Faculty of Science and Engineering - Department of Mathematics and Statistics

STAT-811

(Generalized linear models)

ASSIGNMENT-1

(Semester 2- 2019)

Student ID: 45665761

Name: Sukhjeet Kaur

Master of Applied Statistics

Question 1.

Description of the dataset

The dataset **plasma** which is in the R package **gamlss** containing 315 observations and 14 variables including demographic and dietary factors was studied. For the analysis 8 predictors-age, sex, smokstat, bmi, vituse, fiber, alcohol, betadiet and one response variable -betaplasma were considered.

Aim of the analysis was to investigate the relationship between the levels of beta-carotene in the blood plasma and 8 predictors/covariates – age, sex, smokstat, bmi, vituse, fiber, alcohol and betadiet. Multiple linear Regression method was used to build a model following a backward approach in which one covariate is eliminated from the model at a time, which seems least promising. (using significance level 5% i.e. p-value as 0.05)

- A subset "sub_plasma" was created from the original dataset using only 8 predictors and one outcome variable(betaplasma) after deleting a record (257) from the data set as value for betaplasma was coded as 0 which is implausible.
- Model building process used to construct the final model is as follows:
    1. First of all, the distribution of the response variable was examined to see any asymmetry. As the distribution of the variable betaplasma was non-normal (see figure1), then log transformation was applied to get a normal distribution plot. (Figure 2)

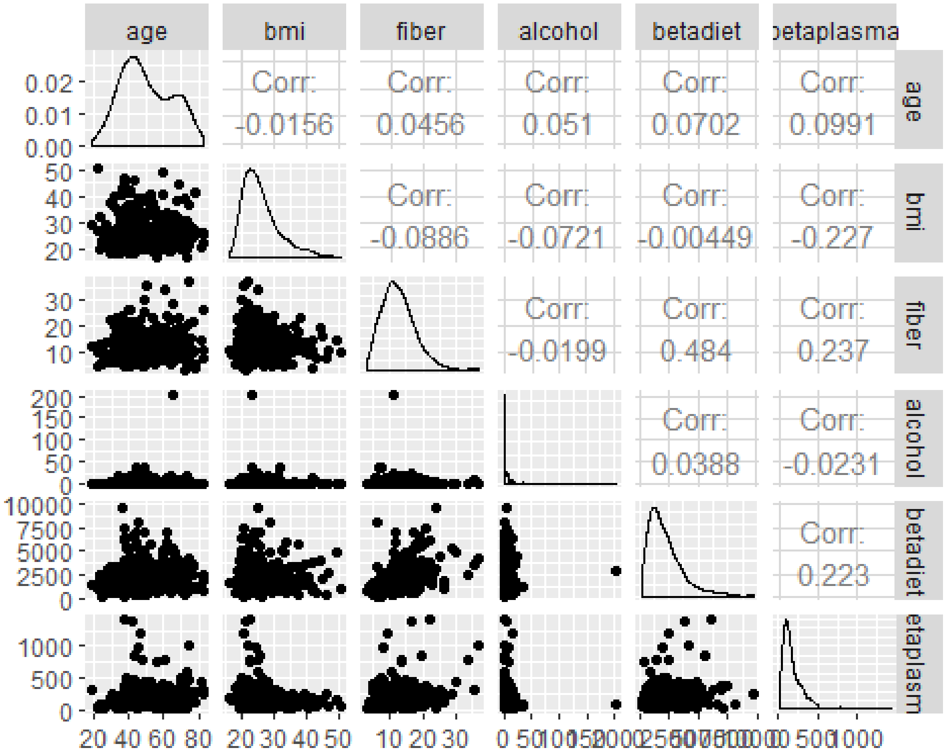


Figure 1.

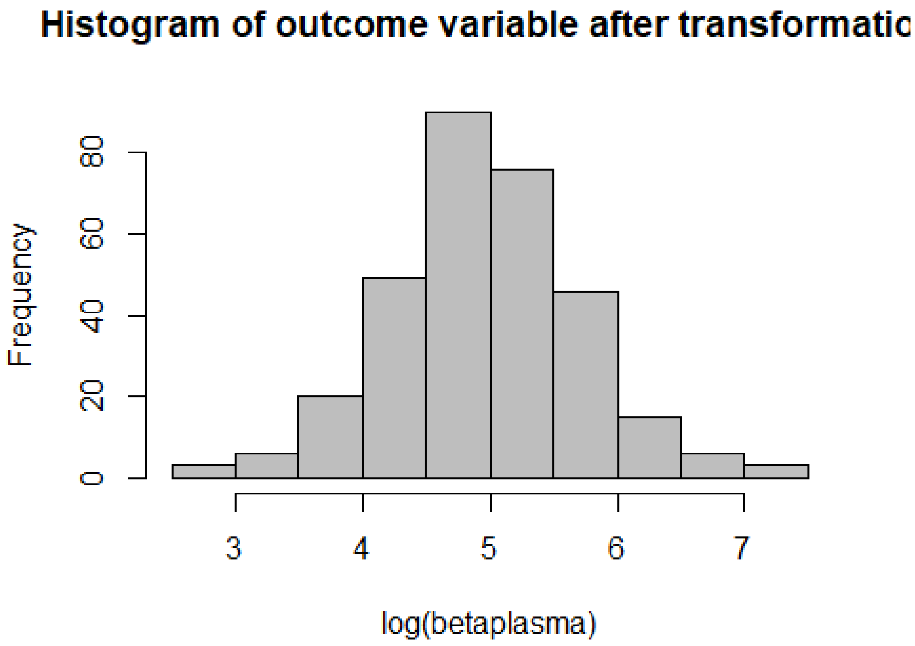


Figure 2)

2. Then, scatterplot of each of the covariates with response variable was examined to see any non-linear pattern so as to apply any transformation if necessary (see figure 1). On visualizing the scatterplots, all the plots were approximately linear and did not show any undesired pattern. So, no transformation was applied to covariates.
3. The distribution of alcohol was extremely right skewed with maximum value 203 drinks consumed per week which seems to be implausible and it was excluded from the model. (see figure 2)

