# Investigation Plasma Concentration of Beta-Carotene


| | |
|---|---|
| Yi Zhou (Zoey) | Zoezhou@ucdavis.edu |
| Yanlin Li (Alice) | Lylli@ucdavis.edu |
| Jing Li | Jinli@ucdavis.edu |

## Abstract

The relationship between plasma concentration of Beta-Carotene and 11 personal variables were studied in 315 surgical, but non-cancerous patients. The dataset is from CMU Statistics Lab. We are interested in which factors in personal characteristics or dietary intake will influence the plasma concentration of Beta-carotene. Our final model suggest that plasma concentration of Beta-carotene is found to be positively related to daily beta-carotene intake in mcg, vitamin usage frequency, smoke status and gram of fiber consumed per day; it is also negative related calories consumed per day and Quetelet index. However, since the data was collected from surgical patients instead of general public, a better understanding of this relationship need further study.


## Introduction

Beta-carotene and retinol are among the most widely studied compounds in people's daily intake. Some epidemiologic studies have shown that beta-carotene has an inverse association with the risk of cancer. This means beta-carotene may can reduce the risk of some types of cancer, such as lung cancer. [1,2]

Some studies also find that sex, alcohol consumption and smoking habit have some influence on plasma concentration of beta-carotene. [3] Scientists want to figure out what other factors may also influence plasma concertation in beta-carotene.

Rather than predicting the plasma concentration of beta-carotene, this report will mainly focus on investigating the relationship between personal characteristics, dietary intake and the plasma concentration of beta-plasma for the given dataset.

## Data Processing

The dataset does not have "NA" or missing value. It consists 3 qualitative variables and 11 quantitative variables including 2 response variables. Looking at the scatter plot matrix (Fig. 1) and statistics summary (Fig. 2) of the quantitative variables, two outliers were spotted: maximum value in alcohol consumption per week (203 drinks), the minimum value in plasma concentrations of beta-carotene (0ng/ml). Since the top 10% man in American drinkers consume, on average, 74 alcoholic drinks per week, it is very unlikely for this patient to have 203 drinks per week. Moreover, given the dietary beta-carotene consumed for the person is 1,028 mcg per day, it is unlikely for the patient to have exactly 0 plasma concentrations of beta-carotene (0ng/ml); therefore, these two values are removed from the dataset.

Histograms (Fig. 3) are performed on the 10 quantitative variables including our chosen response variable "BETAPLASMA." All 10 histograms appear more or less right skewed. The right skew-ness of the response variable indicates the need of transformation. The histogram of log transform gives the most normal-like distribution. The box-cox plot (Fig. 4) also confirms this result, as log-likelihood is maximized when $\lambda$ is around zero.

Pie charts are drawn on the qualitative variables: Sex, Vitamin Use and Smoke Level (Fig. 5). Based on chart of Sex, there are much more female than male in the dataset. Side-by-side box plots (Fig. 6) are then drawn to show how response variable, plasma concentrations of beta-carotene, is distributed in each categories variable. The distribution of plasma concentrations of beta-carotene is not symmetric in each 3 qualitative variables. Moreover, mean of plasma concentration of beta-carotene differs by gender is confirmed by One-Way Anova test, as the CI interval for the difference at 95% level is [0.048, 0.538]. Anova test were

also conducted on Vitamin Use, and Smoke Level, but not all pairs are different in mean (Fig.10).

Dataset is then standardized (only X variables) and randomly divided into training set and testing set on 3:1 ratio. The distributions of these two sets are approximately the same. This is confirmed by preforming side-by-side box plots on each variable (Fig. 7).

There is no obvious nonlinearity in the pairwise scatter plots and correlation matrix (Fig. 8) in training dataset, besides a slight quadratic relation in Quetelet (weight/height^2) and response variables. However, positive linear relationship is obvious between fat consumption and calories consumption. The correlation value between this pair reaches almost 0.90. VIF (variance inflation factor, Fig. 9) value of fat (15.80) and calories (11.64) confirm this finding, indicating that multicollinearity is high between these two variables.

## Model selection

In this part, we used forward stepwise procedure based on both AIC and BIC criterion and got two first-order model and one non-additive model.

**1. First-order model selection:**

- Model 1: first-order full model

We first fitted a model with all first-order effects as our model 1. As shown in table 1-1, model 1 has 15 regression coefficients, with $R^2 = 0.238$ and ${R_a}^2 = 0.1896$. Residual vs. fitted value plot shows slightly nonlinearity and Q-Q plot is a little bit heavy tailed (figure 2-2). But it is acceptable. Box-cox plot (figure 2-3) shows no further transformation is needed, since $\lambda$ is near 1.

From the ANOVA table, we see the p value of ALCOHOL (Pr(>F) = 0.993494), CHOLESTEROL (Pr(>F) = 0.758135) and RETDIET (Pr(>F) = 0.512612) are very large which indicates that these variables are probably insignificant. Further analysis is conducted.

- Best subsets selection:

The best subsets selection is used here to find out the best model based on different criterions: $SSE,\ R^2,\ {R_a}^2,\ C_p,\ AIC,\ BIC$. These models are used for comparison.

The variables contained in each model and its regression coefficients numbers are shown below. (see p-criterion figures in figure 2-11)

$SSE,\ R^2$: full model (p = 14)

$C_p,\ AIC$: AGE, SEX, SMOKESTAT, QUETELET, VITUSE, BETADIET (p = 8)

$BIC$: QUETELET, VITUSE, CALORIES, FIBER (p = 4)

${R_a}^2$: AGE, SEX, SMOKE, QUETELET, VITUSE, BETADIET, CALORIES, FIBER, BETDIET (p = 10)

The smallest $C_p$ is 6.847292 and p = 8, which indicates the in model bias of the model 1 is small.

This result is coincident with what we found from model 1. None of those insignificant coefficients (ALCOHOL, CHOLESTEROL and RETDIET) are selected. Also, the two highly correlated variables: CALORIES and FAT are also not included.

- Forward stepwise procedure using criterion AIC

We started from the null-model and ended up with Model 1 (first-order full model). Then we got our Model 2:

$$\log(Y^*) = {\beta_0}^* + {\beta_1}^* {X_1}^* + {\beta_2}^* {X_2}^* + \cdots + {\beta_8}^* {X_8}^*$$

where $\beta_i^*$, $i = 0, \dots, 8$ is regression coefficients. $X_i^*, i = 1, \dots, 8$ represents standardized QUETELET, VITUSENOT OFTEN, VITUSEOFTEN, SMOKSTATFORMER, SMOKSTATNEVER, BETADIET, CALORIES, FIBER, respectively.

The p-value of each coefficient in model 2 is relatively small (figure 2-4). They are all significant at level 0.1. In the residual vs. fitted value plot, slightly nonlinearity is observed. Q-Q plot shows slightly heavy tailed (figure 2-5). In sum, the model seems reasonable and adequate.

$AIC = -180.19$ in model 2 is slightly larger than the best model based on AIC (-180.39) criterion. This indicate the model 2 is a suboptimal model. Compared with model 1 (table 1), MSE of model 2 is smaller and ${R_a}^2$ is larger, also, less x variables in model 2 lead to a smaller in model variance

$$Var_{in}(M2) \approx \sigma^2 p \approx MSE * 9 = 4.03$$

Therefore, model 2 is better than model 1.

- Forward stepwise procedure using criterion BIC

The same with the previous step, we use null-model as started model and model 1 as ended model. We got model 3:

$$\log(Y^*) = \beta_0{}^* + \beta_1{}^* X_1{}^* + \beta_2{}^* X_2{}^* + \beta_3{}^* X_3{}^* + \beta_4{}^* X_4{}^*$$

where $\beta_i^*$, $i = 0, \dots, 4$ is regression coefficients. $X_i^*, i = 1, \dots, 4$ represents standardized QUETELET, BETADIET, VITUSENOT OFTEN, VITUSEOFTEN, respectively.

Based on the p-value from ANOVA table (figure 2-7), all coefficients are significant at level 0.05 which is better than model 2. Its residual vs. fitted value plot and Q-Q plot are pretty much the same with model 2.

Only 3 coefficients are chosen in model 3. They're the subset of model 2. This is reasonable since BIC puts more penalty on model complexity and thus, tends to choose smaller models than AIC.

Compared with model 2 (table 1), model 3 has much less variables which causes a larger MSE and lower ${R_a}^2$, the in-model bias $Var_{in}(M3) \approx \sigma^2 p \approx MSE * 3 = 1.41$ will increase as well. The in model variance 1.41 is much smaller than model 2 (4.03). Due to the trade-off between the bias-variance, we cannot simply say which model is better. Further model validation is needed for both model 2 and model 3.

**2. Selection of first order and 2-way interaction effects:**

We first fitted a full model with all first-order and 2-way interaction effects (model 4). It has 104 coefficients in total, which is larger compared to our dataset. Model 4 has almost the same MSE with model 1.

Its residual vs. fitted value plot looks better: no nonlinearity is observed. Q-Q plot also shows slight heavy tailed.

- Forward stepwise procedure using criterion AIC

Starting from null-model and ending up with model 4, we got model 5 using forward stepwise procedure based on AIC criterion.

Model 5:

$$\log(Y^*) = \beta_0{}^* + \beta_1{}^* X_1{}^* + \cdots + \beta_7{}^* X_7{}^* + \beta_8{}^* X_2{}^* * X_4{}^* + \beta_9{}^* X_3{}^* * X_4{}^* + \beta_{10}{}^* X_2{}^* * X_5{}^* + \beta_{11}{}^* X_3{}^* * X_5{}^* + \beta_{12}{}^* X_4{}^* * X_7{}^* + \beta_{13}{}^* X_5{}^* * X_7{}^*$$

where $\beta_i^*$, $i = 0, \dots, 7$ is regression coefficients. $X_i^*, i = 1, \dots, 7$ represents standardized QUETELET, VITUSENOT OFTEN, VITUSEOFTEN, SMOKSTATFORMER, SMOKSTATNEVER, BETADIET, CALORIES, respectively.

p-value of the ANOVA table shows that all the coefficients are significant at level 0.05. The residual vs. fitted value is not as good as its of model 4 but is better than first order model. QQ plot is also a little bit heavy tailed.

Two interaction terms included in model 5. Total has 14 variables. MSE = 0.4221. This result is much better than model 4 and model 1 which has 104 and 15 coefficients, respectively. Its AIC = -189.14 is slightly smaller than model 2 (AIC = -180.19)

- Forward stepwise procedure using criterion BIC

Using the same start and end models as in previous step we got the exactly same model as model 3.

