NutriPLasmaPredict_Proj

Load Packages

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

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

Import Data

```
#import data
Plasma <- read.table("Plasma.txt", header = TRUE)

# Generate indices for train and test data
set.seed(23)
n = nrow(Plasma)
train_index = sample.int(n, size = 0.8*n)
train = Plasma[train_index, ]
valid = Plasma[-train_index, ]
```

1). Exploratory Data Analysis

Check variable types

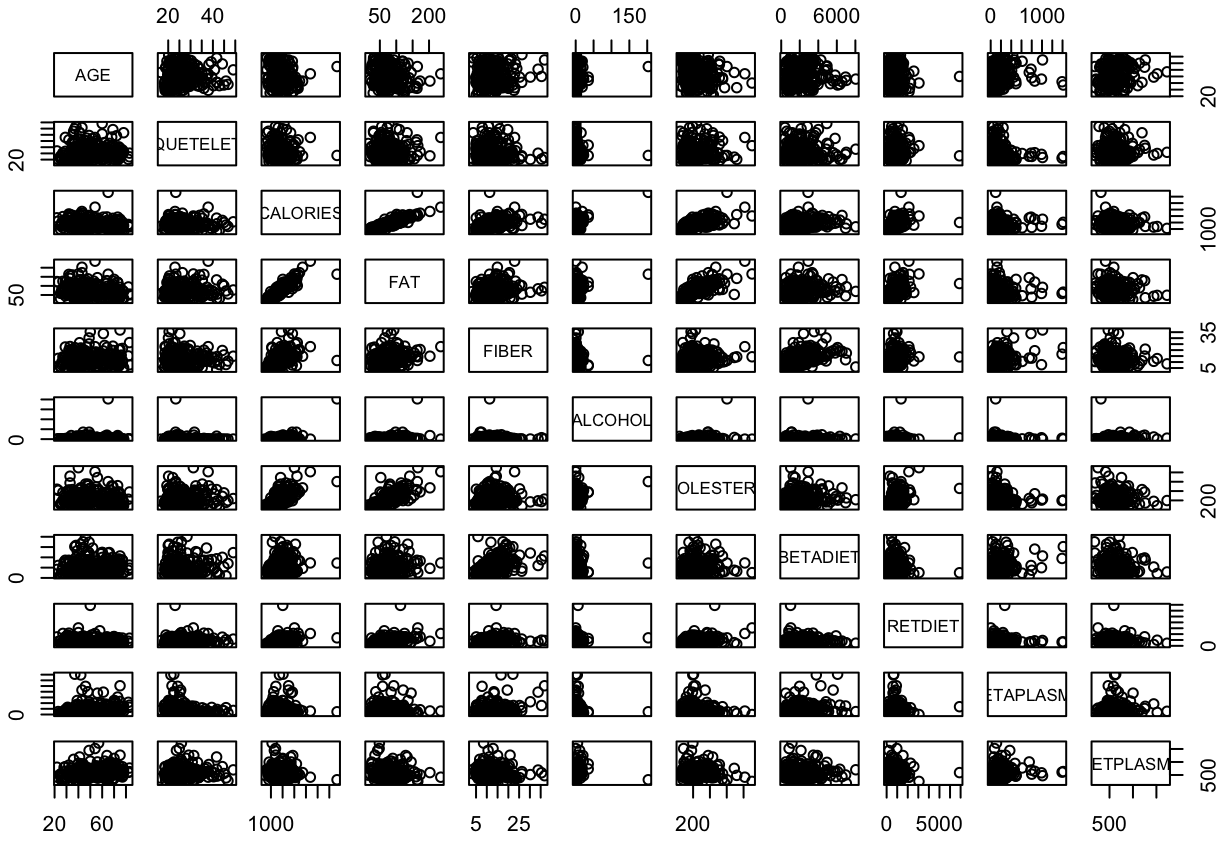
```
# check variable type
cplasm <- sapply(Plasma, class); cplasm
```

```
##         AGE         SEX    SMOKSTAT    QUETELET      VITUSE    CALORIES
##   "integer" "character" "character"   "numeric" "character"   "numeric"
##         FAT       FIBER     ALCOHOL CHOLESTEROL    BETADIET     RETDIET
##   "numeric"   "numeric"   "numeric"   "numeric"   "integer"   "integer"
##  BETAPLASMA   RETPLASMA
##   "integer"   "integer"
```

```
sumplasm <- sapply(Plasma, summary)

# transform categorical var into factor type
Plasma$SEX <- as.factor(Plasma$SEX)
Plasma$SMOKSTAT <- as.factor(Plasma$SMOKSTAT)
Plasma$VITUSE <- as.factor(Plasma$VITUSE)
```

Inspect multicollinearity

```
num = subset(train, select = -c(SEX, SMOKSTAT, VITUSE) )
cor_num <- cor(num)
pairs(num)
```



Inspect Categorical Variable

```
#check categorical var
table(Plasma$SEX)
```

```
##
## FEMALE   MALE
##    273     42
```

```
table(Plasma$SMOKSTAT)
```

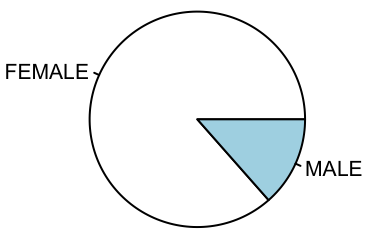
```
##
## CURRENT  FORMER   NEVER
##      43     115     157
```

```
table(Plasma$VITUSE)
```

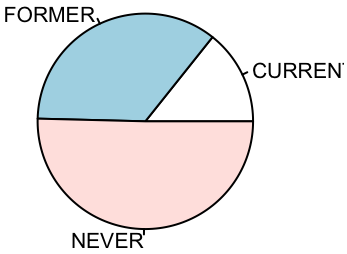
```
##
##        NO NOT OFTEN     OFTEN
##       111        82       122
```

```
# check pie chart of categorical variables
plasma_cat = train[,(names(train)%in%c("SEX","SMOKSTAT", "VITUSE"))]
par(mfrow=c(1,3))
for (j in 1:3){
  n <- nrow(plasma_cat[,j])
  lbls <- names(table(plasma_cat[,j]))
  lab <- paste(lbls)
  pie(table(plasma_cat[,j]), labels = lab, main = paste("Pie chart of",colnames(plasma_c
at[j])))
}
```

