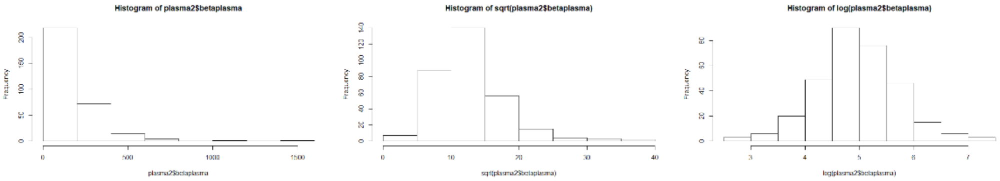
The *plasma* dataset consists of 14 variables – *age, sex, smokstat, bmi, vituse, calories, fat, fiber, alcohol, cholesterol, betadiet, retdiet, betaplasma*, *retplasma* – and has observations recorded from 315 subjects. The aim of this analysis is to determine any relationship between *betaplasma,* the response variable, and other covariates. For this analysis, we will be considering only 8 covariates – *age, sex, smokstat, bmi, vituse, fiber, alcohol, betadiet*. It is to be noted that one observation with 0 value for *betaplasma* is removed from the analysis in order to prevent any effect on the variance.

Response Variable

The histogram of *betadiet*, the response variable, shows a right skewed distribution which calls for transformation. Upon applying square root transformation, the distribution is still slightly right skewed. But the logarithmic transformation produces a normal distribution.

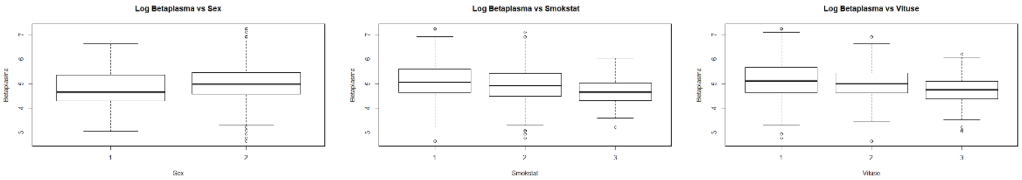


Categorical Covariates

Out of the 8 covariates, *sex, smokstat and vituse* are categorical variables. While the categories are not evenly distributed, there are also not any extremes.

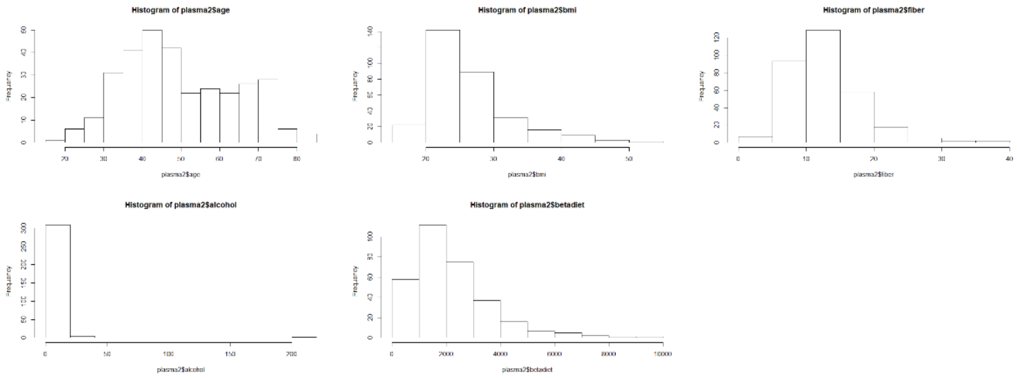
Variables	1	2	3
sex	42	272	na
smokstat	156	115	43
vituse	121	82	111

The boxplots of logarithmically transformed *betaplasma*, the response variable, against these categorical variables - *sex, smokstat and vituse* – seems to have the same median but also few outliers which seems to be negligible.

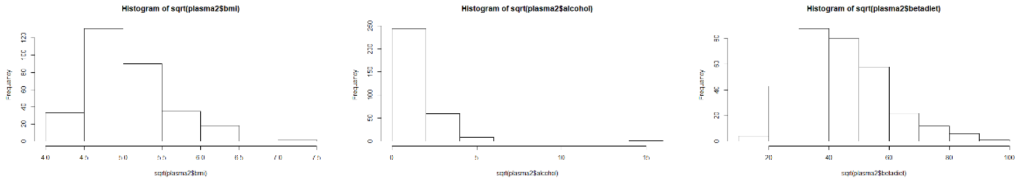


Continuous Covariates

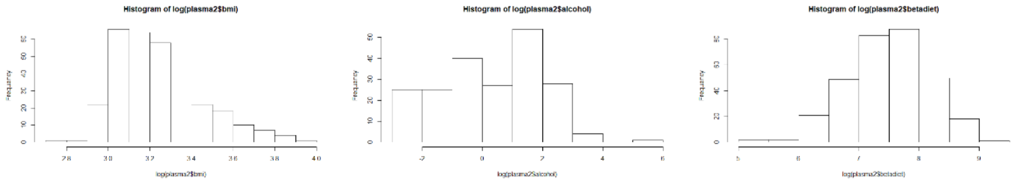
The rest of the covariates – *age, bmi, fiber, alcohol, betadiet* – are the continuous variables and their histograms are shown below. The histograms of *age* and *fiber* show normal distributions, whereas the distribution is right skewed for *alcohol* and slightly right skewed for *bmi* and *betadiet*.



