Factors Affecting Plasma Retinol and Beta-Carotene Levels

Statistics 206 Final Project

Christina De Cesaris
cmdecesaris@ucdavis.edu

Abstract

Blood plasma levels of retinol and beta-carotene have been found to have an inverse association with cancer development. The analyzed data was collected from 315 patients who previously underwent a biopsy or removal procedure of a lesion found noncancerous from the lung, uterus, colon, breast, skin, or ovary. In particular, this study sought to determine the effect consuming alcohol and smoking had on beta-carotene and retinol concentrations. Which factors could be used to best predict plasma levels and whether the inclusion of interactions improved predictive ability was also questioned. Models of both responses contained age as a significant predictor. Retinol was found to have a significant positive relationship with alcohol consumption and age between both models. The addition of interactions to the reintol model did not cause notable improvement. Smoking status became significant in the beta-carotene model when interactions were included. Sex, age, vitamin use, fiber, quetelet, and cholesterol were significant predictors between both the beta-carotene models. Overall, the R_a^2 value for retinol was 0.09312 and 0.1046 for the first order and interaction model respectively, and 0.2037 and 0.3074 for the beta-carotene first order and interaction model.

1. Introduction

Low levels of retinol, commonly known as vitamin A, and beta-carotene in the blood have been identified as risk factors for the development of cancers [7]. In the human body, beta-carotene is converted into retinol which is necessary for vision and the maintenance of cellular mucus membranes among other benefits in the body [4]. Traces of the two agents has also been linked to a decrease in risk of cardiovascular disease and aliments of the eyes such as cateracts [5].

The impacts of personal intake habits, age, and sex on blood plasma levels of beta-carotene and retinol are significantly less studied. Understanding the factors which affect the presence of beta-carotene and retinol levels in the blood could be key for identifying risk factors for major diseases, and beneficial towards inventing practices for preventative health [2].

This study seeks to determine if the consumption of carcinogens such as alcohol and smoking habits affect the metabolic intake of beta-carotene and retionl. Whether levels of beta-carotene in the blood influences the blood levels of retinol, and the impact of beta-carotene and retinol ingested on the plasma levels was also questioned. Finally, attempts to see which factors and potential interactions between factors were effective in predicting the respective levels of beta-carotene and retinol were explored.

Multiple linear regression analysis was performed on a data set consisting of 315 patients who previously under went a three-year procedure for the removal or biopsy of a non-cancerous lesion discovered non-cancerous. Along with plasma levels of beta-carotene and retinol, information collected from each patient included, smoking and drinking habits, age, quetelet ($weight/height^2$), vitamin use, and break down of dietary intake. For the following analysis, beta-carotene blood plasma levels (BETAPLASMA) and retinol blood plasma levels (RETPLASMA) were treated as independent response variables. Patient age (AGE), sex (SEX), dietary retinol (RETDIET), dietary beta-carotene (BETDIET), quetelet (QUETELET), fiber intake (FIBER), smoking habits (SMOKSTAT), alcohol consumption per week (ALCOHOL), fat consumption (FAT), calories (CALORIES), cholesterol (CHOLESTEROL), and vitamin use (VITUSE) were treated as predictor variables.

2. Methods and Results

2.1 EDA

Exploratory data and subsequent multiple linear regression analysis was carried out using R Studio code. First, a structural break down of the data was used to search for missing values, and separate the predictors into qualitative and quantitative categories (Listing 1). No missing values were present in the data. Histograms of the response variables indicated a right skewed for both, and alluded to a need for a later transformation (Figure 1). The distributions of the quantitative predictors showed a similar right skewed pattern. (Figure 2,3,4,5,6). Boxplots of the response variables indicated there were no gross outliers among response data (Figure 7), however the boxplots from the predictors exposed case 62 as a gross outlier, primarily affecting alcohol (Figure 8, 9).

The relationships between quantitative data was analyzed through a scatter plot matrix and corresponding correlation values (Figure 12). There was no evidence of strong non-linearity within the data, and the strongest linear pattern was between calories and fat which also had the largest correlation coefficient of 0.9 (Figure 12). From both its boxplot and scatter plot comparison, case 62 was identified as a gross outlier and removed. The removal of case 62 improved the distribution of many response variables, most notably alcohol. Case 171, an outlier for dietary retinol, was also considered for removal as it implied the patient was ingesting toxic levels of retinol daily, but this point was kept after further research deemed the case possible and not terribly uncommon [3].

Sex, smoking status, and vitamin use, the determined quantitative predictors, were organized as pie charts. Eighty-seven percent of cases were female, half had never smoked, and about forty percent used vitamins often (Figure 10,11). The qualitative variables were plotted by level against each predictor for search of pattern (Figure 13,14). From the plots, males were found to have a slightly higher concentration of retinol on average in the blood compared to their female counterparts (Figure 13). This was unsurprising as men have been recorded as having higher retinol levels on average in previous studies [1]. There was no other clear difference between the averages of vitamin use, smoking status, and sex plotted against beta-plasma, or vitamin use and smoking status against retinol (Figure 13,14). However, this does not mean these variables are not influential in predicting the responses because potential interactions have not yet been taken into account.

The boxcox function from the R MASS package and subsequent histograms implied a transform on both responses would be beneficial for model analysis (Figure 17). In the case of beta-plasma, case 276 was valued at zero, resulting in an error when transformed. To amend this, the numeric data was transformed by log(1 + data) to ensure the point could be kept. This is a commonly used technique used by data scientists when dealing with transformations when zeros are present in the data [6]. The original data set was copied before the transformation was applied to the respective response to ensure the data remained consistent–that is, so when fitting a model with transformed beta-carotene, the predictor of retinol would not also be transformed (Listing 2). Histograms of the potential transformations: log, sqaure root, inverse, were plotted to determine the optimal transformation. In both cases, a log transformation was selected (Figure 15,16).

2.2 Model Selection Process

For each model, the following assumptions were made: a linear relationship between the predictors and response exists, residual values are evenly distributed and normal, error terms are equally distributed across predictors, and multicoliniarity does not largely occur.

The transformed data was split into a 50/50 training and validation set for each response respectively. A first order model without interactions was selected using backwards selection, backwards step-wise, forward selection, and forward step-wise methods. These methods were carried out through the stepAIC() function from the R MASS package (Listing 3). As expected, there was overlap between the chosen 'best' models. The methods were performed with AIC and BIC as evaluating criteria. BIC is bias towards smaller models and was included in an attempt to reduce model size (Listing 4).

Output models from stepAIC() were evaluated for accuracy and potential overfitting through validation analysis (Listing 4). The sum of squared estimate of errors SSE and R_a^2 values were calculated for each model under the training and validation data sets and compared. Multiple R^2 values were not considered in model selection as the R^2 value increases with model size and is therefore not reliable. Models which had a large difference between the R_a^2 values for its training and validation data were less likely to be considered, as an ideal model would have a close R_a^2 value for both sets. Then the $training : SSE/n$ was compared to the model's mean squared prediction error $MSPE$ value to expose over-fitting.[] The closer the $training : SSE/n$ and $MSPE$ values, the less over fitting in the model. In all cases the final model was chosen based on minimal over-fitting, maximum R_a^2, number of predictors, and how well the model fit our initial assumptions.

The chosen final model was then subject to outlier analysis. If a case was both identified as a significant outlier, exceeding the Bonferroni's Threshold, or a high leverage case from Cook's distance, it was subject to removal in the final model (Listing **??**). Case removal was kept to a minimum with only one or two point eligible. In most cases, only one point was able to be removed.

2.3 Retinol Results

The best first order model for retinol contained the predictors ALCOHOL and AGE. Case 81 was removed after being identified as a significant leverage point and outlier. When fit on the full data set, the final model had a multiple $R^2 = 0.09894$ and $R_a^2 = 0.09312$.It had a mean squared error $MSE = 0.10498$ (Listing 6). The fitted coefficients were all deemed significant to a level of 0.001. The model residuals were normally distributed and even, but the QQ plot indicated heavy tails and a non-even distribution of error variance across terms (Figure 18). The final model is as follows:

$$log(Y_{Retinol}) = 6.038717 + 0.005501 * X_{Age} + 0.013693 * X_{Alcohol}$$

The final interaction model for retinol contained AGE, CALORIES, FAT, FIBER, ALCOHOL and the interaction FIBER:FAT as predictors. The original selected model also contained CALORIES, CHOLESTEROL, AGE:CALORIES, and AGE:CHOLESTEROL but these terms were found to be insignificant under the summary output and ANOVA fit, and were dropped. Dropping these terms did not affect the R^2or R_a^2 values but it did reduce the MSE. No cases were subject to outlier removal in the final model. When fit on the full data set, the final model had a multiple $R^2 = 0.1217$ and $R_a^2 = 0.1046$.It had a $MSE = 0.10295$ (Listing 7). The model residuals showed non-normal distribution but otherwise appeared even. The QQ plot indicated heavy tails and a non-even distribution of error variance across terms (Figure 19). The final model is as follows:

$$log(Y_{Retinol_int}) = 5.906 + .0056001 * X_{Age} + 0.0001580 * X_{Calories} - 0.0004602 * X_{Fat} + 0.006400 * X_{Fiber} + 0.01181 * X_{Alcohol} - 0.0001922 * X_{Fat} * X_{Fiber}$$

Overall, dietary retinol consumption had no effect on the prediction of retinol in the blood plasma. Smoking status as well had no correlation with the retinol response. In both the interaction and first order model, age and alcohol consumption had a positive correlation with plasma retinol.

2.4 Beta-Carotene Results

The best selected first order model for beta-carotene contained AGE, SEX, QUETELT, VITUSE, FIBER, and CHOLESTEROL as predictors. Case 257 was removed from the model after being identified as a high leverage point and an outlier. The final model had a multiple $R^2 = 0.2215$ and $R_a^2 = 0.2037$. The was $MSE = 0.4296$ (Listing 8). The model residuals were normal and even across the fitted values. As in previous models fitted to this data, the QQ plot indicated heavy tails and a non-even distribution of error variance across terms (Figure 20). The final model is as follows: $log(beta-carotene) = 4.9435924 + 0.0077376 * X_{Age} - 0.2649471 * X_{Sex} - 0.0286197 * X_{Quetelet} + 0.3170306 * X_{VituseOften} + 0.2745545 * X_{VitusNotOften} + 0.0299246 * X_{Fiber} - 0.0006471 * X_{Cholesterol}$

where when $X_{Sex} = 1$ it represents MALE and when $X_{Sex} = 0$ it represents FEMALE, and when $X_{VituseOften} = 0$ and $X_{VituseNotOften} = 0$ it represents VITUSENEVER.