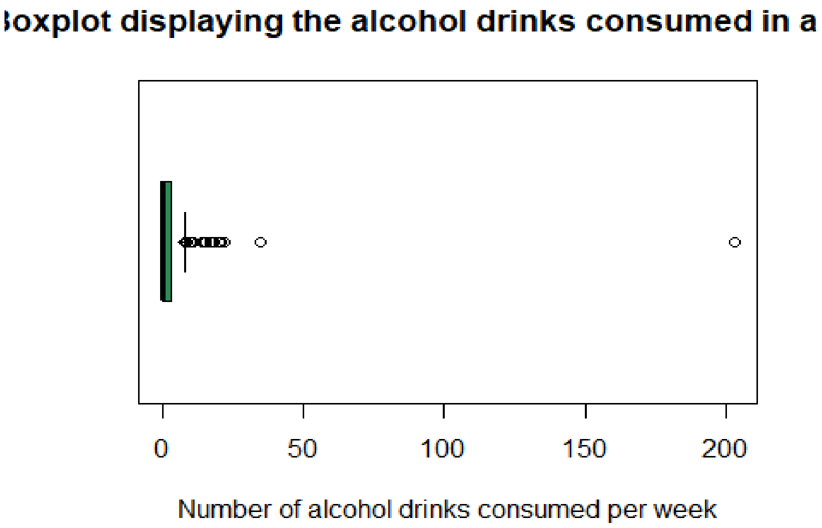
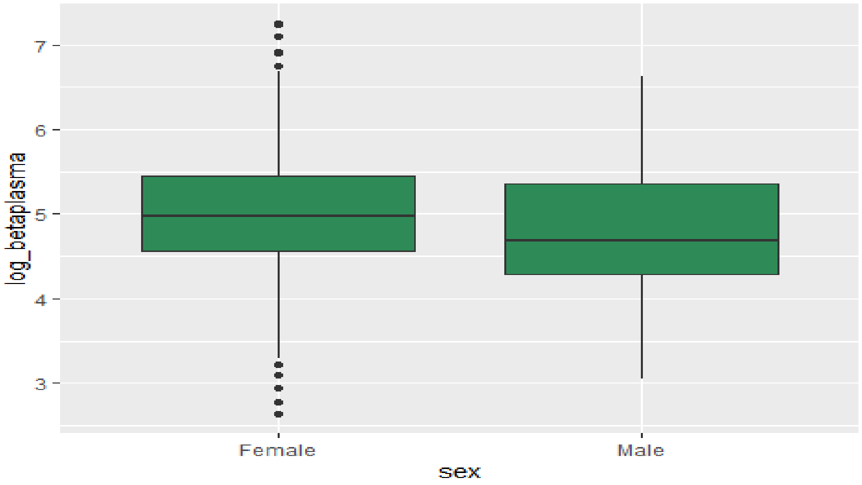
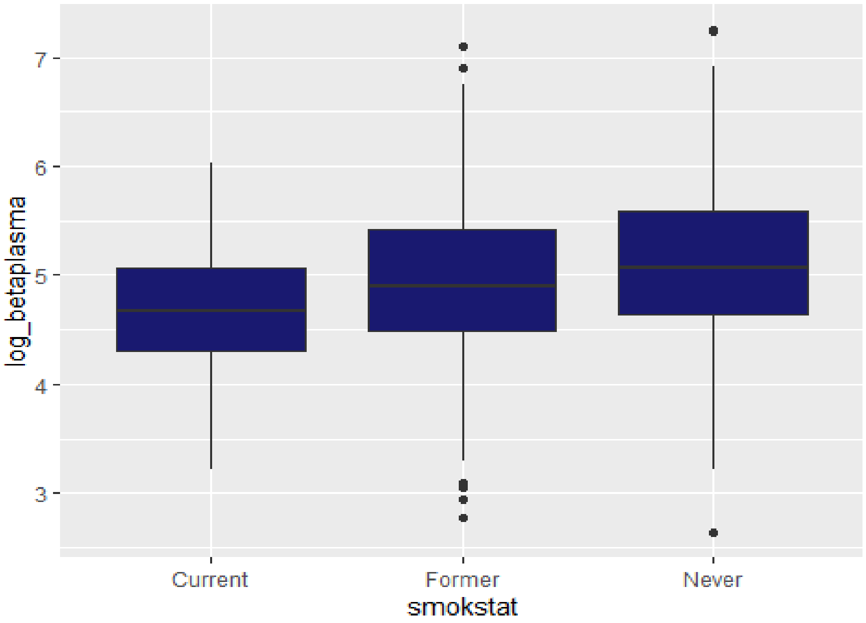
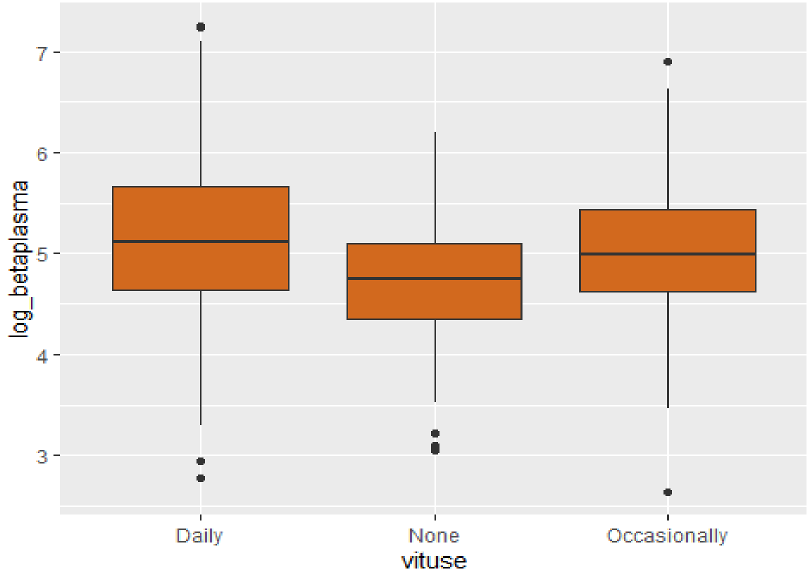


Figure 3)

4. The boxplots of categorical predictors were plotted with outcome variable and frequency of the categorical predictors were checked. The levels of each of the categorical variable were randomly and sensibly divided and did not require any manipulations. (description of the boxplots is given in Appendix)







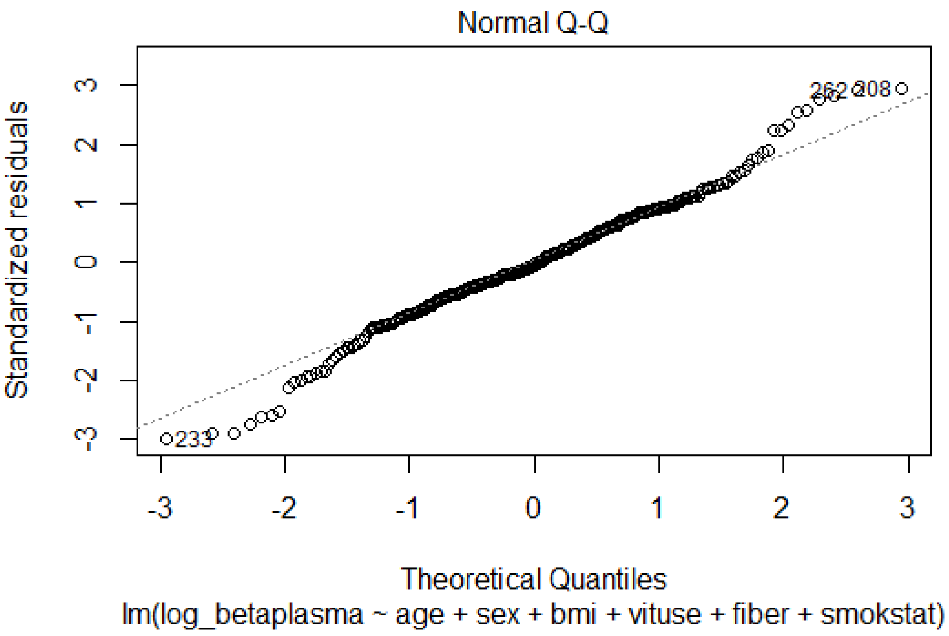
5. After that the categorical variables were recoded as:
   Sex variable - 1 = Male and 2 = Female
   Smokstat - 1 = Never, 2 = Former and 3= Current
   Vituse - 1 = Daily, 2= Occasionally and 3= None
   And the variables were set as factor to treat them categorical in R software.
6. The reference category was set as Never and None for variables smokstat and vituse respectively.
7. Initially, the model with all the 8 predictors was constructed and after executing the model it was noticed that alcohol was most insignificant variable after taking into account all the other predictors except betadiet, however a model was rebuilt and the slopes of each covariates were tested using anova taking into account all the predictors before alcohol and again alcohol was insignificant. R-squared obtained was 23.44%. So, the alcohol variable was excluded from the model.
8. Next the model after excluding alcohol was built and on executing the regression analysis for this model two variables were found insignificant smokstatFormer and betadiet, so the anova was carried out with the model to decide which is least promising between the two and R-squared obtained in this model was 23.44%. As a result, betadiet was found insignificant. Hence, betadiet was removed from the model.
9. Finally, the model was built using 6 predictors after excluding betadiet and alcohol variable and in the result of regression model one indicator variable was found insignificant for smokstat and remaining were significant but joint significance of the categorical variable was tested through anova and as a result the variable was significant. So, keeping the variable smokstat in the final model was a better choice as it provides some additional information in determining

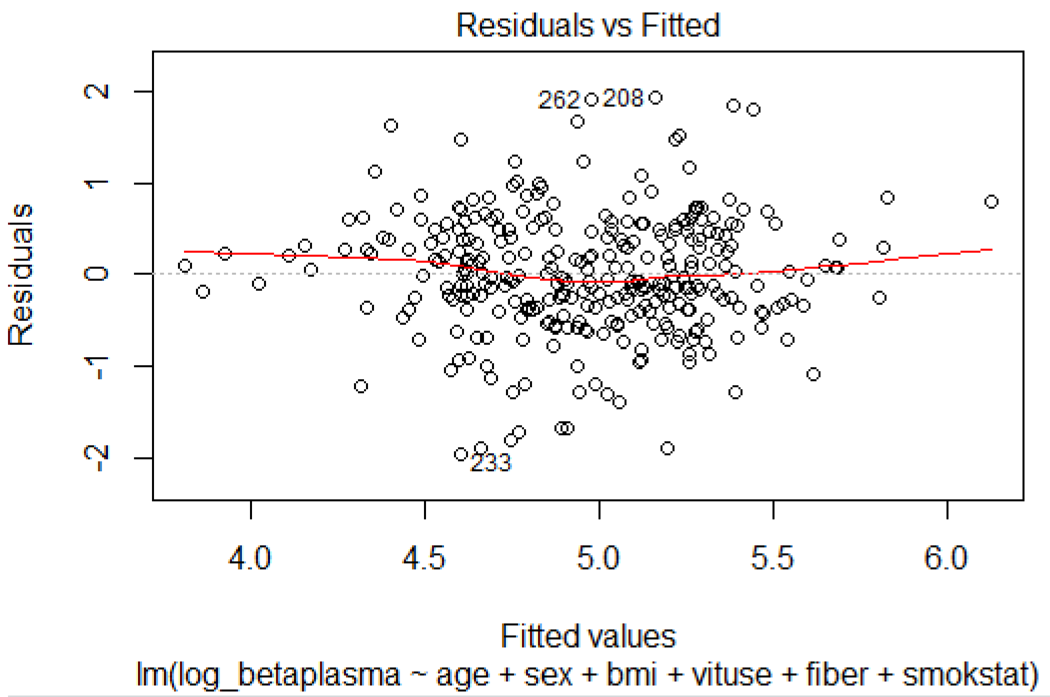
the levels of beta-carotene in blood plasma and the model provides R-squared value = 22.87%.