We also used forward selection procedure, the results are the same with forward stepwise. This is coincident with our multicollinearity analysis (multicollinearity only exists in the two features, but is not prevalent in our dataset), since high multicollinearity will influence the performance of forward selection. Besides forward procedure, we also used backward selection to do the model fitting. However, in this particular dataset, the result in backward selection is not as good as the results in forward stepwise, so we use forward stepwise procedure in our model selection process to keep the model more consistently.

For now, our preliminary models are determined: model 2, model 3 and model 5. None of them contains both two high interaction terms: CALORIES and FAT. The summary of these three models are shown in table 2-5

## Model validation

In this part, internal validation and external validation are conducted to check the validation of our preliminary models: model 2, model 3 and model 5. Model 2 seems better.

### 1. Internal validation

For model internal validation of model 2, model 3 and model 5, $C_p$ and $Press_p$ are calculated. To calculate them, we need an unbiased estimator of the error variance $\sigma^2$. The largest model we considered is model 4. However, the number of regression coefficients of model 4 is too large (104) relative to the sample size. This makes its parameter estimation unreliable due to large sampling variability. Therefore, we use a smaller model consisting of all predictors identified by Model 2, as well all the 2-way interaction terms among these predictors. Denote this model by Model 6. Note that, all the variables in model 3 and model 5 are included in model 6.

The number of regression coefficients is 35 for model 6. MSE = 0.4254.

The results are shown in table 2 (p is the number of coefficients).

*Table 1 internal validation comparison*

| | p | $Press_p$ | $C_p$ | $SSE_p$ | $Press_p - SSE_p$ |
|---|---|---|---|---|---|
| Model 2 | 9 | 109.6038 | 20.6848 | 101.1113 | 8.493 |
| Model 3 | 5 | 112.9551 | 28.7789 | 107.9577 | 4.997 |
| Model 5 | 14 | 105.7838 | 12.2756 | 93.27997 | 12.503 |

From $C_p$, we see model 5 is better since its $C_p < 14$, indicates little or no bias. Although its $Press_p$ and $SSE_p$ But the difference between $Press_p - SSE_p$ is large of model 5 which means model 5 may overfitting.

## 2. External validation

Three preliminary models are fitted on the validation dataset. The comparison between fitted regression coefficients from the training data and those from the validation data are shown in table 3-5. For model 2 and 3, the sign of the standardized regression coefficients between training data and validation data didn't change. But the differences are large. In model 5, the sign also changed.

Using the validation data, the mean squared prediction error (MSPE) are calculated to measure the predictive ability of three models. Calculating MSPE - $SSE_p/n$, Model 2 is the smallest (0.0235207), model 5 is largest (0.1156075). This is consistent with our guessing that model 5 is overfitting. Model 2 has good predictive ability.

Combining with both external and internal validation. We chose model 2 as our final model. We transform the regression coefficients back and get our final model:

$$\log(\mathrm{Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_8 X_8$$

$$\beta_0 = 5.49, \beta_1 = \text{-}0.03566, \beta_2 = 0.1825, \beta_3 = 0.3169, \beta_4 = 0.2851, \beta_5 = 0.3309, \beta_6 = 6.607 * 10^{-5}, \beta_7 = \text{-}2.148 * 10^{-4}, \beta_8 = 0.01946$$

where $\beta_{\mathrm{i}}$, $i = 0, \ldots, 8$ is regression coefficients. $X_{\mathrm{i}}$, $i = 1, \ldots, 8$ represents variable QUETELET, VITUSENOT OFTEN, VITUSEOFTEN, SMOKSTATFORMER, SMOKSTATNEVER, BETADIET, CALORIES, FIBER, respectively.

Then we use our final model fit whole dataset. The residual plot and QQ plot are similar with others. MSE = 0.4410, ${R_a}^2 = 0.213$ and $R^2 = 0.2331$ which seems adequately.

## Model diagnostic: Outlying and influential cases

We calculate the studentized deleted residuals to check the outlying Y observations.

To identify the outlying Y observation, we use the Bonferroni outlier test procedure at α = 0.1. The Bonferroni's threshold is $\mathrm{t}(1 - \frac{\alpha}{2n}; \mathrm{n} - \mathrm{p} - 1)$ = 3.57794. The Y observations corresponding to those studentized deleted residuals ($\mathrm{t}_i$) which are greater than Bonferroni's threshold can be deemed as significant outlying observation. In this dataset, we do not have any obvious outlying Y observations.

We then obtain the leverage for each case to check potential outlying X observations. Compare the leverages with the value of $\frac{2p}{n} = 0.1106282$. Any case with $h_{ii} > \frac{2p}{n}$ will be defined as outlying X observations. In this dataset, there are 4 outlying X observations whose index are 37, 164, 193 and 230. The Residuals vs, Leverage Plot figure 2-13 convinces this result, as we can clearly see that there are 4 cases on the right side of the plot, and they are far away from the other cases.

In order to check whether the outlying cases are influential in the regression function fitting, we use Cook's distance ($\mathrm{D_i}$) by equation:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p * MSE}, \qquad i = 1, \ldots, n$$

According to the Cook's distance plot figure 2-14, case 36 has the biggest Cook's distance. $\mathrm{D_{36}}$=0.06033. If we do a F test on this case, we can find that $p_{36} = \mathrm{P}(\mathrm{F}_{p,n-p} < 0.06033)$ = 9.974726e-07. Since the p-value is very small, the case 36 has little aggregated influence on all the fitted values.

We also need to check the if outlying X observations are influential in the regression function fitting. We do the same F test on all of the 4 outliers, and the p-value of them are much

smaller than 0.05. Therefore, there is no influential cases in the regression function fitting according to this measure.

## Partial coefficients and Added Variable Plots

Given our final model contains 6 features: VITUSE (X1), SMOKSTAT (X2), BETADIET (X3), CALORIES (X4), FIBER (X5) and QUETELET (X6). Partial coefficients and added variable plots are calculated and plotted for each variable.

By adding X6 into the model containing X1-X5, SSE decreases marginally by 8.97%.
By adding X1 into the model containing X2-X6, SSE decreases marginally by 3.78%.
By adding X4 into the model containing X12356, SSE decreases marginally by 2.58%.
By adding X2 into the model containing X13456, SSE decreases marginally by 2.43%.
By adding X3 into the model containing X12456, SSE decreases marginally by 1.59%.
By adding X5 into the model containing X12346, SSE decreases marginally by 1.3%.

## Conclusions and Discussion

Based on data collected from 315 surgical but not cancerous patients, it was found that plasma concentration of Beta-carotene were positively related to daily beta-carotene intake in mcg, vitamin usage frequency, smoke status and gram of fiber consumed per day; it is negatively related to calories consumed per day and Quetelet index. Among these 6 variables in our best-selected model, 'Quetelet' is the most influential feature, as SSE decreases marginally by 8.97%, given the model contains all other 5 variables. Mean-value in plasma concentration of Beta-carotene tend to differ by gender. Moreover, there are some outliers in both Y and X observations, but most of them are proven to be not influential. Thus we recommend to the general public to take care of their calories consumption and be aware of obesity.

However, this results actually are hard to generalize in the general population, since the data were collected from 315 surgical patients. This sampling of the dataset is not representative of the general population. Moreover, the data collection process is not very clear to the investigator. How variables, such as calories consumed, fiber consumed or fat consumed, are measured is not clearly stated. Besides, patients may hide or exaggerate their real responses when being asked smoking status and drinking habit. Therefore, the dataset is not so reliable. Preprocesses of data, such us transformation, outlier checking and standardization are necessary before trying to do further analysis.

# Appendix

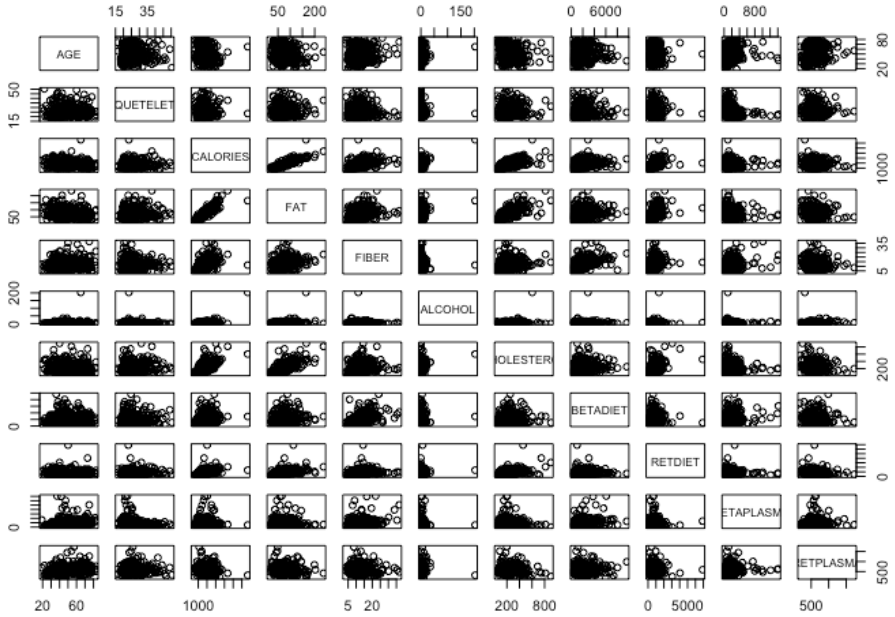


Figure 1 part1: Scatter Plot Matrix

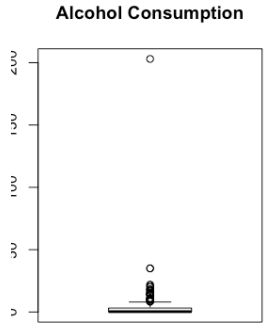


Figure1 part2: Boxplot of Alcohol Consumption

```
> summary(plasma[,-c(2,3,5)])
      AGE           QUETELET        CALORIES          FAT             FIBER          ALCOHOL       
 Min.   :19.00   Min.   :16.33   Min.   : 445.2   Min.   : 14.40   Min.   : 3.10   Min.   :  0.000  
 1st Qu.:39.00   1st Qu.:21.80   1st Qu.:1338.0   1st Qu.: 53.95   1st Qu.: 9.15   1st Qu.:  0.000  
 Median :48.00   Median :24.74   Median :1666.8   Median : 72.90   Median :12.10   Median :  0.300  
 Mean   :50.15   Mean   :26.16   Mean   :1796.7   Mean   : 77.03   Mean   :12.79   Mean   :  3.279  
 3rd Qu.:62.50   3rd Qu.:28.85   3rd Qu.:2100.4   3rd Qu.: 95.25   3rd Qu.:15.60   3rd Qu.:  3.200  
 Max.   :83.00   Max.   :50.40   Max.   :6662.2   Max.   :235.90   Max.   :36.80   Max.   :203.000  
  CHOLESTEROL      BETADIET        RETDIET        BETAPLASMA      RETPLASMA     
 Min.   : 37.7   Min.   : 214   Min.   :  30.0   Min.   :   0.0   Min.   : 179.0  
 1st Qu.:155.0   1st Qu.:1116   1st Qu.: 480.0   1st Qu.:  90.0   1st Qu.: 466.0  
 Median :206.3   Median :1802   Median : 707.0   Median : 140.0   Median : 566.0  
 Mean   :242.5   Mean   :2186   Mean   : 832.7   Mean   : 189.9   Mean   : 602.8  
 3rd Qu.:308.9   3rd Qu.:2836   3rd Qu.:1037.0   3rd Qu.: 230.0   3rd Qu.: 716.0  
 Max.   :900.7   Max.   :9642   Max.   :6901.0   Max.   :1415.0   Max.   :1727.0
```