The interaction model for beta-carotene showed improvement in predictive ability compared to the first order model. The best selected, beta-carotene interaction model contained QUETELET CHOLESTEROL, FIBER, RETPLASMA, VITUSE, AGE, SEX, SMOKSTAT, CHOLESTEROL:RETPLASMA, SEX:BETADIET, VITUSE:BETADIET,and BETADIET:SMOKSTAT as significant predictors. The final model had a multiple $R^2 = 0.345$ and $R_a^2 = 0.3074$. The was $MSE = 0.4340$ (Listing 9). The model residuals were normal and even across the fitted values. As in previous models fitted to this data, the QQ plot indicated heavy tails, and a non-even distribution of error variance across terms (Figure 21). The final model is as follows:

$log(beta-carotene) = 5.488040 - 0.0003079646 * X_{Quetelet} - 0.002188216 * X_{Cholesterol} + 0.0002217834 * X_{Fiber} - 0.0001784889 * X_{Retplasma} + 0.1628632 * X_{VituseNotOften} - 0.08178284 * X_{VituseOften} + 0.005729410 * X_{AGE} - 0.007726995 * X_{Sex} - 0.0002366277 * X_{Betadiet} - 0.1848441 * X_{SmokstatFormer} - 0.1027800 * X_{SmokestatNever} + 0.000002621534 * X_{Cholesterol} * X_{Retplasma} - 0.0001083952 X_{Sex} * X_{Betadiet} + 0.00004060739 * X_{VituseNotOften} * X_{Betadiet} + 0.0001632967 * X_{VituseOften} * X_{Betadiet} + 0.0002277851 * X_{Betadiet} * X_{SmokestatFormer} + 0.0002462635 * X_{Betadiet} * X_{SmokstatNever}$

where when $X_{Sex} = 1$ it represents MALE and when $X_{Sex} = 0$ it represents FEMALE, and when $X_{VituseOften} = 0$ and $X_{VituseNotOften} = 0$ it represents VITUSENEVER, and when $X_Smokestat Never = 0$ and $X_Smokestat Former = 0$ it represents SMOKSTATCURRENT.

Unlike retinol, beta-carotene showed no correlation with alcohol consumption. The beta-carotene models also both contained vitamin use as a predictor indicating that vitamin use might improve beta-carotene absorption in some manner. Dietary intake of beta-carotene was also significant in the interaction model along with its interaction with vitamin usage. The interaction between vitamin usage and dietary beta-carotene may be a result of beta-carotene being contained within ingested vitamins.

3. Conclusions and Discussion

In the case of retinol, the age and alcohol were consistent predictors. The addition of interactions int he retinol model did not cause notable improvement. As well, the data in general was highly variable and not easily predicted by a linear model. The overall R squared values for the retinol models were small, indicating there might not be a strong relationship between retinol and this data.

Age was also identified as a significant predictor for both models concerning beta-carotene. This discovery implies that age is influential in predicting levels of nutrients absorbed in the blood. It is possible that this relationship is a result of how metabolism changes with age. Even though smoking status was included in the beta-carotene interaction model, it was not notably important among any of the models. This leads to the conclusion that smoking status does not greatly affect the levels of beta-carotene or retinol in the blood.

Dietary beta-carotene's relationship with plasma beta-carotene is unsurprising, because beta-carotene levels depend on nutrient consumption. The same argument applies to why vitamin use was found to be significant for beta-carotene.

On the other hand, retinol is derived from beta-carotene, so it is sensible that vitamin use and dietary retinol did not have an affect on plasma retinol.

Finally, the variability of the data made it ill-suited for our assumptions, and the best fitted models did not have a significant linear relationship. It was possible to fit models with higher R^2 values but these models were found to be severely over fitted, large, and inaccurate when applied to the entire data set.

Appendix 1.

Figure 1:

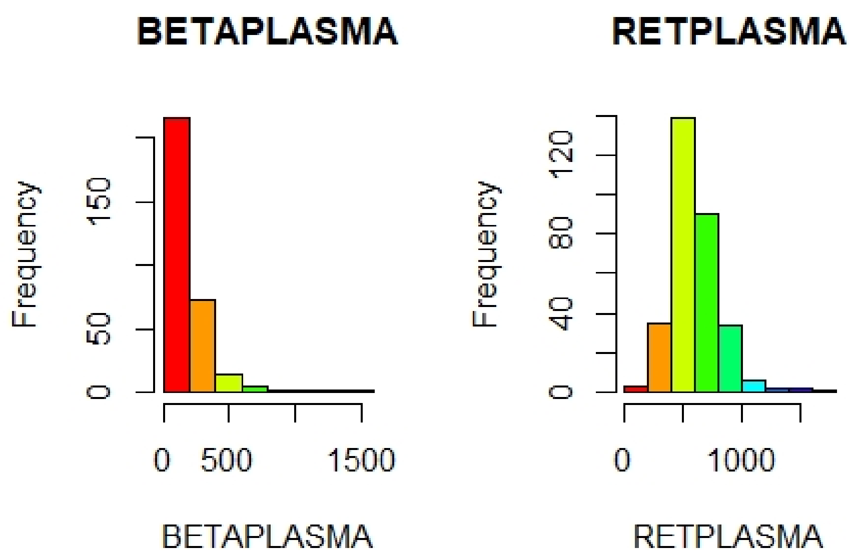


Figure 2:

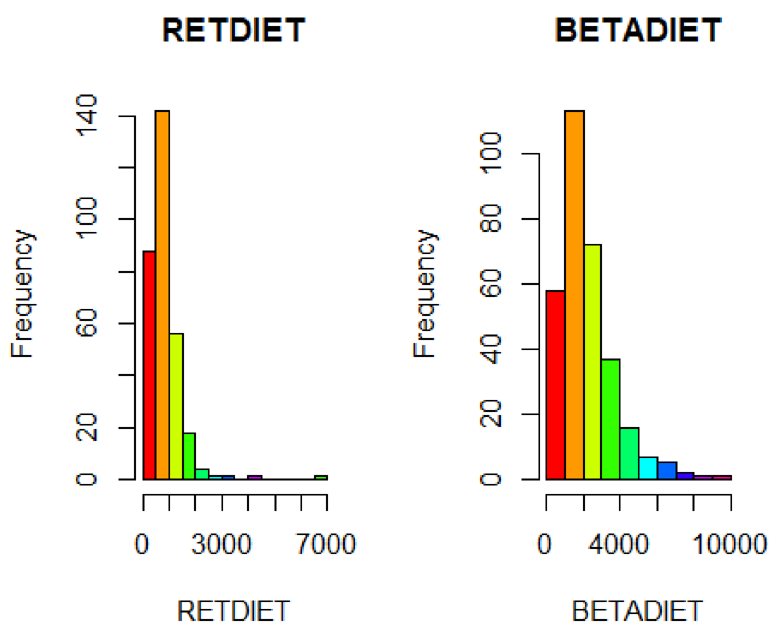


Figure 3:

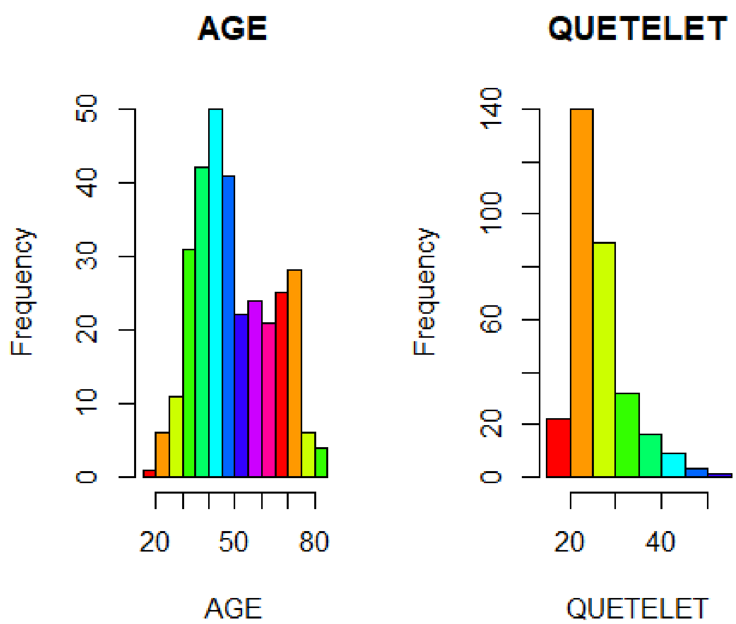


Figure 4:

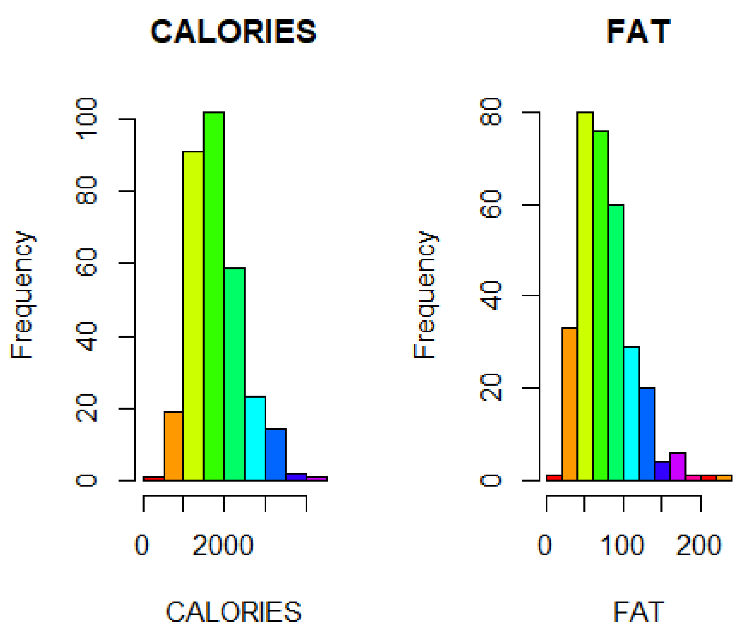


Figure 5:

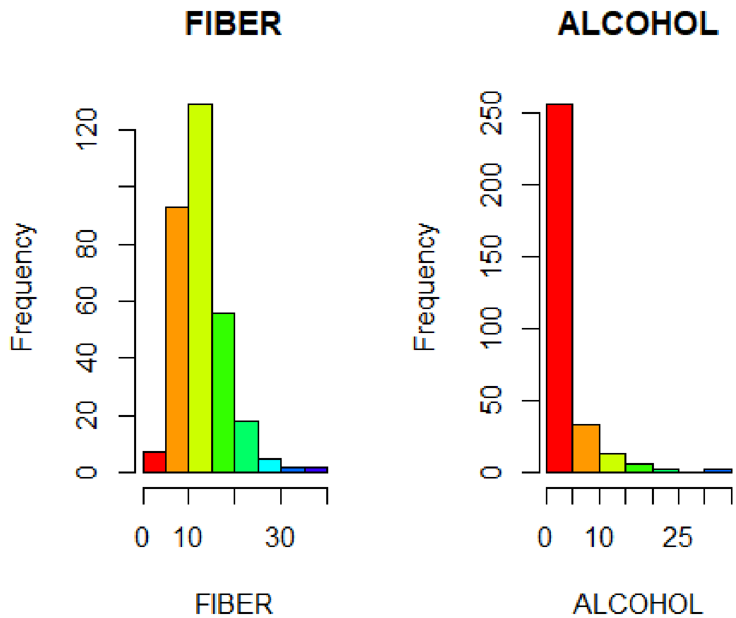


Figure 6:

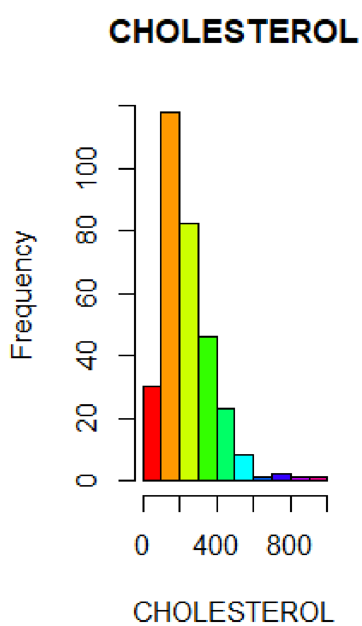


Figure 7:

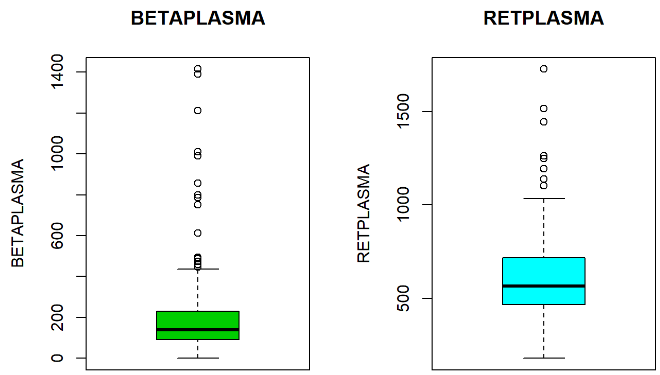


Figure 8:

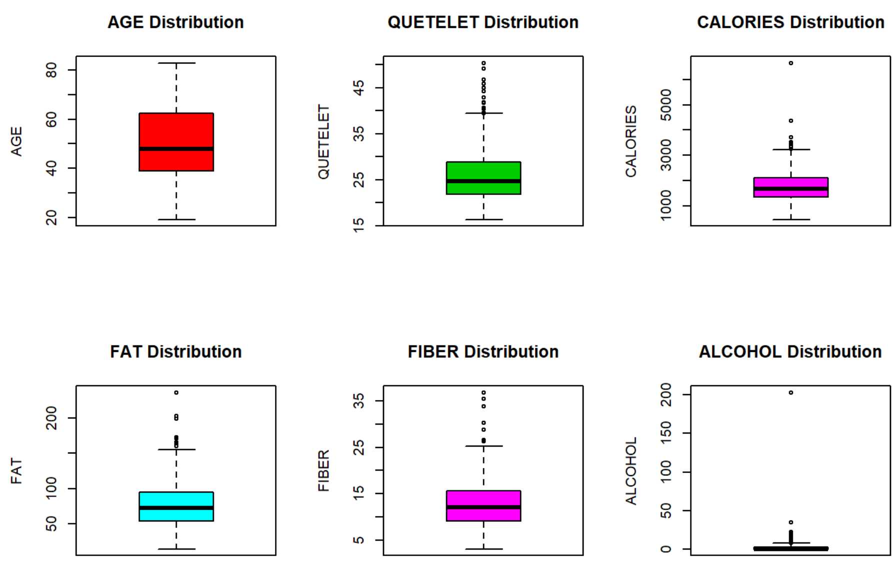


Figure 9:

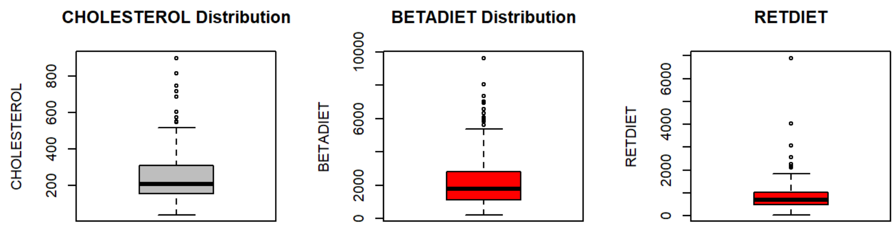


Figure 10:

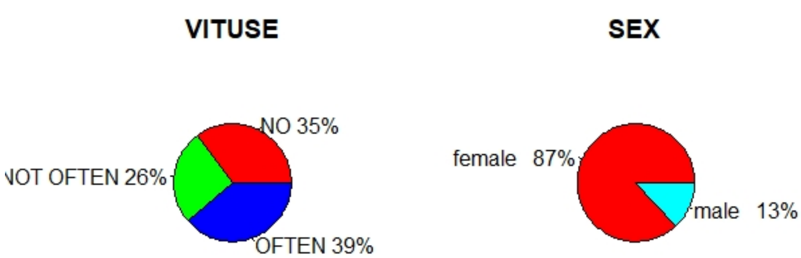


Figure 11:

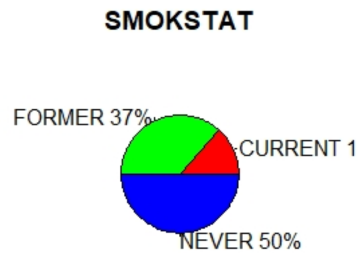


Figure 12:

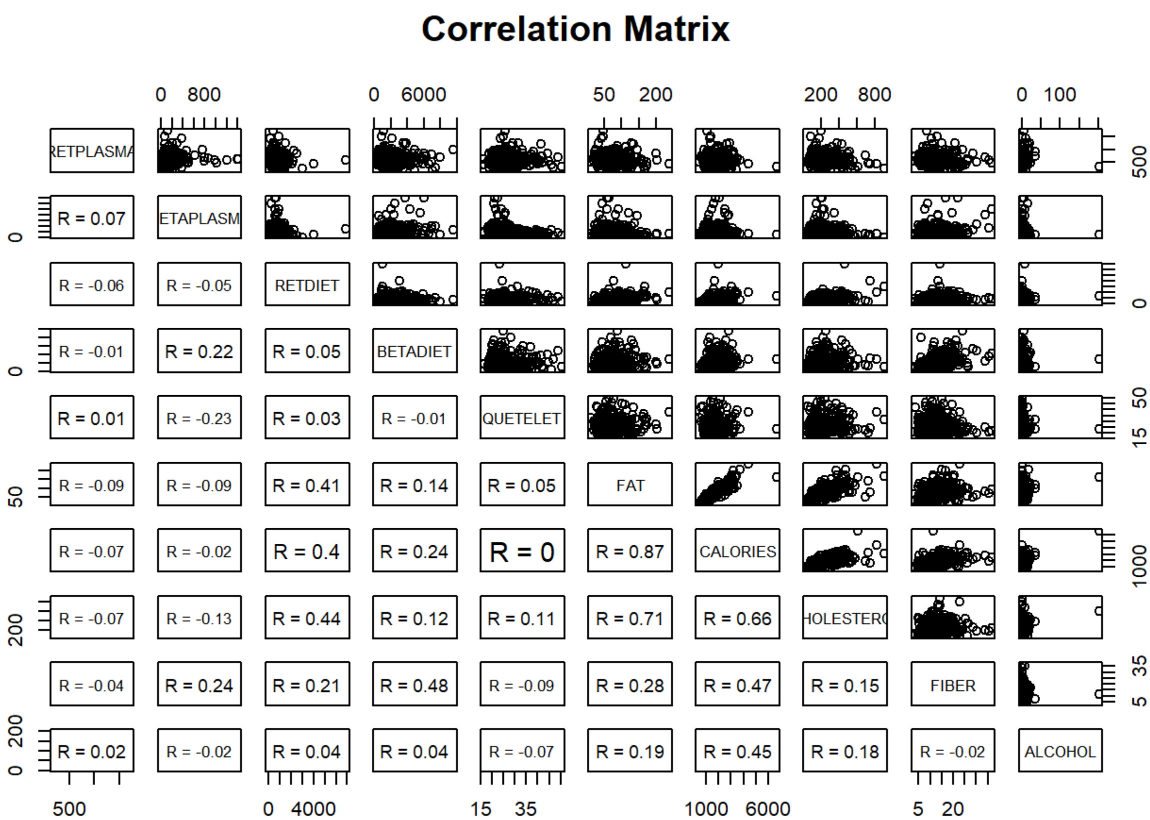


Figure 13:

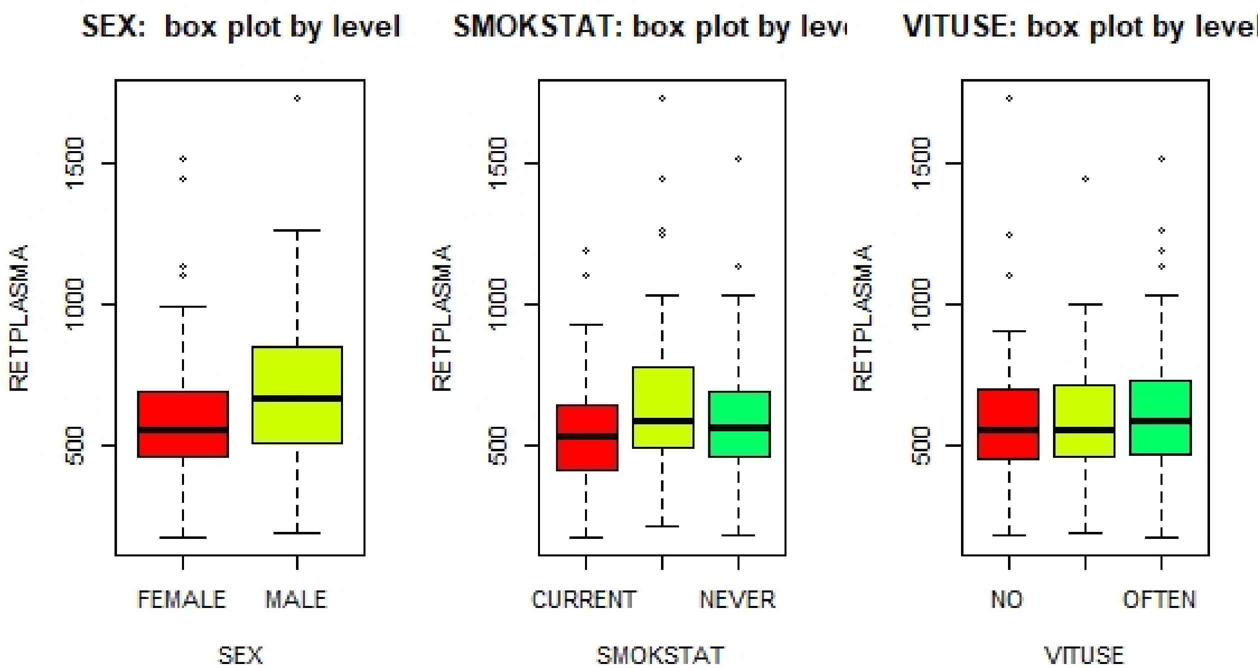


Figure 14:

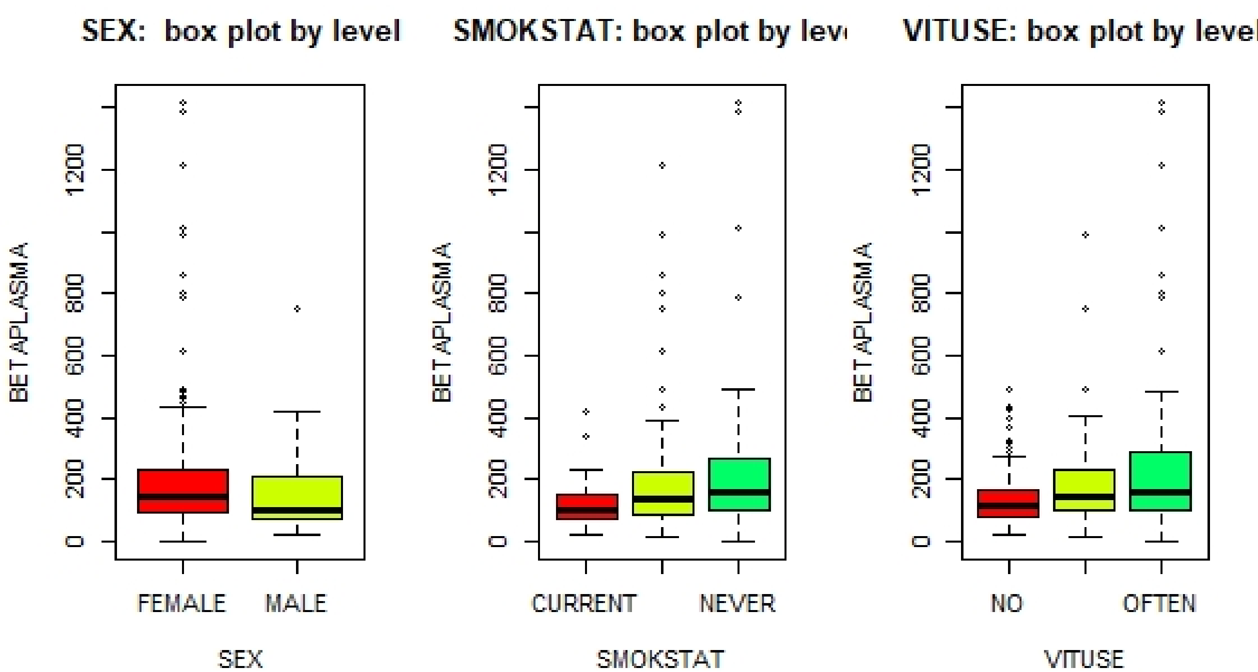


Figure 15:

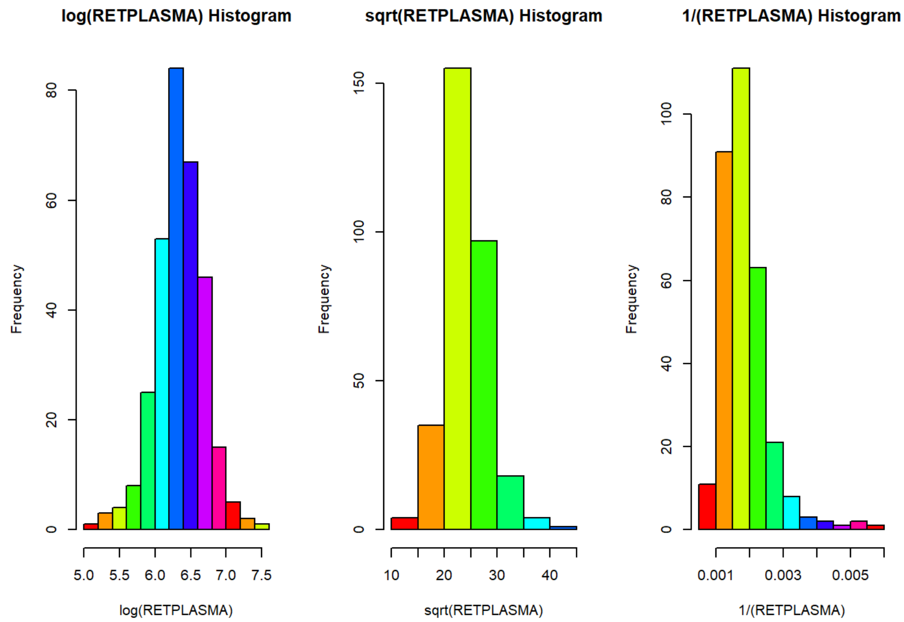
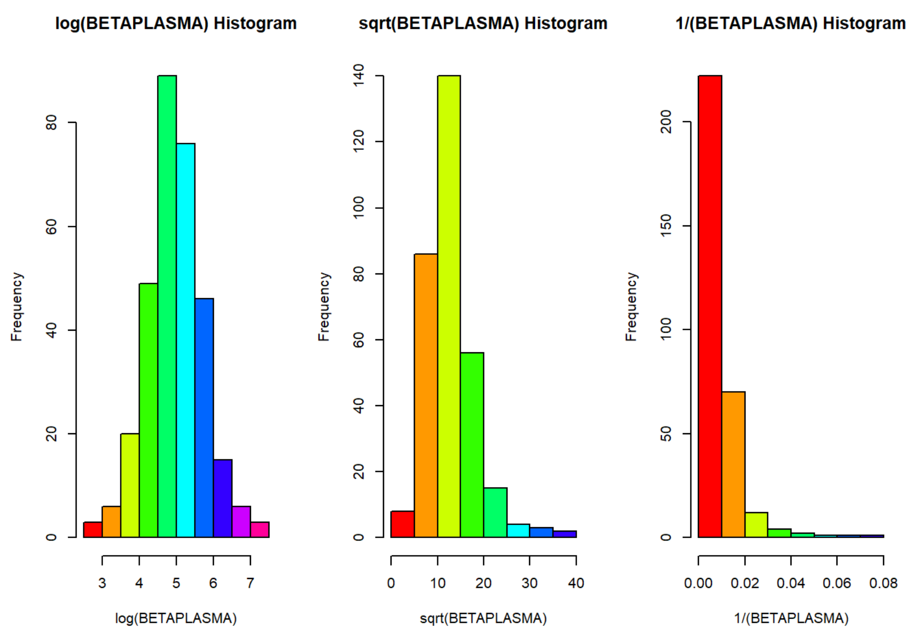


Figure 16:



Appendix 2.