Summary of the models

Name of the model	**Significant variables**	**Insignificant variables**	**Variable excluded**	**Coefficient of determination(R^2)**
model_all	Age, sex, smokstat, fiber, vituse, bmi	Alcohol, betadiet	None	23.44%
Model1 (changing the order of alcohol)	Age, sex, smokstat, fiber, vituse, bmi	Alcohol, betadiet	none	23.44%
Model2	Age, sex, smokstatCurrent, fiber, vituse, bmi	Betadiet, smokstatFormer	alcohol	23.44%
Model_2 (changing the order of smokstat)	Age, sex, fiber, vituse, bmi, smokstatCurrent	Betadiet, smokstatFormer	alcohol	23.44%
Model3	Age, sex, fiber, vituse, bmi, smokstatCurrent	smokstatFormer	Alcohol, betadiet	22.87%
Model4 (changing the order of smokstat)	Age, sex, smokstatCurrent, vituse, fiber, bmi	smokstatFormer	Alcohol, betadiet	22.87%

Diagnostic checking for the final model





Two assumptions need to be satisfied for normal linear model

- Residuals should have normal distribution
- Residuals should have homogeneous variance

By looking at QQPlot it can be stated that there is fairly normal distribution among residuals though there is slight deviation at both ends which may be due to some outliers. Otherwise the plot looks linear approximately.

Also, there is no distinct pattern in residuals vs fitted plot, the residuals seem to be randomly and evenly scattered around horizontal line at zero which indicates the homogeneous variance although there is slight less variation in both ends.

Overall, we can say that assumptions are verified considering the number of observations in the dataset.

Output for the final model is:

```
Model4<-lm(log_betaplasma~age+sex+bmi+vituse+fiber+smokstat,data=sub_plasma)
summary(Model4)

##
## Call:
## lm(formula = log_betaplasma ~ age + sex + bmi + vituse + fiber +
##     smokstat, data = sub_plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96395 -0.37140 -0.03385  0.41977  1.94276
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         5.060877   0.262596  19.272  < 2e-16 ***
## age                 0.007684   0.002763   2.781 0.005764 ** 
## sexMale            -0.309327   0.119290  -2.593 0.009972 ** 
## bmi                -0.033149   0.006372  -5.202 3.62e-07 ***
## vituseDaily         0.289762   0.090182   3.213 0.001454 ** 
## vituseOccasionally  0.260064   0.099370   2.617 0.009311 ** 
## fiber               0.024905   0.007225   3.447 0.000646 ***
## smokstatCurrent    -0.327436   0.121527  -2.694 0.007445 ** 
## smokstatFormer     -0.088719   0.083397  -1.064 0.288258    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.666 on 304 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.2084 
## F-statistic: 11.27 on 8 and 304 DF,  p-value: 5.601e-14
```

```
anova(Model4)
```

```
## Analysis of Variance Table
## 
## Response: log_betaplasma
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## age         1   3.325  3.3251  7.4965 0.0065464 ** 
## sex         1   5.477  5.4768 12.3477 0.0005087 ***
## bmi         1  13.322 13.3225 30.0359 8.915e-08 ***
## vituse      2   7.845  3.9223  8.8429 0.0001850 ***
## fiber       1   6.775  6.7755 15.2755 0.0001146 ***
## smokstat    2   3.240  1.6198  3.6520 0.0270838 *  
## Residuals 304 134.839  0.4436                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final model equation:

$$\log\_\widehat{betaplasma} = 5.061 + 0.008\,age - 0.309\,sexMale - 0.033bmi + 0.290vituseDaily + 0.26vituseOccasionally + 0.025fiber - 0.327smokstatCurrent - 0.089smokstatFormer$$

$$\widehat{betaplasma} = \exp(5.061 + 0.008\,age - 0.309\,sexMale - 0.033bmi + 0.290vituseDaily + 0.26vituseOccasionally + 0.025fiber - 0.327smokstatCurrent - 0.089smokstatFormer\ )$$

$$\begin{aligned}\widehat{betaplasma} &= \exp(5.061) * \exp(0.008\,age) * \exp(-0.309\,sexMale) \\ &\quad * \exp(-0.033bmi) * \exp(0.290vituseDaily) \\ &\quad * \exp(0.26vituseOccasionally) * \exp(0.025fiber) \\ &\quad * \exp(-0.327smokstatCurrent) * \exp(-0.089smokstatFormer)\end{aligned}$$

Interpretation of the parameters

- one-year increase in age multiplies the expected value of betaplasma by $e^{0.008}$ = 1.00units(ng/ml).
- If a patient is male then compared to female betaplasma multiplies by $e^{-0.309}$ = 0.73units.
- A unit (weight/height^2) increase in bmi, betaplasma multiplies by $e^{-0.033}$ = 0.97units.
- If a person takes vitamins daily compared to those who do not take, their betaplasma multiplies by $e^{0.290}$ = 1.34 units and if vitamin intake is occasionally compared to patient who do not take vitamin supplements then levels of betaplasma multiplies by $e^{0.26}$ = 1.30 units.
- One gram of fiber intake multiplies betaplasma by $e^{0.025}$ = 1.03 units.
- If a patient is current smoker then compared to non-smoker the betaplasma multiplies by $e^{-0.327}$ = 0.72units and if person is former smoker then compared to non-smoker their betaplasma multiplies by $e^{-0.089}$ = 0.91 units but this is not statistically significant under 5% significance level.

The above statements are true while all other predictors or factors are held constant on average.

Report

The analysis was performed over 313 subjects to investigate the association between levels of beta-carotene in blood plasma and the demographic and dietary factors (age, sex, smoking status, vitamin intake, fiber intake, number of alcohol drinks consumed per week, and dietary intake of beta-carotene and Body mass index).

The relation of each of the factors with levels of beta-carotene was graphically examined to study the nature of change in the levels of beta-carotene. On visualizing, females tend to have higher levels of beta-carotene as compared to males. It was noticed that relation between BMI and levels of beta-carotene is negatively associated whereas positive association was noticed between fiber intake, betadiet intake with levels of beta-carotene independently. The smoking status of the subject had slight negative impact on levels of beta-carotene whereas intake of vitamin supplements showed a positive association.

However, to get significant results, the analysis was carried out over 313 subjects to study the relationship between levels of beta-carotene and the 8 factors. The analysis was conducted without considering any interaction effect among gender, smoking status and vitamins intake of the subjects. So, simply the main factors were used to study the relationship that how those factors influence the levels of beta-carotene in blood plasma.

After performing the study, it was noticed that consumption of alcohol and betadiet was extremely unrelated in determining levels of beta-carotene in blood plasma. These two factors does not impact much on levels of beta-carotene. In contrast, BMI and fiber intake are highly associated with levels of beta-carotene. However, higher the BMI lower is the level of beta-carotene and higher the fiber intake higher will be the levels of beta-carotene in blood plasma.

In terms of vitamin use the subjects whether they use supplements occasionally or daily compared to those who do not use, tend to have higher levels of beta-carotene. Smoking status of the subject has different observations as current smokers tend to have lower levels compared to non-smokers, and the former smokers also have lower levels compared to non-smokers but statistically it is insignificant means non-smokers and former smokers have approximately similar levels of beta-carotene. Lastly, age of the subject indicates a slight positive association between the levels of beta-carotene. These results are quite consistent with graphical observations.

Overall, the factors that determine levels of beta-carotene are age, sex, smoking status, BMI, fiber intake and vitamin use. However, the information explained by these factors is just 22.87%. There could be some other factors responsible in determining relationship with the levels of beta-carotene in blood plasma such as medical drugs taken or physical activity or may be the numbers of subjects under study are small and subjects may have not been randomly selected. Moreover, these factors if tested jointly or considering the interaction might give different results. In conclusion, another study could be done to know better association.

Question 2.

The geometric distribution has the following probability function:

$$f(y) = (1-\pi)^{y}\,\pi \qquad y = 0,1,\ldots\ 0 < \pi < 1$$

i. To show that geometric distribution is a member of exponential family, the pdf has to be expressed in the form:

$$f(y;\theta,\boldsymbol{\phi}) = \exp\left\{\frac{y\theta-b(\theta)}{\phi}+c(y,\phi)\right\}$$

Solution: $f(y) = (1-\pi)^{y}\,\pi$

We can express this as: exp {y log (1-π) + log (π)}

Further we suppose $\theta = \log(1-\pi)$

Therefore, $e^{\theta} = (1-\pi)$

$\pi = 1 - e^{\theta}$ and $\log\pi = \log(1-e^{\theta})$

So, the pdf can be written as:

$$f(y;\theta,\boldsymbol{\phi}) = \exp\left\{\frac{y\log(1-\pi)+\log(\pi)}{1}\right\}$$

$$= \exp\left\{\frac{y\theta-(-\log(1-e^{\theta})}{1}\right\}$$

where $\theta = \log(1-\pi)$ is the Natural parameter.

ϕ=1 is the scale parameter.

$b(\theta) = -\log(1-e^{\theta})$

and $c(y,\phi) = 0$

ii. Give the natural parameter and the scale parameter.

Solution: $\log(1-\pi)$ is the natural parameter and ϕ=1 is the scale parameter.

iii. Hence, show that the mean and variance of the geometric distribution are

$$E(Y) = \frac{1-\pi}{\pi} \text{ and } Var(Y) = \frac{1-\pi}{\pi^{2}}$$

Solution: From part i. $b(\theta) = -\log(1-e^{\theta})$ ……….(1)

Taking derivative w.r.t θ we get, $b'(\theta) = \frac{-(-e^{\theta})}{(1-e^{\theta})} = \frac{(e^{\theta})}{(1-e^{\theta})}$ ……......(2)

On substituting the value $e^{\theta} = (1-\pi)$ and $\pi = 1 - e^{\theta}$, we get

$$E(Y) = \frac{1-\pi}{1-(1-\pi)} = \frac{1-\pi}{\pi}$$

Var(Y) = $b''(\theta)$ as ϕ=1

Taking derivative of equation (2) w.r.t θ again, we get

$$= \frac{e^{\theta}(1-e^{\theta})-e^{\theta}(-e^{\theta})}{(1-e^{\theta})^2} = \frac{e^{\theta}-e^{2\theta}+e^{2\theta}}{(1-e^{\theta})^2} = \frac{e^{\theta}}{(1-e^{\theta})^2}$$

on substituting value of e^{θ} = $(1-\pi)$ and $\pi = 1 - e^{\theta}$we get,

Var(Y) = $\frac{1-\pi}{\pi^2}$. Hence, proved.

iv. Using the notation μ = E (Y) and σ^2 = Var (Y) show that

$\sigma^2 = \mu^2 + \mu$.

Solution: $\mu^2 = E\,(Y)^2 = \left\{\frac{1-\pi}{\pi}\right\}\left\{\frac{1-\pi}{\pi}\right\} = \left\{\frac{1+\pi^2-2\pi}{\pi^2}\right\}$

Now, $\mu^2 + \mu = \left\{\frac{1+\pi^2-2\pi}{\pi^2}\right\} + \left\{\frac{1-\pi}{\pi}\right\} = \left\{\frac{1+\pi^2-2\pi+\pi-\pi^2}{\pi^2}\right\} = \frac{1-\pi}{\pi^2} = \sigma^2$

Hence, proved.

v. Does the geometric distribution have homogeneous variance (homoscedasticity)? If yes, explain why. If no, describe the nature of the non-homogeneous variance.

Solution: Yes, the geometric distribution has the homogeneous variance as the dispersion parameter is constant = 1 and variance function is just the function of mean.

Appendix

ASSIGNMENT_1_45665761

Sukhjeet Kaur

15/08/2019

