

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" = "DMEDEDUC2",
                    "marry" = "DMDMARTL",
                    "race" = "RIDRETH3",
                    "height" = "WHDO10",
                    "weight" = "WHDO20",
                    "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)

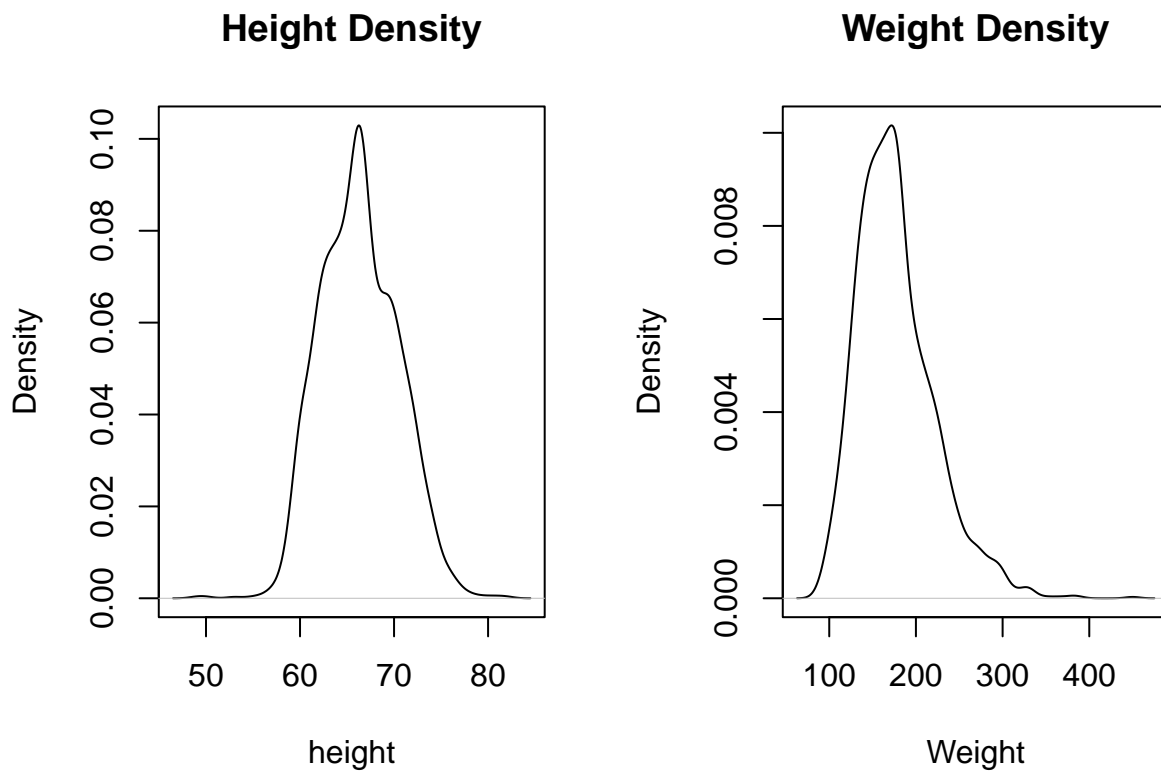
summary(df)
```

```
##      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
```

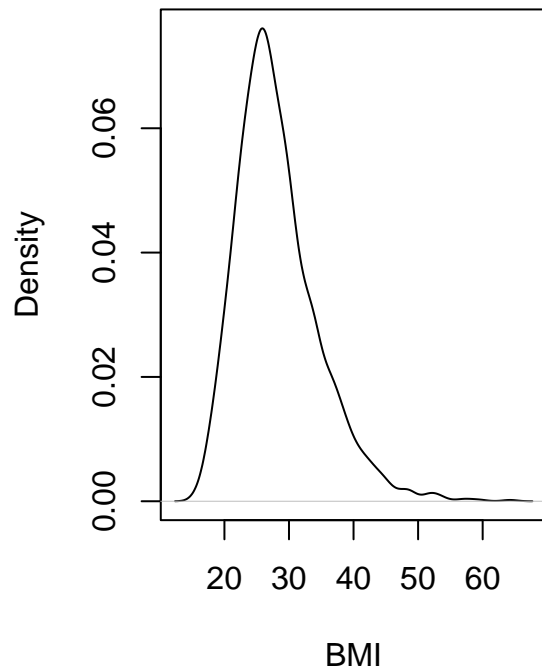
```
# Set up the plotting window with two plots side-by-side
```

```
par(mfrow = c(1, 2))
plot(density(df$height), main = "Height Density", xlab = "height")
plot(density(df$weight), main = "Weight Density", xlab = "Weight")
```

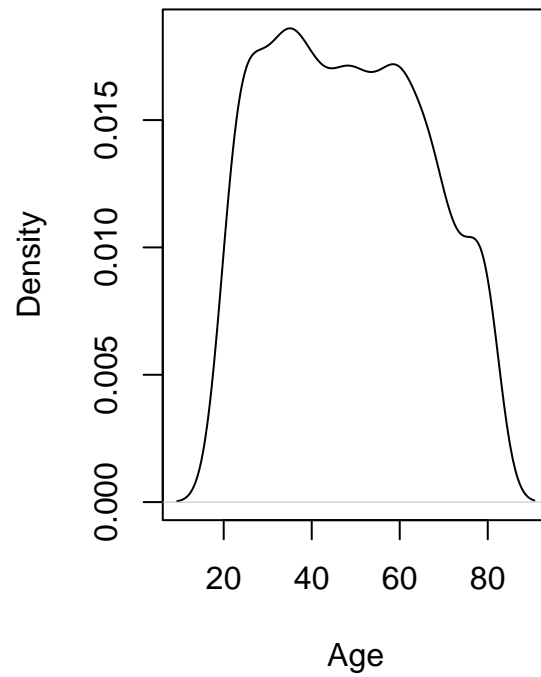