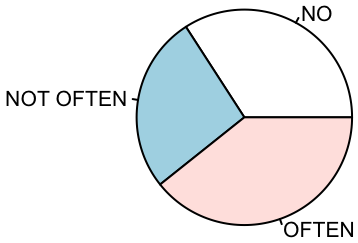
Pie chart of SEX



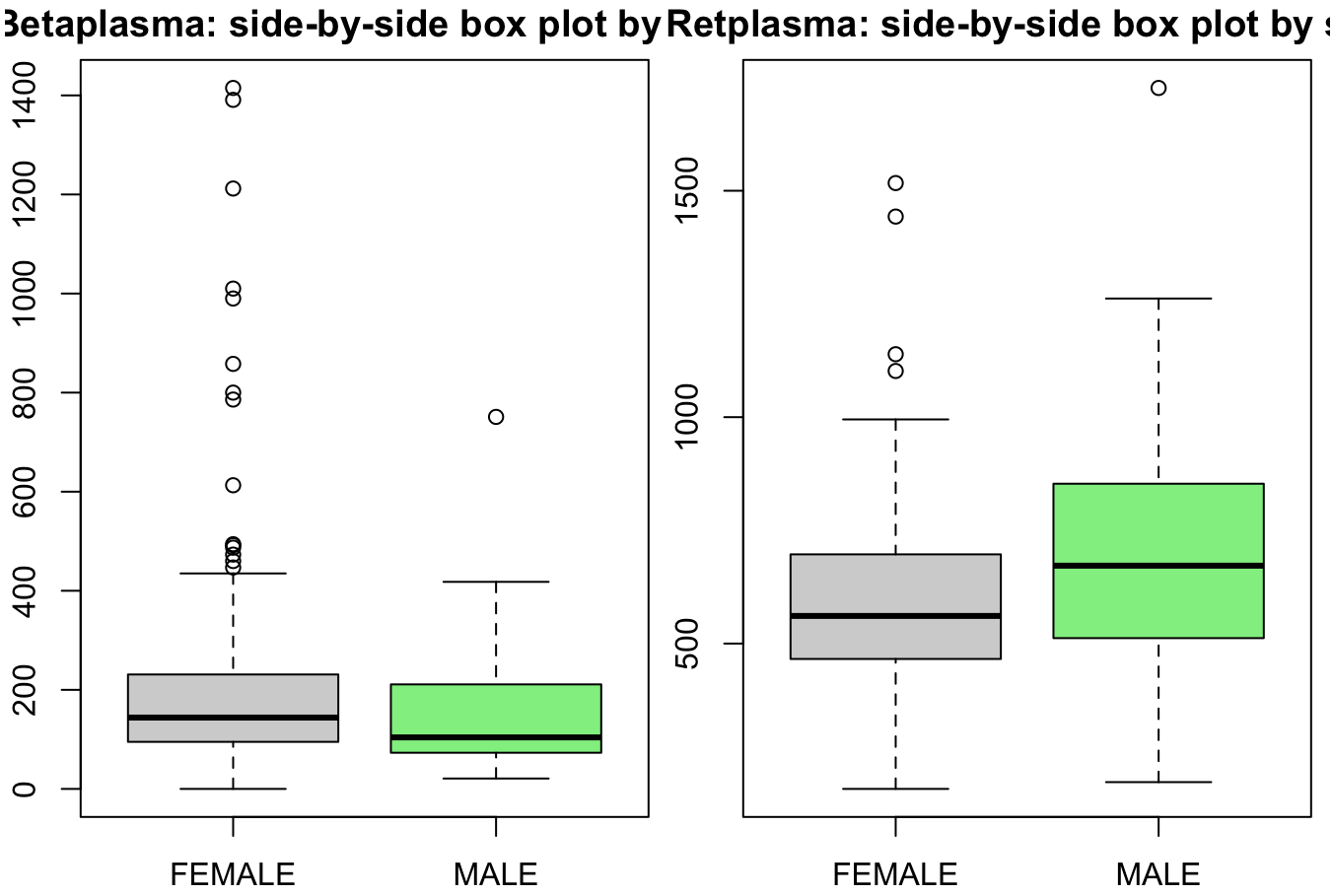
Pie chart of SMOKSTAT



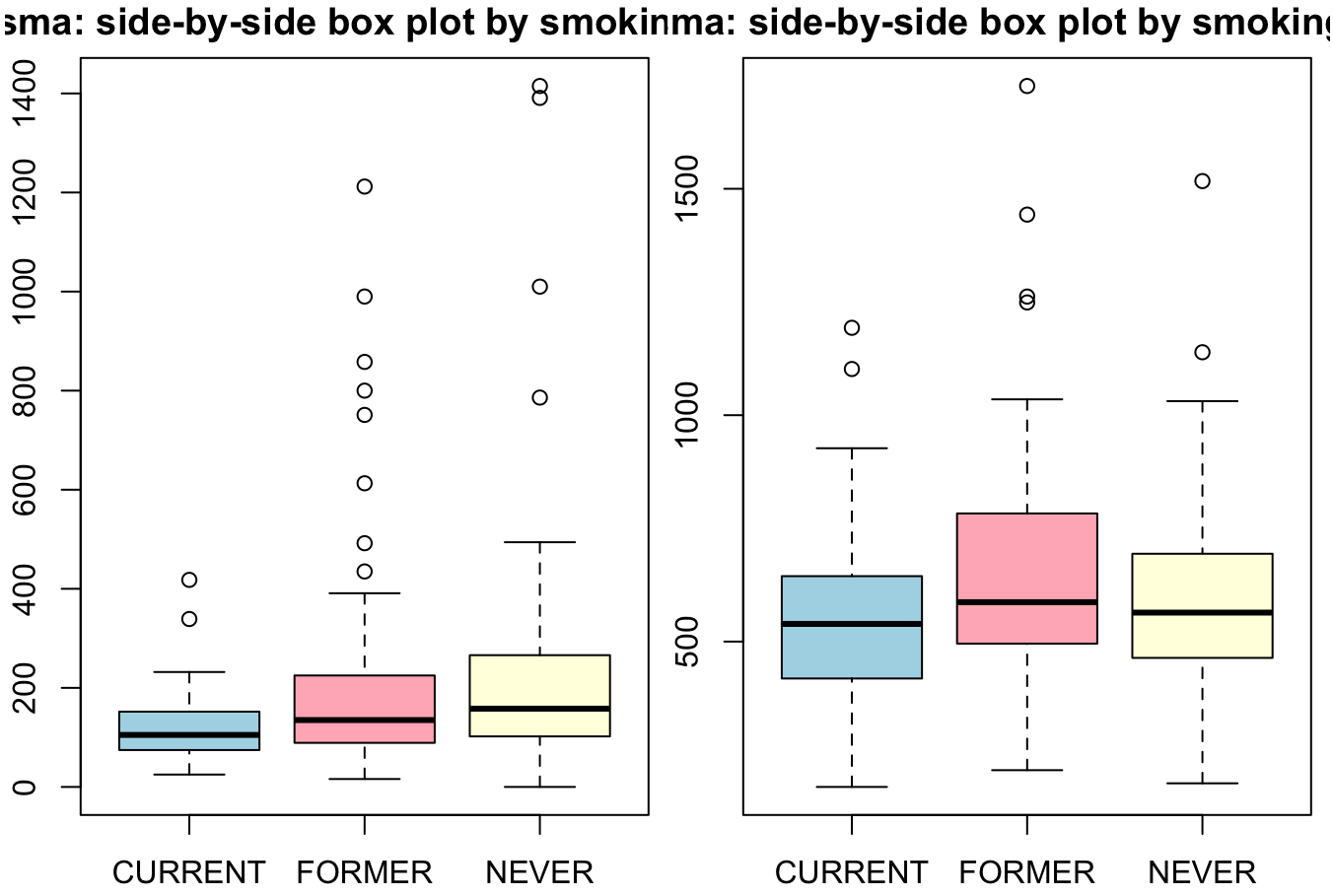
Pie chart of VITUSE



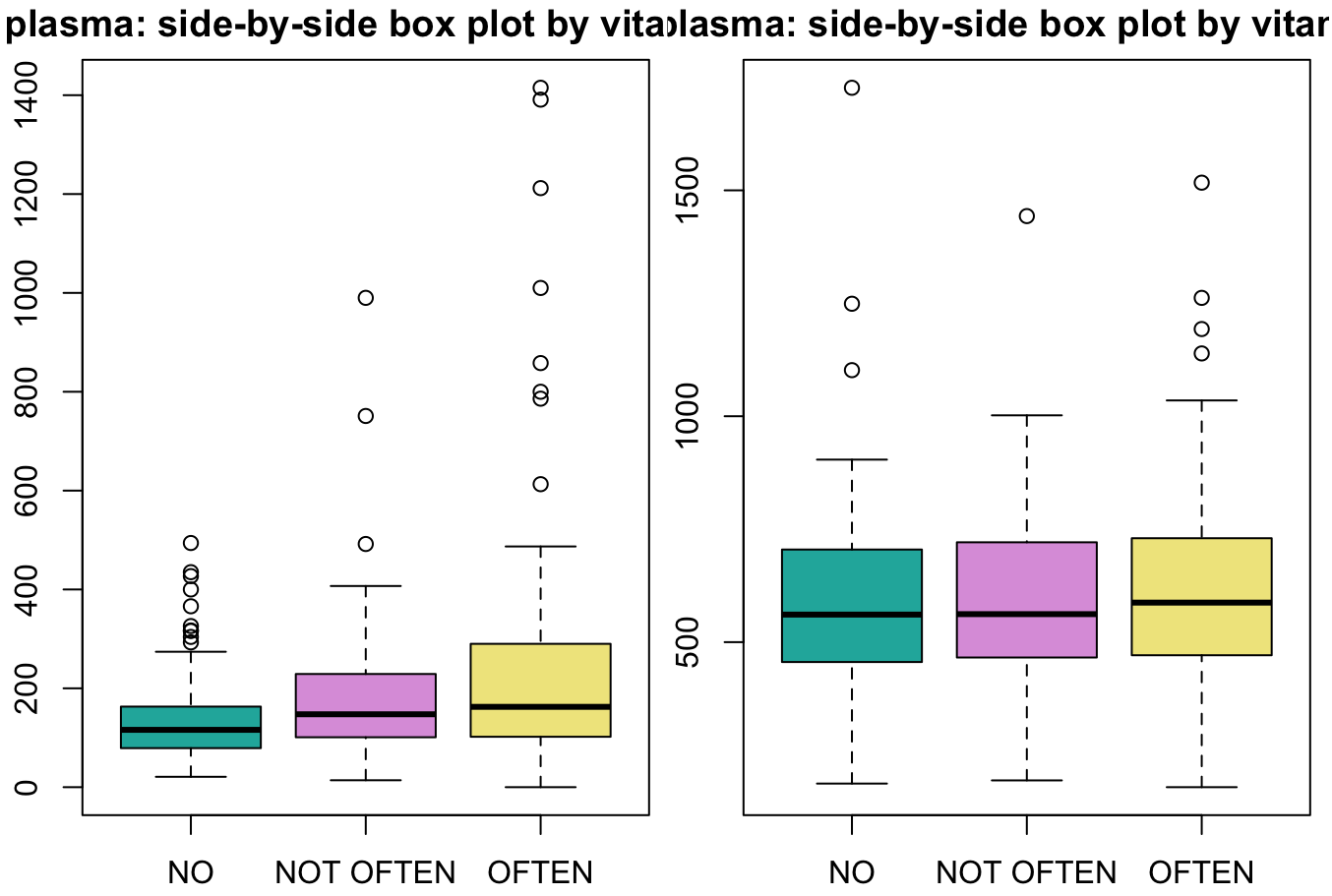

```
# checking boxplot of categorical var
par(mfrow=c(1,2), mar=c(3,2,2,.5), mgp=c(3,1,0))
boxplot(Plasma$BETAPLASMA~Plasma$SEX,main='Betaplasma: side-by-side box plot by sex',xla
b='sex',ylab='betaplasma', col = c("lightgrey","lightgreen"))
boxplot(Plasma$RETPLASMA~Plasma$SEX,main='Retplasma: side-by-side box plot by sex',xlab
='sex',ylab='Retplasma', col =c("lightgrey","lightgreen"))
```



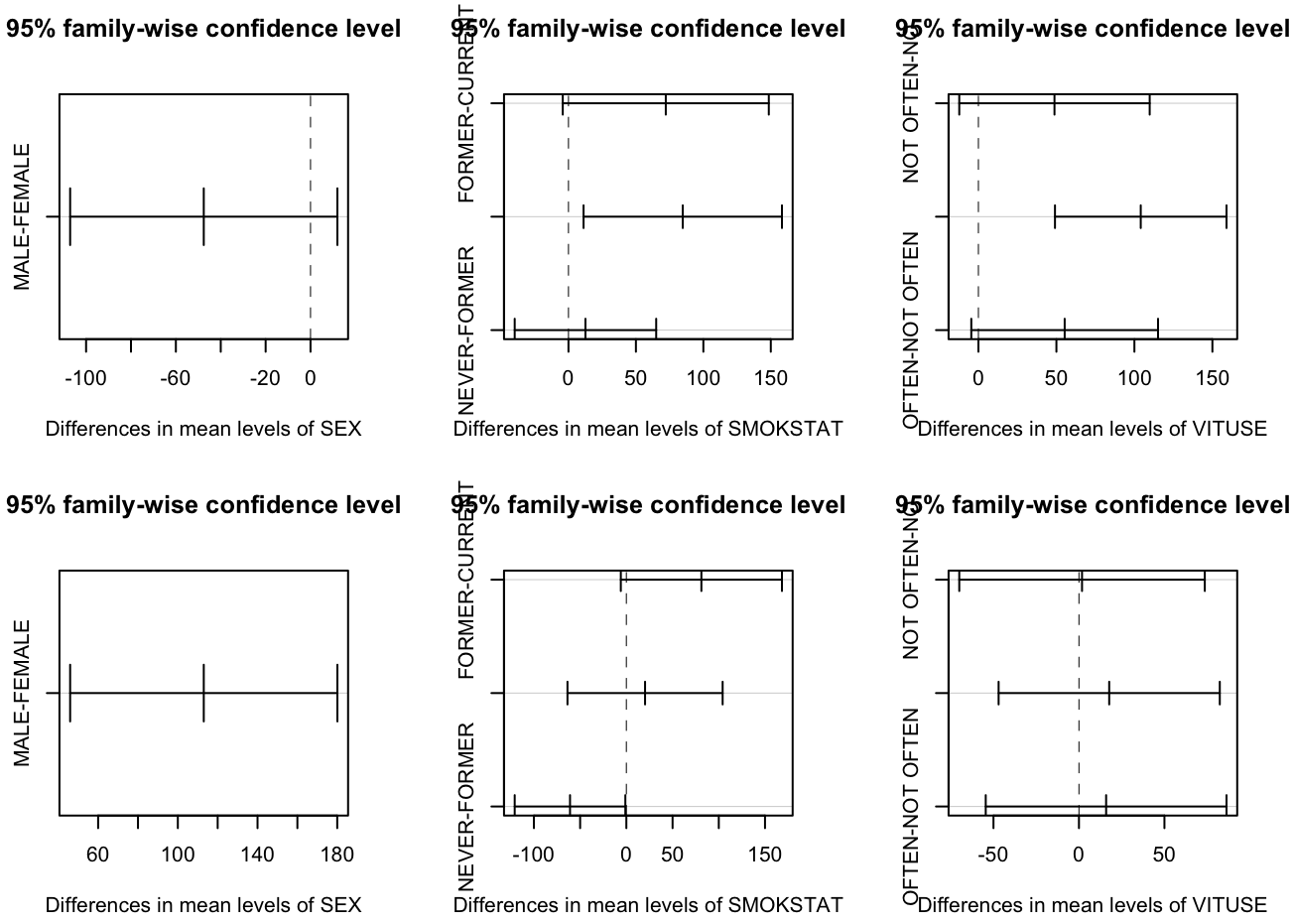

```
boxplot(Plasma$BETAPLASMA~Plasma$SMOKSTAT,main='Betaplasma: side-by-side box plot by smo
king activities',
xlab='sex',ylab='betaplasma', col = c("lightblue", "lightpink","lightyellow"))
boxplot(Plasma$RETPLASMA~Plasma$SMOKSTAT,main='Retplasma: side-by-side box plot by smoki
ng activities',
xlab='sex',ylab='Retplasma', col = c("lightblue", "lightpink","lightyellow"))
```




```
boxplot(Plasma$BETAPLASMA~Plasma$VITUSE,main='Betaplasma: side-by-side box plot by vitam
in use',
xlab='sex',ylab='betaplasma', col = c("lightseagreen", "plum","khaki"))
boxplot(Plasma$RETPLASMA~Plasma$VITUSE,main='Retplasma: side-by-side box plot by vitamin
use',
xlab='sex',ylab='Retplasma', col = c("lightseagreen", "plum","khaki"))
```