```
library(gamlss)

data(plasma)  #### reading the dataset ####
View(plasma)
table(which(plasma$betaplasma==0)) ## checking which record has value 0 for
betaplasma ##

##
## 257
##   1
```

Creating a subset of the data plasma using 8 predictors and one outcome variable-“betaplasma” and removing record(257) as there is an implausible value coded as 0 under “betaplasma”

```
sub_plasma<-plasma[-c(257),c(1,2,3,4,5,8,9,11,13)]
```

Visual representation of the numerical data to see any undesired patterns(non-linear) before building a model

```
#### graphical examination of the variables ####
library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

geov=ggpairs(sub_plasma,columns = c(1,4,6,7,8,9))
geov
```

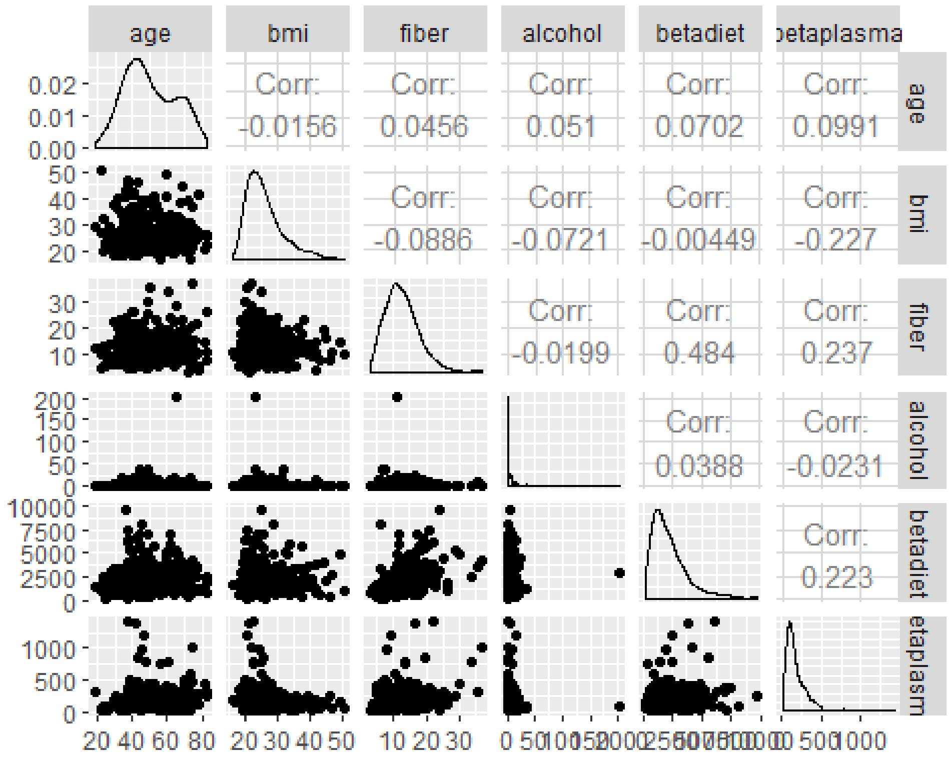


Fig a)

```
### as the distribution of the outcome variable is non-normal we will use
log-transformation to get a normal distribution.Though the distribution of
 numerical covariates is also non-normal but at this stage we do not apply
 any transformation as there is no assumption of normality of the covariat
es under multiple linear regression model.###

####transformaing outcome variable as it is right skewed ####
sub_plasma$log_betaplasma<-log(sub_plasma$betaplasma)
hist(sub_plasma$log_betaplasma,col="grey",xlab = "log(betaplasma)",ylab = "F
requency",
     main = "Histogram of outcome variable after transformation")
```

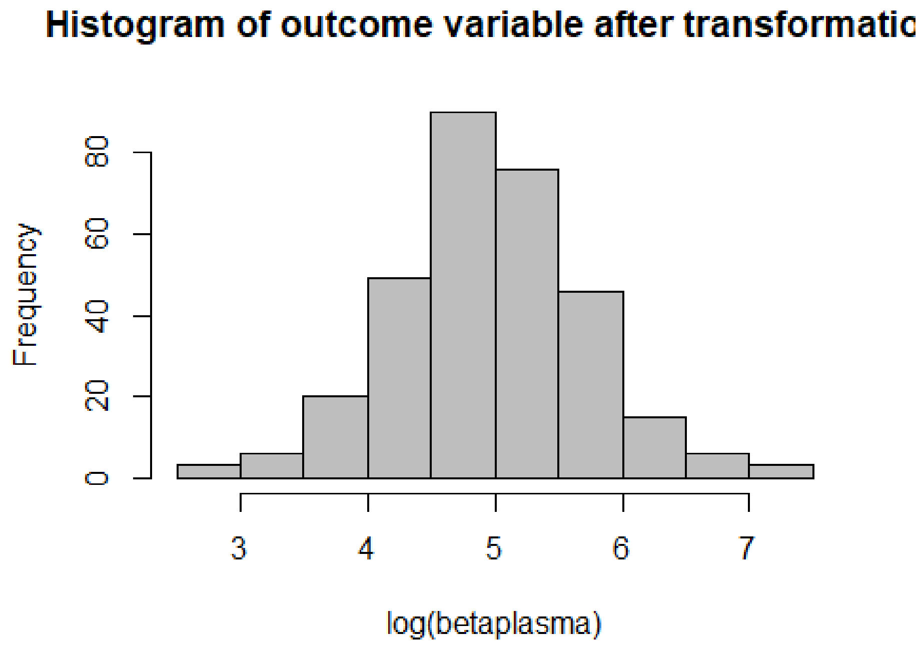


Fig b)