```
par(mfrow = c(1, 2))
plot(density(df$bmi), main = "BMI Density", xlab = "BMI")
plot(density(df$age), main = "Age Density", xlab = "Age")
```

BMI Density

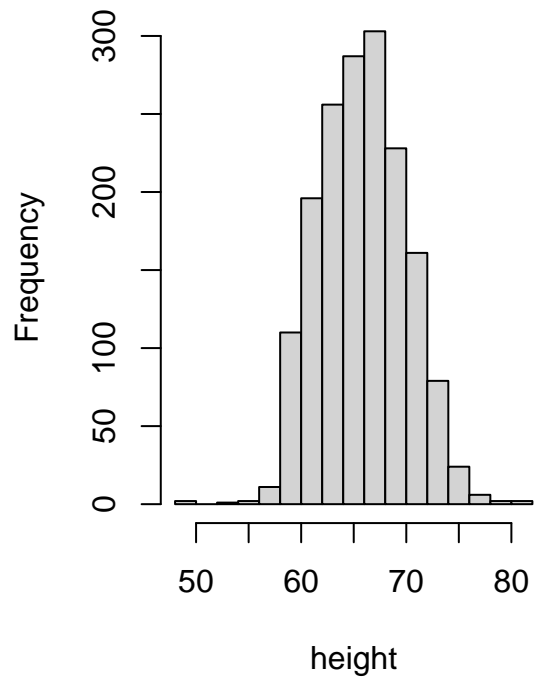


Age Density

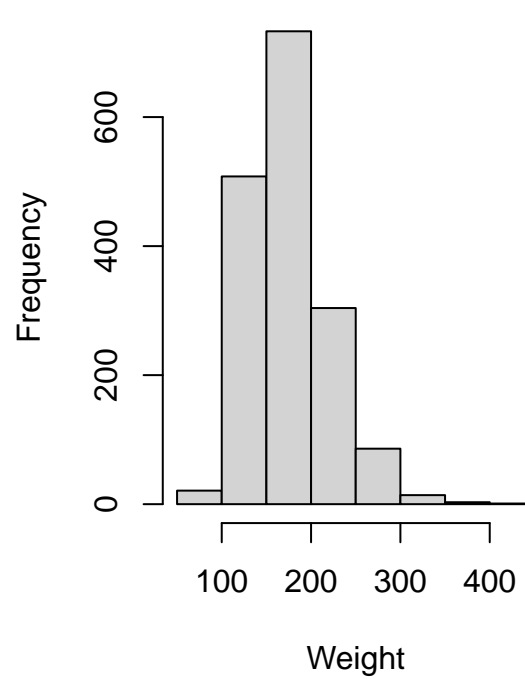


```
par(mfrow = c(1, 2))  
hist(df$height, main = "Height Hist", xlab = "height")  
hist(df$weight, main = "Weight Hist", xlab = "Weight")
```

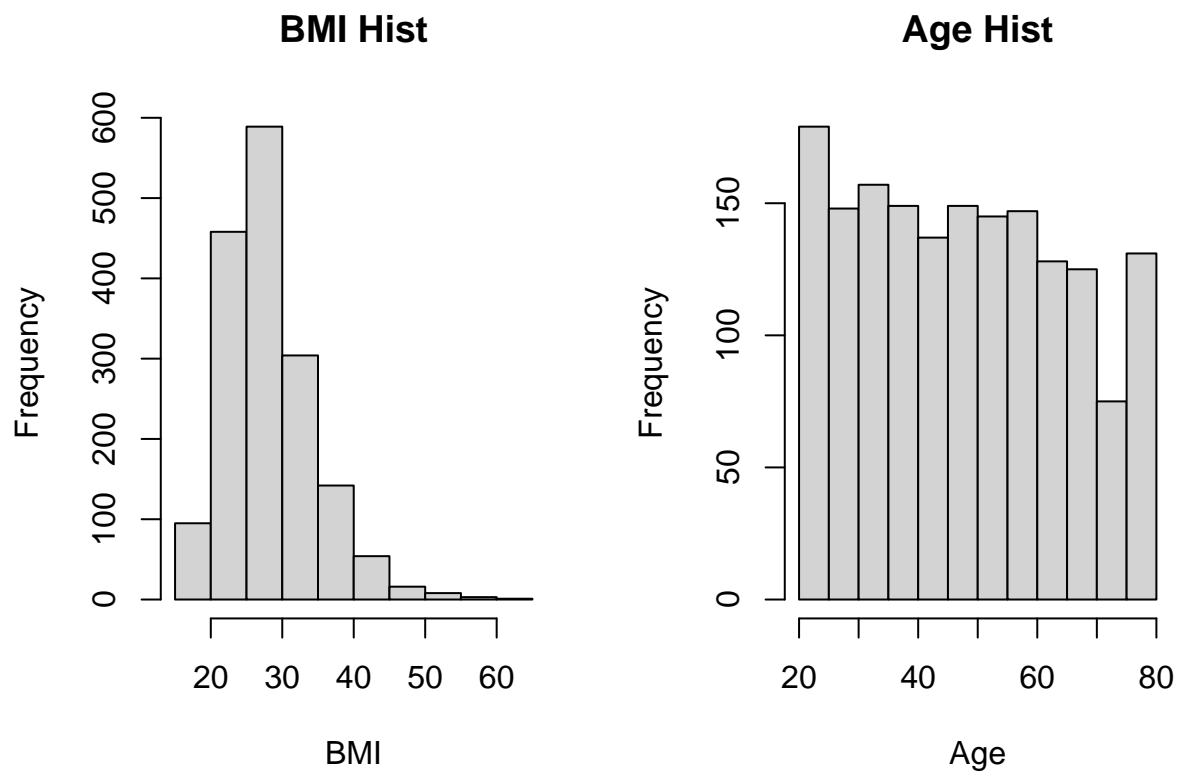
Height Hist



Weight Hist

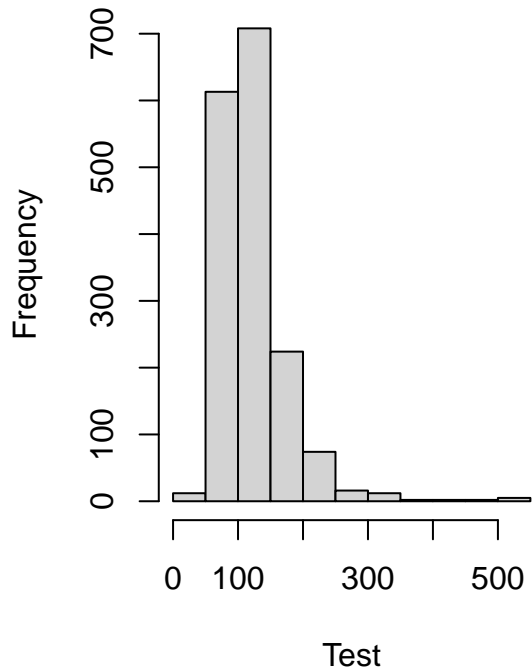