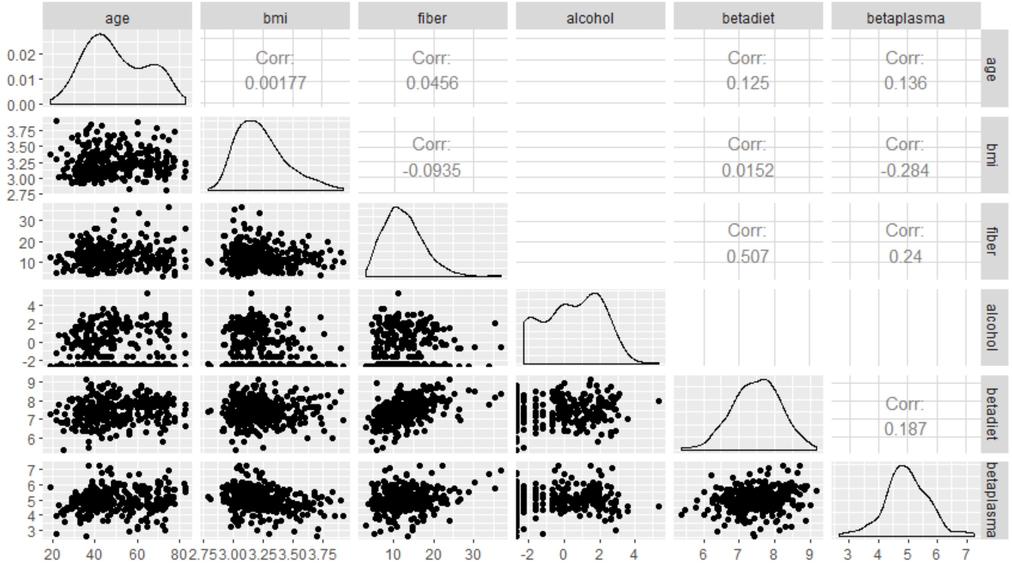
The square root transformation has negligible effect on *alcohol, bmi and betadiet,* still staying largely right skewed.



But the logarithmic transformation does produce an almost normal distributions for *alcohol, bmi and betadiet*.



The scatterplots of the logarithmically transformed response variable *betaplasma* against the continuous covariates largely show random scatter.



The covariate *fiber* against the logarithmically transformed response variable *betaplasma* shows slightly positive linear relationship and there seems to be collinearity between covariates *fiber* and *betadiet*.

Regression Models of Covariates

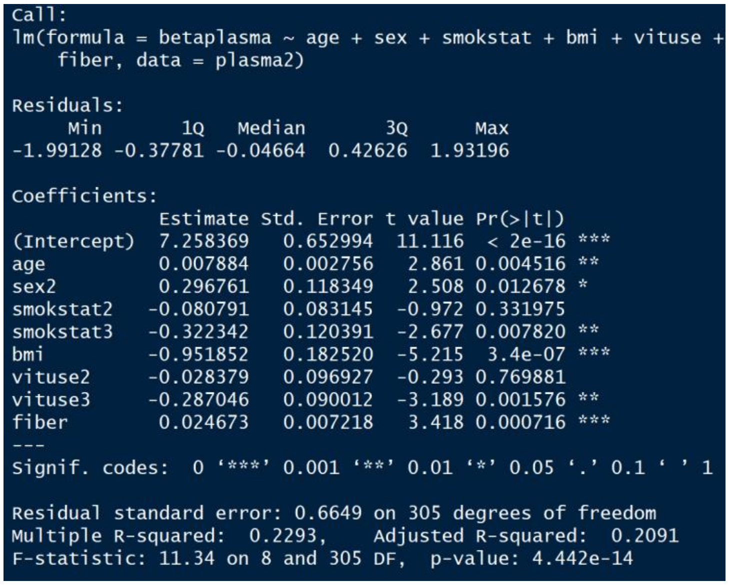
A linear regression model is fit between the logarithmically transformed response variable *betaplasma* and the 8 covariates - *age, sex, smokstat, (log) bmi, vituse, fiber, (log) alcohol, (log) betadiet*.

Covariate	**P Value**
age	0.01596
sex	0.01656
smokstat	0.001694
(log) bmi	3.097e-07
vituse	3.775e-05
fiber	1.695e-05
(log) alcohol	NA
(log) betadiet	0.0008509

A linear model was not possible between the logarithmically transformed response variable *betaplasma* and the logarithmically transformed *alcohol* as the latter has -Inf values. From the above table, we can infer that all the p values of the covariates are less than 0.2 and can be considered for the multivariate regression model.