```
### The distribution of alcohol is extremely right skewed, there seems an
extreme value or outlier under this variable so we can remove that value #
##
summary(sub_plasma$alcohol)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.30    3.29    3.20  203.00

boxplot(sub_plasma$alcohol,horizontal = T,col = "seagreen",xlab="Number of a
lcohol drinks consumed per week",main="Boxplot displaying the alcohol drin
ks consumed in a week")
```

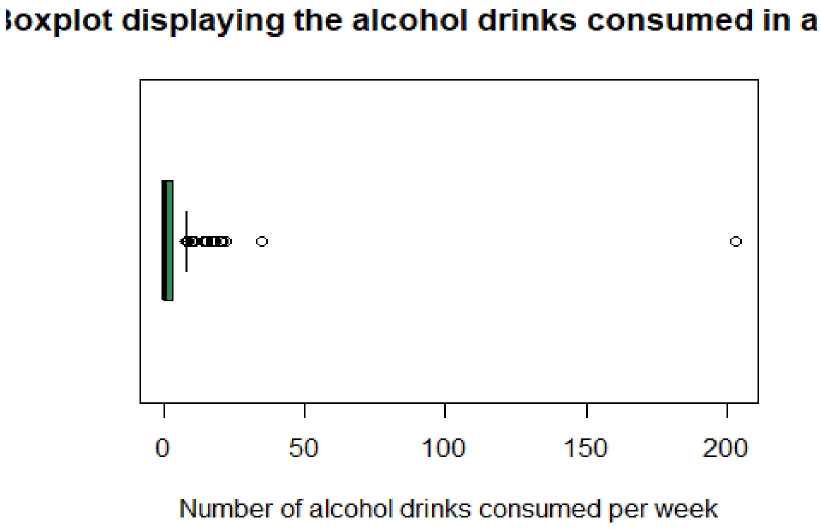
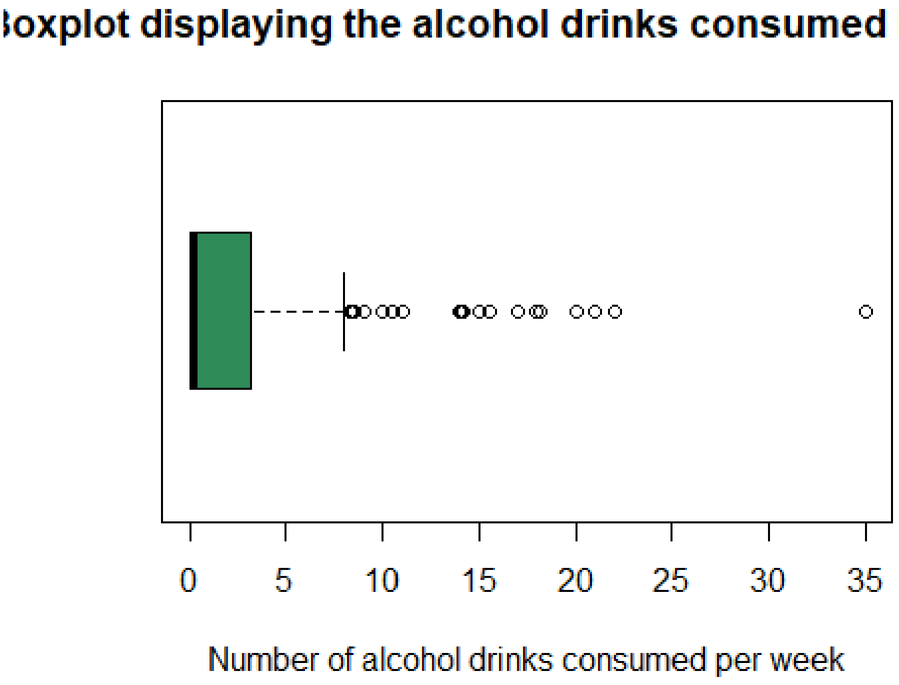


Fig c)

```
## as it is clear from the boxplot that there is an extreme value possibly
 greater than 40 which can be considered as implausible so excluding it fr
om the data is a good choice ##

table(which(sub_plasma$alcohol>40)) ## to see which record has extreme value
 under alcohol variable

##
## 62
##  1

sub_plasma<-sub_plasma[-62,] ## removing a record with value for alcohol=20
3 ##
boxplot(sub_plasma$alcohol,horizontal = T,col = "seagreen",xlab="Number of a
lcohol drinks consumed per week",main="Boxplot displaying the alcohol drin
ks consumed in a week")
```



Recoding the categorical variables

```
library(car)

## Loading required package: carData

sub_plasma$sex<-recode(sub_plasma$sex,'1 ="Male";2 ="Female"')
sub_plasma$sex<-as.factor(sub_plasma$sex)
sub_plasma$smokstat<-recode(sub_plasma$smokstat,'1 ="Never";2 ="Former";3="C
urrent"')
sub_plasma$smokstat<-as.factor(sub_plasma$smokstat)
sub_plasma$vituse<-recode(sub_plasma$vituse,'1 ="Daily";2 ="Occasionally";3=
"None"')
sub_plasma$vituse<-as.factor(sub_plasma$vituse)
```

```
### looking at the boxplots of categorical covariates with log_betaplasma
###
ggplot(sub_plasma,aes(x=sex,y=log_betaplasma))+geom_boxplot(fill="seagreen")
```

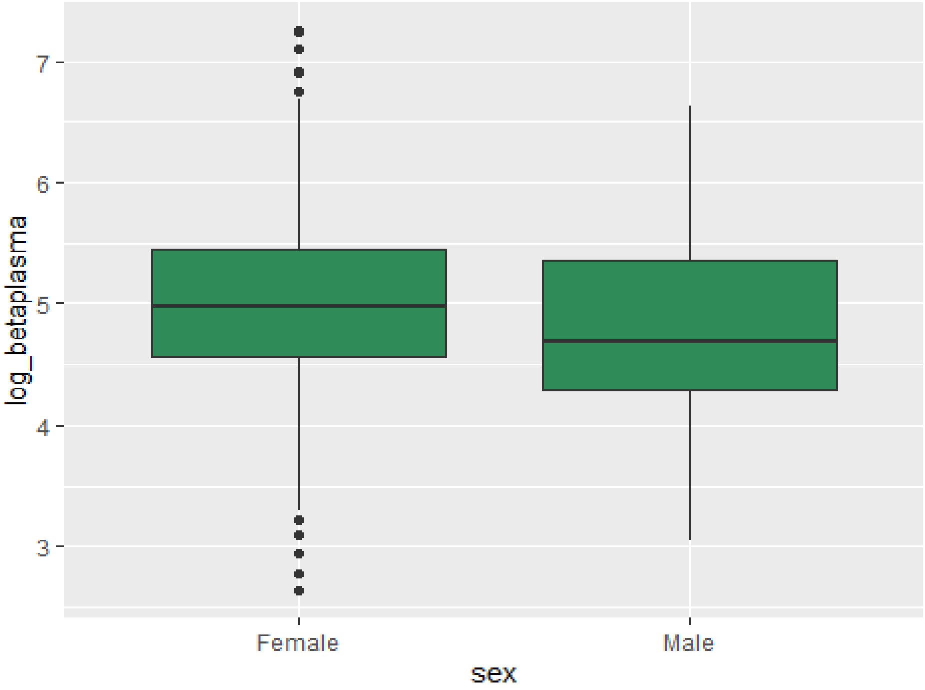


Fig d)

```
ggplot(sub_plasma,aes(x=smokstat,y=log_betaplasma))+geom_boxplot(fill="midnig
htblue")
```

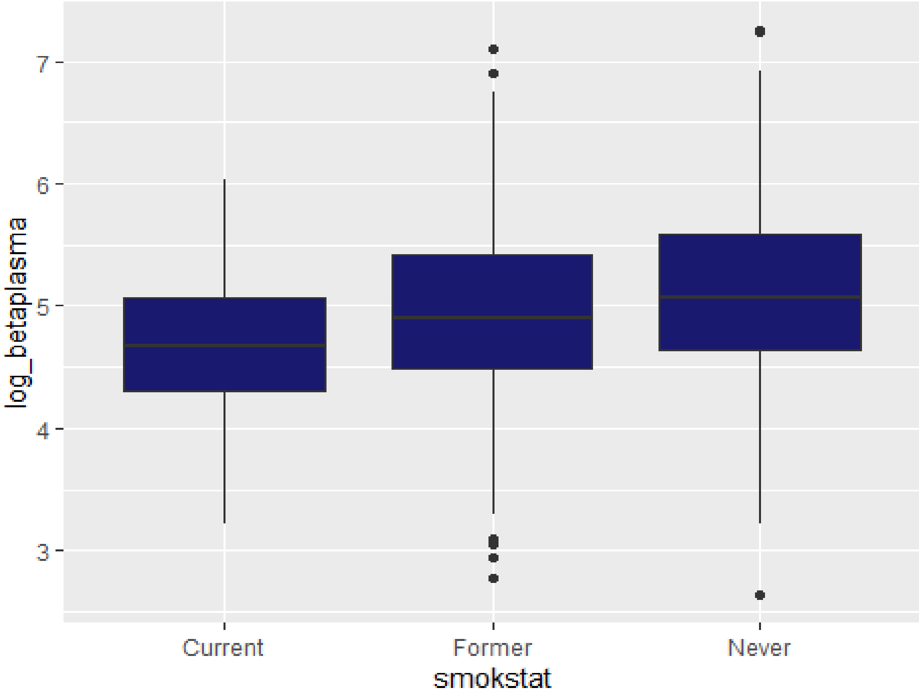


Fig e)