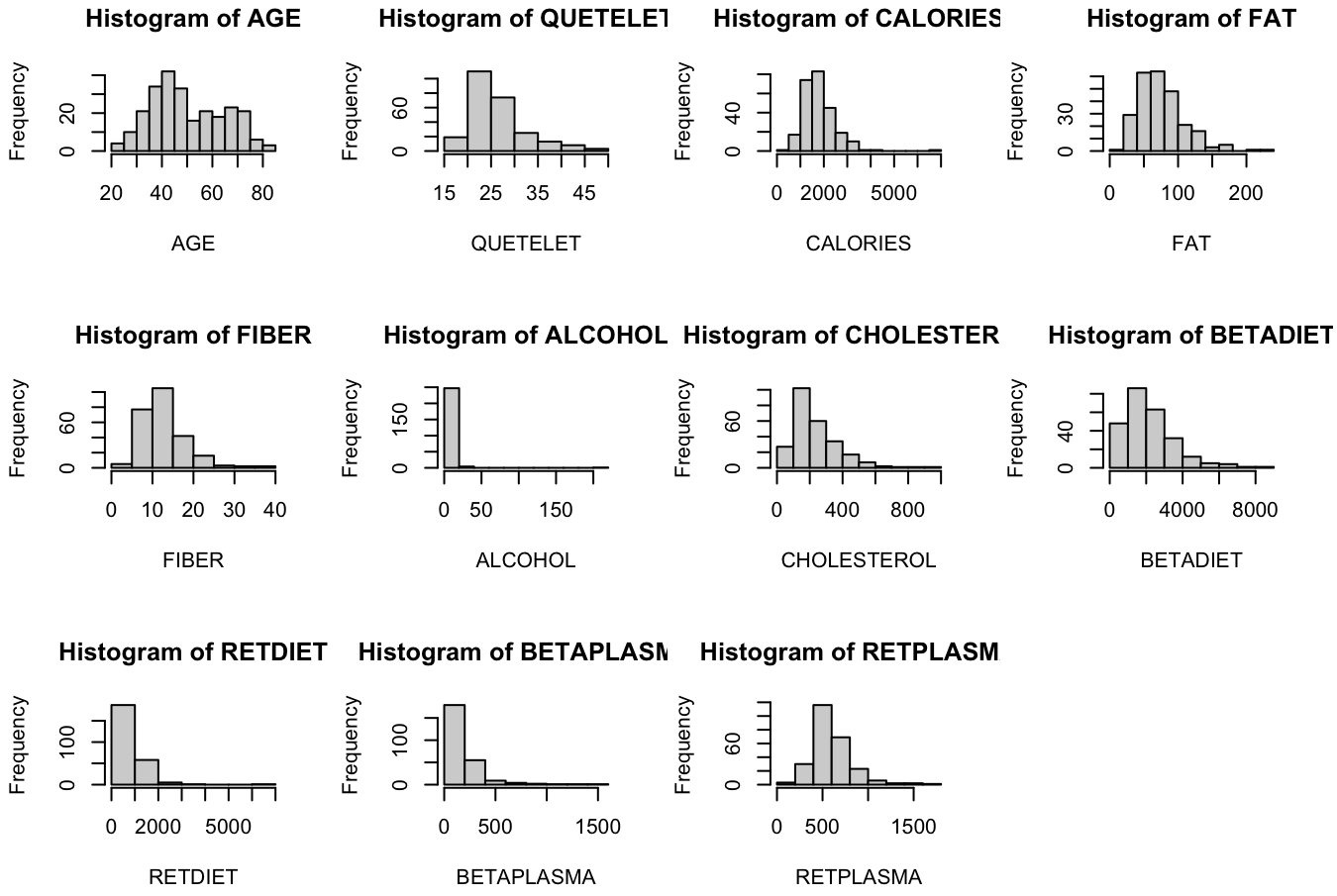

```
# Check if there is significant differences between groups
# Use Tukey's HSD- honest significant difference
T1 <- TukeyHSD(aov(BETAPLASMA ~ SEX, Plasma))
T2 <- TukeyHSD(aov(BETAPLASMA ~ SMOKSTAT, Plasma))
T3 <- TukeyHSD(aov(BETAPLASMA ~ VITUSE, Plasma))

T4 <- TukeyHSD(aov(RETPLASMA ~ SEX, Plasma))
T5 <- TukeyHSD(aov(RETPLASMA ~ SMOKSTAT, Plasma))
T6 <- TukeyHSD(aov(RETPLASMA ~ VITUSE, Plasma))

par(mfrow=c(2,3))
plot(T1)
plot(T2)
plot(T3)
plot(T4)
plot(T5)
plot(T6)
```



Check Numerical Variable

```
# check numeric var
sumplasm <- sapply(num, summary)
describe(num,fast=TRUE)
knitr::kable(round(describe(num,fast=TRUE),3), format = "html")
```

```
# check hist plot of numeric variables
par(mfrow=c(3,4))
for (i in 1:11){
  hist(num[,i], main = paste("Histogram of",colnames(num[i])), xlab = paste(colnames(num
[i])))
}
```



2. Model Fitting

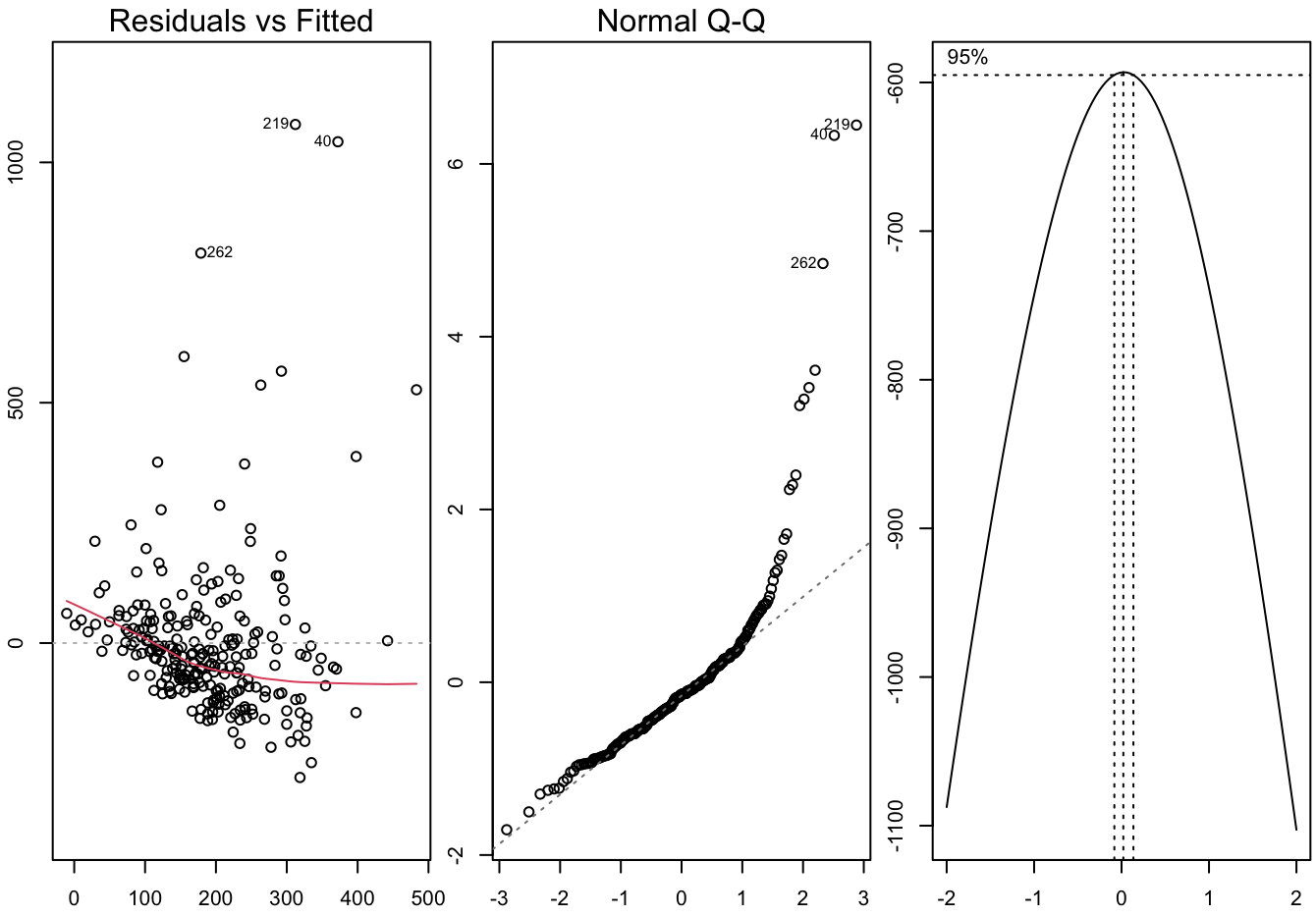
A. Betaplasma

First order model

```
# Use stepwise regression to find first order model
beta = subset(train, select = -c(RETPLASMA) )
fit.0=lm(BETAPLASMA~1, data=beta) ##initial model, only intercept
step.aic=stepAIC(fit.0, scope=list(upper=~AGE+SEX+SMOKSTAT+QUETELET+VITUSE+CALORIES+FAT+
FIBER+ALCOHOL+CHOLESTEROL+BETADIET+RETDIET, lower= ~1), direction="both", k=2, trace=FAL
SE)
step.aic$anova
```

Result : BETAPLASMA ~ FIBER + QUETELET + VITUSE + FAT + BETADIET + AGE

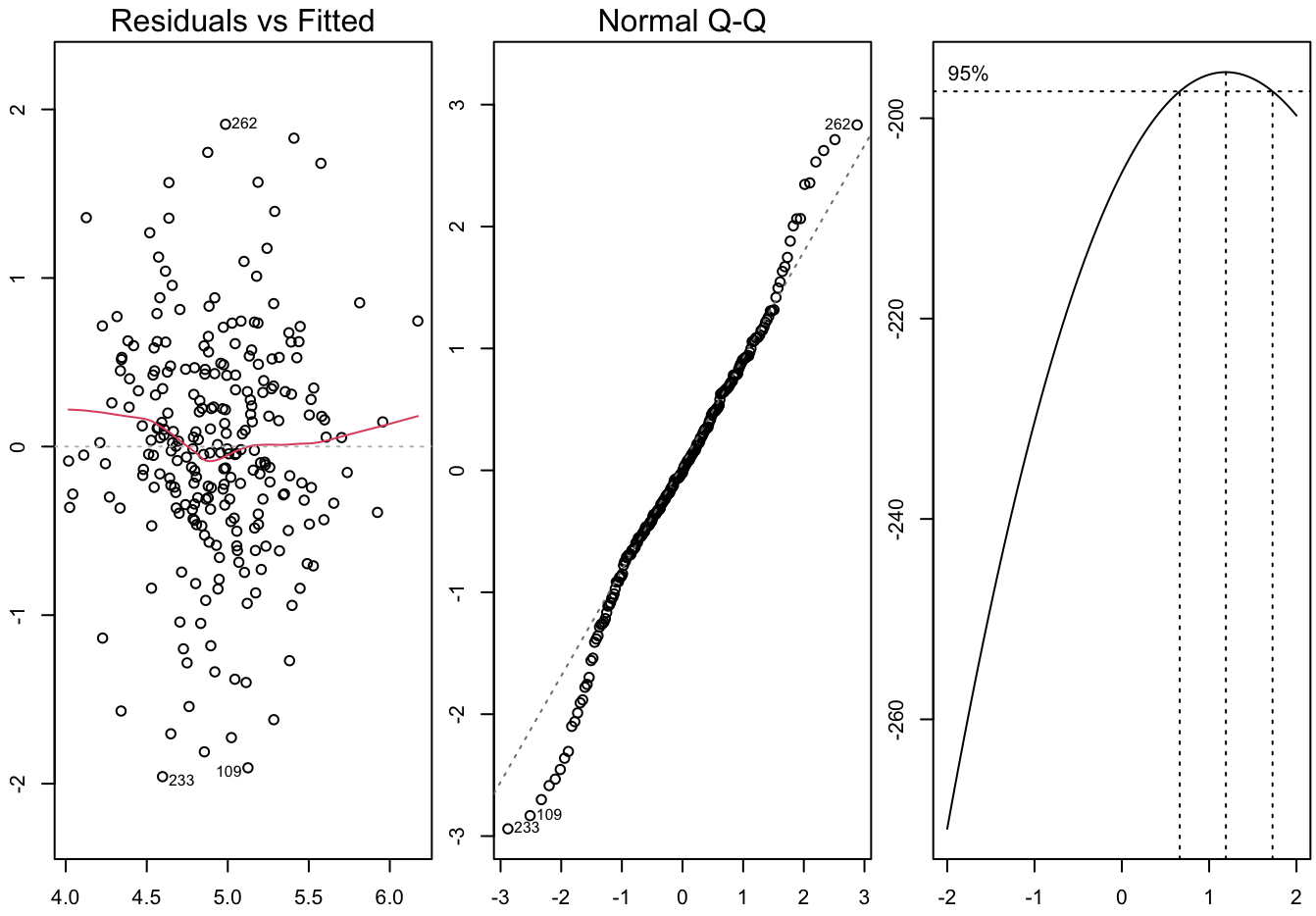
```
# Diagnostic of BETA model1
train_pos = beta[beta$BETAPLASMA > 0, ]
betamodel1 <- lm(BETAPLASMA ~ FIBER + QUETELET + VITUSE + FAT + BETADIET + AGE, data=tra
in_pos)
par(mfrow = c(1,3), mar=c(3,2,2,.5), mgp=c(3,1,0))
plot(betamodel1, which=c(1,2))
MASS::boxcox(betamodel1)
```