Listing 1: Data Structure Exploration

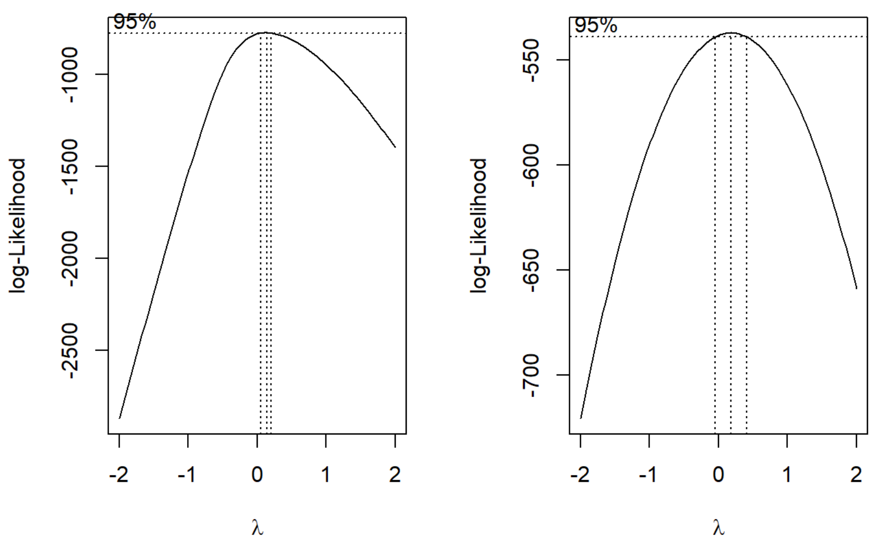
```
plas = read.table("Plasma.txt", header=TRUE)
head(plas)
-------------------------------------------------------------------

Output:
    AGE   SEX SMOKSTAT QUETELET VITUSE CALORIES FAT FIBER
  1  64 FEMALE FORMER 21.48380 OFTEN 1298.8 57.0  6.3
  2  76 FEMALE NEVER 23.87631  OFTEN 1032.5 50.1  15.8
  3  38 FEMALE FORMER 20.01080 NOT OFTEN 2372.3 83.6 19.1
  4  40 FEMALE FORMER 25.14062    NO  2449.5 97.5   26.5
  5  72 FEMALE NEVER 20.98504  OFTEN 1952.1 82.6  16.2
  6  40 FEMALE FORMER 27.52136    NO  1366.9 56.0    9.6
      CHOLESTEROL BETADIET RETDIET BETAPLASMA RETPLASMA
          ALCOHOL
  1        170.3      1945      890         200        915       0.0
  2         75.8      2653      451         124        727       0.0
  3        257.9      6321      660         328        721      14.1
  4        332.6      1061      864         153        615       0.5
  5        170.8      2863     1209          92        799       0.0
  6        154.6      1729     1439         148        654       1.3
-------------------------------------------------------------------

dim(plas)
-------------------------------------------------------------------

Output:
 [1]  315  14
-------------------------------------------------------------------

str(plas) #no missing values
-------------------------------------------------------------------

Output:
 'data.frame':     315 obs.  of   14  variables:
  $ AGE        :  int   64 76 38 40 72 40 65 58 35 55  ...
  $ SEX        :  Factor w/ 2 levels "FEMALE","MALE": 1 1 1 ...
  $ SMOKSTAT : Factor w/ 3 levels "CURRENT","FORMER",..: 2 3 2 3...
  $ QUETELET : num 21.5 23.9 20 25.1 21  ...
```

```
$ VITUSE    : Factor w/ 3 levels "NO","NOT OFTEN",..: 3 3 2 1 3...
$ CALORIES : num 1299 1032 2372 2450 1952 ...
$ FAT       : num 57 50.1 83.6 97.5 82.6 56 52...
$ FIBER     : num 6.3 15.8 19.1 26.5 16.2 9.6 ...
$ ALCOHOL : num 0 0 14.1 0.5 0 1.3 0 0 0.6 0...
$ CHOLESTEROL: num 170.3 75.8 257.9 332.6 170.8...
$ BETADIET : int 1945 2653 6321 1061 2863 1729...
$ RETDIET : int  890 451 660 864 1209 1439...
$ BETAPLASMA : int 200 124 328 153 92 148...
$ RETPLASMA : int 915 727 721 615 799 6542...
-----------------------------------------------------------------
```

Listing 2: Transformations

```
plas=plas[-c(62),] #62 gross outlier removal

library(MASS)

#Retplasma
plasrt=plas #from now on plasrt is the transformed for response retplasma.
#This is done to avoid confusion when fitting the model on betaplasma
fit.rt = lm(RETPLASMA~.,data=plasrt)
plasrt$RETPLASMA = log((plasrt$RETPLASMA))
boxcox(fit.rt)

#Betaplasma
plasbt=plas
plasbt[sapply(plasbt, is.numeric)]= plasbt[sapply(plasbt, is.numeric)]+1 #
    applied only to numeric values
fit.bt = lm(BETAPLASMA~., data=(plasbt))
boxcox(fit.bt)
```

Figure 17: Right: Beta-Carotene, Left: Retinol



Listing 3: Training and Validation Split

```
set.seed(1000)
n = nrow(plasrt)/2
ind = sample(1:(2*n), n, replace=FALSE)
train = plasrt[ind, ] #training set for retplas
valid = plasrt[-ind, ] #validation/test set for retplas

train_b = plasbt[ind, ] #train_b  set for betaplasma
valid_b = plasbt[-ind, ] #test set for betaplas
```

Listing 4: Selection Process Example

```
##model with only intercept
none_mod = lm(RETPLASMA~1, data=train)
##first order model with all predictors
full_mod = lm(RETPLASMA~., data=train)

#forward selection based on AIC:
------------------------------------------------------------

 Output:
stepAIC(none_mod, scope=list(upper=full_mod, lower = ~1),
        direction="forward", k=2, trace = FALSE)

   Call:
   lm(formula = RETPLASMA ~ ALCOHOL + AGE, data = train)

   Coefficients :
   (Intercept)      ALCOHOL       AGE
      6.087539     0.013684    0.004764
------------------------------------------------------------

#backward elimination based on AIC

stepAIC(full_mod, scope=list(upper=full_mod, lower = ~1),
        direction="backward", k=2, trace = FALSE)
------------------------------------------------------------

 Output:
   Call:
   lm(formula = RETPLASMA ~ AGE + ALCOHOL + BETAPLASMA, data
      = train)
```

```
Coefficients :
 (Intercept)           AGE      ALCOHOL BETAPLASMA
  5.974061e+00 6.253808e-03 1.690739e-02 8.726458e-05
```

#forward stepwise based on AIC

```
 Output:
stepAIC(none_mod, scope=list(upper=full_mod, lower = ~1),
        direction="both", k=2, trace = FALSE)

  Call:
  lm(formula = RETPLASMA ~ ALCOHOL + AGE, data = train)

  Coefficients :
 (Intercept)      ALCOHOL      AGE
    6.087539     0.013684     0.004764
```

```
#backward stepwise based on AIC
stepAIC(full_mod, scope=list(upper=full_mod, lower = ~1),
        direction="both", k=2, trace = FALSE)
```

```
  Output:
  Call:
  lm(formula = RETPLASMA ~ AGE + ALCOHOL + BETAPLASMA, data
     = train)

  Coefficients :
 (Intercept)           AGE      ALCOHOL BETAPLASMA
  5.974061e+00 6.253808e-03 1.690739e-02 8.726458e-05
```

```
#selection based on BIC: set option "k=log(n)"
stepAIC(none_mod, scope=list(upper=full_mod, lower = ~1),
        direction="forward", k=log(n), trace = FALSE)
```

Output:

```
Call:
lm(formula = RETPLASMA ~ ALCOHOL + AGE, data = train)

Coefficients:
(Intercept)      ALCOHOL          AGE
   6.087539     0.013684     0.004764
```

```
stepAIC(full_mod, scope=list(upper=full_mod, lower = ~1),
        direction="backward", k=log(n), trace = FALSE)
```

Output:

```
Call:
lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = train)

Coefficients:
(Intercept)          AGE      ALCOHOL
   6.087539     0.004764     0.013684
```

```
stepAIC(none_mod, scope=list(upper=full_mod, lower = ~1),
        direction="both", k=log(n), trace = FALSE)
```

Output:

```
Call:
lm(formula = RETPLASMA ~ ALCOHOL + AGE, data = train)

Coefficients:
(Intercept)      ALCOHOL          AGE
   6.087539     0.013684     0.004764
```

```
stepAIC(full_mod, scope=list(upper=full_mod, lower = ~1),
        direction="both", k=log(n), trace = FALSE)
```

Output:

```
Call:
lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = train)

Coefficients:
(Intercept)          AGE      ALCOHOL
   6.087539     0.004764     0.013684
```

```
##Validate 1st order selected

#AIC and BIC selections output many duplicates,
#these are the two produced unique models.
ar_1t = lm(RETPLASMA~AGE+ALCOHOL, data = train)
ar_1v = lm(RETPLASMA~AGE+ALCOHOL, data = valid)
br_1t = lm(RETPLASMA~AGE + ALCOHOL+ BETAPLASMA, data = train)
br_1v = lm(RETPLASMA~AGE + ALCOHOL+ BETAPLASMA, data = valid)



#compare the estimates and standard error between the two sets
ar_1_sum = cbind(coef(summary(ar_1t))[,1], coef(summary(ar_1v))[,1],
coef(summary(ar_1t))[,2], coef(summary(ar_1v))[,2])
colnames(ar_1_sum) = c("ar_1:Train Est","Valid Est","Train s.e.","Valid s.e.")

br_1_sum = cbind(coef(summary(br_1t))[,1], coef(summary(br_1v))[,1],
coef(summary(br_1t))[,2], coef(summary(br_1v))[,2])
colnames(br_1_sum) = c("br_1:Train Est","Valid Est","Train s.e.","Valid s.e.")
ar_1_sum
```

Output:

```
            ar_1:Train Est  Valid Est  Train s.e.  Valid s.e.
(Intercept)    6.087538956 6.00154440 0.091739775 0.097046514
AGE            0.004764294 0.00582498 0.001790546 0.001795436
ALCOHOL        0.013683894 0.01370914 0.004489622 0.006331594
```

```
br_1_sum
```

Output:

```
             br_1:Train Est    Valid Est    Train s.e.    Valid s.e.
(Intercept)  6.0643980377 5.9915995020 0.0934303061 0.0984456766
AGE          0.0046587012 0.0055447091 0.0017893044 0.0018505770
ALCOHOL      0.0135051663 0.0140287449 0.0044837803 0.0063629508
BETAPLASMA 0.0001376332 0.0001389059 0.0001102281 0.0002153363
```

```
# Both models have close estimates,
#it is clear that model br including betaplasma
#has greater standard error overall.
sse_ar_1t = sum(ar_1t$residuals^2)
sse_ar_1v = sum(ar_1v$residuals^2)
Radj_ar_1t = summary(ar_1t)$adj.r.squared
Radj_ar_1v = summary(ar_1v)$adj.r.squared
train_ar_1t_sum = c(sse_ar_1t,Radj_ar_1t)
valid_ar_1v_sum = c(sse_ar_1v,Radj_ar_1v)
criteria_ar_1 = rbind(train_ar_1t_sum,valid_ar_1v_sum)
colnames(criteria_ar_1) = c("ar_1SSE","R2_adj")
criteria_ar_1
```