Multivariate Regression Model

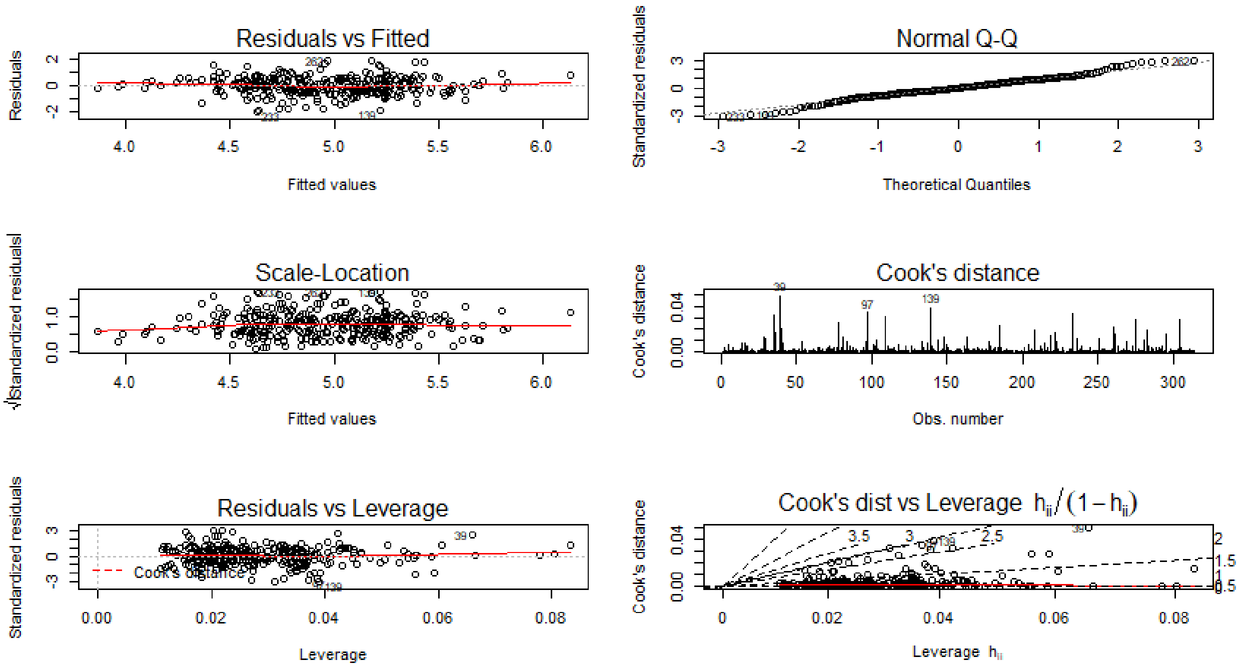
A multivariate regression model is being fit between the logarithmically transformed response variable *betaplasma* and the covariates. Since there is collinearity between *fiber* and *betadiet*, they cannot be infused into the multivariate regression model and only one of them can be considered. The covariate *fiber* is considered due to its slightly positive linear relationship with the logarithmically transformed response variable *betaplasma.* The logarithmically transformed *alcohol* is also excluded due to its -Inf values.



```
Call:
lm(formula = betaplasma ~ age + sex + smokstat + bmi + vituse +
    fiber, data = plasma2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99128 -0.37781 -0.04664  0.42626  1.93196

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.258369   0.652994  11.116  < 2e-16 ***
age          0.007884   0.002756   2.861 0.004516 **
sex2         0.296761   0.118349   2.508 0.012678 *
smokstat2   -0.080791   0.083145  -0.972 0.331975
smokstat3   -0.322342   0.120391  -2.677 0.007820 **
bmi         -0.951852   0.182520  -5.215  3.4e-07 ***
vituse2     -0.028379   0.096927  -0.293 0.769881
vituse3     -0.287046   0.090012  -3.189 0.001576 **
fiber        0.024673   0.007218   3.418 0.000716 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6649 on 305 degrees of freedom
Multiple R-squared:  0.2293,    Adjusted R-squared:  0.2091
F-statistic: 11.34 on 8 and 305 DF,  p-value: 4.442e-14
```

The output shows that $R^2 = 0.2293$ indicating that the model explains 22.93% of the variation in *betaplasma* which is not high but is normal in medical research.

Model Diagnosis



The Residuals vs Fitted plot shows that the points are random around the line at 0, indicating conformity with the assumption of equal variance. The Normal Q-Q plot also shows that the points are normally distributed.

The Cook's distance shows large leverages at observations 39, 97 and 139. The cutoff for large h_i is 2p/n, where p = 9 (intercept + age + sex2 + smokstat2 + smokstat3 + bmi + vituse2 + vituse3 + fiber), which provides (2*9)/314 = 0.057. There are some C_i near the cutoff at 1 indicating that they might be influential in the model fit.

Model Equation

(log betaplasma y_i) = (7.258369 intercept) + (0.007884 age) + (0.296761 sex2) - (0.080791 smokstat2) – (0.322342 smokstat3) - (log 0.951852 bmi) - (0.028379 vituse2) - (0.287046 vituse3) + (0.024673 fiber)