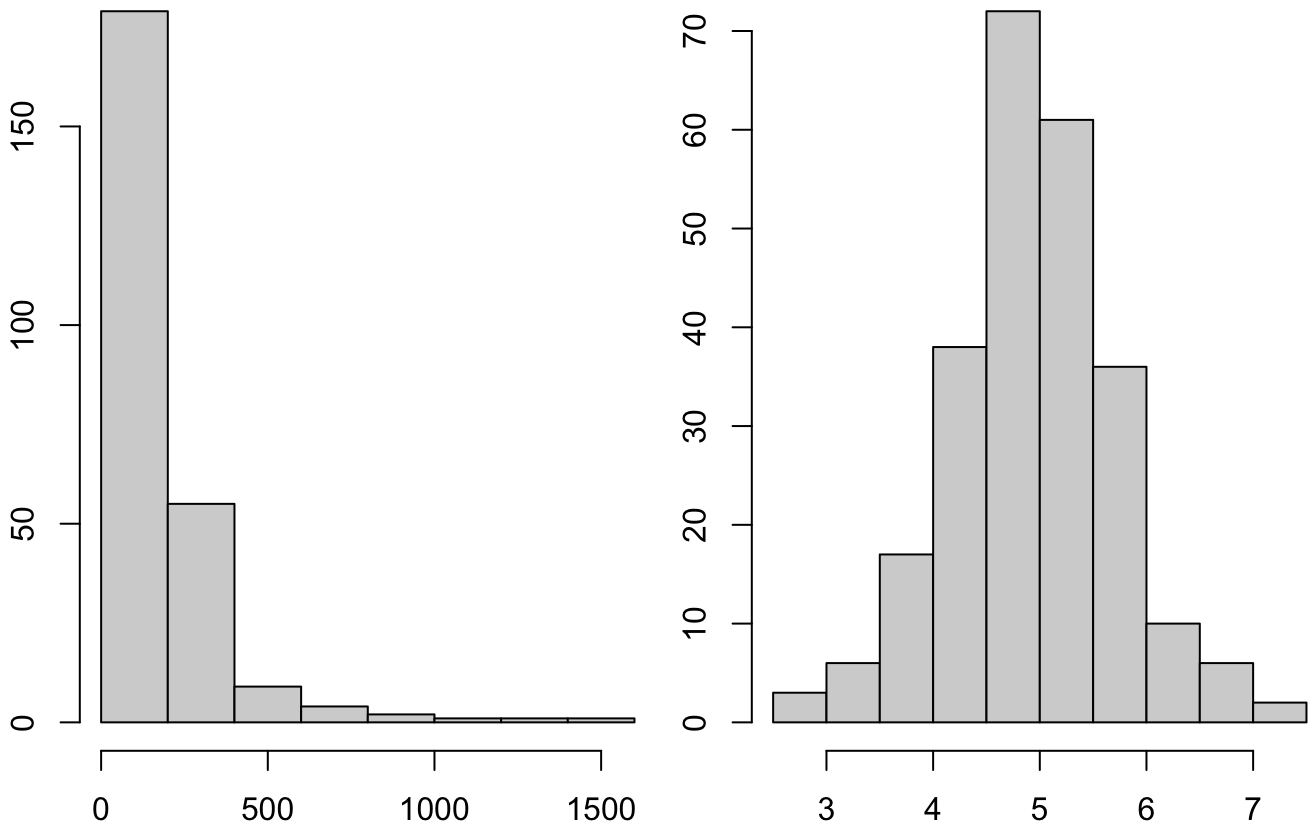

```
# log-transformation
log_betamodel1 <- lm(log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT + BETADIET + AGE,
data=train_pos)
summary(log_betamodel1)
```

```
##
## Call:
## lm(formula = log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT +
##     BETADIET + AGE, data = train_pos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95811 -0.35377 -0.01352  0.42705  1.91169
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.002e+00  2.895e-01  17.274  < 2e-16 ***
## FIBER            3.262e-02  9.329e-03   3.496  0.00056 ***
## QUETELET        -3.273e-02  7.267e-03  -4.503 1.04e-05 ***
## VITUSENOT OFTEN  3.449e-01  1.132e-01   3.047  0.00256 **
## VITUSEOFTEN      3.340e-01  1.019e-01   3.277  0.00120 **
## FAT             -4.085e-03  1.395e-03  -2.929  0.00373 **
## BETADIET         4.186e-05  3.556e-05   1.177  0.24028
## AGE              7.261e-03  3.101e-03   2.342  0.02000 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6817 on 243 degrees of freedom
## Multiple R-squared:  0.2354, Adjusted R-squared:  0.2134
## F-statistic: 10.69 on 7 and 243 DF,  p-value: 1.002e-11
```

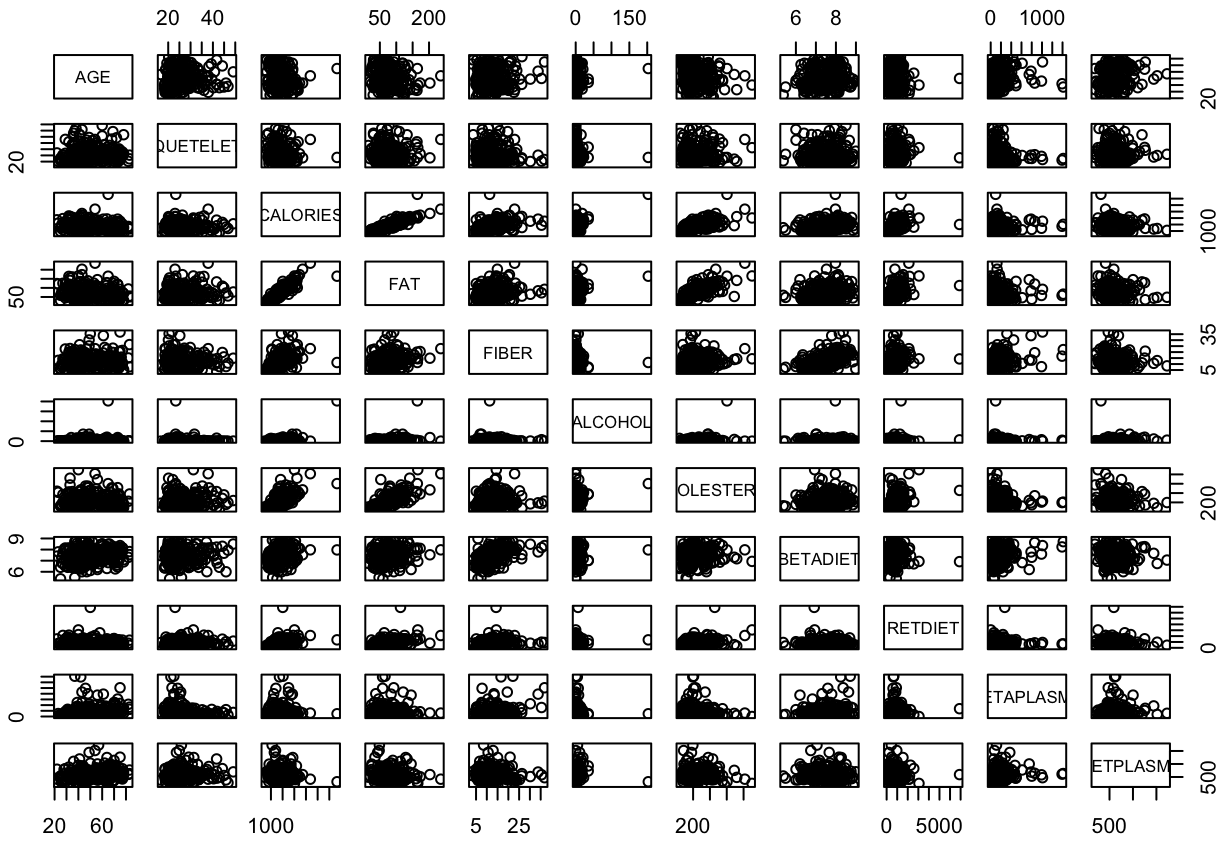
```
# diagnosis of log
par(mfrow = c(1,3), mar=c(3,2,2,.5), mgp=c(3,1,0))
plot(log_betamodel1, which= c(1,2))
MASS::boxcox(log_betamodel1)
```




```
# hist plot
par(mfrow = c(1,2), mar=c(3,2,2,.5), mgp=c(3,1,0))
hist(train$BETAPLASMA, main = paste("Histogram of betaplasma", xlab = "betaplasma"))
hist(log(train$BETAPLASMA), main = paste("Histogram of log betaplasma", xlab = "log beta
plasma"))
```




```
par(mfrow=c(1,1), mar=c(3,2,2,.5), mgp=c(3,1,0))
# linearity check
num$BETADIET <- log(num$BETADIET)
pairs(num)
```