```
par(mfrow = c(1, 2))
hist(df$bmi, main = "BMI Hist", xlab = "BMI")
hist(df$age, main = "Age Hist", xlab = "Age")
```

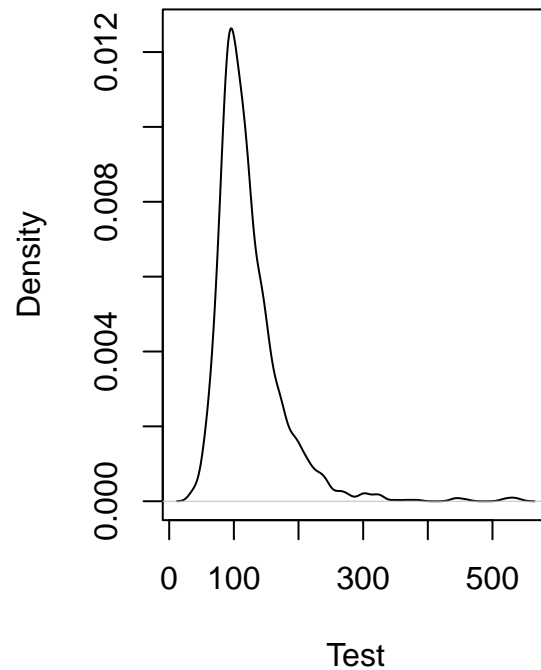


```
par(mfrow = c(1, 2))
hist(df$ogtt, main = "Oral Test Hist", xlab = "Test")
plot(density(df$ogtt), main = "Oral Test Density", xlab = "Test")
```

Oral Test Hist



Oral Test Density

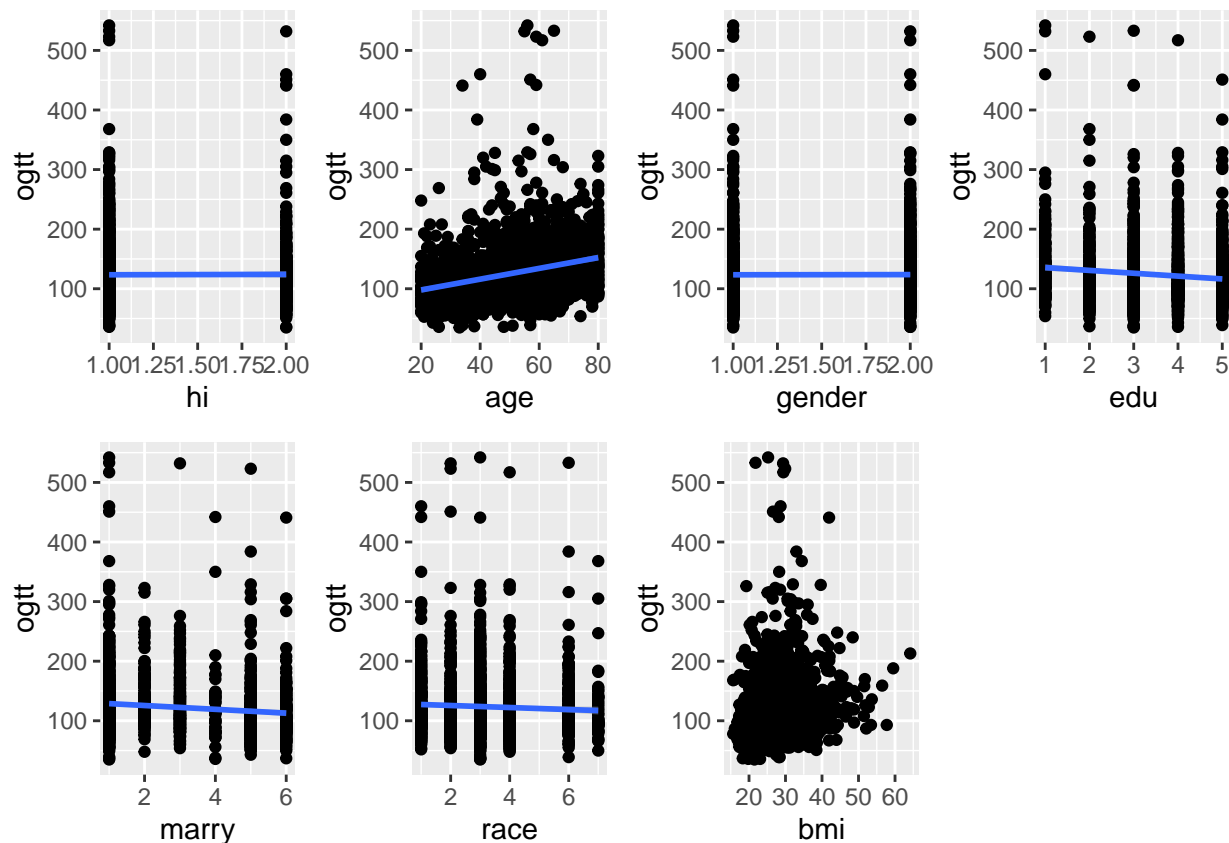