Output:

```
                  ar_1SSE     R2_adj
train_ar_1t_sum 14.35971 0.08279178
valid_ar_1v_sum 18.18816 0.08108855
```

```
sse_br_1t = sum(br_1t$residuals^2)
sse_br_1v = sum(br_1v$residuals^2)
Radj_br_1t = summary(br_1t)$adj.r.squared
Radj_br_1v = summary(br_1v)$adj.r.squared
train_br_1t_sum = c(sse_br_1t,Radj_br_1t)
valid_br_1v_sum = c(sse_br_1v,Radj_br_1v)
criteria_br_1 = rbind(train_br_1t_sum,valid_br_1v_sum)
colnames(criteria_br_1) = c("br_1SSE","R2_adj")
criteria_br_1
```

Output:

```
                        br_1SSE      R2_adj
train_br_1t_sum 14.21486 0.08610942
valid_br_1v_sum 18.13883 0.07759123
```

```
#Model ar has closer R adjusted values
#while model br shows a larger difference between its R adjusted values.
#Model ar appears favorable here.

#Get MSPE_v from new data
###
newdata = valid[, -14] #remove predictor
n=dim(valid)[1]

RETPLAS.hat_ar = predict(ar_1t, newdata)

MSPE_ar_1 = mean((valid$RETPLAS - RETPLAS.hat_ar)^2)
MSPE_ar_1
```

Output:

```
[1]  0.1170858
```

```
sse_ar_1t/n
```

Output:

```
[1]  0.09146314
```

```
RETPLAS.hat_br = predict(br_1t, newdata)

MSPE_br_1 = mean((valid$RETPLAS - RETPLAS.hat_br)^2)
MSPE_br_1
```

Output:

```
[1]  0.1163897
```

---

sse_br_1t/n

Output:

```
[1] 0.09054054
```

#it is difficult to determine which model prevails in this case.
#Both do not have severe overfitting.
#An anova fit ultimatly tells us that the betaplasma in model br is insignificant,
#therefore, model ar is selected as the best model here.

anova(ar_1**t**)

Output:

```
Analysis of Variance Table

Response: RETPLASMA
          Df  Sum Sq Mean Sq F value   Pr(>F)
AGE        1  0.6333 0.63329  6.7917 0.010057 *
ALCOHOL    1  0.8662 0.86621  9.2897 0.002713 **
Residuals 154 14.3597 0.09324
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

---

anova(br_1**t**)

Output:

```
Analysis of Variance Table

Response: RETPLASMA
          Df  Sum Sq Mean Sq F value   Pr(>F)
AGE        1  0.6333 0.63329  6.8163 0.009931 **
ALCOHOL    1  0.8662 0.86621  9.3234 0.002669 **
```

```
BETAPLASMA 1 0.1448 0.14485 1.5591 0.213710
Residuals   153 14.2149 0.09291
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
----------------------------------------------------------------
```

Listing 5: Outlier Removal Example

```
final_ret_1st=lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = plasrt
    )



summary(final_ret_1st)
-------------------------------------------------------------------

Output:

  Call:
  lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = plasrt)

  Residuals:
       Min       1Q   Median       3Q      Max
  -1.15651 -0.18320 -0.00469 0.19695 1.03798

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) 6.047014   0.066309  91.194  < 2e-16 ***
  AGE         0.005240   0.001257   4.170 3.95e-05 ***
  ALCOHOL     0.014092   0.003700   3.808 0.000169 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.324 on 311 degrees of freedom
  Multiple R-squared: 0.09276,     Adjusted R-squared: 0.08692
  F-statistic:  15.9 on 2 and 311 DF, p-value: 2.666e-07
-------------------------------------------------------------------


anova(final_ret_1st)
-------------------------------------------------------------------

Output:
  Analysis of Variance Table

  Response: RETPLASMA
             Df Sum Sq Mean Sq F value  Pr(>F)
  AGE         1  1.816 1.81568  17.295 4.144e-05 ***
```

```
ALCOHOL 1   1.522 1.52247   14.502 0.0001687 ***
Residuals 311 32.649 0.10498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#MSE
anova(final_ret_1st)['Residuals',3]
```

Output:

```
[1] 0.1049819
```

```
e=final_ret_1st$residuals ##ordinary residuals
h=influence(final_ret_1st)$hat
de=e/(1-h) ##deleted residuals
summary(h)
```

Output:

```
     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003201 0.004813 0.006621 0.009554 0.010040 0.140254
```

```
n=dim(plasrt)[1]
p=length(final_ret_1st$coefficients)

stu.res.del = studres(final_ret_1st)
head(sort(abs(stu.res.del), decreasing=TRUE))
```

Output:

```
      81       36       20      293      276      311
3.653123 3.430452 3.260331 3.118909 2.988398 2.972008
```

```
qt(1-.1/(2*n), n-p-1) #Bonferroni's Threshold (alpha=0.1, n=sample size, p=3)
    = 3.64381
```

Output:

```
[1] 3.640728
```

#point s again identified 81

```
h = influence(final_ret_1st)$hat #leverage

sort(h[which(h>2*p/n)], decreasing = TRUE)
```

Output:

```
       308        145        140         80         16         78         23
0.14025426 0.13975546 0.05228896 0.04713755 0.04544958 0.04273419 0.03777876
        95        283        305        256        296        208        251
0.03467477 0.03128063 0.02826298 0.02656060 0.02431657 0.02323525 0.02295591
        17          3        273        111        135         71         50
0.02247162 0.02247067 0.02098173 0.02035542 0.02035542 0.02016289 0.02000762
```

```
res  = final_ret_1st$residuals
mse = anova(final_ret_1st)["Residuals", 3]
cook.d = res^2*h/(p*mse*(1-h)^2)

sort(cook.d[which(cook.d>4/(n-p))], decreasing = TRUE)
```

Output:

```
        75        140         16         81        296        201        276
0.04544497 0.03460413 0.03437471 0.03200903 0.03027704 0.02665687 0.02490188
       145         50         78         80         18         36         20
0.02395083 0.01966931 0.01936733 0.01831579 0.01779668 0.01774978 0.01598592
        97        293        235
0.01382902 0.01369184 0.01288771
```

```
#81 is both influential and an outlier, it is subject to removal
best_r_out=lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = plasrt,
    subset=setdiff(rownames(plasrt), "81"))
```

---

summary(final_ret_1st)

Output:

```
Call:
lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = plasrt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15651 -0.18320 -0.00469 0.19695 1.03798

Coefficients :
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.047014   0.066309  91.194  < 2e-16 ***
AGE         0.005240   0.001257   4.170 3.95e-05 ***
ALCOHOL     0.014092   0.003700   3.808 0.000169 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01  '*' 0.05  '.' 0.1  ' ' 1

Residual standard error: 0.324 on 311 degrees of freedom
Multiple R-squared: 0.09276,     Adjusted R-squared: 0.08692
F-statistic:  15.9 on 2 and 311 DF, p-value: 2.666e-07
```

---

summary(best_r_out)

Output:

```
Call:
lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = plasrt, subset =
    setdiff(rownames(plasrt),
    "81"))

Residuals:
     Min       1Q   Median       3Q      Max
-1.09337 -0.19026 -0.00976 0.19706 1.03341

Coefficients :
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.038717   0.065071  92.802  < 2e-16 ***
AGE         0.005501   0.001234   4.456 1.17e-05 ***
ALCOHOL     0.013693   0.003631   3.771 0.000195 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3178 on 310 degrees of freedom
Multiple R-squared: 0.09894,     Adjusted R-squared: 0.09312
F-statistic: 17.02 on 2 and 310 DF, p-value: 9.706e-08
----------------------------------------------------------------------
```

Listing 6: Retinol: Best 1st Order Model

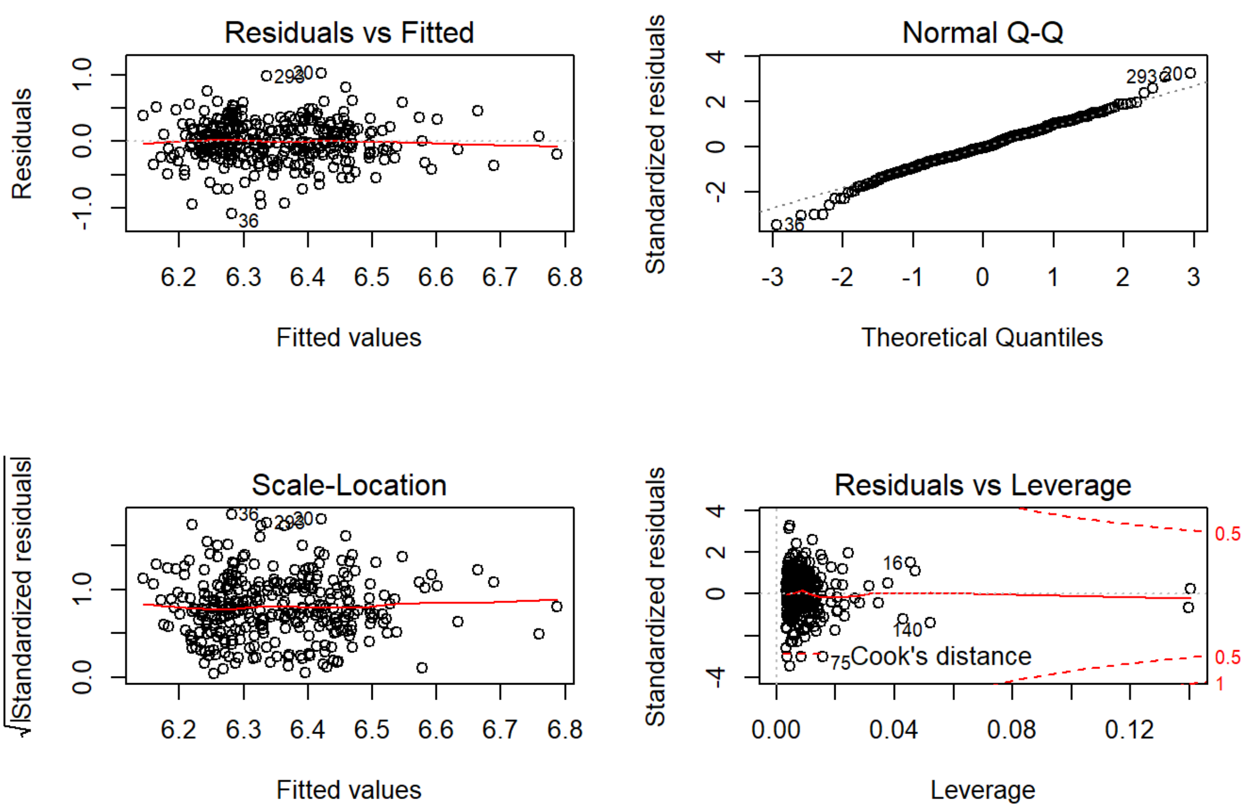
```
summary(best_r_out)
--------------------------------------------------------------

Output:
  Call:
  lm(formula = RETPLASMA ~ AGE + ALCOHOL, data = plasrt, subset =
     setdiff(rownames(plasrt),
     "81"))

  Residuals:
       Min        1Q    Median        3Q       Max
  -1.09337 -0.19026 -0.00976 0.19706 1.03341

  Coefficients :
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) 6.038717   0.065071  92.802  < 2e-16 ***
  AGE         0.005501   0.001234   4.456 1.17e-05 ***
  ALCOHOL     0.013693   0.003631   3.771 0.000195 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.3178 on 310 degrees of freedom
  Multiple R-squared: 0.09894,     Adjusted R-squared: 0.09312
  F-statistic: 17.02 on 2 and 310 DF, p-value: 9.706e-08
--------------------------------------------------------------

anova(best_r_out)
--------------------------------------------------------------

Output:
  Analysis of Variance Table

  Response: RETPLASMA
             Df  Sum Sq Mean Sq F value  Pr(>F)
  AGE         1  2.0009 2.00089  19.816 1.191e-05 ***
  ALCOHOL 1   1.4361 1.43608  14.222 0.0001945 ***
  Residuals 310 31.3018 0.10097
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------------------------
```

```
#MSE
anova(best_r_out)['Residuals',3]
```

Output:

```
[1] 0.1009737
```

Figure 18: Retinol:Best First Order Diagnostic Graphs



Listing 7: Retinol: Best 1st Order Interaction Model

```
summary(best_r_int)
----------------------------------------------------------------

Output:

  Call:
  lm(formula = RETPLASMA ~ AGE + CALORIES + FAT + FIBER
   +ALCOHOL + FAT:FIBER, data = plasrt)

  Residuals:
       Min        1Q    Median        3Q       Max
  -1.17059 -0.18400 -0.00333 0.20303 1.04354

  Coefficients :
                  Estimate Std. Error t value Pr(>|t|)
  (Intercept)   5.906e+00 1.462e-01 40.384  < 2e-16 ***
  AGE           5.600e-03 1.325e-03  4.226 3.14e-05 ***
  CALORIES      1.580e-04 9.415e-05  1.678  0.09441 .
  FAT          -4.602e-04 2.191e-03 -0.210  0.83377
  FIBER         6.400e-03 9.700e-03  0.660  0.50989
  ALCOHOL       1.181e-02 3.988e-03  2.962  0.00329 **
  FAT:FIBER    -1.922e-04 1.091e-04 -1.763 0.07898 .
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.3209 on 307 degrees of freedom
  Multiple R-squared: 0.1217, Adjusted R-squared: 0.1046
  F-statistic: 7.091 on 6 and 307 DF, p-value: 4.349e-07
----------------------------------------------------------------


anova(best_r_int)
----------------------------------------------------------------

Output:
   Analysis of Variance Table

   Response: RETPLASMA
              Df  Sum Sq Mean Sq F value  Pr(>F)
   AGE         1  1.8157 1.81568 17.6358 3.509e-05 ***
```

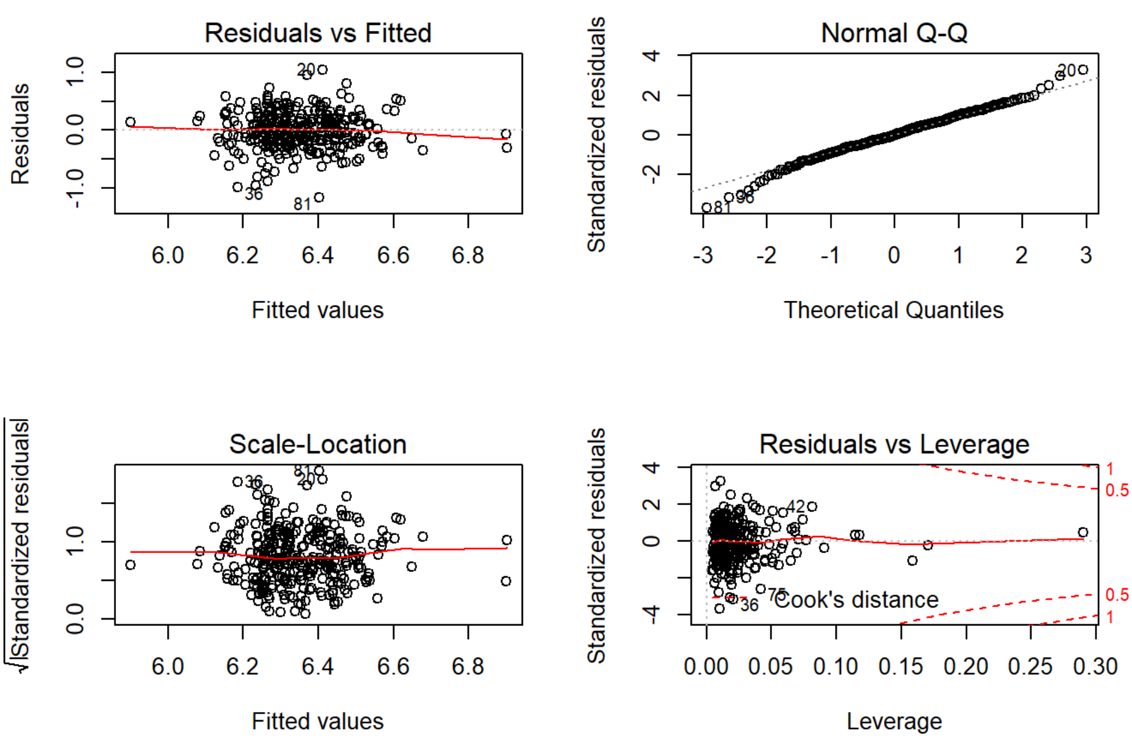
```
CALORIES 1  0.0161 0.01606  0.1560  0.693187
FAT      1  0.4811 0.48108  4.6727  0.031418 *
FIBER    1  0.7641 0.76410  7.4218  0.006813 **
ALCOHOL  1  0.9837 0.98373  9.5550  0.002177 **
FAT:FIBER 1 0.3198 0.31982  3.1065  0.078977 .
Residuals 307 31.6070 0.10295
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##MSE
anova(best_r_int)['Residuals',3]
```

Output:

```
[1] 0.10295
```

Figure 19: Retinol:Best First Order Interaction Diagnostic Graphs



Listing 8: Beta-Carotene: Best 1st Order Model

```
summary(best_beta_1)
---------------------------------------------------------------------------

Output:
Call
 lm(formula = BETAPLASMA ~ AGE + factor(SEX) + QUETELET factor(
     VITUSE) +
   FIBER + CHOLESTEROL, data = plasbt, subset setdiff(rownames(plasbt),
     c("257")))

 Residuals:
      Min        1Q    Median       3Q      Max
 -1.95433 -0.35372 -0.04004 0.42215 1.91918

  Coefficients :
                              Estimate Std. Error t value Pr(>|t|)
 (Intercept)                 4.9358391  0.2589097  19.064  < 2e-1***
 AGE                         0.0077976  0.0027598   2.825 0.00503**
 factor(SEX)MALE            -0.2676666 0.1225256  -2.185 0.02968*
 QUETELET                   -0.0288925 0.0063128  -4.577 6.88e-0***
 factor(VITUSE)NOT OFTEN 0.2769421 0.0986257 2.808 0.00530**
 factor(VITUSE)OFTEN 0.3189478  0.0887626   3.593 0.00038***
 FIBER                       0.0301743  0.0071957   4.193 3.61e-0***
 CHOLESTEROL                -0.0006532 0.0003225  -2.026 0.04367*
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 0.6618 on 305 degrees of freedom
 Multiple R-squared: 0.2211, Adjusted R-squared: 0.2032
 F-statistic: 12.37 on 7 and 305 DF, p-value: 6.106e-14
---------------------------------------------------------------------------


anova(best_beta_1)
---------------------------------------------------------------------------

Output:
Analysis of Variance Table

Response: BETAPLASMA
                   Df  Sum Sq Mean Sq F value  Pr(>F)
```

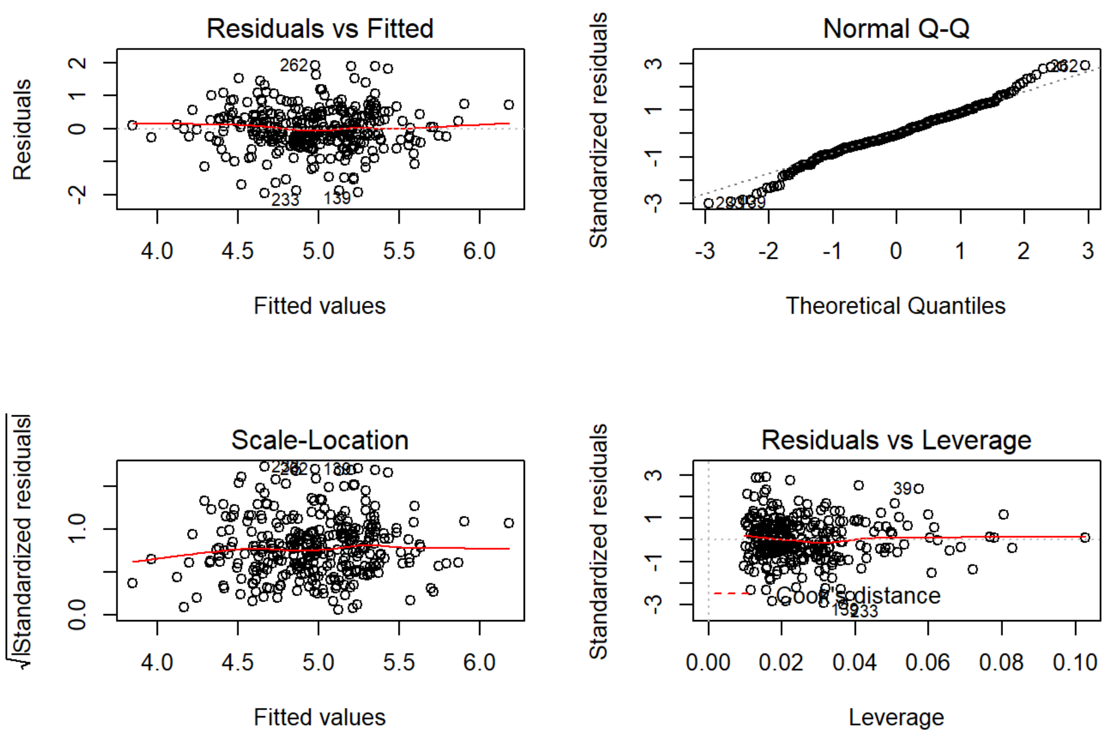
```
AGE               1     3.278   3.2782   7.4853 0.0065851 **
factor(SEX)       1     5.371   5.3711  12.2639 0.0005308 ***
QUETELET          1    13.073  13.0731  29.8504 9.704e-08 ***
factor(VITUSE)    2     7.737   3.8687   8.8335 0.0001865 ***
FIBER             1     6.662   6.6624  15.2124 0.0001182 ***
CHOLESTEROL       1     1.797   1.7970   4.1031 0.0436760 *
Residuals       305   133.576   0.4380
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1
  1
```

```
anova(best_beta_1)['Residuals',3]
```

Output:

```
[1] 0.4380
```

Figure 20: Beta Carotene: Best First Order Diagnostic Graphs



Listing 9: Beta-Carotene: Best 1st Order Interaction Model

```
summary(best_beta_int)
-----------------------------------------------------------------

Output:

 Call:
 lm(formula = BETAPLASMA ~ QUETELET + CHOLESTEROL + FIBER +
     RETPLASMA +
     VITUSE + AGE + SEX + BETADIET + SMOKSTAT + CHOLESTEROL:
         RETPLASMA +
     SEX:BETADIET + VITUSE:BETADIET + BETADIET:SMOKSTAT, data
         = plasbt)

 Residuals:
      Min       1Q   Median       3Q      Max
 -2.62524 -0.35843 -0.01008 0.40161 1.91458

 Coefficients :
                          Estimate Std. Error t value Pr(>|t|)
 (Intercept)            5.997e+00 3.625e-01 16.544  < 2e-16 ***
 QUETELET                  -3.239e-02 6.422e-03 -5.043 8.01e-07 ***
 CHOLESTEROL               -4.364e-03 8.240e-04 -5.296 2.31e-07 ***
 FIBER                      2.332e-02 8.285e-03  2.814  0.00521 **
 RETPLASMA                 -8.345e-04 3.876e-04 -2.153  0.03213 *
 VITUSENOT OFTEN            2.168e-01 1.735e-01  1.249  0.21250
 VITUSEOFTEN               -1.298e-01 1.623e-01 -0.800  0.42450
 AGE                        5.583e-03 2.828e-03  1.974  0.04933 *
 SEXMALE                    7.669e-02 2.572e-01  0.298  0.76576
 BETADIET                  -2.517e-04 9.420e-05 -2.672  0.00795 **
 SMOKSTATFORMER    -2.310e-01 2.083e-01 -1.109  0.26833
 SMOKSTATNEVER      -1.461e-01 2.032e-01 -0.719  0.47276
 CHOLESTEROL:RETPLASMA 5.659e-06 1.402e-06 4.035 6.94e-05 ***
 SEXMALE:BETADIET  -1.302e-04 1.028e-04 -1.267  0.20605
 VITUSENOT OFTEN:BETADIET 2.410e-05 6.989e-05 0.345 0.73042
 VITUSEOFTEN:BETADIET 1.814e-04 6.304e-05 2.878 0.00429 **
 BETADIET:SMOKSTATFORMER 2.571e-04 9.211e-05 2.792 0.00558 **
 BETADIET:SMOKSTATNEVER 2.715e-04 9.067e-05 2.995 0.00298 **
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 0.6588 on 296 degrees of freedom
```

```
Multiple R-squared: 0.345,   Adjusted R-squared: 0.3074
F-statistic:  9.172 on 17 and 296 DF, p-value: < 2.2e-16
```

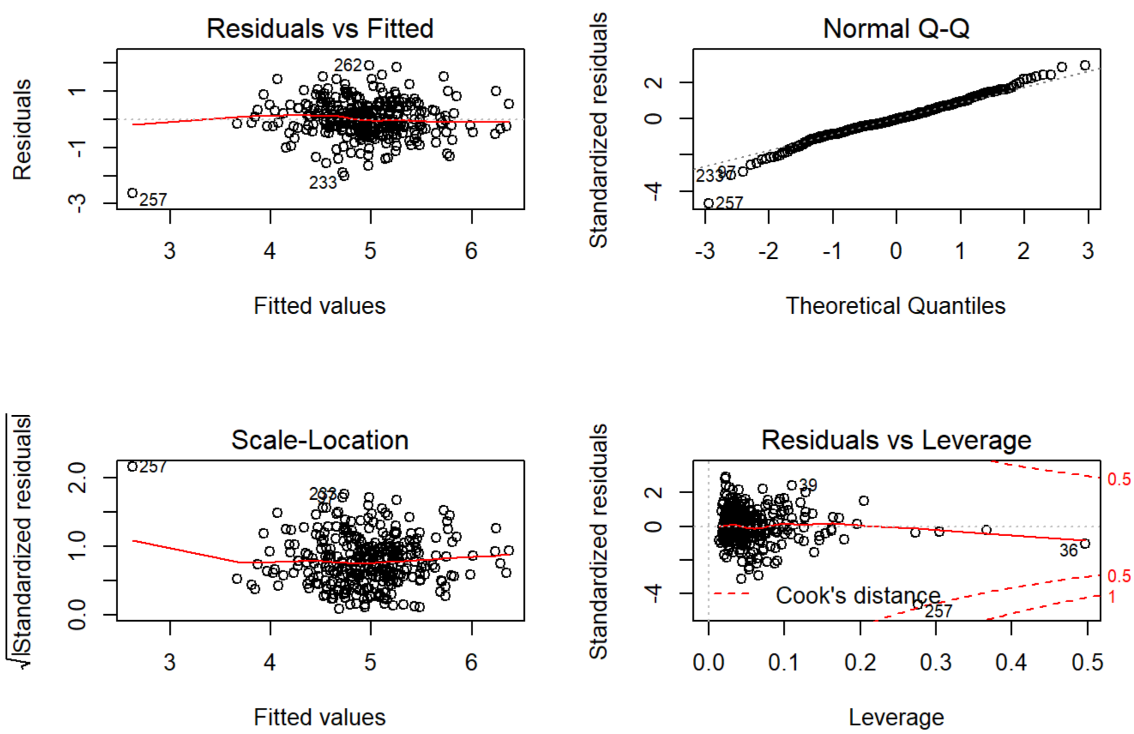
```
anova(best_beta_int)
```

```
Output:
Analysis of Variance Table

Response: BETAPLASMA
                       Df  Sum Sq Mean Sq F value  Pr(>F)
QUETELET                1  15.306 15.3059 35.2699 8.037e-09 ***
CHOLESTEROL             1   8.633  8.6326 19.8925 1.165e-05 ***
FIBER                   1  10.719 10.7191 24.7003 1.135e-06 ***
RETPLASMA               1   4.069  4.0687  9.3757 0.0024008 **
VITUSE                  2   5.824  2.9118  6.7096 0.0014131 **
AGE                     1   1.312  1.3116  3.0223 0.0831662 .
SEX                     1   1.821  1.8214  4.1972 0.0413722 *
BETADIET                1   2.216  2.2160  5.1065 0.0245633 *
SMOKSTAT                2   1.982  0.9911  2.2838 0.1036934
CHOLESTEROL:RETPLASMA 1 6.355 6.3547 14.6433 0.0001585 ***
SEX:BETADIET            1   1.351  1.3511  3.1135 0.0786786 .
VITUSE:BETADIET         2   4.023  2.0113  4.6347 0.0104241 *
BETADIET:SMOKSTAT 2     4.055  2.0274  4.6718 0.0100563 *
Residuals             296 128.454  0.4340
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
    1
```

```
anova(best_beta_int)['Residuals',3]
```

```
Output:
   [1]  0.4340
```

Figure 21: Beta Carotene: Best First Order Diagnostic Interaction Graphs



References

[1] R. A. B. S. G. P. H. S. F. A. Faure H, Preziosi P. Factors influencing blood concentration of retinol, alpha-tocopherol, vitamin c, and beta-carotene in the french participants of the su.vi.max trial. *Eur J Clin Nutr*, 60:706–717, 2006.

[2] D. S. Goodman. Vitamin a and retinoids in health and disease. *N Engl J Med*, 310:1023–1031, 1984.

[3] J. M. O. M. A. A. A. Goyal. Vitamin a toxicity. *N Engl J Med*, 310:1023–1031, 1984.

[4] A. J. Bioconversion of dietary provitamin a carotenoids to vitamin a in humans. *The American Journal of Clinical Nutrition*, 91:1468S–1473S, 2010.

[5] I. of Medicine (US) Panel on Dietary Antioxidants and R. Compounds. Dietary reference intakes for vitamin c, vitamin e, selenium, and carotenoids. *National Academies Press*, 8, 2000.

[6] C. Ryu. cran.r project data transformation.

[7] B. W and M. A. From vitamin a to retinoids in experimental and clinical oncology: Achievements, failures and outlook. *Ann N Y Acad Sci*, 351:9–23, 1981.