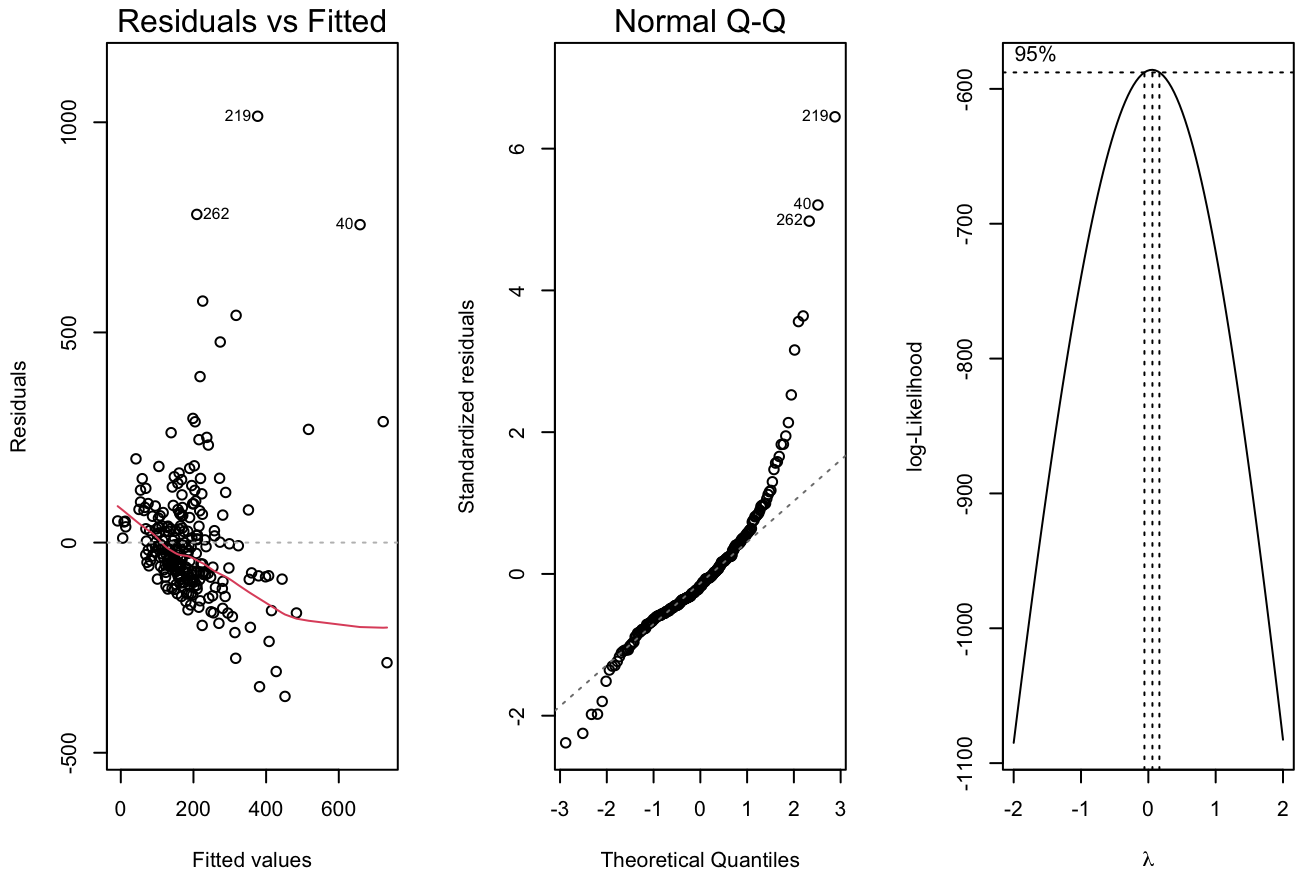


Model with interactive terms

```
#Adding interaction effects
#betamodel1 = lm(BETAPLASMA ~ FIBER + QUETELET + VITUSE + FAT + BETADIET, data = beta)
stepbeta <- step(betamodel1, scope = FIBER + QUETELET + VITUSE + FAT + BETADIET + AGE ~
.^2, direction = 'both')
```

Result : BETAPLASMA ~ FIBER + QUETELET + VITUSE + FAT + BETADIET + AGE + FIBER:VITUSE + FIBER:BETADIET + FAT:BETADIET + BETADIET:AGE + FAT:AGE

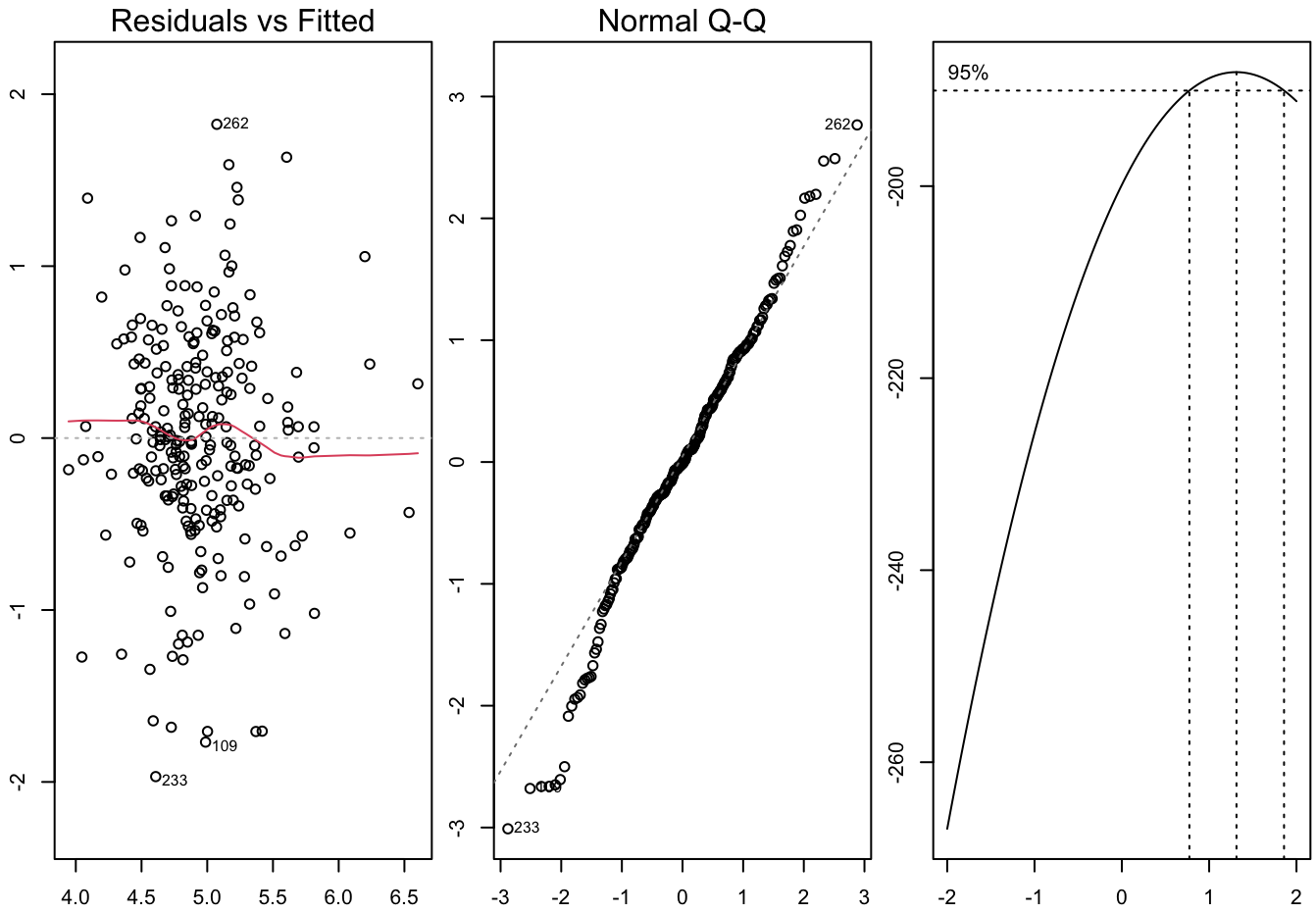
```
# Diagnostic of BETA model2
betamodel2 <- lm(formula = BETAPLASMA ~ FIBER + QUETELET + VITUSE + FAT + BETADIET + AGE
+
          FIBER:VITUSE + FIBER:BETADIET + FAT:BETADIET + BETADIET:AGE + FAT:AGE, data =
train_pos)

par(mfrow = c(1,3))
plot(betamodel2, which=c(1,2))
MASS::boxcox(betamodel2)
```




```
# log-transformation
log_betamodel2 <- lm(formula = log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT + BETAD
IET + AGE + FIBER:VITUSE + FIBER:BETADIET + FAT:BETADIET + BETADIET:AGE + FAT:AGE, data
= train_pos)
summary(log_betamodel2)
```

```
##
## Call:
## lm(formula = log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT +
##     BETADIET + AGE + FIBER:VITUSE + FIBER:BETADIET + FAT:BETADIET +
##     BETADIET:AGE + FAT:AGE, data = train_pos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96893 -0.35025 -0.00864  0.41613  1.82490
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.005e+00  5.613e-01   8.916  < 2e-16 ***
## FIBER                  9.299e-05  2.090e-02   0.004   0.9965
## QUETELET              -2.947e-02  7.234e-03  -4.074 6.31e-05 ***
## VITUSENOT OFTEN        3.001e-01  3.149e-01   0.953   0.3415
## VITUSEOFTEN           -1.970e-01  2.480e-01  -0.794   0.4278
## FAT                   -5.929e-03  5.561e-03  -1.066   0.2874
## BETADIET               3.477e-04  1.787e-04   1.946   0.0529 .
## AGE                    7.021e-03  8.706e-03   0.806   0.4208
## FIBER:VITUSENOT OFTEN  3.177e-03  2.411e-02   0.132   0.8953
## FIBER:VITUSEOFTEN      4.107e-02  1.783e-02   2.303   0.0221 *
## FIBER:BETADIET         4.859e-06  6.013e-06   0.808   0.4198
## FAT:BETADIET          -2.015e-06  1.114e-06  -1.809   0.0716 .
## BETADIET:AGE          -4.499e-06  2.747e-06  -1.638   0.1028
## FAT:AGE                1.338e-04  1.006e-04   1.330   0.1848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6716 on 237 degrees of freedom
## Multiple R-squared:  0.2762, Adjusted R-squared:  0.2365
## F-statistic: 6.957 on 13 and 237 DF,  p-value: 2.268e-11
```

```
# diagnosis of log
par(mfrow = c(1,3), mar=c(3,2,2,.5), mgp=c(3,1,0))
plot(log_betamodel2, which=c(1,2))
MASS::boxcox(log_betamodel2)
```




```
summary(log_betamodel2)
```

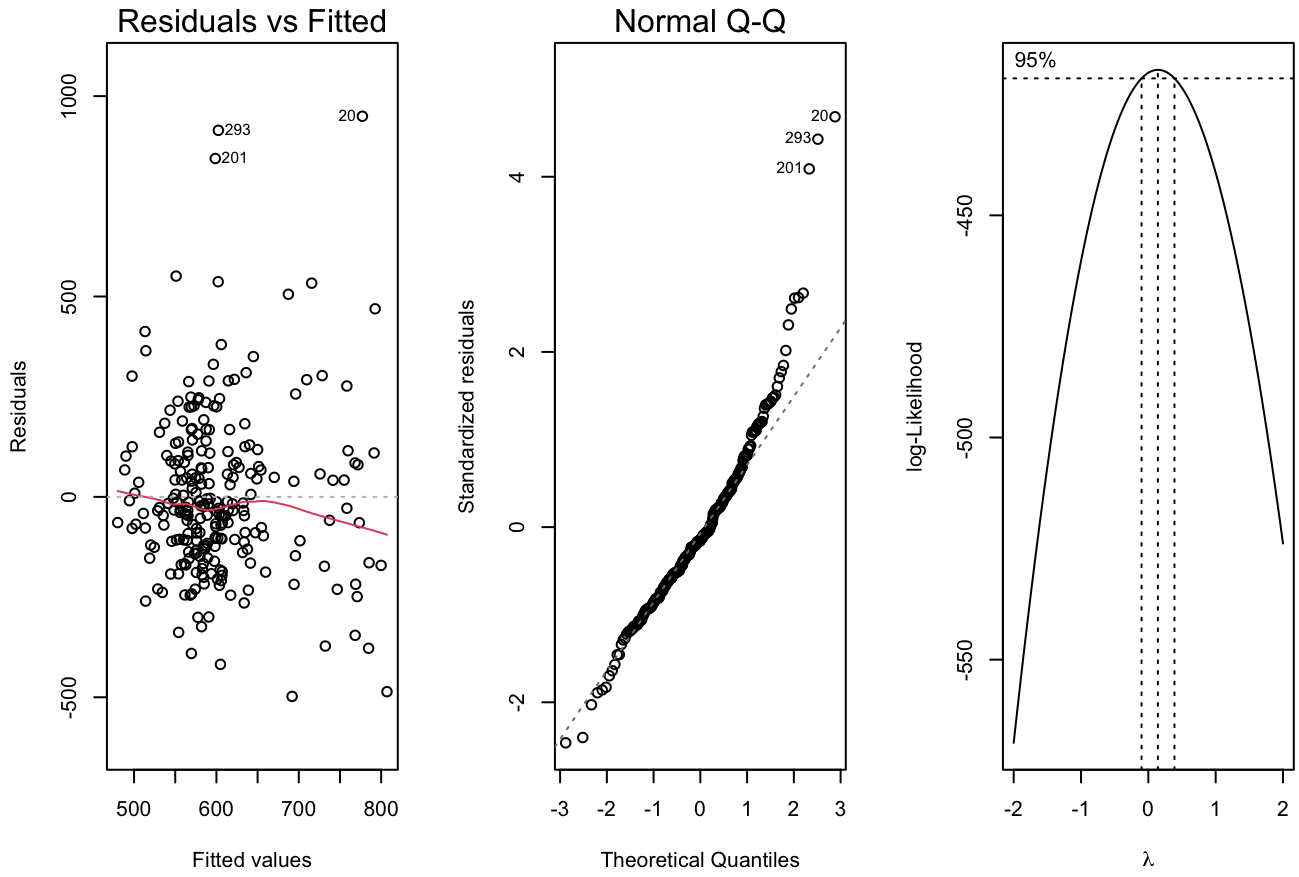
B. Retoplasma

First order model

```
# Stepwise Regression for first order
ret = subset(train, select = -c(BETAPLASMA) )
fit.02=lm(RETPLASMA~1, data=ret) ##initial model, only intercept
step.aic2=stepAIC(fit.02, scope=list(upper=~AGE+SEX+SMOKSTAT+QUETELET+VITUSE+CALORIES+FA
T+FIBER+ALCOHOL+CHOLESTEROL+BETADIET+RETDIET, lower= ~1), direction="both", k=2, trace=F
ALSE)
step.aic2$anova
```