```
library("gridExtra")

##
## 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)
```

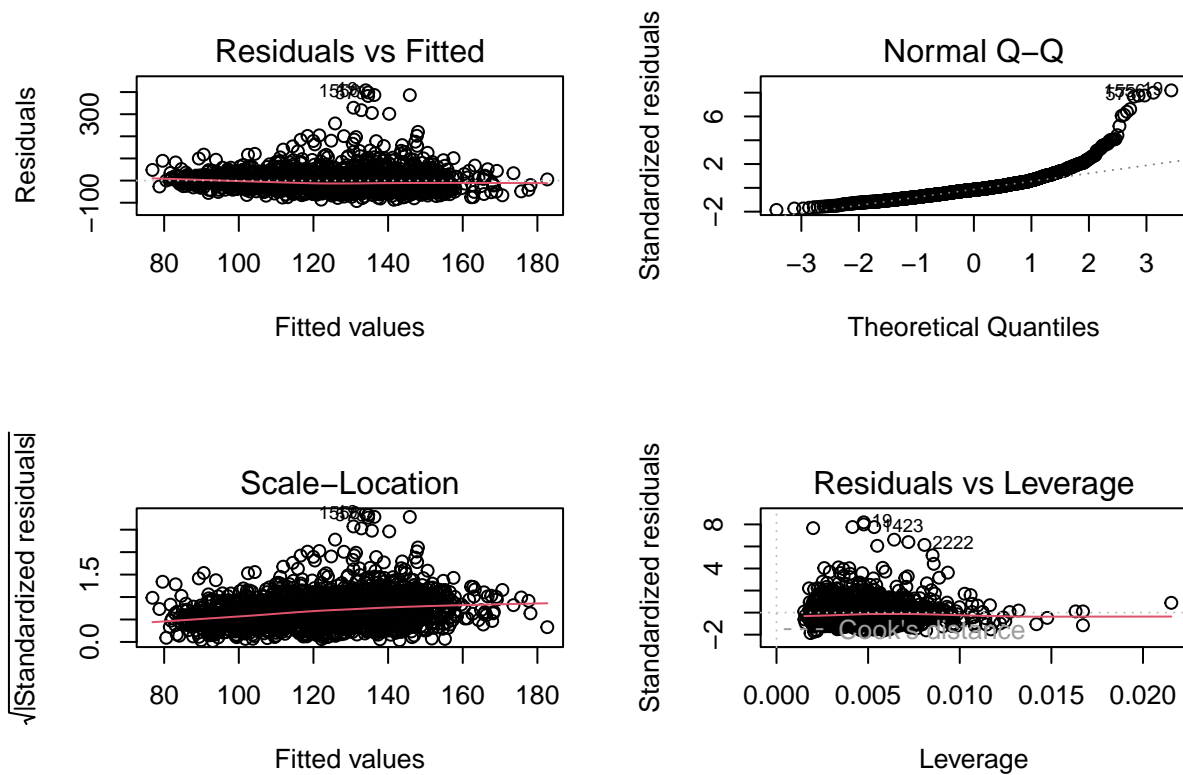
```
##
## Call:
## lm(formula = ogtt ~ . - weight - height, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -92.58  -28.78   -7.93   16.32  408.00
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  40.47390   11.29572   3.583    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: < 0.000000000000000022
```

```
#Producing diagnostics 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")
```

```
##
```

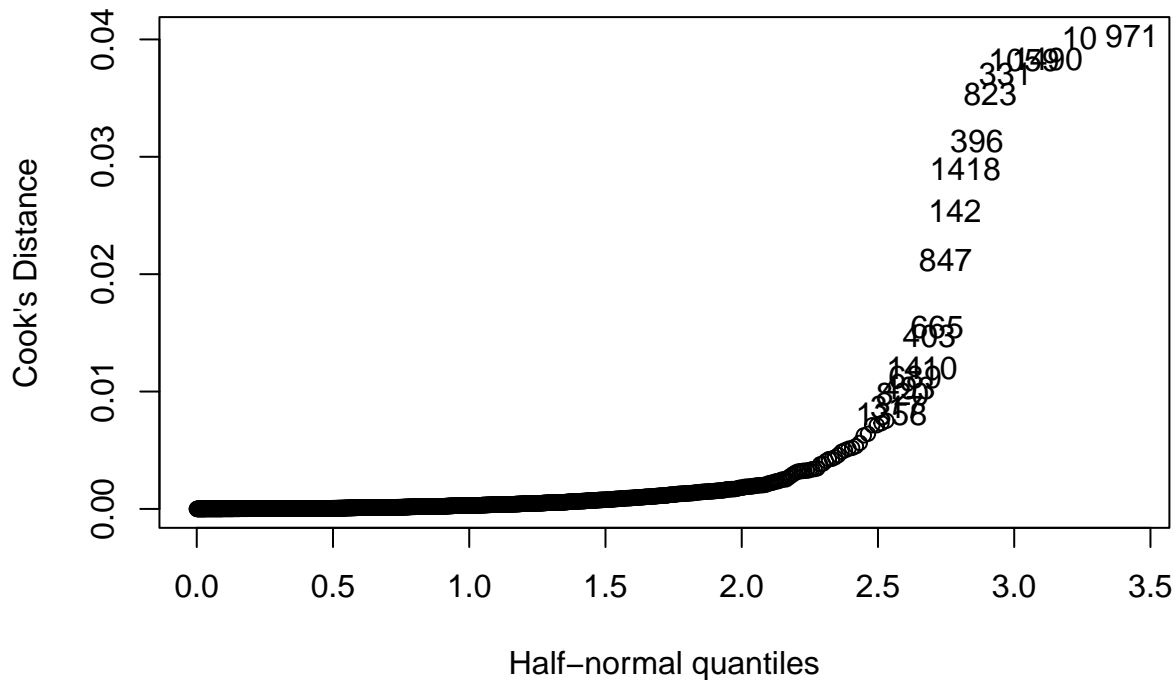
```
## Attaching package: 'faraway'
```