```
ggplot(sub_plasma,aes(x=vituse,y=log_betaplasma))+geom_boxplot(fill="chocolat
e")
```

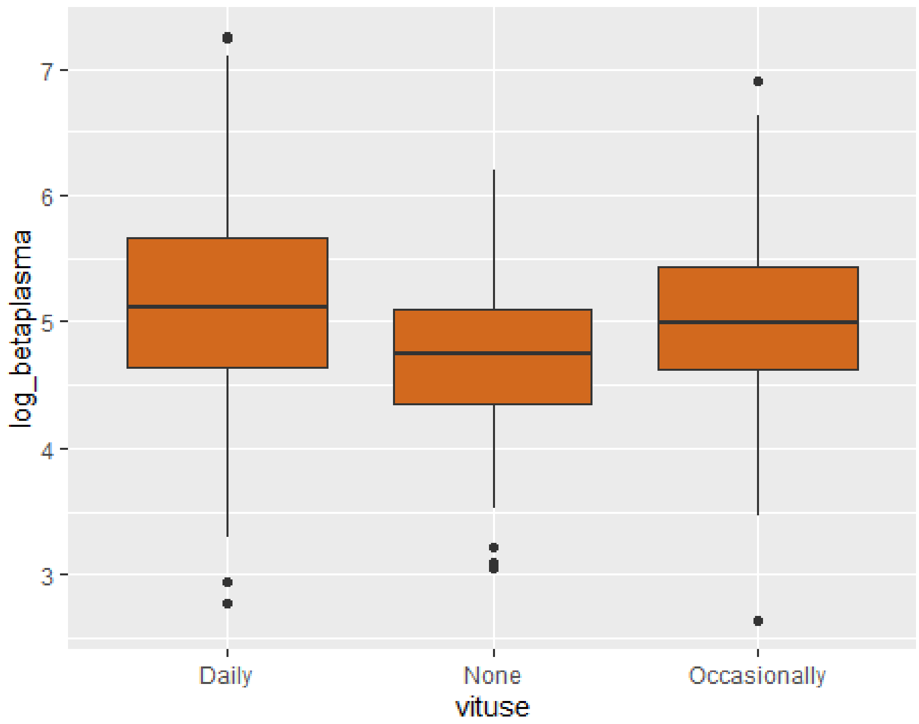


Fig f)

a) By visualizing the comparative boxplots, we can say that the levels of plasma beta-carotene are relatively higher among females than males but there are few outliers in the lower and upper tails in case of females.

b) By looking at the comparative boxplots for smoking status it can be stated that on average beta-carotene is lowest among current smokers and high among ones who do not smoke with two outliers one at each end. For the cases or patients those under former smoking status has approximately similar levels of beta-carotene as that of Non- smokers although there are few outliers in both ends.

c) From the comparative boxplots it is clear that patient with regular vitamin intake has higher levels of beta-carotene than those who do not take vitamins often as well as who donot take at all, on average.But interestingly the levels of plasma beta-carotene are not relatively much lower for ones who donot take vitamins rather there are few cases with low levels who consume vitamins quite often.

setting reference level never for smoking status and none for vituse

```
sub_plasma$vituse<-relevel(factor(sub_plasma$vituse),ref="None")
sub_plasma$smokstat<-relevel(factor(sub_plasma$smokstat),ref="Never")
```

Building a Multiple regression model using all predictors

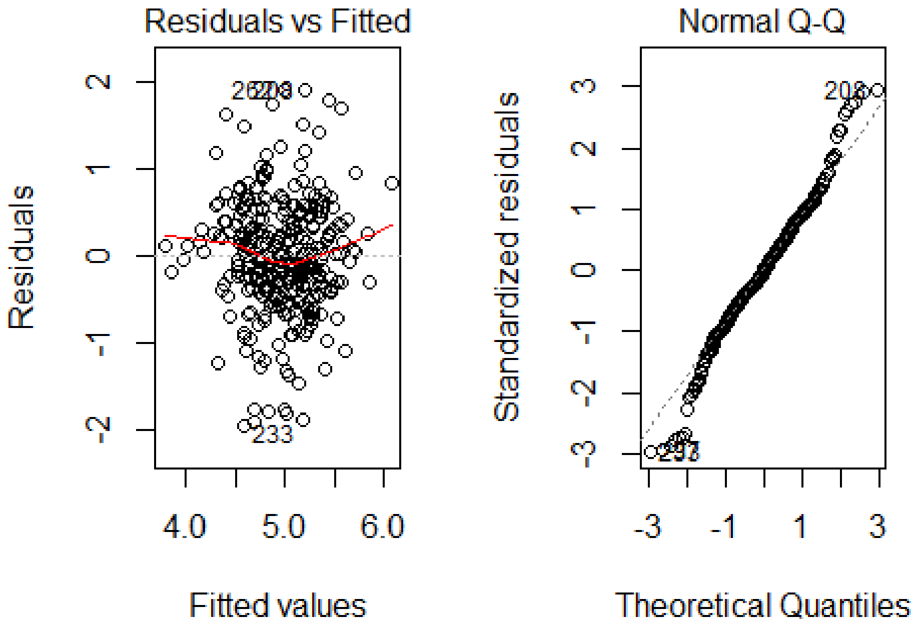
```
model_all<-lm(log_betaplasma~age+sex+smokstat+bmi+vituse+fiber+alcohol+betadi
et,data=sub_plasma)
summary(model_all)

##
## Call:
```

```
## lm(formula = log_betaplasma ~ age + sex + smokstat + bmi + vituse + 
##     fiber + alcohol + betadiet, data = sub_plasma)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.93603 -0.36104 -0.01756  0.41423  1.91079 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          5.067e+00  2.675e-01  18.939  < 2e-16 ***
## age                  7.467e-03  2.772e-03   2.694  0.00746 ** 
## sexMale             -3.022e-01  1.225e-01  -2.467  0.01418 *  
## smokstatCurrent     -3.242e-01  1.216e-01  -2.668  0.00805 ** 
## smokstatFormer      -9.852e-02  8.452e-02  -1.166  0.24468    
## bmi                 -3.361e-02  6.438e-03  -5.221 3.32e-07 ***
## vituseDaily          2.810e-01  9.099e-02   3.088  0.00220 ** 
## vituseOccasionally   2.581e-01  9.941e-02   2.596  0.00989 ** 
## fiber                1.910e-02  8.208e-03   2.327  0.02062 *  
## alcohol              3.092e-04  8.110e-03   0.038  0.96962    
## betadiet             4.418e-05  2.953e-05   1.496  0.13562    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6657 on 302 degrees of freedom
## Multiple R-squared:  0.2344, Adjusted R-squared:  0.2091 
## F-statistic: 9.248 on 10 and 302 DF,  p-value: 2.31e-13
```

```
anova(model_all)
```

```
## Analysis of Variance Table
## 
## Response: log_betaplasma
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## age         1   3.325  3.3251  7.5028 0.0065263 ** 
## sex         1   5.477  5.4768 12.3582 0.0005064 ***
## smokstat    2   4.439  2.2193  5.0077 0.0072519 ** 
## bmi         1  15.333 15.3326 34.5972 1.075e-08 ***
## vituse      2   6.140  3.0700  6.9273 0.0011441 ** 
## fiber       1   5.271  5.2709 11.8935 0.0006434 ***
## alcohol     1   0.008  0.0084  0.0189 0.8906022    
## betadiet    1   0.992  0.9922  2.2389 0.1356215    
## Residuals 302 133.839  0.4432                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
plot(model_all,which=(1:2))
```



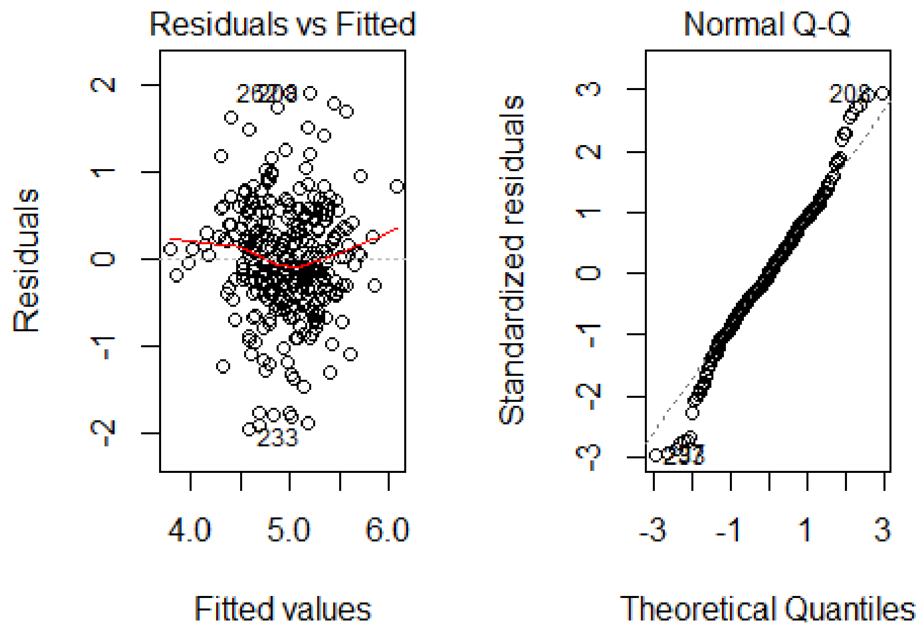
we can see that in model_all, taking into account all other predictors except betadiet alcohol has highest p-value which is insignificant means it has no relation with levels of betaplasma in the blood.But we will rerun the model by taking into account all the predictors first