Result : RETPLASMA ~ SEX + CALORIES + AGE

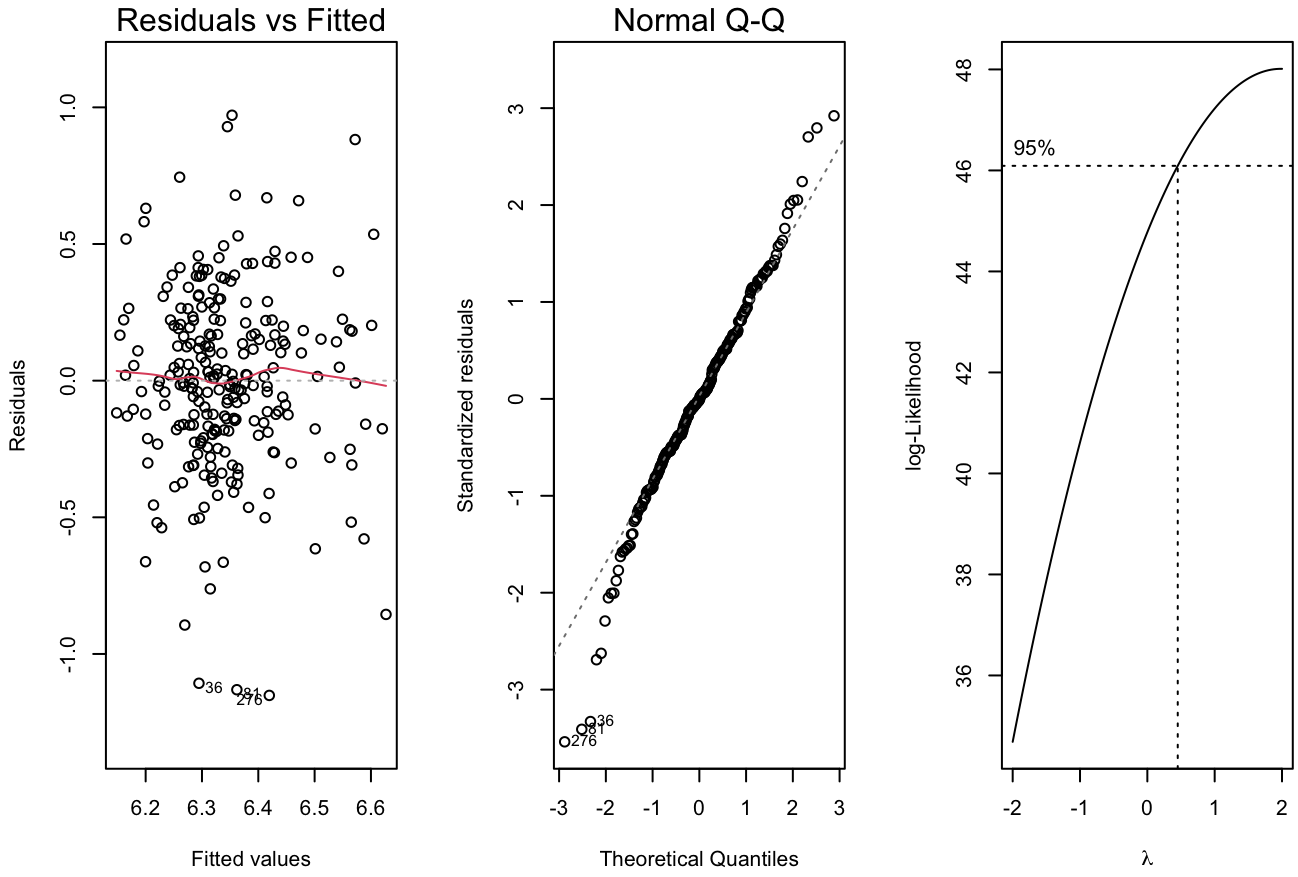
```
# Diagonistic of RET model1
retmodel1 <- lm(formula = RETPLASMA ~ SEX + CALORIES + AGE, data = ret)
par(mfrow = c(1,3))
plot(retmodel1, which = c(1,2))
MASS::boxcox(retmodel1)
```



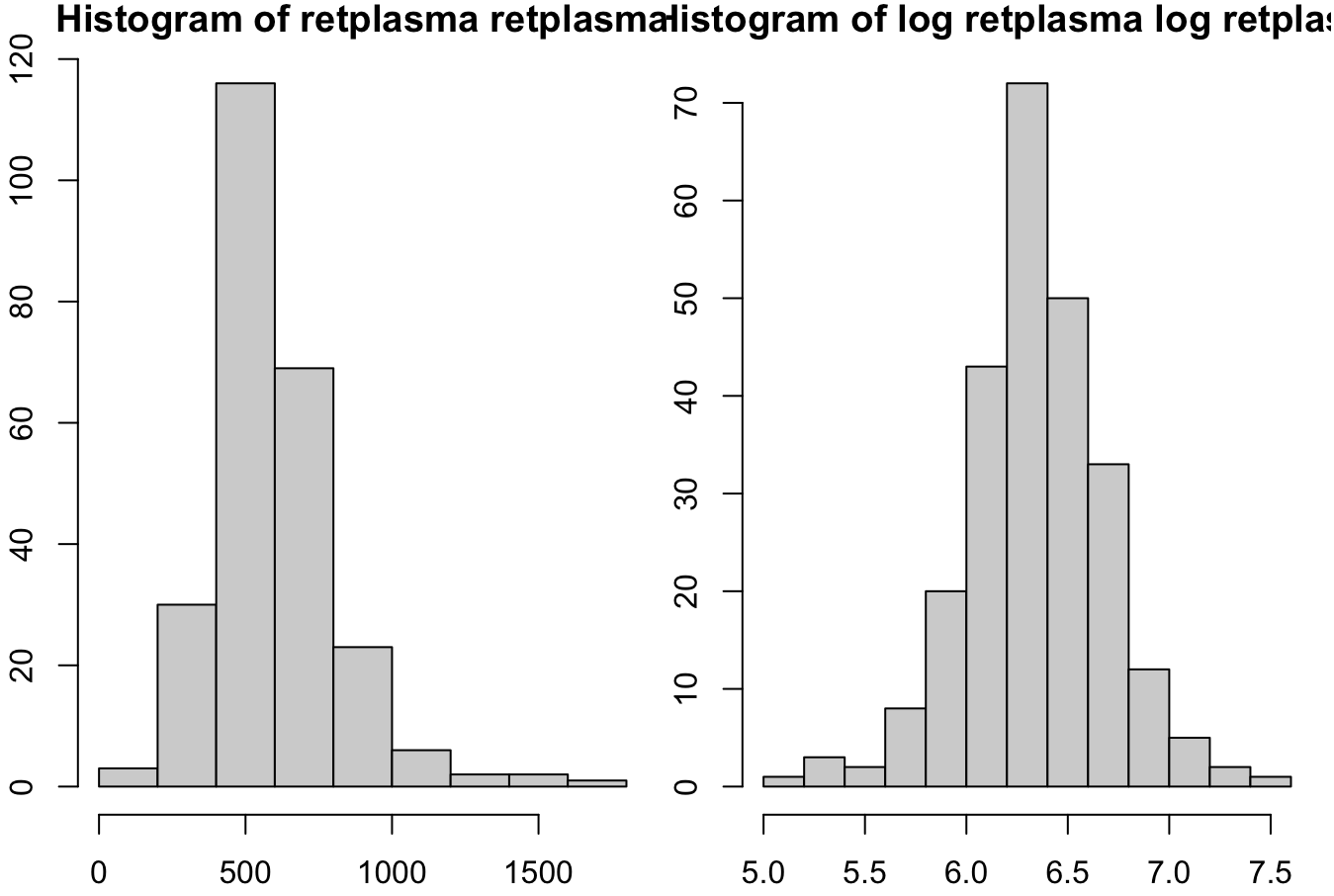

```
# log-transformation
log_retmodel1 <- lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE, data = ret)
summary(log_retmodel1)
```

```
##
## Call:
## lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE, data = ret)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15167 -0.18104 -0.00226  0.20174  0.97114
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.287e+00  1.077e-01  58.349  < 2e-16 ***
## SEXMALE      1.896e-01  6.730e-02   2.817  0.00524 **
## CALORIES    -6.869e-05  3.227e-05  -2.129  0.03427 *
## AGE          3.002e-03  1.575e-03   1.906  0.05782 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3334 on 248 degrees of freedom
## Multiple R-squared:  0.07803,    Adjusted R-squared:  0.06688
## F-statistic: 6.996 on 3 and 248 DF,  p-value: 0.0001552
```

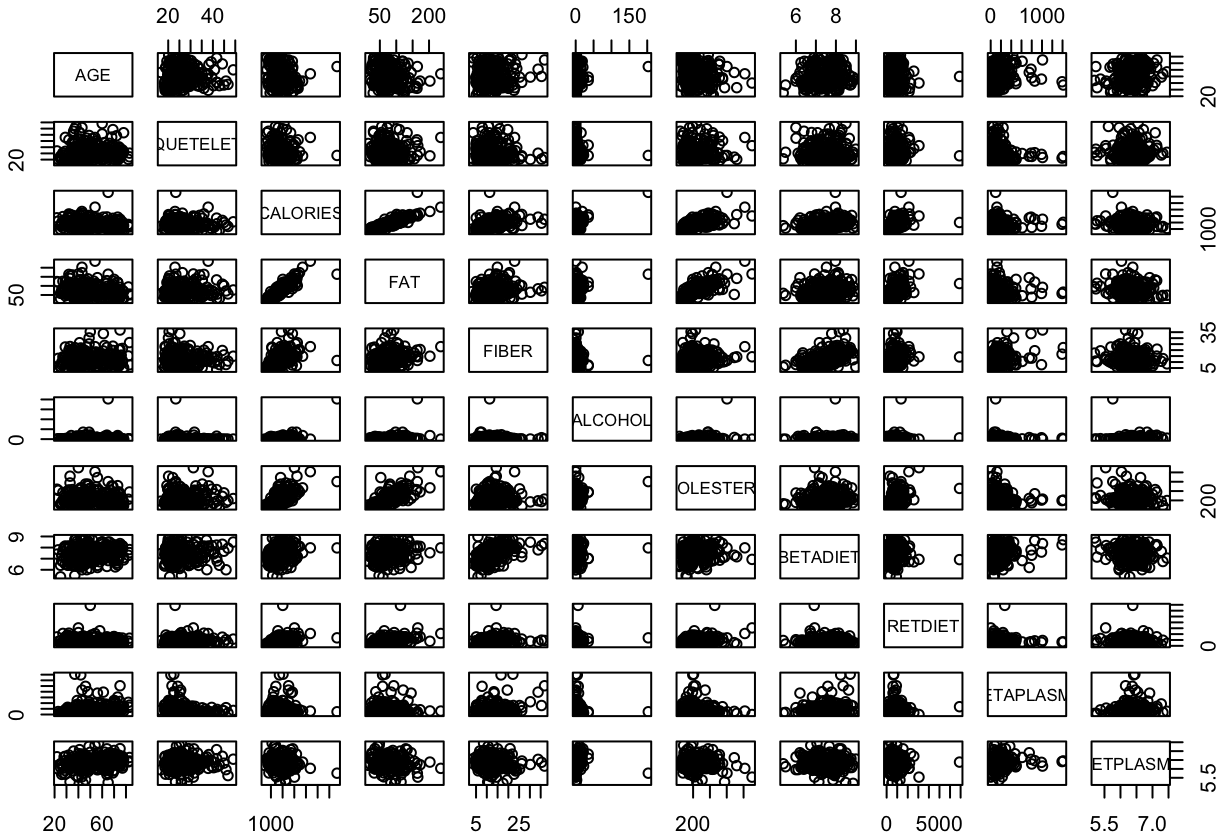
```
# diagnosis of log
par(mfrow = c(1,3))
plot(log_retmodel1, which=c(1,2))
MASS::boxcox(log_retmodel1)
```




```
summary(log_retmodel1)
# hist plot
par(mfrow = c(1,2), mar=c(3,2,2,.5), mgp=c(3,1,0))
hist(train$RETPLASMA, main = paste("Histogram of retplasma", xlab = "retplasma"))
hist(log(train$RETPLASMA), main = paste("Histogram of log retplasma", xlab = "log retpla
sma"))
```




```
par(mfrow=c(1,1), mar=c(3,2,2,.5), mgp=c(3,1,0))
# linearity check
num$RETPLASMA <- log(num$RETPLASMA)
pairs(num)
```