```
## 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
```

#observation 10 and 971 has the high distance

```
# Detect outliers using studentized residuals. Use the Bonferroni correction.
```

```
stud <- rstudent(lmod)
stud[which.max(abs(stud))]
```

19

```
## 8.357222
```

```
p <- 5
n <- nrow (model.matrix(lmod))
qt(1-.05 / (n*2), n-p-1)
```

```
## [1] 4.185502
```

```
which(abs(stud) > qt(1-.05/(n*2),n-p-1))
```

##	19	221	488	573	583	1205	1242	1423	1556	2108	2222
##	10	142	331	396	403	823	847	971	1059	1418	1490

```
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	-0.04	-0.05	-0.05	-0.68	0.13	0.03	0.10	-0.04
## weight	0.46	1.00	-0.05	0.14	-0.01	-0.29	0.01	-0.04	-0.09	0.86
## hi	-0.04	-0.05	1.00	-0.03	-0.27	-0.04	-0.24	0.17	-0.13	-0.03
## ogtt	-0.05	0.14	-0.03	1.00	0.34	0.01	-0.10	-0.14	-0.06	0.18
## age	-0.05	-0.01	-0.27	0.34	1.00	0.02	-0.07	-0.35	-0.08	0.03


```
## gender -0.68 -0.29 -0.04 0.01 0.02 1.00 0.02 0.01 -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 -0.14 -0.35 0.01 -0.04 1.00 0.01 -0.06
## race 0.10 -0.09 -0.13 -0.06 -0.08 -0.05 0.30 0.01 1.00 -0.16
## bmi -0.04 0.86 -0.03 0.18 0.03 0.05 -0.06 -0.06 -0.16 1.00
```

```
vif(lmod)
```

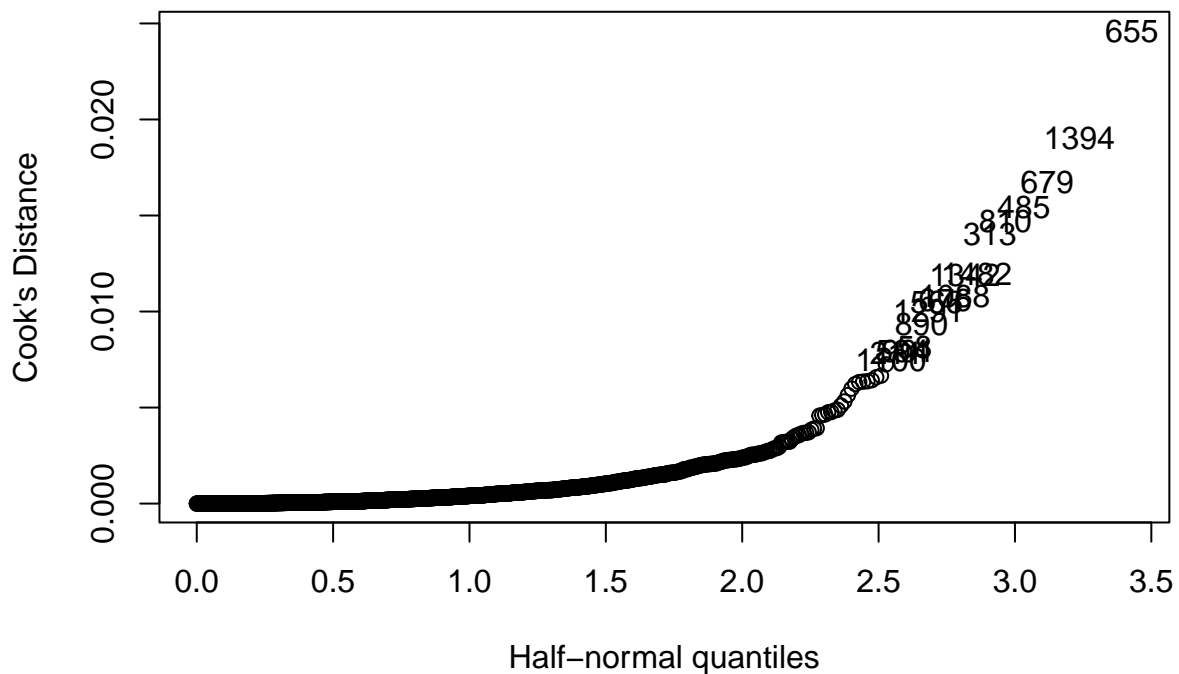
```
## hi age gender edu marry race bmi
## 1.179021 1.231975 1.007441 1.170569 1.153592 1.144568 1.033909
```

```
#new model without outliers
```