```
Model1<-lm(log_betaplasma~age+sex+smokstat+bmi+vituse+fiber+betadiet+
              alcohol,data=sub_plasma)
summary(Model1)

##
## Call:
## lm(formula = log_betaplasma ~ age + sex + smokstat + bmi + vituse +
##     fiber + betadiet + alcohol, data = sub_plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93603 -0.36104 -0.01756  0.41423  1.91079
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.067e+00  2.675e-01  18.939  < 2e-16 ***
## age                  7.467e-03  2.772e-03   2.694  0.00746 **
## sexMale             -3.022e-01  1.225e-01  -2.467  0.01418 *
## smokstatCurrent     -3.242e-01  1.216e-01  -2.668  0.00805 **
## smokstatFormer      -9.852e-02  8.452e-02  -1.166  0.24468
## bmi                 -3.361e-02  6.438e-03  -5.221 3.32e-07 ***
## vituseDaily          2.810e-01  9.099e-02   3.088  0.00220 **
## vituseOccasionally   2.581e-01  9.941e-02   2.596  0.00989 **
## fiber                1.910e-02  8.208e-03   2.327  0.02062 *
## betadiet             4.418e-05  2.953e-05   1.496  0.13562
## alcohol              3.092e-04  8.110e-03   0.038  0.96962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6657 on 302 degrees of freedom
## Multiple R-squared:  0.2344, Adjusted R-squared:  0.2091
## F-statistic: 9.248 on 10 and 302 DF,  p-value: 2.31e-13
```

```
anova(Model1)
```

```
## Analysis of Variance Table
##
## Response: log_betaplasma
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## age         1   3.325  3.3251  7.5028 0.0065263 **
## sex         1   5.477  5.4768 12.3582 0.0005064 ***
## smokstat    2   4.439  2.2193  5.0077 0.0072519 **
## bmi         1  15.333 15.3326 34.5972 1.075e-08 ***
## vituse      2   6.140  3.0700  6.9273 0.0011441 **
## fiber       1   5.271  5.2709 11.8935 0.0006434 ***
## betadiet    1   1.000  1.0000  2.2564 0.1341087
## alcohol     1   0.001  0.0006  0.0015 0.9696176
## Residuals 302 133.839  0.4432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
plot(Model1,which = (1:2))
```



Still the p-value for alcohol is highest (>0.05) means insignificant to determine the relation of alcohol drinks consumed with beta-carotene levels in the blood.Hence, alcohol would be excluded from the model.

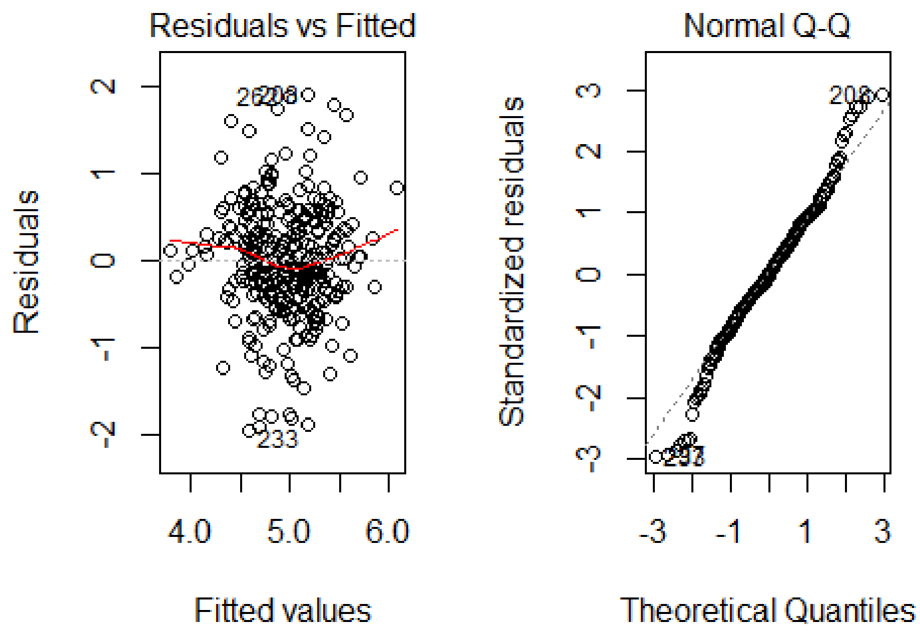
```
Model2<-lm(log_betaplasma~age+sex+smokstat+bmi+vituse+fiber+betadiet,data=sub
_plasma)
summary(Model2)
```

```
##
## Call:
## lm(formula = log_betaplasma ~ age + sex + smokstat + bmi + vituse +
##     fiber + betadiet, data = sub_plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93614 -0.36177 -0.01883  0.41402  1.91438
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.069e+00  2.621e-01  19.339  < 2e-16 ***
## age                  7.460e-03  2.762e-03   2.701  0.00729 **
## sexMale             -3.011e-01  1.192e-01  -2.527  0.01201 *
## smokstatCurrent     -3.241e-01  1.213e-01  -2.672  0.00795 **
## smokstatFormer      -9.805e-02  8.346e-02  -1.175  0.24097
## bmi                 -3.365e-02  6.367e-03  -5.284 2.42e-07 ***
## vituseDaily          2.805e-01  9.020e-02   3.110  0.00205 **
## vituseOccasionally   2.579e-01  9.917e-02   2.601  0.00976 **
## fiber                1.908e-02  8.182e-03   2.332  0.02034 *
## betadiet             4.426e-05  2.941e-05   1.505  0.13347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6646 on 303 degrees of freedom
## Multiple R-squared:  0.2344, Adjusted R-squared:  0.2117
## F-statistic: 10.31 on 9 and 303 DF,  p-value: 6.917e-14
```

```
anova(Model2)
```

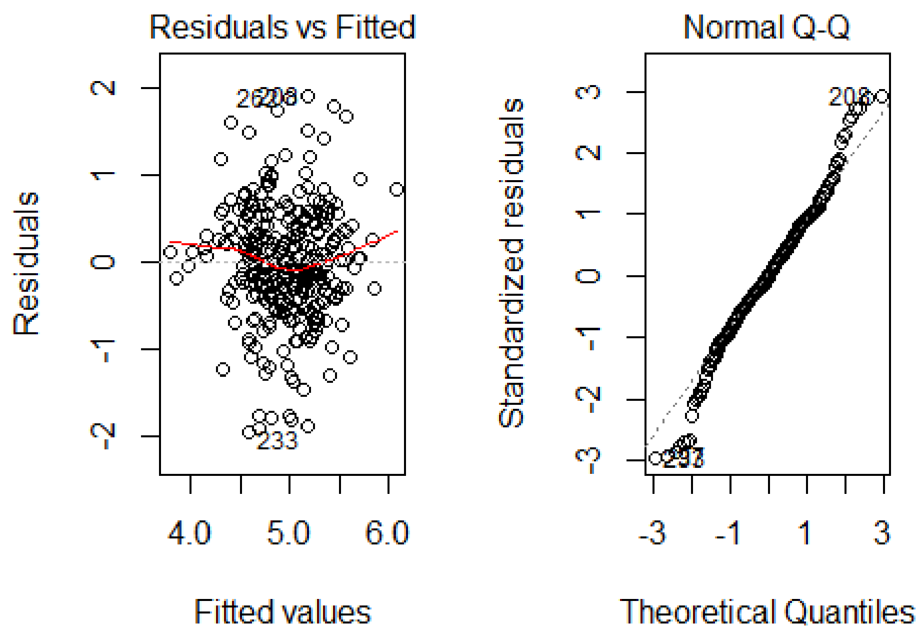
```
## Analysis of Variance Table
##
## Response: log_betaplasma
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## age         1   3.325  3.3251  7.5276 0.0064380 **
## sex         1   5.477  5.4768 12.3990 0.0004956 ***
## smokstat    2   4.439  2.2193  5.0243 0.0071347 **
## bmi         1  15.333 15.3326 34.7116 1.016e-08 ***
## vituse      2   6.140  3.0700  6.9502 0.0011188 **
## fiber       1   5.271  5.2709 11.9328 0.0006302 ***
## betadiet    1   1.000  1.0000  2.2639 0.1334660
## Residuals 303 133.839  0.4417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
plot(Model2,which=(1:2))
```



```
Model_2<-lm(log_betaplasma~age+sex+bmi+vituse+fiber+betadiet+smokstat,data=su
b_plasma)
summary(Model_2)