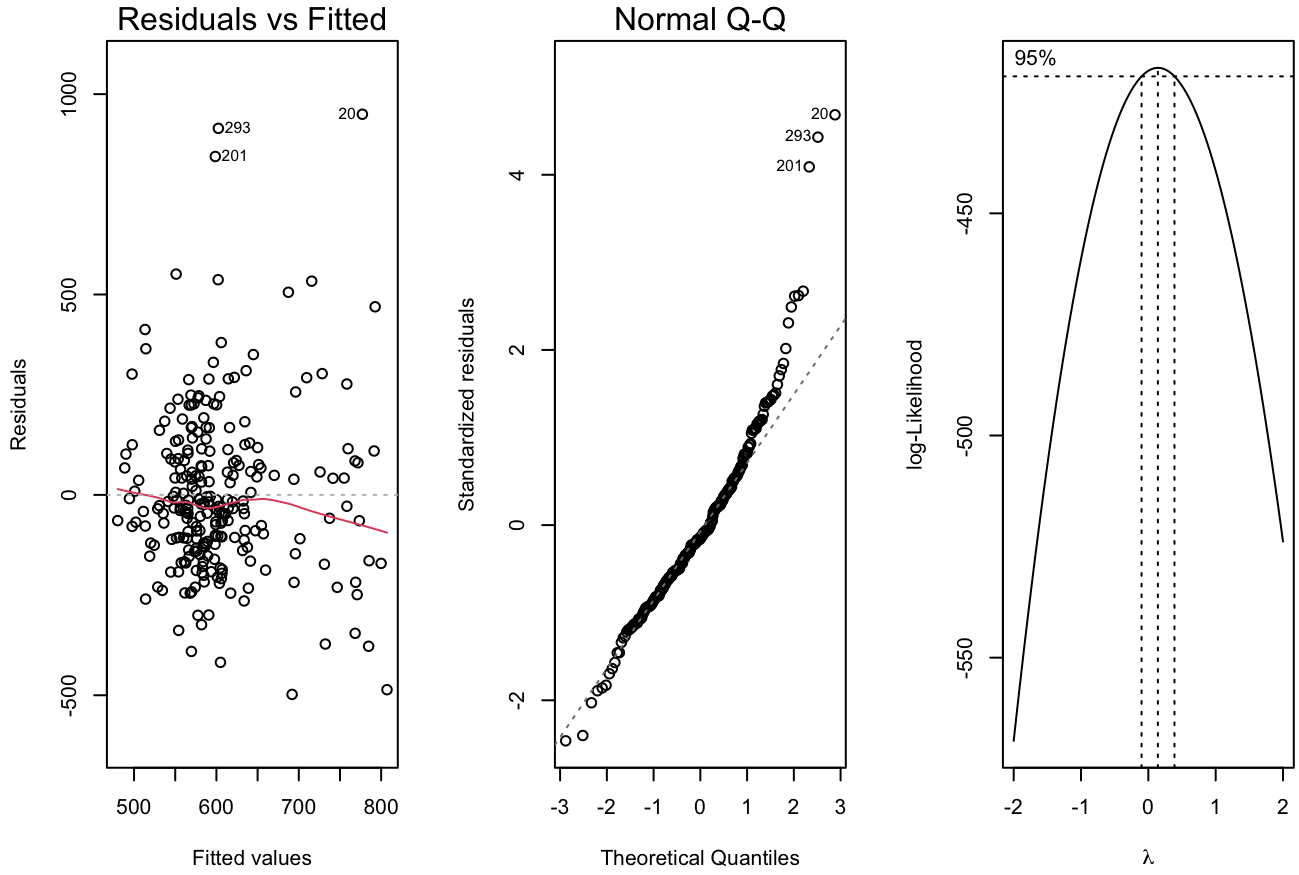
Model with interactive terms

```
#Adding interactive terms
#init_mod2 <- lm(RETPLASMA ~ SEX + CALORIES + AGE, data = ret)
stepret <- step(retmodel1, scope = SEX + CALORIES + AGE ~ .^2, direction = 'both')
```

```
## Start:  AIC=2691.92
## RETPLASMA ~ SEX + CALORIES + AGE
##
##                Df Sum of Sq      RSS    AIC
## + SEX:AGE       1    111978 10525632 2691.2
## <none>                      10637610 2691.9
## + CALORIES:AGE  1     78885 10558725 2692.0
## - AGE           1    108584 10746194 2692.5
## + SEX:CALORIES  1     54903 10582708 2692.6
## - CALORIES      1    179399 10817009 2694.1
## - SEX           1    614579 11252190 2704.1
##
## Step:  AIC=2691.25
## RETPLASMA ~ SEX + CALORIES + AGE + SEX:AGE
##
##                Df Sum of Sq      RSS    AIC
## + SEX:CALORIES  1     94089 10431543 2691.0
## <none>                      10525632 2691.2
## - SEX:AGE       1    111978 10637610 2691.9
## + CALORIES:AGE  1     51469 10474163 2692.0
## - CALORIES      1    185969 10711602 2693.7
##
## Step:  AIC=2690.99
## RETPLASMA ~ SEX + CALORIES + AGE + SEX:AGE + SEX:CALORIES
##
##                Df Sum of Sq      RSS    AIC
## <none>                      10431543 2691.0
## - SEX:CALORIES  1     94089 10525632 2691.2
## - SEX:AGE       1    151164 10582708 2692.6
## + CALORIES:AGE  1      3575 10427968 2692.9
```

Result : RETPLASMA ~ SEX + CALORIES + AGE + SEX:AGE + SEX:CALORIES

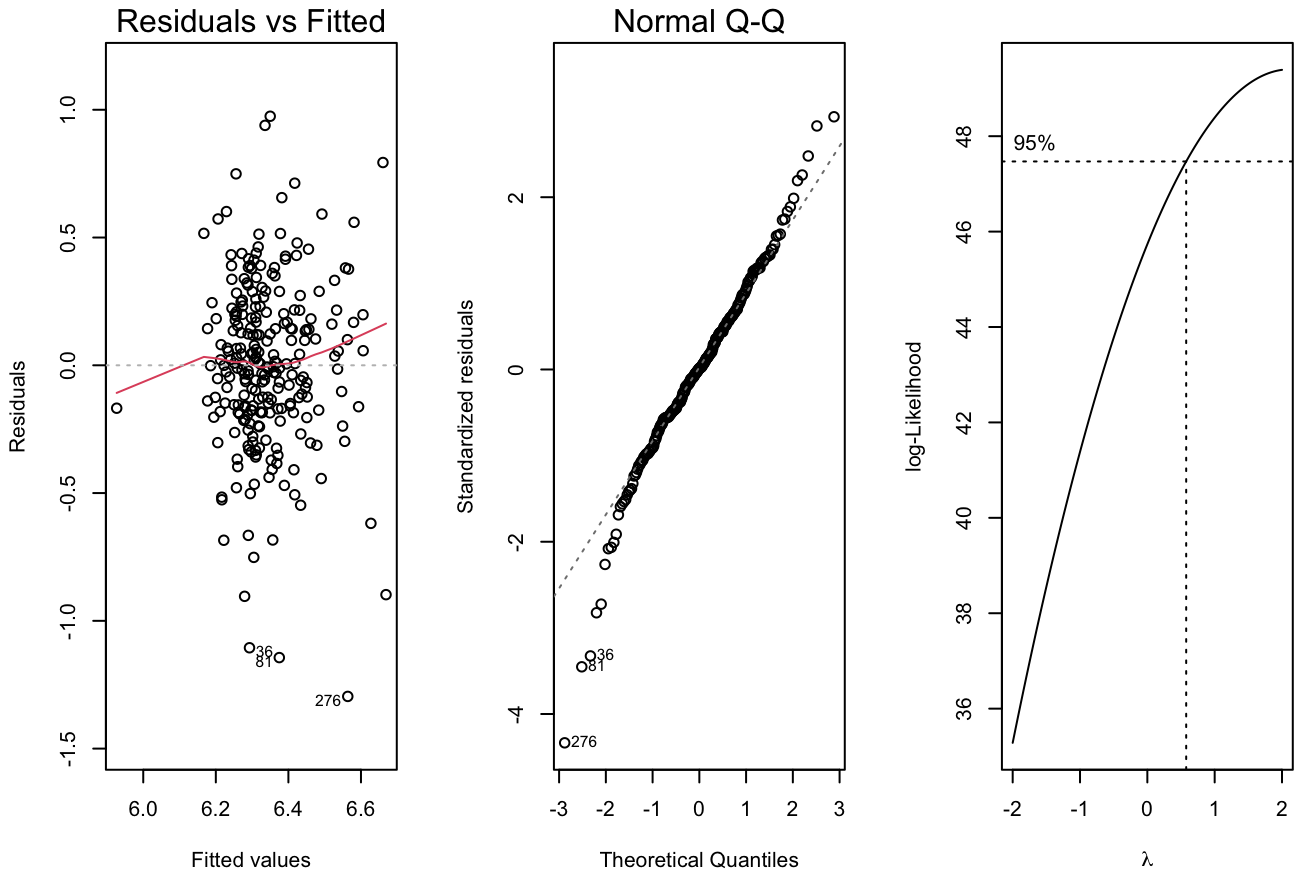
```
# Diagonistic of RET model2
retmodel2 <- lm(formula = RETPLASMA ~ SEX + CALORIES + AGE + SEX:AGE + SEX:CALORIES, dat
a = ret)
par(mfrow = c(1,3))
plot(retmodel1, which = c(1,2))
MASS::boxcox(retmodel1)
```




```
# log-trans and diagonistic
log_retmodel2 <- lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE + SEX:AGE + SEX:CALO
RIES, data = ret)
summary(log_retmodel2)
```

```
##
## Call:
## lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE + SEX:AGE +
##     SEX:CALORIES, data = ret)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29577 -0.17809  0.00404  0.19936  0.97462
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.210e+00  1.195e-01  51.958   <2e-16 ***
## SEXMALE           7.194e-01  3.692e-01   1.948   0.0525 .
## CALORIES         -4.625e-05  3.804e-05  -1.216   0.2252
## AGE               3.775e-03  1.674e-03   2.256   0.0250 *
## SEXMALE:AGE      -6.010e-03  5.053e-03  -1.189   0.2355
## SEXMALE:CALORIES -8.250e-05  7.177e-05  -1.149   0.2515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3332 on 246 degrees of freedom
## Multiple R-squared:  0.08651,    Adjusted R-squared:  0.06794
## F-statistic: 4.659 on 5 and 246 DF,  p-value: 0.0004409
```

```
par(mfrow = c(1,3))
plot(log_retmodel2, which = c(1,2))
MASS::boxcox(log_retmodel2)
```