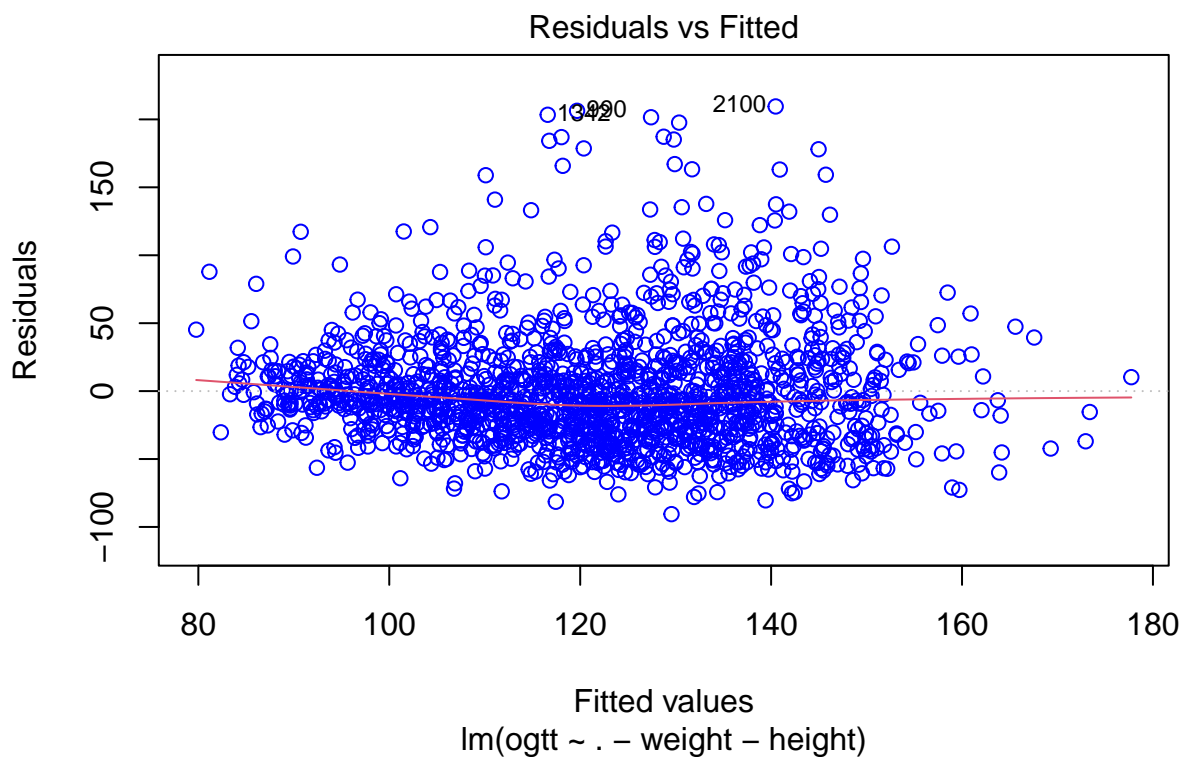
```
model2 <- lm(ogtt ~ .-weight-height, subsetdf)
summary(model2)
```

```
##
## Call:
## lm(formula = ogtt ~ . - weight - height, data = subsetdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -90.567 -27.070  -6.727  17.465  209.521
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  46.44578    9.49170   4.893 0.000001088828390 ***
## hi           5.81777    2.76504   2.104    0.03553 *
## age          0.86729    0.06543  13.255 < 0.0000000000000002 ***
## gender      -0.11337    2.05443  -0.055    0.95600
## edu         -2.27663    0.86579  -2.630    0.00863 **
## marry       -0.47407    0.57457  -0.825    0.40945
## race         0.56829    0.69143   0.822    0.41125
## bmi          1.19310    0.16364   7.291 0.0000000000000476 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.57 on 1642 degrees of freedom
## Multiple R-squared:  0.1486, Adjusted R-squared:  0.145
## F-statistic: 40.95 on 7 and 1642 DF, p-value: < 0.00000000000000022

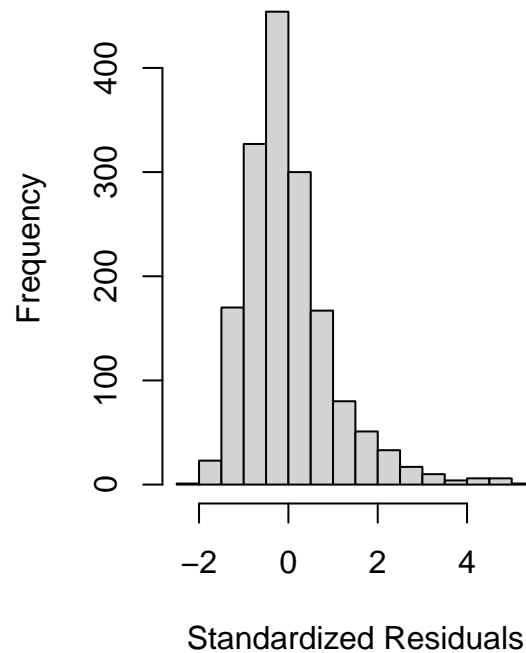
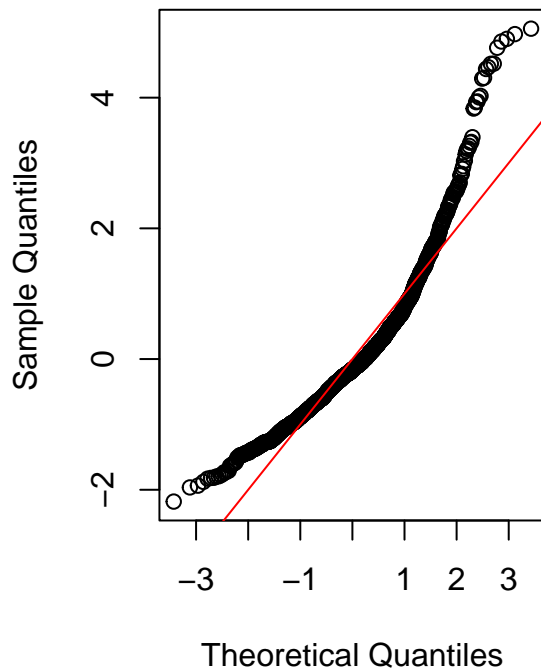
cook2 <- cooks.distance(model2)
halfnorm(cook2, 18, ylab = "Cook's Distance")
```



```
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