Figure 2: Summary Statistics of Quantitative Variables

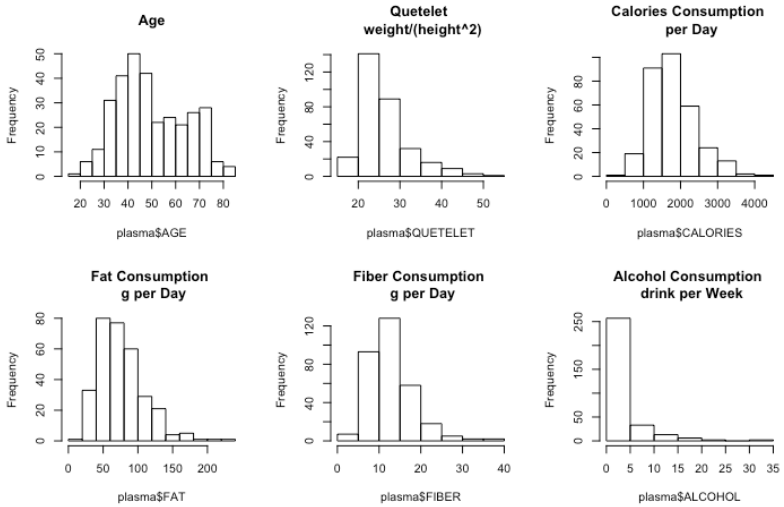


Figure 3 (Part 1): Histogram of Qualitative Variable

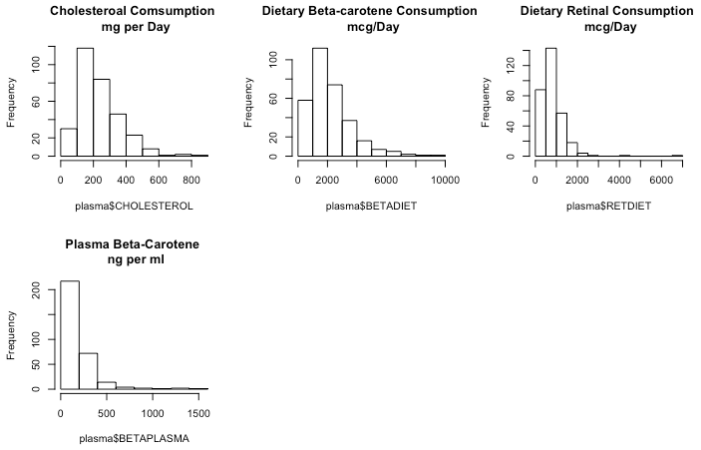


Figure 3 (Part 2): Histogram of Qualitative Variable

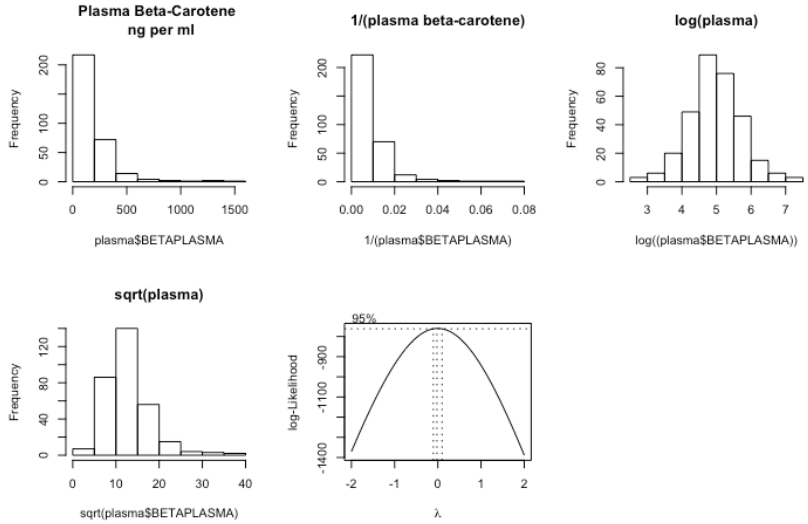


Figure 4: Histogram of Y and Its Tranformation

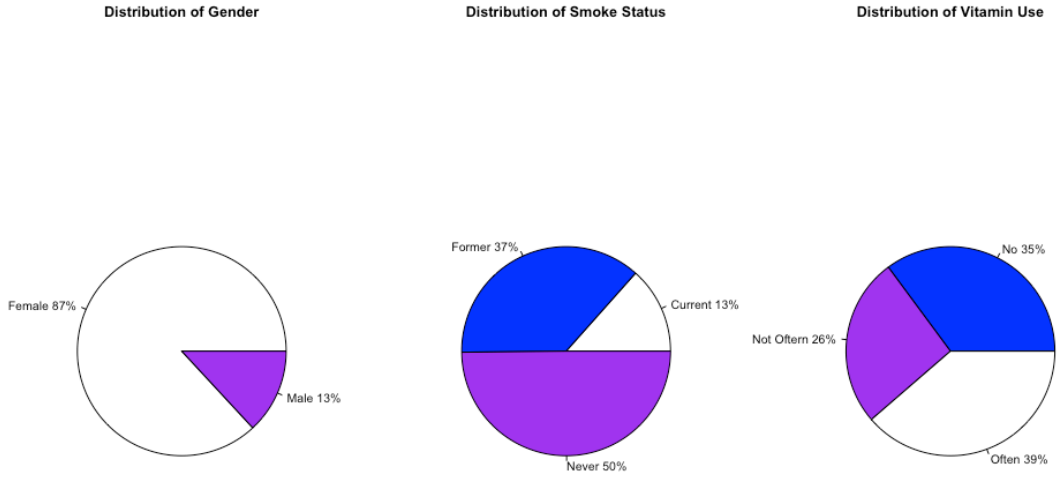


Figure 5: Pie of Categorical Variables

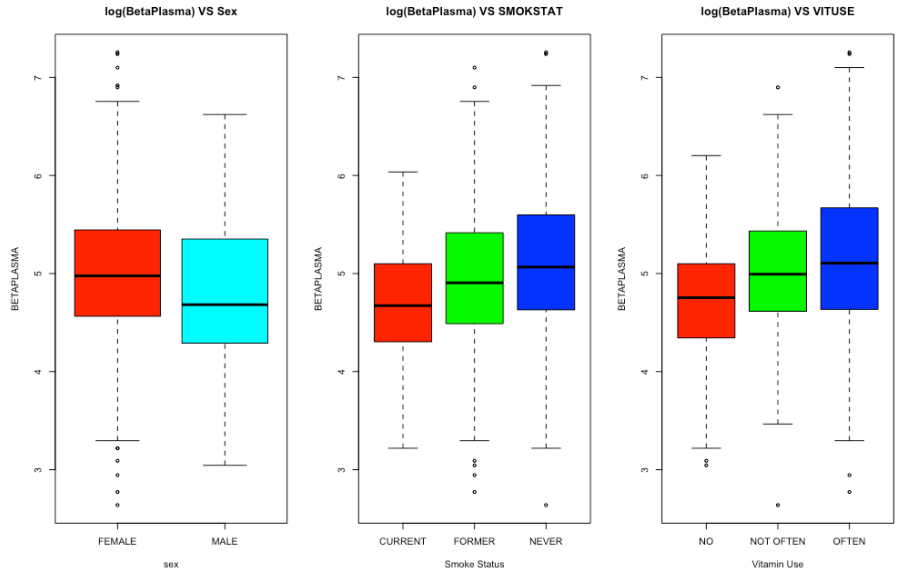


Figure 6: Boxplot for BetaPlasma VS Categorical Variables

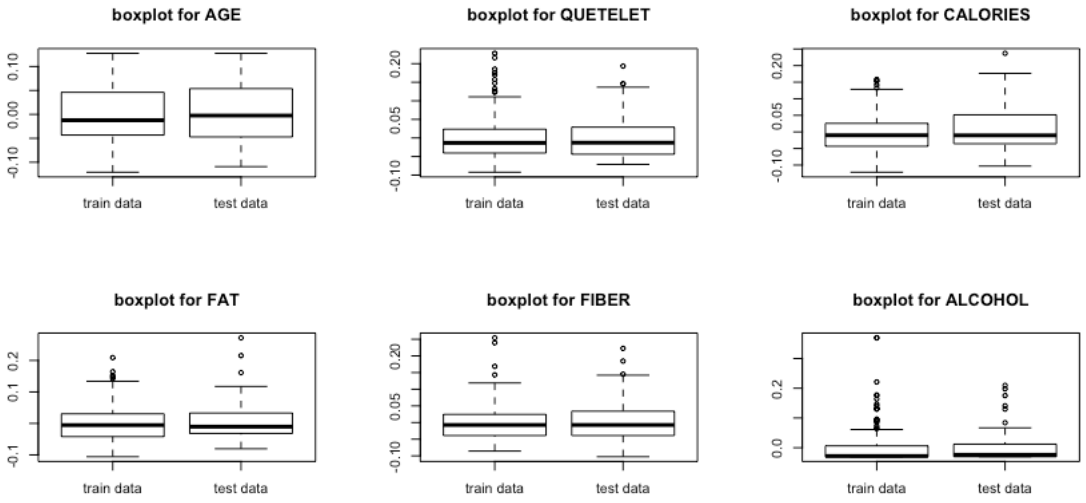


Figure 7: Comparison between Testing and Training in Quantitative Variables (part 1)