3. Model Selection: Criterion

```
# get rsq, radj, bic, aic
sumb1 <- glance(log_betamodel1)
sumb2 <- glance(log_betamodel2)
sumr1 <- glance(log_retmodel1)
sumr2 <- glance(log_retmodel2)

# get sse
anvb1 <- anova(log_betamodel1)
anvb2 <- anova(log_betamodel2)
anvr1 <- anova(log_retmodel1)
anvr2 <- anova(log_retmodel2)
```

```
# Get cp
# Full model for models with interaction
fullmod1 <- lm(log(BETAPLASMA)~.^2, data=train_pos)
fullmod2 <- lm(log(RETPLASMA)~.^2, data= ret)
anvf1 <- anova(fullmod1)
anvf2 <- anova(fullmod2)
MSEf1 <- anvf1$`Mean Sq`[79]
MSEf2 <- anvf2$`Mean Sq`[79]
```

```
#Full model for first order model
fullmod11 <- lm(log(BETAPLASMA)~., data=train_pos)
fullmod22 <- lm(log(RETPLASMA)~., data=ret)
anvf11 <- anova(fullmod11)
anvf22 <- anova(fullmod22)
MSEf11 <- anvf11$`Mean Sq`[13]
MSEf22 <- anvf22$`Mean Sq`[13]
```

```
# Calculate cp
cpb1 <- (anvb1$`Sum Sq`[7]/MSEf11) - (252-(2*7))
cpb2 <- (anvb2$`Sum Sq`[12]/MSEf1) - (252-(2*12))
cpr1 <- (anvr1$`Sum Sq`[4]/MSEf22) - (252-(2*4))
cpr2 <- (anvr2$`Sum Sq`[6]/MSEf2) - (252-(2*6))
```

```
#Get pressp
# 1. log_betamodel1
eb1 = log_betamodel1$residuals
hb1 = influence(log_betamodel1)$hat
de_b1 = eb1/(1-hb1)
pressb1 = sum((de_b1)^2)

# 2. log_betamodel2
eb2 = log_betamodel2$residuals
hb2 = influence(log_betamodel2)$hat
de_b2 = eb2/(1-hb2)
pressb2 = sum( de_b2^2 )

# 3. log_retmodel1
er1 = log_retmodel1$residuals
hr1 = influence(log_retmodel1)$hat
de_r1 = er1/(1-hr1)
pressr1 = sum( de_r1^2 )

# 4. log_retmodel2
er2 = log_retmodel2$residuals
hr2 = influence(log_retmodel2)$hat
de_r2 = er2/(1-hr2)
pressr2 = sum( de_r2^2 )
```

```
mod <- c("Beta1","Beta2", "Ret1", "Ret2")
sse <- round(c(anvb1$`Sum Sq`[7], anvb2$`Sum Sq`[12], anvr1$`Sum Sq`[4], anvr2$`Sum Sq`
[6]),3)
rsq <- round(c(sumb1$r.squared, sumb2$r.squared, sumr1$r.squared, sumr2$r.squared ),3)
rsqa <- round(c(sumb1$adj.r.squared, sumb1$adj.r.squared, sumr1$adj.r.squared, sumr2$ad
j.r.squared),3)
cp <- round(c(cpb1, cpb2, cpr1, cpr2),3)
bic <- round(c(sumb1$BIC, sumb2$BIC, sumr1$BIC, sumr2$BIC),3)
aic <- round(c(sumb1$AIC, sumb2$AIC, sumr1$AIC, sumr2$AIC),3)
pressp <- round(c(pressb1, pressb2, pressr1, pressr2),3)
res_sub <- cbind(mod, sse, rsq, rsqa, cp, bic, aic, pressp)
colnames(res_sub)<-c("Model", "sse", "R^2", "R^2_a", "Cp","bic", "aic", "press_p")
as_tibble(res_sub)
```

```
## # A tibble: 4 × 8
##   Model sse     `R^2` `R^2_a` Cp     bic     aic     press_p
##   <chr> <chr>   <chr> <chr>   <chr>  <chr>   <chr>   <chr>
## 1 Beta1 112.923 0.235 0.213   10.649 561.551 529.822 120.868
## 2 Beta2 106.896 0.276 0.213   31.822 580.935 528.053 120.271
## 3 Ret1  27.569  0.078 0.067   1.967  185.187 167.54  28.7
## 4 Ret2  27.316  0.087 0.068   16.046 193.917 169.211 29.836
```

4. Data Validation

```
# Data Validation (Beta)
beta_train <- log_betamodel1
beta_valid <- lm(formula = log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT + BETADIET
+ AGE, data = valid)

mod_sum <- cbind(coef(summary(beta_train))[,1], coef(summary(beta_valid))[,1],
coef(summary(beta_train))[,2], coef(summary(beta_valid))[,2])
colnames(mod_sum) <- c("Train Est","Valid Est","Train s.e.","Valid s.e.")

mod_sum
```

```
##                    Train Est     Valid Est   Train s.e.   Valid s.e.
## (Intercept)      5.001642e+00  5.3552053216 2.895453e-01 5.612374e-01
## FIBER            3.261764e-02 -0.0074651847 9.328685e-03 2.026965e-02
## QUETELET        -3.272665e-02 -0.0214975270 7.266998e-03 1.339022e-02
## VITUSENOT OFTEN  3.449115e-01  0.2734590333 1.131850e-01 2.134941e-01
## VITUSEOFTEN      3.340306e-01  0.3623420929 1.019183e-01 1.825662e-01
## FAT             -4.085030e-03 -0.0014876503 1.394810e-03 2.468100e-03
## BETADIET         4.186106e-05  0.0000998519 3.556058e-05 5.318189e-05
## AGE              7.261043e-03  0.0011271804 3.100550e-03 5.385298e-03
```

```
#compare the SSE and R2a (Beta)
sse_t <- sum(beta_train$residuals^2)
n_t = nrow(train)
mse_t <- sse_t/(n_t-7)
sse_v <- sum(beta_valid$residuals^2)
n_v = nrow(valid)
mse_v <- sse_v/(n_v-7)
Radj_t <- summary(beta_train)$adj.r.squared
Radj_v <- summary(beta_valid)$adj.r.squared
train_sum <- c(sse_t,mse_t,Radj_t)
valid_sum <- c(sse_v,mse_v,Radj_v)
criteria <- rbind(train_sum,valid_sum)
colnames(criteria) <- c("SSE","MSE","R2_adj")
criteria
```

```
##                 SSE       MSE    R2_adj
## train_sum 112.92296 0.4609100 0.2133766
## valid_sum  20.32715 0.3629847 0.1383134
```

```
#Get MSPE_v from new data (beta)
beta_newdata <- data.frame(valid[, 1:12])
y.hat <- predict(beta_train, beta_newdata)
MSPE <- mean((log(valid$BETAPLASMA) - y.hat)^2)
criteria <- cbind(MSPE,sse_t/n_t)
colnames(criteria) <- c("MSPE","SSE/n")
criteria
```

```
##           MSPE    SSE/n
## [1,] 0.4022245 0.448107
```

```
# Data Validation (Ret)
ret_train <- log_retmodel1
ret_valid <- lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE, data = valid)

mod_sum <- cbind(coef(summary(ret_train))[,1], coef(summary(ret_valid))[,1],
coef(summary(ret_train))[,2], coef(summary(ret_valid))[,2])
colnames(mod_sum) <- c("Train Est","Valid Est","Train s.e.","Valid s.e.")

mod_sum
```

```
##                Train Est     Valid Est   Train s.e.   Valid s.e.
## (Intercept)  6.286854e+00  5.961337e+00 1.077451e-01 2.008045e-01
## SEXMALE      1.895982e-01 -2.314629e-01 6.730454e-02 1.208523e-01
## CALORIES    -6.868637e-05  3.400575e-05 3.226816e-05 6.366515e-05
## AGE          3.002228e-03  7.511085e-03 1.575247e-03 2.755141e-03
```

```
#compare the SSE and R2a (Ret)
sse_t <- sum(ret_train$residuals^2)
n_t = nrow(train)
mse_t <- sse_t/(n_t-7)
sse_v <- sum(ret_valid$residuals^2)
n_v = nrow(valid)
mse_v <- sse_v/(n_v-7)
Radj_t <- summary(ret_train)$adj.r.squared
Radj_v <- summary(ret_valid)$adj.r.squared
train_sum <- c(sse_t,mse_t,Radj_t)
valid_sum <- c(sse_v,mse_v,Radj_v)
criteria <- rbind(train_sum,valid_sum)
colnames(criteria) <- c("SSE","MSE","R2_adj")
criteria
```

```
##                 SSE        MSE     R2_adj
## train_sum 27.569465 0.11252843 0.06687572
## valid_sum  5.548297 0.09907672 0.09195875
```

```
#Get MSPE_v from new data (ret)
ret_newdata <- data.frame(valid[, 1:12])
y.hat <- predict(ret_train, ret_newdata)
MSPE <- mean((log(valid$RETPLASMA) - y.hat)^2)
criteria <- cbind(MSPE,sse_t/n_t)
colnames(criteria) <- c("MSPE","SSE/n")
criteria
```

```
##           MSPE     SSE/n
## [1,] 0.1088285 0.1094026
```

5. Rebuild model use whole dataset

```
Plasma_pos = Plasma[Plasma$BETAPLASMA > 0, ]
log_betamodel_final <- lm(log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT + BETADIET +
AGE, data = Plasma_pos)
summary(log_betamodel_final)
```

```
##
## Call:
## lm(formula = log(BETAPLASMA) ~ FIBER + QUETELET + VITUSE + FAT +
##     BETADIET + AGE, data = Plasma_pos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99928 -0.35983  0.00063  0.38345  1.88948
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.101e+00  2.576e-01  19.800  < 2e-16 ***
## FIBER            2.666e-02  8.419e-03   3.167 0.001696 **
## QUETELET        -3.024e-02  6.355e-03  -4.758 3.02e-06 ***
## VITUSENOT OFTEN  2.986e-01  9.876e-02   3.024 0.002709 **
## VITUSEOFTEN      3.337e-01  8.909e-02   3.745 0.000215 ***
## FAT             -3.489e-03  1.203e-03  -2.901 0.003986 **
## BETADIET         5.018e-05  2.948e-05   1.702 0.089752 .
## AGE              5.163e-03  2.685e-03   1.923 0.055420 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6693 on 306 degrees of freedom
## Multiple R-squared:  0.2167, Adjusted R-squared:  0.1988
## F-statistic: 12.09 on 7 and 306 DF,  p-value: 1.225e-13
```

```
log_retmodel_final <- lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE, data = Plasma)
summary(log_retmodel_final)
```

```
##
## Call:
## lm(formula = log(RETPLASMA) ~ SEX + CALORIES + AGE, data = Plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16277 -0.19099 -0.00168  0.21739  0.96259
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.213e+00  9.561e-02  64.980  < 2e-16 ***
## SEXMALE      9.992e-02  5.946e-02   1.681  0.09386 .
## CALORIES    -4.653e-05  2.902e-05  -1.603  0.10985
## AGE          4.039e-03  1.380e-03   2.926  0.00369 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3314 on 311 degrees of freedom
## Multiple R-squared:  0.06016,    Adjusted R-squared:  0.05109
## F-statistic: 6.636 on 3 and 311 DF,  p-value: 0.000234
```