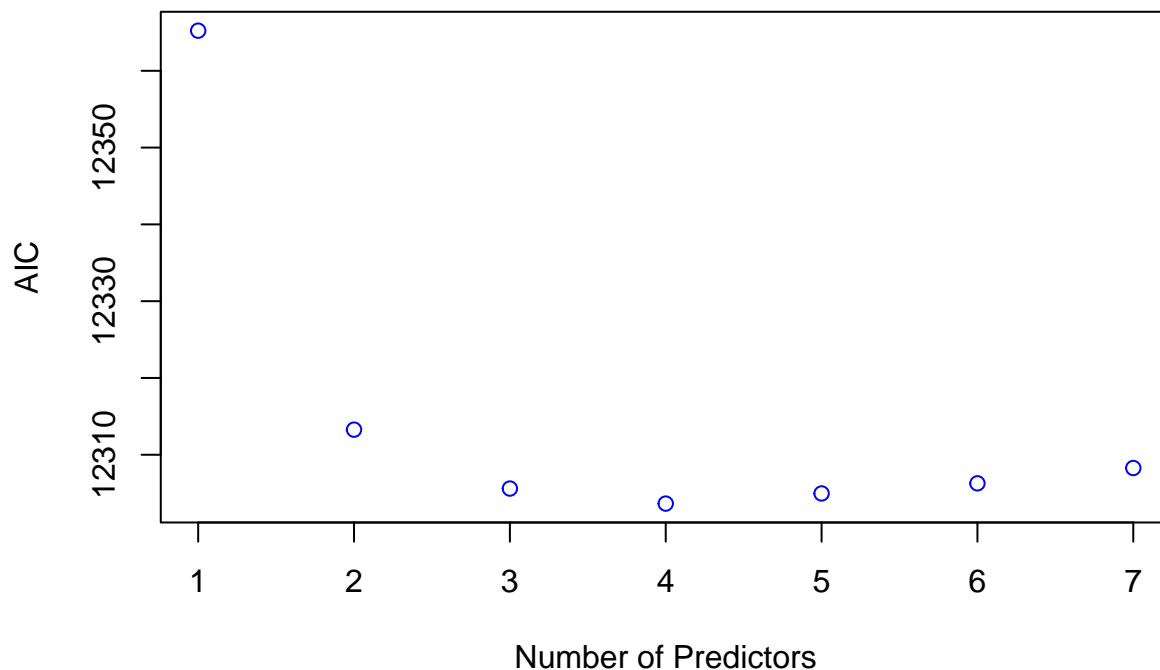
amod <- regsubsets(ogtt ~ .-weight-height, subsetdf)
rs <- summary(amod)
rs$which

##      (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 FALSE  TRUE   FALSE  TRUE  FALSE FALSE  TRUE
## 4      TRUE  TRUE  TRUE   FALSE  TRUE  FALSE FALSE  TRUE
## 5      TRUE  TRUE  TRUE   FALSE  TRUE  TRUE  FALSE  TRUE
## 6      TRUE  TRUE  TRUE   FALSE  TRUE  TRUE  TRUE  TRUE
## 7      TRUE  TRUE  TRUE    TRUE  TRUE  TRUE  TRUE  TRUE

rs$rss

## [1] 2958468 2863285 2846553 2839737 2838565 2837386 2837381

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
## -90.038 -27.130  -6.739  17.671 208.116
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  46.28888   8.14072   5.686 0.00000001535162 ***
## hi           5.44893   2.74227   1.987    0.0471 *
## age          0.88013   0.06178  14.247 < 0.0000000000000002 ***
## edu         -2.05719   0.83375  -2.467    0.0137 *
## bmi          1.17932   0.16114   7.319 0.000000000000039 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.55 on 1645 degrees of freedom
## Multiple R-squared:  0.1479, Adjusted R-squared:  0.1458
## F-statistic: 71.39 on 4 and 1645 DF,  p-value: < 0.00000000000000022
```

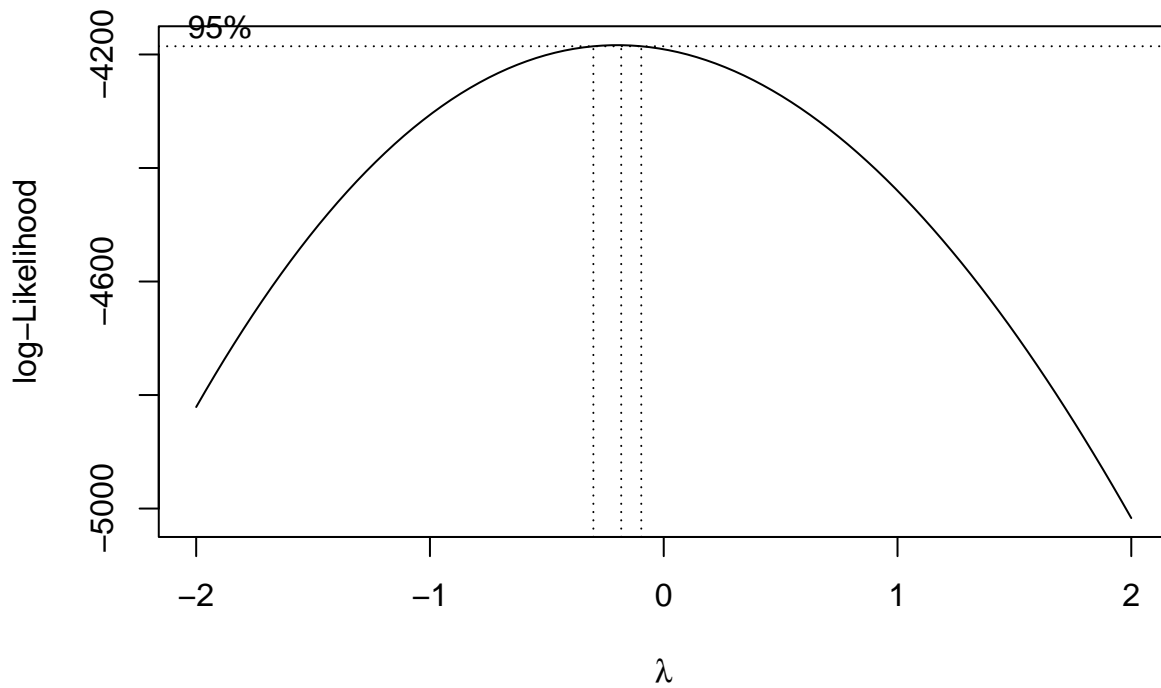
```
require(MASS)
```

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

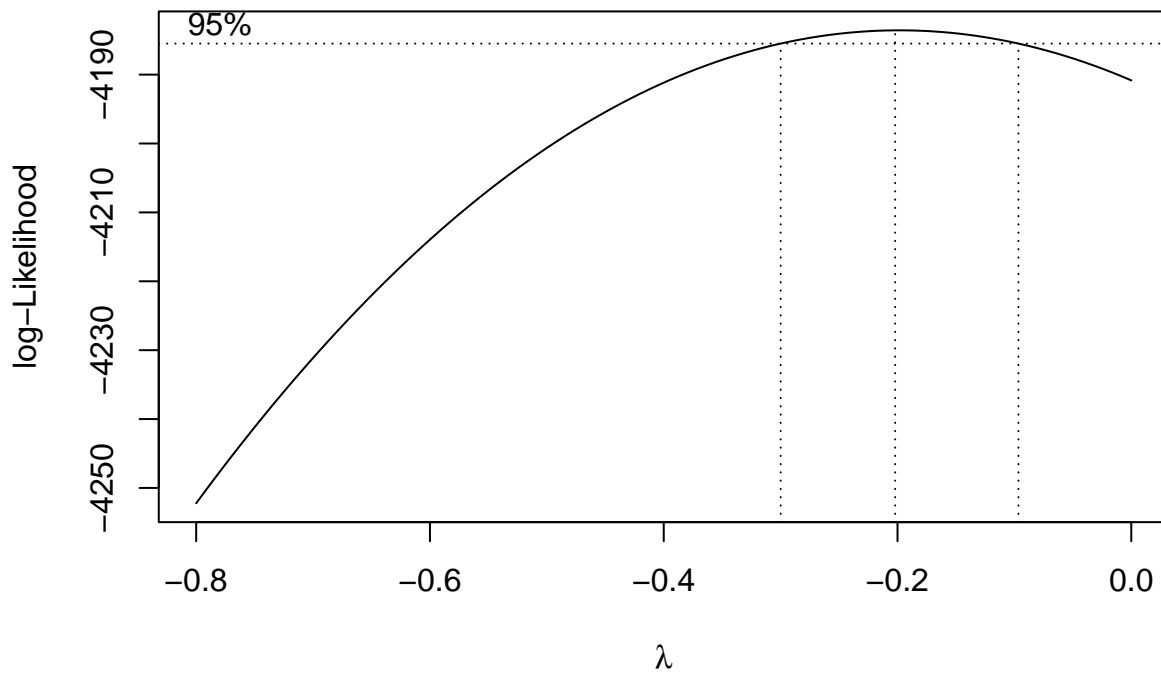
```
##
```

```
## 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))
```



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

```
## [1] -0.2020202
```

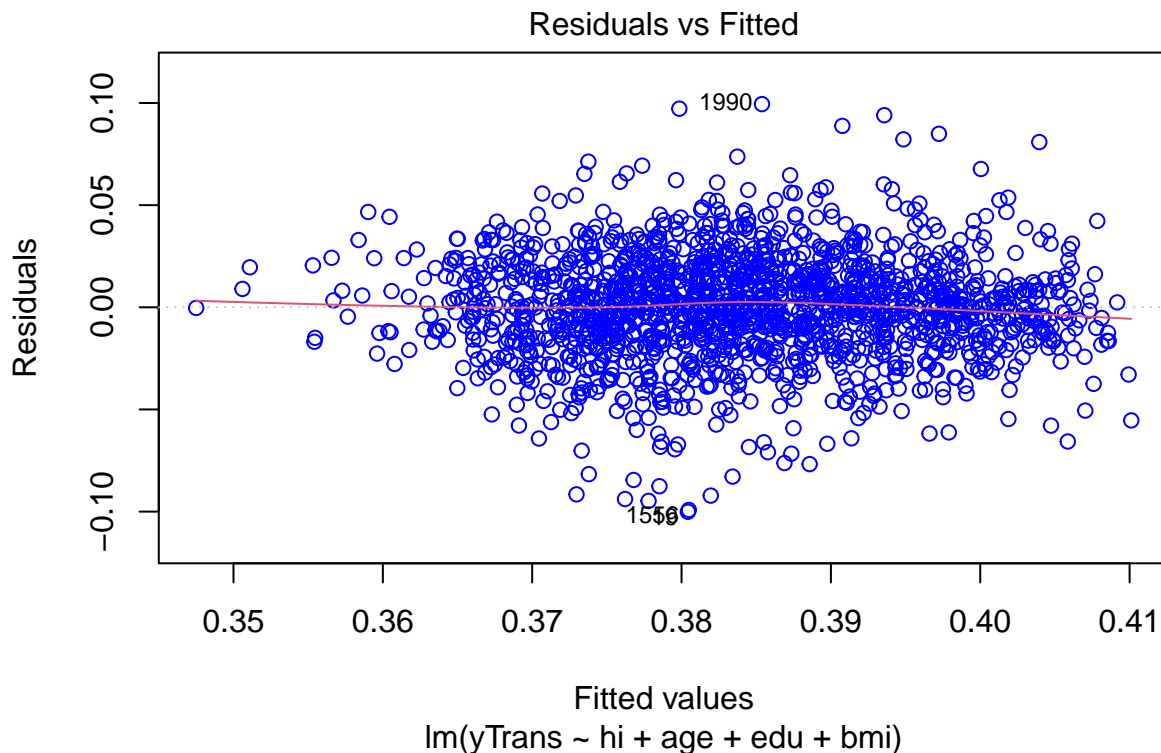
```

yTrans <- df$ogtt^(mLambda) # 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
## -0.100092 -0.014606  0.000447  0.015193  0.099451
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  0.43568807  0.00488297  89.226 < 0.0000000000000002 ***
## hi          -0.00490832  0.00164002  -2.993      0.0028 **
## age         -0.00055486  0.00003727 -14.886 < 0.0000000000000002 ***
## edu           0.00112185  0.00050058   2.241      0.0252 *
## bmi         -0.00080263  0.00009702  -8.273 0.000000000000000265 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02516 on 1665 degrees of freedom
## Multiple R-squared:  0.1598, Adjusted R-squared:  0.1578
## F-statistic: 79.19 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")

```