## 
## Call:
## lm(formula = log_betaplasma ~ age + sex + bmi + vituse + fiber + 
##     betadiet + smokstat, data = sub_plasma)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.93614 -0.36177 -0.01883  0.41402  1.91438 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          5.069e+00  2.621e-01  19.339  < 2e-16 ***
## age                  7.460e-03  2.762e-03   2.701  0.00729 ** 
## sexMale             -3.011e-01  1.192e-01  -2.527  0.01201 *  
## bmi                 -3.365e-02  6.367e-03  -5.284 2.42e-07 ***
## vituseDaily          2.805e-01  9.020e-02   3.110  0.00205 ** 
## vituseOccasionally  2.579e-01  9.917e-02   2.601  0.00976 ** 
## fiber                1.908e-02  8.182e-03   2.332  0.02034 *  
## betadiet             4.426e-05  2.941e-05   1.505  0.13347    
## smokstatCurrent     -3.241e-01  1.213e-01  -2.672  0.00795 ** 
## smokstatFormer      -9.805e-02  8.346e-02  -1.175  0.24097    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6646 on 303 degrees of freedom
## Multiple R-squared:  0.2344, Adjusted R-squared:  0.2117 
## F-statistic: 10.31 on 9 and 303 DF,  p-value: 6.917e-14
```

```
anova(Model_2)

## Analysis of Variance Table
##
## Response: log_betaplasma
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## age           1   3.325  3.3251  7.5276 0.0064380 **
## sex           1   5.477  5.4768 12.3990 0.0004956 ***
## bmi           1  13.322 13.3225 30.1608 8.427e-08 ***
## vituse        2   7.845  3.9223  8.8796 0.0001788 ***
## fiber         1   6.775  6.7755 15.3390 0.0001111 ***
## betadiet      1   1.034  1.0342  2.3414 0.1270182
## smokstat      2   3.205  1.6027  3.6284 0.0277192 *
## Residuals   303 133.839  0.4417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(1,2))
plot(Model_2,which=(1:2))
```



In Model2 p-value for betadiet is insignificant after taking into account all other predictors and excluding alchohol from the model which indicates that dietary intake of beta-carotene is not helpful in determinig the beta-carotene levels in blood plasma. Therefore, we will remove variable beta-diet also from the model.

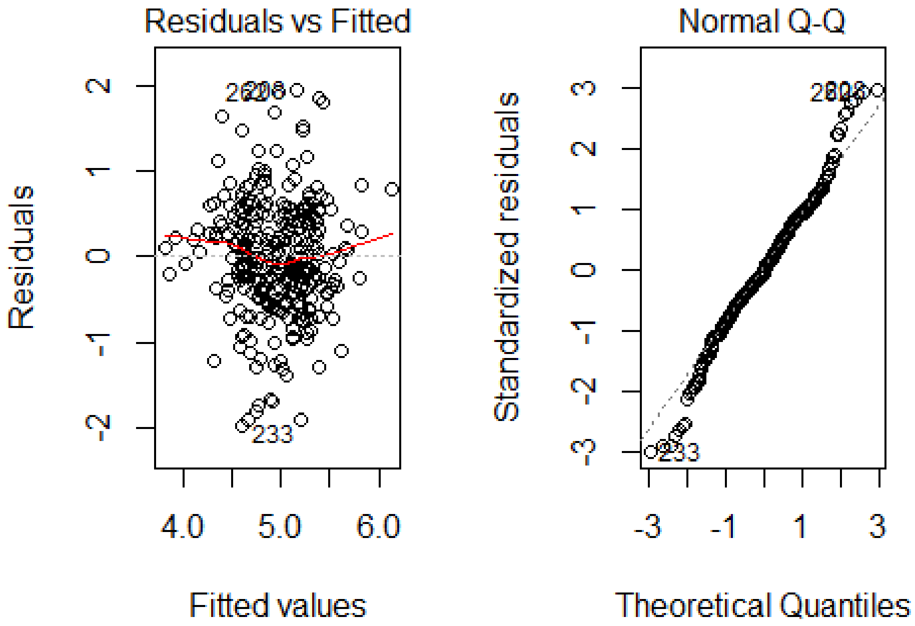
```
Model3<-lm(log_betaplasma~age+sex+smokstat+bmi+vituse+fiber,data=sub_plasma)
summary(Model3)

##
## Call:
## lm(formula = log_betaplasma ~ age + sex + smokstat + bmi + vituse +
##     fiber, data = sub_plasma)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.96395 -0.37140 -0.03385  0.41977  1.94276 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         5.060877   0.262596  19.272  < 2e-16 ***
## age                 0.007684   0.002763   2.781 0.005764 ** 
## sexMale            -0.309327   0.119290  -2.593 0.009972 ** 
## smokstatCurrent    -0.327436   0.121527  -2.694 0.007445 ** 
## smokstatFormer     -0.088719   0.083397  -1.064 0.288258    
## bmi                -0.033149   0.006372  -5.202 3.62e-07 ***
## vituseDaily         0.289762   0.090182   3.213 0.001454 ** 
## vituseOccasionally  0.260064   0.099370   2.617 0.009311 ** 
## fiber               0.024905   0.007225   3.447 0.000646 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.666 on 304 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.2084 
## F-statistic: 11.27 on 8 and 304 DF,  p-value: 5.601e-14
```

```
anova(Model3)
```

```
## Analysis of Variance Table
## 
## Response: log_betaplasma
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## age         1   3.325  3.3251  7.4965 0.0065464 ** 
## sex         1   5.477  5.4768 12.3477 0.0005087 ***
## smokstat    2   4.439  2.2193  5.0035 0.0072780 ** 
## bmi         1  15.333 15.3326 34.5679 1.083e-08 ***
## vituse      2   6.140  3.0700  6.9215 0.0011494 ** 
## fiber       1   5.271  5.2709 11.8834 0.0006462 ***
## Residuals 304 134.839  0.4436                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

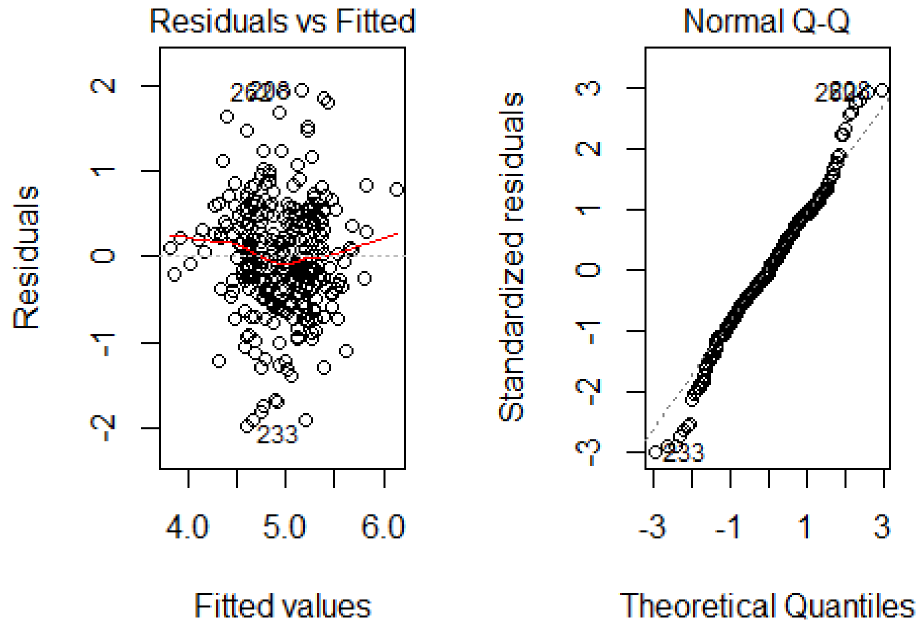
```
par(mfrow=c(1,2))
plot(Model3,which=(1:2))
```



```
Model4<-lm(log_betaplasma~age+sex+bmi+vituse+fiber+smokstat,data=sub_plasma)
summary(Model4)

##
## Call:
## lm(formula = log_betaplasma ~ age + sex + bmi + vituse + fiber +
##     smokstat, data = sub_plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96395 -0.37140 -0.03385  0.41977  1.94276
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.060877   0.262596  19.272  < 2e-16 ***
## age                   0.007684   0.002763   2.781 0.005764 **
## sexMale              -0.309327   0.119290  -2.593 0.009972 **
## bmi                  -0.033149   0.006372  -5.202 3.62e-07 ***
## vituseDaily           0.289762   0.090182   3.213 0.001454 **
## vituseOccasionally    0.260064   0.099370   2.617 0.009311 **
## fiber                 0.024905   0.007225   3.447 0.000646 ***
## smokstatCurrent      -0.327436   0.121527  -2.694 0.007445 **
## smokstatFormer       -0.088719   0.083397  -1.064 0.288258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.666 on 304 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.2084
## F-statistic: 11.27 on 8 and 304 DF,  p-value: 5.601e-14

anova(Model4)
```

```
## Analysis of Variance Table
##
## Response: log_betaplasma
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## age           1   3.325  3.3251  7.4965 0.0065464 **
## sex           1   5.477  5.4768 12.3477 0.0005087 ***
## bmi           1  13.322 13.3225 30.0359 8.915e-08 ***
## vituse        2   7.845  3.9223  8.8429 0.0001850 ***
## fiber         1   6.775  6.7755 15.2755 0.0001146 ***
## smokstat      2   3.240  1.6198  3.6520 0.0270838 *
## Residuals 304 134.839  0.4436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
plot(Model4,which = (1:2))
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles