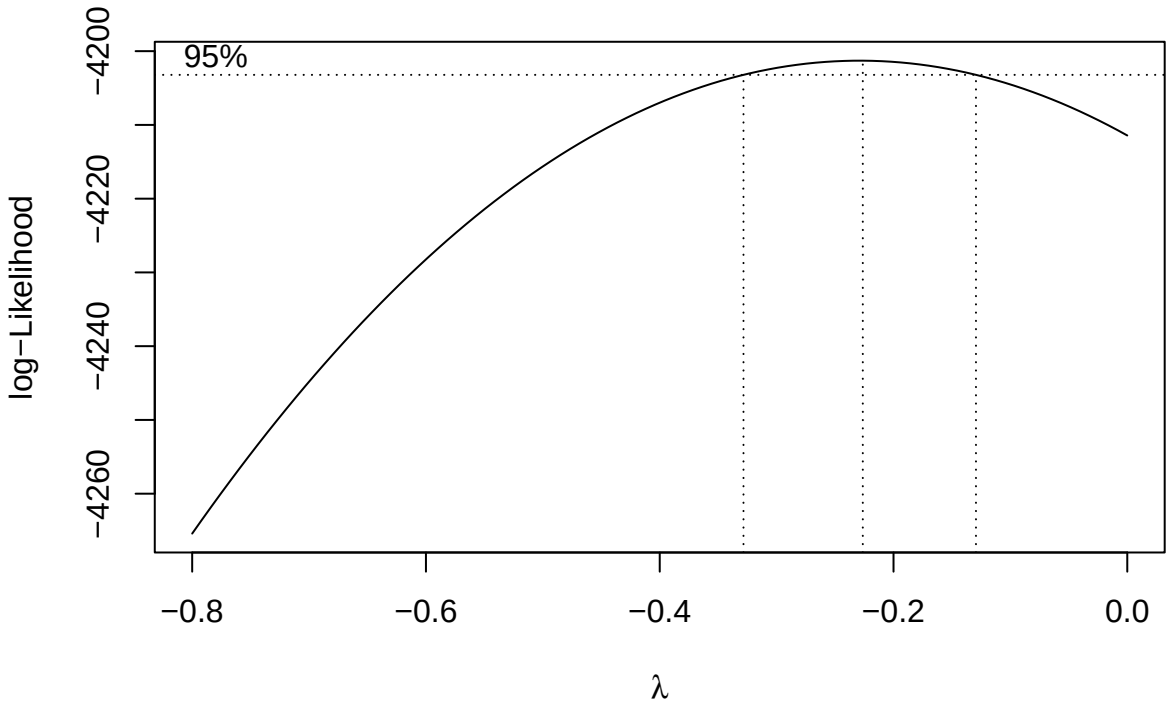
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
obj <- boxcox(model1, plotit = TRUE)
```




```
obj <- boxcox(model1, plotit = TRUE, lambda=seq(-0.8,0,by=0.1))
```




```
mlLambda <- obj$x[which.max(obj$y)]
mlLambda
```

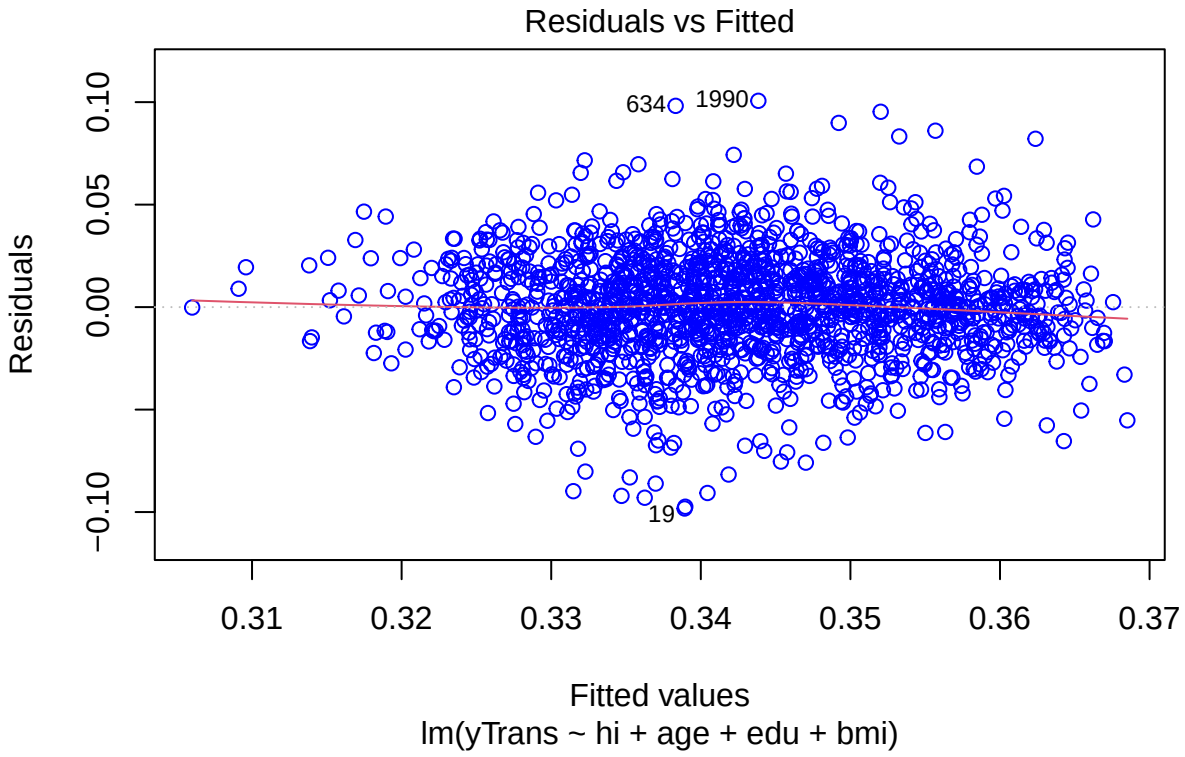
```
## [1] -0.2262626
```

```
yTrans <- df$ogtt^(mlLambda) # Transform the response
lmodTrans <- lm(yTrans ~ hi + age + edu + bmi,data=df) # Fit a new regression

summary(lmodTrans)
```

```
##
## Call:
## lm(formula = yTrans ~ hi + age + edu + bmi, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.098260 -0.014655  0.000371  0.015144  0.100649
##
## Coefficients:
##               Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)  0.39408050  0.00486944  80.929 < 0.0000000000000002 ***
## hi          -0.00488677  0.00163547  -2.988              0.00285 **
## age         -0.00055345  0.00003717 -14.889 < 0.0000000000000002 ***
## edu          0.00110872  0.00049919   2.221              0.02648 *
## bmi         -0.00080221  0.00009675  -8.291 0.00000000000000228 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02509 on 1665 degrees of freedom
## Multiple R-squared:  0.1599, Adjusted R-squared:  0.1579
## F-statistic: 79.25 on 4 and 1665 DF,  p-value: < 0.00000000000000022
```

```
plot(lmodTrans, which = 1, col = "blue")
```




```
par (mfrow = c (1,2))
qqnorm(rstandard(lmodTrans), main = "")
abline(0,1, col = "red")
hist (rstandard(lmodTrans), main = "", xlab = "Standardized Residuals")
```