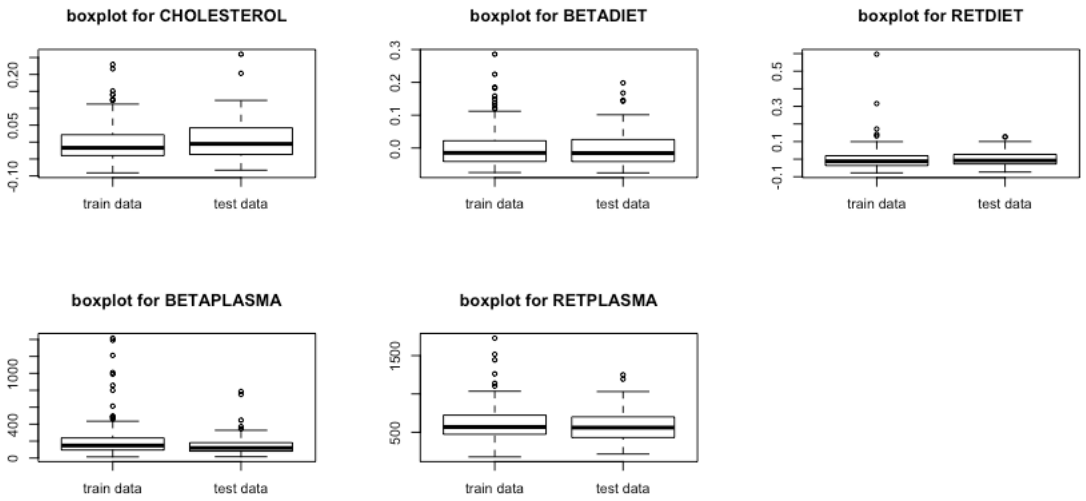


Figure 7: Comparison between Testing and Training in Quantitative Variables (part 2)

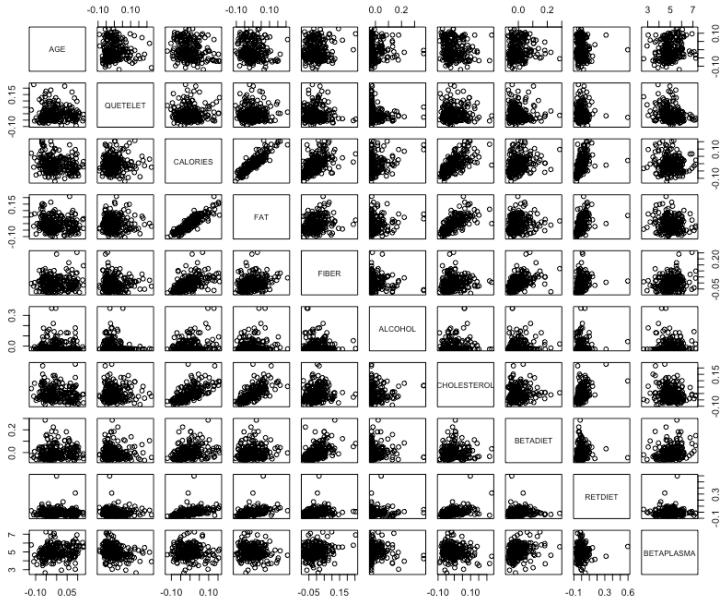


Figure 8 par1: Scatter Plot for Training

```
> round(cor(plasma.t[,-c(2,3,5,14)]),3)
             AGE QUETELET CALORIES    FAT  FIBER ALCOHOL CHOLESTEROL BETADIET RETDIET BETAPLASMA
AGE        1.000   -0.069   -0.215 -0.170  0.059   0.002      -0.125    0.070   0.007      0.150
QUETELET  -0.069    1.000    0.031  0.035 -0.065  -0.147       0.045   -0.037   0.021     -0.293
CALORIES  -0.215    0.031    1.000  0.897  0.481   0.225       0.647    0.259   0.401     -0.098
FAT       -0.170    0.035    0.897  1.000  0.266   0.083       0.693    0.125   0.395     -0.134
FIBER      0.059   -0.065    0.481  0.266  1.000   0.004       0.124    0.507   0.201      0.183
ALCOHOL    0.002   -0.147    0.225  0.083  0.004   1.000       0.087    0.071   0.011     -0.047
CHOLESTEROL -0.125  0.045    0.647  0.693  0.124   0.087       1.000    0.081   0.426     -0.142
BETADIET   0.070   -0.037    0.259  0.125  0.507   0.071       0.081    1.000   0.051      0.205
RETDIET    0.007    0.021    0.401  0.395  0.201   0.011       0.426    0.051   1.000     -0.002
BETAPLASMA 0.150   -0.293   -0.098 -0.134  0.183  -0.047      -0.142    0.205  -0.002      1.000
>
```

Figure 8: Correlation Matrix for Training Data

```
> round(VIF,2)
  rep(1, n)         AGE    QUETELET    CALORIES         FAT
       0.00        1.60        1.46       15.80       11.64
      FIBER     ALCOHOL CHOLESTEROL    BETADIET     RETDIET
       3.30        1.69        3.13        1.83        1.50
```

Figure 9: VIF output

```
$SMOKSTAT
                     diff         lwr       upr     p adj
FORMER-CURRENT  0.3263144  0.01379855 0.6388302 0.0383713
NEVER-CURRENT   0.4536540  0.15232465 0.7549833 0.0013113
NEVER-FORMER    0.1273396 -0.08570433 0.3403835 0.3381838
$VITUSE
                      diff         lwr       upr     p adj
NOT OFTEN-NO     0.2919769  0.04217055 0.5417833 0.0171759
OFTEN-NO         0.4308057  0.20523993 0.6563714 0.0000289
OFTEN-NOT OFTEN 0.1388287 -0.10608016 0.3837376 0.3768725
```

Figure 10: Tukey Interval

```
Call:
lm(formula = plasma.t$BETAPLASMA ~ ., data = plasma.t)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97470 -0.38834  0.01513  0.44223  1.82180

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.62255    0.13946  33.146  < 2e-16 ***
AGE              1.25465    0.91391   1.373  0.17120
SEXMALE         -0.24684    0.15474  -1.595  0.11210
SMOKSTATFORMER   0.27563    0.14516   1.899  0.05890 .
SMOKSTATNEVER    0.30474    0.14248   2.139  0.03356 *
QUETELET        -3.74428    0.82651  -4.530 9.66e-06 ***
VITUSENOT OFTEN  0.18407    0.11694   1.574  0.11692
VITUSEOFTEN      0.30321    0.11123   2.726  0.00693 **
CALORIES        -2.46559    2.71475  -0.908  0.36476
FAT              0.51577    2.33736   0.221  0.82556
FIBER            1.68750    1.23261   1.369  0.17238
ALCOHOL         -0.05032    0.92491  -0.054  0.95666
CHOLESTEROL      0.03098    1.25513   0.025  0.98033
BETADIET         1.66624    0.91614   1.819  0.07031 .
RETDIET          0.54322    0.82828   0.656  0.51261
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6711 on 220 degrees of freedom
Multiple R-squared:  0.238,     Adjusted R-squared:  0.1896
F-statistic: 4.909 on 14 and 220 DF,  p-value: 6.672e-08


Response: plasma.t$BETAPLASMA
             Df Sum Sq Mean Sq F value    Pr(>F)
AGE           1  2.915  2.9151  6.4734  0.011636 *
SEX           1  4.081  4.0808  9.0620  0.002915 **
SMOKSTAT      2  3.064  1.5319  3.4019  0.035071 *
QUETELET      1 12.099 12.0986 26.8668 4.928e-07 ***
VITUSE        2  4.513  2.2565  5.0109  0.007445 **
CALORIES      1  0.028  0.0283  0.0628  0.802378
FAT           1  0.315  0.3149  0.6993  0.403926
FIBER         1  2.294  2.2936  5.0933  0.025002 *
ALCOHOL       1  0.000  0.0000  0.0001  0.993494
CHOLESTEROL   1  0.043  0.0428  0.0951  0.758135
BETADIET      1  1.405  1.4053  3.1207  0.078688 .
RETDIET       1  0.194  0.1937  0.4301  0.512612
Residuals   220 99.070  0.4503
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 1 R-output of model 1*

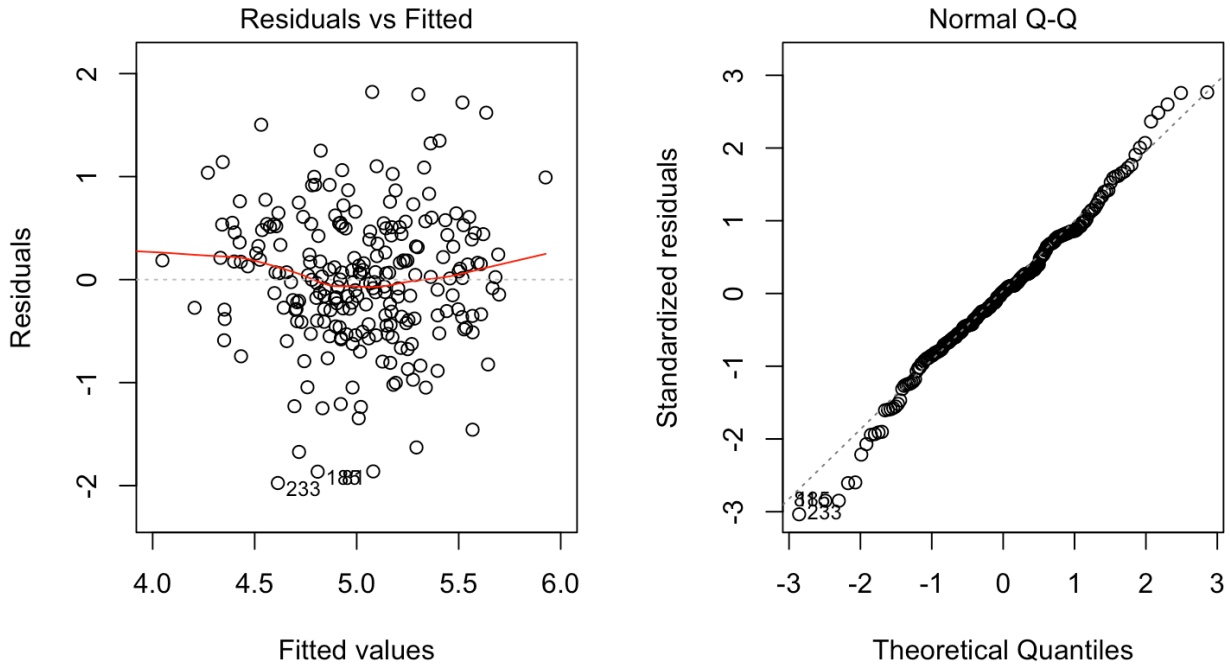


*Figure 2 residual plot and Q-Q plot for model 1*

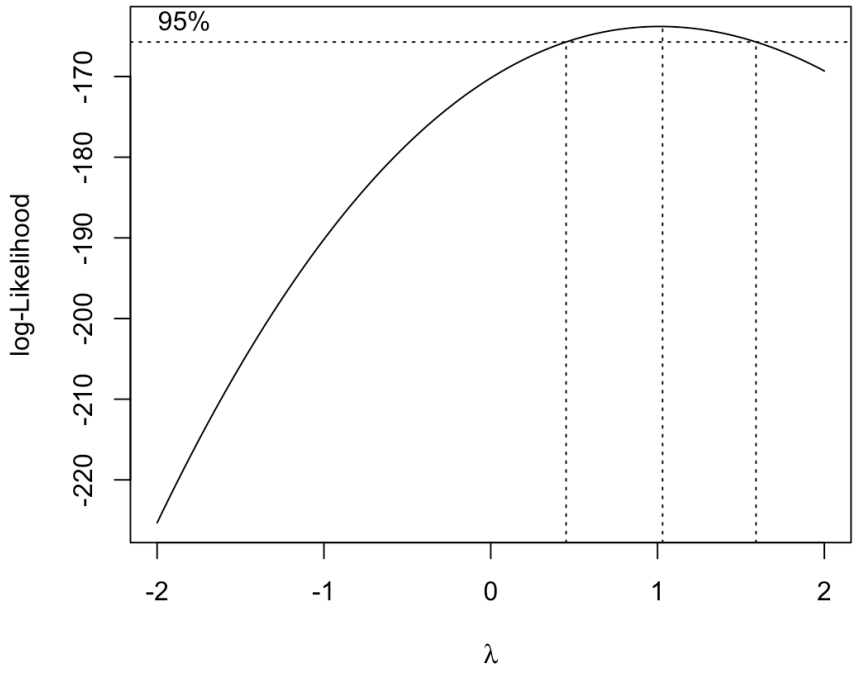


*Figure 3 Box-cox plot for model 1*

```
Call:
lm(formula = BETAPLASMA ~ QUETELET + VITUSE + SMOKSTAT + BETADIET +
    CALORIES + FIBER, data = plasma.t)

Residuals:
     Min       1Q   Median       3Q      Max
-2.00113 -0.42168 -0.01748  0.41489  1.82773

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.5692     0.1346  33.953  < 2e-16 ***
QUETELET         -3.7943     0.8043  -4.718 4.18e-06 ***
VITUSENOT OFTEN   0.1825     0.1111   1.642  0.10198
VITUSEOFTEN       0.3169     0.1065   2.975  0.00324 **
SMOKSTATFORMER    0.2851     0.1423   2.003  0.04633 *
SMOKSTATNEVER     0.3309     0.1408   2.350  0.01962 *
BETADIET          1.7232     0.9008   1.913  0.05700 .
CALORIES         -2.3542     0.9631  -2.445  0.01527 *
FIBER             1.8377     1.0648   1.726  0.08573 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6689 on 226 degrees of freedom
Multiple R-squared:  0.2223,    Adjusted R-squared:  0.1948
F-statistic: 8.077 on 8 and 226 DF,  p-value: 1.396e-09
Analysis of Variance Table

Response: BETAPLASMA
           Df  Sum Sq Mean Sq F value    Pr(>F)
QUETELET    1  11.142 11.1418 24.9037 1.203e-06 ***
VITUSE      2   7.148  3.5742  7.9889 0.0004442 ***
SMOKSTAT    2   4.978  2.4892  5.5638 0.0043782 **
BETADIET    1   2.745  2.7454  6.1364 0.0139753 *
CALORIES    1   1.562  1.5617  3.4906 0.0630129 .
FIBER       1   1.333  1.3327  2.9787 0.0857328 .
Residuals 226 101.111  0.4474
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 4 R-output of model 2*

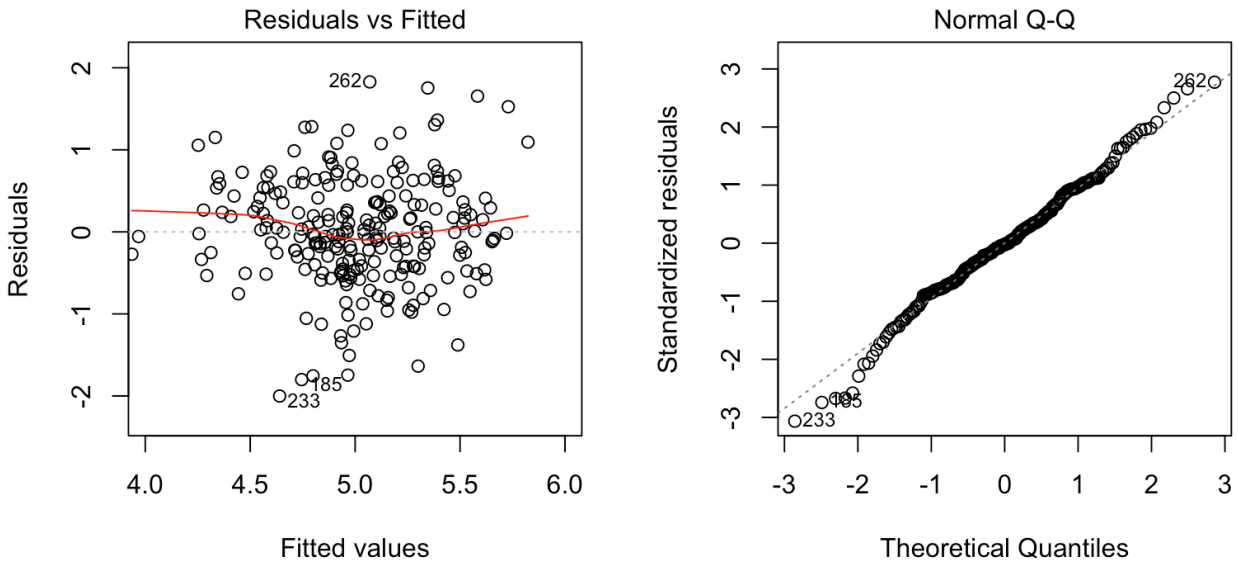


*Figure 5 residual plot and Q-Q plot of model 2*

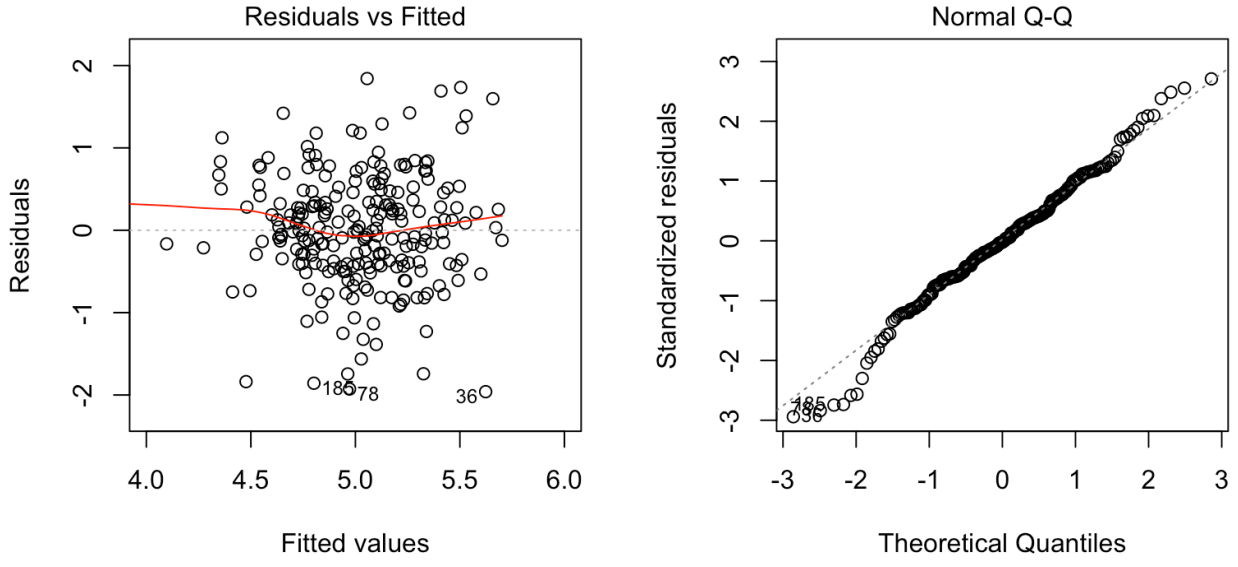


*Figure 6 Residual plot and Q-Q plot of model 3*

```
Call:
lm(formula = BETAPLASMA ~ QUETELET + BETADIET + VITUSE, data = plasma.t)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95975 -0.40789 -0.01432  0.43737  1.84154

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.81086    0.07779  61.847  < 2e-16 ***
QUETELET       -3.66647    0.80926  -4.531 9.45e-06 ***
BETADIET        2.25620    0.79592   2.835 0.004995 **
VITUSENOT OFTEN 0.19753    0.11330   1.743 0.082611 .
VITUSEOFTEN     0.38451    0.10734   3.582 0.000416 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6851 on 230 degrees of freedom
Multiple R-squared:  0.1697,    Adjusted R-squared:  0.1552
F-statistic: 11.75 on 4 and 230 DF,  p-value: 1.06e-08
Analysis of Variance Table

Response: BETAPLASMA
           Df  Sum Sq Mean Sq F value    Pr(>F)
QUETELET    1  11.142 11.1418 23.7372 2.056e-06 ***
BETADIET    1   4.896  4.8955 10.4298  0.001421 **
VITUSE      2   6.025  3.0123  6.4176  0.001940 **
Residuals 230 107.958  0.4694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 7 R-output of model 3*

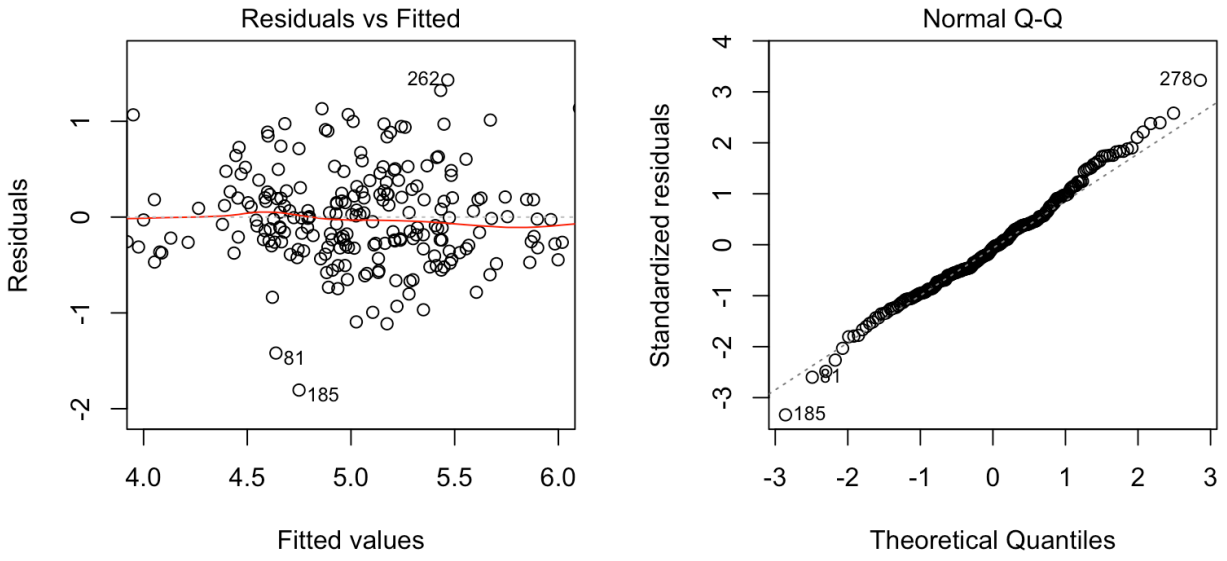


*Figure 8 Residual plot and QQ plot of model 4*

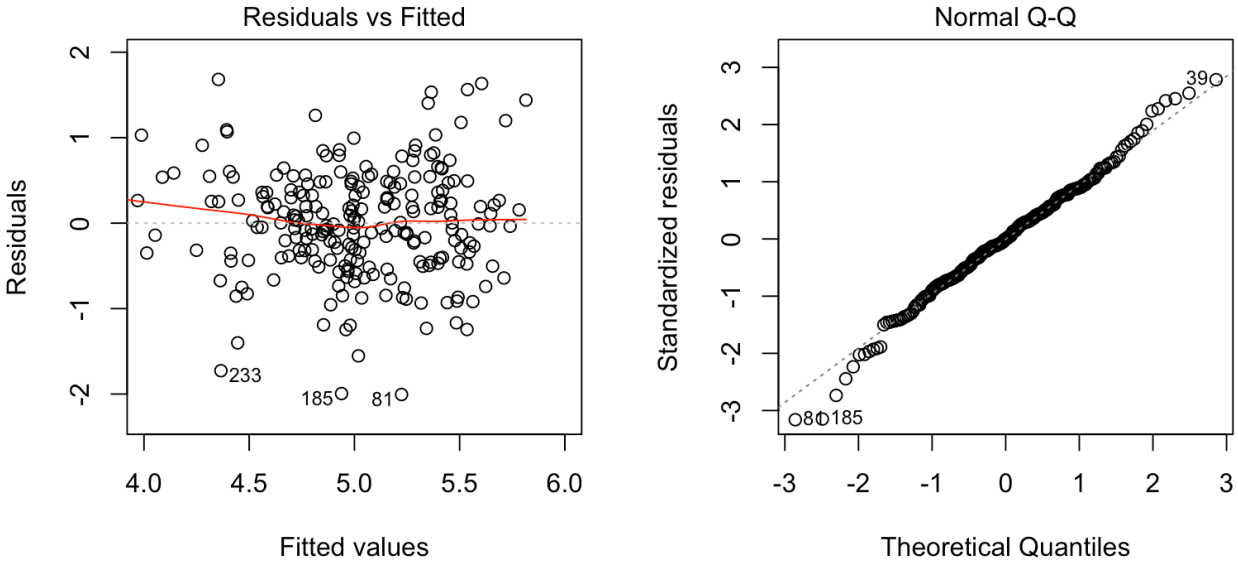


*Figure 9 Residual plot and QQ plot of model 5*

```
Call:
lm(formula = BETAPLASMA ~ QUETELET + VITUSE + SMOKSTAT + BETADIET +
    CALORIES + VITUSE:SMOKSTAT + SMOKSTAT:CALORIES, data = plasma.t)

Residuals:
     Min       1Q   Median       3Q      Max
-2.00520 -0.40717  0.00314  0.40576  1.68179

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.73749    0.17797  26.619  < 2e-16 ***
QUETELET                      -3.76034    0.77887  -4.828 2.57e-06 ***
VITUSENOT OFTEN                0.10446    0.27128   0.385  0.70056
VITUSEOFTEN                   -0.35446    0.29333  -1.208  0.22818
SMOKSTATFORMER                -0.09157    0.21285  -0.430  0.66746
SMOKSTATNEVER                  0.30977    0.21240   1.458  0.14616
BETADIET                       2.54418    0.80303   3.168  0.00175 **
CALORIES                      -3.77784    1.89811  -1.990  0.04779 *
VITUSENOT OFTEN:SMOKSTATFORMER 0.44112    0.32240   1.368  0.17263
VITUSEOFTEN:SMOKSTATFORMER     0.96449    0.33726   2.860  0.00465 **
VITUSENOT OFTEN:SMOKSTATNEVER -0.19771    0.31536  -0.627  0.53135
VITUSEOFTEN:SMOKSTATNEVER      0.58636    0.32692   1.794  0.07425 .
SMOKSTATFORMER:CALORIES       -0.13389    2.30016  -0.058  0.95363
SMOKSTATNEVER:CALORIES         4.94837    2.26258   2.187  0.02979 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6497 on 221 degrees of freedom
Multiple R-squared:  0.2826,    Adjusted R-squared:  0.2404
F-statistic: 6.696 on 13 and 221 DF,  p-value: 8.905e-11

Analysis of Variance Table

Response: BETAPLASMA
                   Df Sum Sq Mean Sq F value    Pr(>F)
QUETELET            1 11.142 11.1418 26.3973   6.1e-07 ***
VITUSE              2  7.148  3.5742  8.4680 0.0002861 ***
SMOKSTAT            2  4.978  2.4892  5.8974 0.0031973 **
BETADIET            1  2.745  2.7454  6.5045 0.0114370 *
CALORIES            1  1.562  1.5617  3.6999 0.0556998 .
VITUSE:SMOKSTAT     4  5.132  1.2831  3.0399 0.0181884 *
SMOKSTAT:CALORIES   2  4.032  2.0158  4.7759 0.0093199 **
Residuals         221 93.280  0.4221
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 10 R-output of model 5*

```
####Appendix 2
setwd("~/Dropbox/0_STA206/STA206 Group Project/Yanlin")
plasma = read.table("Plasma.txt",header=TRUE)

#########################################################################
#
#1.1Preliminary investigation on the Whole Dataset
#########################################################################
#
sum((is.na(plasma))) #NO NA
par(mfrow = c(1,2))
sapply(plasma,class)
pairs(plasma[,-c(2,3,5)])
round(cor(plasma[,-c(2,3,5)]),2) ## pairwise correlations
par(mfrow = c(1,2))
boxplot(plasma$ALCOHOL)
boxplot(plasma$ALCOHOL,main="Alcohol Consumption")
max(plasma$ALCOHOL)
summary(plasma[,-c(2,3,5)])
min(plasma$BETAPLASMA)
plasma[which(plasma$BETAPLASMA==0),]
################################################
#1.2 Delete two outlier
################################################
plasma=plasma[-which.max(plasma$ALCOHOL),]
plasma=plasma[-which(plasma$BETAPLASMA==0),]
dim(plasma) #313


####################################
#1.3 Histogram of quantitative variables
####################################
sapply(plasma,class)
par(mfrow = c(2,3))
#X
hist(plasma$AGE,main="Age")
hist(plasma$QUETELET,main="Quetelet \n weight/(height^2)")
hist(plasma$CALORIES,main="Calories Consumption \n per Day")
hist(plasma$FAT,main='Fat Consumption \n g per Day')
hist(plasma$FIBER,main="Fiber Consumption \n g per Day")
hist(plasma$ALCOHOL,main = "Alcohol Consumption \n drink per Week")

hist(plasma$CHOLESTEROL,main='Cholesteroal Comsumption \n mg per Day')
hist(plasma$BETADIET, main='Dietary Beta-carotene Consumption \n mcg/
Day')
hist(plasma$RETDIET,main="Dietary Retinal Consumption \n mcg/Day")
#y#some transformation of Y are needed
hist(plasma$BETAPLASMA,main = "Plasma Beta-Carotene \n ng per ml")
hist(1/(plasma$BETAPLASMA),main='1/(plasma beta-carotene)')
hist(log((plasma$BETAPLASMA)),main='log(plasma)')
```

```
hist(sqrt(plasma$BETAPLASMA),main='sqrt(plasma)')
boxcox(BETAPLASMA~., data=plasma)

########################################
#1.4 Pie chart on qualitative variables
########################################
par(mfrow = c(1,3))
#2
lbls=c('Female','Male')
pct=round(100*table(plasma$SEX)/sum(table(plasma$SEX)))
lab=paste(lbls,pct)
lab=paste(lab,'%',sep='')
lab
pie(table(plasma$SEX),labels=lab,col=c('white','purple'),
    main = "Distribution of Gender")
#3
table(plasma$SMOKSTAT)
lbls=c('Current','Former',"Never")
pct=round(100*table(plasma$SMOKSTAT)/sum(table(plasma$SMOKSTAT)))
lab=paste(lbls,pct)
lab=paste(lab,'%',sep='')
lab
pie(table(plasma$SMOKSTAT),labels=lab,col=c('white','blue','purple'),
    main = "Distribution of Smoke Status")
#5
table(plasma$VITUSE)
lbls=c('No','Not Oftern',"Often")
pct=round(100*table(plasma$VITUSE)/sum(table(plasma$VITUSE)))
lab=paste(lbls,pct)
lab=paste(lab,'%',sep='')
lab
pie(table(plasma$VITUSE),labels=lab,col=c('blue','purple','white'),
    main = "Distribution of Vitamin Use")


########################################
#1.5 Boxplot on qualitative variables
########################################
#for better prepresentation, we take log of the Y varaibles
par(mfrow = c(1,3))
boxplot(log(plasma$BETAPLASMA)~plasma$SEX,
        main='log(BetaPlasma) VS
Sex' ,xlab='sex',ylab='BETAPLASMA',col=rainbow(2))
boxplot(log(plasma$BETAPLASMA)~plasma$SMOKSTAT,
        main='log(BetaPlasma) VS SMOKSTAT',xlab='Smoke
Status',ylab='BETAPLASMA',col=rainbow(3))
boxplot(log(plasma$BETAPLASMA)~plasma$VITUSE,
        main='log(BetaPlasma) VS VITUSE',xlab='Vitamin
Use',ylab='BETAPLASMA',col=rainbow(3))
```

```
########################################
#1.6 Standardization/Tranformation of non-factor features(X)
########################################
dim(plasma) #235 *14
sapply(plasma, class)
for (i in 1:12){
  if (class(plasma[,i])!='factor'){
    n = length(plasma[,i])
    plasma[,i]= (1/sqrt(n-1))*((plasma[,i]-mean(plasma[,i]))/
sd(plasma[,i]))
  }
}
sapply(plasma,class)


#####################################################################
#1.7 Split into Trainign and Testing, and decide y is BetaPlasma
#####################################################################
r = (dim(plasma))[1]
set.seed(10)
index.s <- sample(1:r, size = round(r/4), replace = FALSE)
plasma.v <- plasma[index.s,] #78
plasma.t <- plasma[-index.s,] #235
dim(plasma.t)

################################################
#1.8 Compare X in Training and Testing dataset
################################################
par(mfrow = c(2,3))
quantitative_vars <- unlist(sapply(1:length(plasma), function(i) {
  if (!is.factor(plasma[[i]])) names(plasma[i])
}))
invisible(lapply(quantitative_vars, function(x) {
  boxplot(plasma.t[[x]], plasma.v[[x]], names=c('train data', 'test
data'), main = paste('boxplot for', x))
}))


########################################################
#1.9 chekcing multicollinearity in training: VIF
########################################################
n=dim(plasma.t)[1]
X_s <- plasma.t[, !(names(plasma.t) %in% c("SEX", "SMOKSTAT",
"VITUSE", "BETAPLASMA"))]
X_sta <- as.matrix(cbind(rep(1,n), X_s))
X_sta_sq <- t(X_sta) %*% X_sta
X_inverse <- solve(X_sta_sq)
VIF = diag(X_inverse)
round(VIF,2)
```

```
################################################
#1.9 chekcing multicollinearity in training: Scatter plot
################################################
pairs(plasma.t[,-c(2,3,5,14)])
round(cor(plasma.t[,-c(2,3,5,14)]),3)

#########################################################################
#####
#3.Model Selection
#########################################################################
######

#############################################################
#3.1 Subset Selection FIrst Order
#############################################################
library(leaps)
library(bestglm)
plasma.tSubset = plasma.t

dim(plasma.t)
plasma.t=plasma.t[,-14]
nc=14
sum_sub <- summary(regsubsets(log(BETAPLASMA) ~ ., data = plasma.t,
nbest = 1, nvmax = nc-1))
n = nrow(plasma.t)
p.m = as.integer(rownames(sum_sub$which))+1
ssto = sum((plasma.t$BETAPLASMA-mean(plasma.t$BETAPLASMA))^2)
sum_sub$sse = (1-sum_sub$rsq)*ssto
sum_sub$aic = n*log(sum_sub$sse/n)+2*p.m
cri_sum_sub <- sapply(c('sse', 'rsq', 'adjr2', 'cp', 'aic', 'bic'),
function(x) {
  sum_sub[[x]]
})
cri_sum_sub
# the best model
apply(cri_sum_sub[, c(1, 4, 5, 6)], 2, which.min) #13 8 8 4
apply(cri_sum_sub[, c(2, 3)], 2, which.max)  #13 10

cri_sum_sub=data.frame(cri_sum_sub)
names(cri_sum_sub)

par(mfrow = c(2,3))
plot(c(1:13),log(cri_sum_sub$sse),main="SSE",type='o')
points(which.min(cri_sum_sub$sse),log(cri_sum_sub
$sse[which.min(cri_sum_sub$sse)]),col='red',pch = 16)

plot(c(1:13),log(cri_sum_sub$rsq),main="rsq",type='o')
points(which.max(cri_sum_sub$rsq),log(cri_sum_sub
```

```
$rsq[which.max(cri_sum_sub$rsq)]),col='red',pch = 16)

plot(c(1:11),log(cri_sum_sub$adjr2)[1:11],main="rsq adj",type='o') #10
points(which.max(cri_sum_sub$adjr2),log(cri_sum_sub
$adjr2[which.max(cri_sum_sub$adjr2)]),col='red',pch = 16)

plot(c(1:9),cri_sum_sub$cp[1:9],main="Cp",type='o') #8
points(which.min(cri_sum_sub$cp),cri_sum_sub$cp[which.min(cri_sum_sub
$cp)],col='red',pch = 16)

plot(c(1:9),cri_sum_sub$aic[1:9],main="AIC",type='o') #8
points(which.min(cri_sum_sub$aic),cri_sum_sub
$aic[which.min(cri_sum_sub$aic)],col='red',pch = 16)

plot(c(1:5),(cri_sum_sub$bic)[1:5],main="BIC",type='o')  #4
points(which.min(cri_sum_sub$bic),cri_sum_sub
$bic[which.min(cri_sum_sub$bic)],col='red',pch = 16)

sum_sub$which[3,]
sum_sub$which[4,]
sum_sub$which[8,]
#########################################################################
#3.2 STEPwise using AICstep ,First Order
#########################################################################
library(MASS)
dim(plasma.t)
plasma_yIsBeta.t = plasma.t[,-14]
dim(plasma_yIsBeta.t)
names(plasma_yIsBeta.t)

full.model =  lm(log(BETAPLASMA) ~ ., data = plasma_yIsBeta.t)
empty.model = lm(log(BETAPLASMA)~ 1 ,data = plasma_yIsBeta.t)
#AIC
best.forward.AIC  = stepAIC(empty.model,scope = list(lower =
empty.model, upper = full.model),direction = "forward",trace =
FALSE,k=2)
best.backward.AIC = stepAIC(full.model,scope = list(lower =
empty.model, upper = full.model),direction = "backward", trace =
FALSE,k=2)
best.FB.AIC = step(empty.model,scope = list(lower = empty.model, upper
= full.model),direction = "both", trace = FALSE,k=2)
best.BF.AIC = step(full.model,scope = list(lower = empty.model, upper
= full.model),direction = "both", criterion = "AIC", trace =
FALSE,k=2)
#BIC, k = log(n), where n is the number of rows in your dataset:
n=dim(plasma_yIsBeta.t)[1]
best.forward.BIC = stepAIC(empty.model,scope = list(lower =
empty.model, upper = full.model),direction = "forward", k = log(n),
trace = FALSE)
best.backward.BIC = stepAIC(full.model,scope = list(lower =
```

```
empty.model, upper = full.model),direction = "backward", k = log(n),
trace = FALSE)
best.FB.BIC = stepAIC(empty.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", k = log(n), trace = FALSE)
best.BF.BIC = stepAIC(full.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", k = log(n), trace = FALSE)

#result
anova(best.forward.AIC)
#QUETELET,VITUSE,FIBER,CALORIES,SMOKSTAT,BETADIET  #6
summary(best.forward.AIC)$adj.r.squared # 0.1948105
anova(best.FB.AIC) #QUETELET,VITUSE,FIBER,CALORIES,SMOKSTAT,BETADIET
(TC)
summary(best.FB.AIC)$adj.r.squared
n=dim(plasma_yIsBeta.t)[1]
SSE = sum((summary(best.forward.AIC)$residuals)^2)
AIC = n*log(SSE/n)+2*7
AIC  #-184.1905


anova(best.backward.AIC)
#AGE,SEX,SMOKSTAT,QUETELET,VITUSE,CALORIES,FIBER,BETADIET   #8
summary(best.backward.AIC)$adj.r.squared # #tran is con
anova(best.BF.AIC)
#AGE,SEX,SMOKSTAT,QUETELET,VITUSE,CALORIES,FIBER,BETADIET
summary(best.BF.AIC)$adj.r.squared #0.2018917 #train is cons
n=dim(plasma_yIsBeta.t)[1]
SSE = sum((summary(best.backward.AIC)$residuals)^2)
AIC = n*log(SSE/n)+2*9
AIC  #-184.3552

anova(best.forward.BIC)  #QUETELET,VITUSE,FIBER  (no cal)
summary(best.forward.BIC)$adj.r.squared #0.1552411
anova(best.FB.BIC) #QUETELET,VITUSE,CALORIES,FIBER  (no cal)
summary(best.FB.BIC)$adj.r.squared  #0.1552411

anova(best.backward.BIC) #train:QUETELET,CALORIES,FIBER (no vit)
summary(best.backward.BIC)$adj.r.squared #0.1932932
anova(best.BF.BIC) #train:QUETELET,CALORIES,FIBER (no vit)
summary(best.BF.BIC)$adj.r.squared  #0.1932932


#####################################################
#3.3 Stepwise Check if Quadratic Term is in the model
#####################################################
par(mfrow = c(1,1))
plot(plasma.t$BETAPLASMA,plasma.t$QUETELET)  ##
plasma_yIsBeta = plasma.t[,-14]
plasma_yIsBeta_QUETELETSquare = plasma_yIsBeta
plasma_yIsBeta_QUETELETSquare$QUETELETSquare = (plasma_yIsBeta
```

```
$QUETELET)^2
dim(plasma_yIsBeta_QUETELETSquare)
full.model =  lm(log(BETAPLASMA) ~ ., data =
plasma_yIsBeta_QUETELETSquare)
empty.model = lm(log(BETAPLASMA)~ 1 ,data =
plasma_yIsBeta_QUETELETSquare)
#AIC
best.forward.AIC  = stepAIC(empty.model,scope = list(lower =
empty.model, upper = full.model),direction = "forward",trace =
FALSE,k=2)
best.backward.AIC = stepAIC(full.model,scope = list(lower =
empty.model, upper = full.model),direction = "backward", trace =
FALSE,k=2)
best.FB.AIC = step(empty.model,scope = list(lower = empty.model, upper
= full.model),direction = "both", trace = FALSE,k=2)
best.BF.AIC = step(full.model,scope = list(lower = empty.model, upper
= full.model),direction = "both", criterion = "AIC", trace =
FALSE,k=2)
#BIC, k = log(n), where n is the number of rows in your dataset:
n=dim(plasma_yIsBeta)[1]
best.forward.BIC = stepAIC(empty.model,scope = list(lower =
empty.model, upper = full.model),direction = "forward", k = log(n),
trace = FALSE)
best.backward.BIC = stepAIC(full.model,scope = list(lower =
empty.model, upper = full.model),direction = "backward", k = log(n),
trace = FALSE)
best.FB.BIC = stepAIC(empty.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", k = log(n), trace = FALSE)
best.BF.BIC = stepAIC(full.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", k = log(n), trace = FALSE)

anova(best.forward.AIC)
#QUETELET,VITUSE,FIBER,CALORIES,SMOKSTAT,BETADIET (TC)
anova(best.FB.AIC) #QUETELET,VITUSE,FIBER,CALORIES,SMOKSTAT,BETADIET
(TC)
anova(best.backward.AIC)
#AGE,SEX,SMOKSTAT,QUETELET,VITUSE,CALORIES,FIBER,BETADIET  (TC)
anova(best.BF.AIC)
#AGE,SEX,SMOKSTAT,QUETELET,VITUSE,CALORIES,FIBER,BETADIET (TC)
anova(best.forward.BIC)  #QUETELET,VITUSE,FIBER (T:w/o CALORIES )
anova(best.FB.BIC) #QUETELET,VITUSE,FIBER  (T:w/o CALORIES )
anova(best.backward.BIC)  #QUETELET,CALORIES,FIBER (T: w/o VITUSE)
anova(best.BF.BIC) #QUETELET,CALORIES,FIBER (T: w/o VITUSE)


################################################
#3.4 Stepwise check is model has Interaction terms
################################################
#stepAIC(lm(glyhb ~ 1, data = diabetes.c), scope = list(upper =
lm(glyhb ~ .^2, data = diabetes.c)), direction = 'both')
```

```
#plasma_yIsBeta.t2 = plasma.t[,-c(7,9,12,14)]
plasma_yIsBeta.t2 = plasma.t[,-14]
names(plasma_yIsBeta.t2)
sapply(plasma_yIsBeta.t2,class)

full.model =  lm(log(BETAPLASMA) ~ .^2, data = plasma_yIsBeta.t2)
empty.model = lm(log(BETAPLASMA)~ 1 ,data = plasma_yIsBeta.t2)

best.forward.AIC2  = stepAIC(empty.model,scope = list(lower =
empty.model, upper = full.model),direction = "forward",trace =
FALSE,k=2)
best.backward.AIC2 = stepAIC(full.model,scope = list(lower =
empty.model, upper = full.model),direction = "backward", trace =
FALSE,k=2)
best.FB.AIC2 = step(empty.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", trace = FALSE,k=2)
best.BF.AIC2 = step(full.model,scope = list(lower = empty.model, upper
= full.model),direction = "both", criterion = "AIC", trace =
FALSE,k=2)
#BIC, k = log(n), where n is the number of rows in your dataset:
n=dim(plasma_yIsBeta)[1]
best.forward.BIC2 = stepAIC(empty.model,scope = list(lower =
empty.model, upper = full.model),direction = "forward", k = log(n),
trace = FALSE)
best.backward.BIC2 = stepAIC(full.model,scope = list(lower =
empty.model, upper = full.model),direction = "backward", k = log(n),
trace = FALSE)
best.FB.BIC2 = stepAIC(empty.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", k = log(n), trace = FALSE)
best.BF.BIC2 = stepAIC(full.model,scope = list(lower = empty.model,
upper = full.model),direction = "both", k = log(n), trace = FALSE)

#result
anova(best.forward.AIC2)
#QUETELET,VITUSE,FIBER,CALORIES,SMOKSTAT,BETADIET
summary(best.forward.AIC2)$adj.r.squared #0.2516287

anova(best.FB.AIC2) #QUETELET,VITUSE,FIBER,CALORIES,SMOKSTAT,BETADIET
summary(best.FB.AIC2)$adj.r.squared # 0.2516287


anova(best.backward.AIC2)
#AGE,SEX,SMOKSTAT,QUETELET,VITUSE,CALORIES,FIBER,BETADIET
summary(best.backward.AIC2)$adj.r.squared #0.2774081

anova(best.BF.AIC2)
#AGE,SEX,SMOKSTAT,QUETELET,VITUSE,CALORIES,FIBER,BETADIET
summary(best.BF.AIC2)$adj.r.squared #0.2825982
```

```
anova(best.forward.BIC2) #QUETELET,VITUSE,CALORIES,FIBER
summary(best.forward.BIC2)$adj.r.squared #0.1932932

anova(best.FB.BIC2) #QUETELET,VITUSE,FIBER
summary(best.FB.BIC2)$adj.r.squared  #0.1932932

anova(best.backward.BIC2) #QUETELET,VITUSE,CALORIES,FIBER
summary(best.backward.BIC2)$adj.r.squared #0.1932932

anova(best.BF.BIC2) #QUETELET,VITUSE,CALORIES,FIBER
summary(best.BF.BIC2)$adj.r.squared  #0.1932932


#########################################################################
########
#4 Model Validation on 3 final Models
#########################################################################
##########
#best.FB.AIC   (first order )
#best.FB.BIC   (first order ,interaction)
#best.FB.AIC2  (interaction)


#################################################################
## Internal validation
#################################################################

data.3 = plasma.t[, c("SMOKSTAT", "QUETELET", "VITUSE", "CALORIES",
                      "FIBER", "BETADIET", "BETAPLASMA")]
fit3.f = lm(BETAPLASMA~.^2,data=data.3)
(nf = length(fit3.f$coefficients))
(msef = anova(fit3.f)["Residuals", 3])

(sse.model2 = anova(fit.AIC.f)["Residuals",2])
(sse.model3 = anova(fit.BIC.f)["Residuals",2])
(sse.model5 = anova(fit2.AIC.f)["Residuals",2])

(mse.model2 = anova(fit.AIC.f)["Residuals",3])
(mse.model3 = anova(fit.BIC.f)["Residuals",3])
(mse.model5 = anova(fit2.AIC.f)["Residuals",3])

(p.model2 = length(fit.AIC.f$coefficients))
(p.model3 = length(fit.BIC.f$coefficients))
(p.model5 = length(fit2.AIC.f$coefficients))

(cp.model2 = sse.model2/msef - (n-2*p.model2))
(cp.model3 = sse.model3/msef - (n-2*p.model3))
(cp.model5 = sse.model5/msef - (n-2*p.model5))

(press.model2 = sum(fit.AIC.f$residuals^2/(1-influence(fit.AIC.f)
```

```
$hat)^2))
(press.model3 = sum(fit.BIC.f$residuals^2/(1-influence(fit.BIC.f)
$hat)^2))
(press.model5 = sum(fit2.AIC.f$residuals^2/(1-influence(fit2.AIC.f)
$hat)^2))

press.model2 - sse.model2
press.model3 - sse.model3
press.model5 - sse.model5

#########################################################################
####
## external validation
#########################################################################
####
# model 2
fit.model2.v <- lm(fit.AIC.f,data=plasma.v)
summary(fit.model2.v)
coef(fit.model2.v)
coef(fit.AIC.f)
round(abs(coef(fit.AIC.f)-coef(fit.model2.v))/
abs(coef(fit.AIC.f))*100,3)

# model 3
fit.model3.v <- lm(fit.BIC.f,data=plasma.v)
summary(fit.model3.v)
coef(fit.model3.v)
coef(fit.BIC.f)
round(abs(coef(fit.BIC.f)-coef(fit.model3.v))/
abs(coef(fit.BIC.f))*100,3)

# model 5
fit.model5.v <- lm(fit2.AIC.f,data=plasma.v)
summary(fit.model5.v)
coef(fit.model5.v)
coef(fit2.AIC.f)
round(abs(coef(fit2.AIC.f)-coef(fit.model5.v))/
abs(coef(fit2.AIC.f))*100,3)

## mean squared prediction error
nv <- ncol(plasma.v)
newdata = plasma.v[,-nv]

# Model 2
pred.m2 <- predict.lm(fit.AIC.f, newdata)
(mspe.m2 <- mean((pred.m2 - plasma.v[,nv])^2))
mspe.m2 - sse.model2/n
mspe.m2 - press.model2/n

# Model 3
```

```
pred.m3 <- predict.lm(fit.BIC.f, newdata)
(mspe.m3 <- mean((pred.m3 - plasma.v[,nv])^2))
mspe.m3 - sse.model3/n
mspe.m3 - press.model3/n

# Model 5
pred.m5 <- predict.lm(fit2.AIC.f, newdata)
(mspe.m5 <- mean((pred.m5 - plasma.v[,nv])^2))
mspe.m5 - sse.model5/n
mspe.m5 - press.model5/n

##########################################
#5. Outlier checking, influence checking
##########################################
# Draw residual vs. fitted value plot and residual Q-Q plot
fit.AIC.f
par(mfrow = c(1,2))
plot(fit.AIC.f, which=1)
plot(fit.AIC.f, which=2)

n = nrow(plasma.t)
p = ncol(plasma.t)
plot(fit.AIC.f,which=4)
res=residuals(fit.AIC.f)# residuals of the final model
h1 = influence(fit.AIC.f)$hat
d.res.std=studres(fit.AIC.f)
qt(1-0.1/(2*n),n-p-1) # bonferronis thresh hold
idx.Y = as.vector(which(abs(d.res.std)>=qt(1-0.1/(2*n),n-p-1)))
idx.Y

#Obtain the leverage and identify any outlying X observations. Draw
residual vs.leverage plot.
idx.X = as.vector(which(h1>(2*p/n)))
idx.X ## 4 outliers
par(mfrow = c(1,1))
plot(h1,res,xlab="leverage",ylab="residuals", main = " Residuals vs.
Leverage Plot")
idx.X


#Draw an influence index plot using Cook's distance. Are there any
influential cases according to this measure?
plot(fit.AIC.f,which=4)
plot(fit.AIC.f,which=5)
#full model MSE: 0.4254
cook.d = res^2*h1/(p*0.4254*(1-h1)^2)
cook.max = cook.d[which(cook.d==max(cook.d))]
pf(cook.max,p,n-p)
idx = c(idx.X,idx.Y)
cook.d[idx]
```

```
pf(cook.d[idx],p,n-p)


#########################################################################
########
#6 Added Variables Analysis
#########################################################################
##########
best.FB.AIC
anova(best.FB.AIC)
library(car)
par(mfrow = c(1,1))
#check Quetelet
fit = lm(log(BETAPLASMA)~VITUSE + SMOKSTAT + BETADIET + CALORIES +
FIBER+ QUETELET ,data = plasma.t)
anova(fit)
SSTO <- sum(anova(fit)[,2])
Rsq_Y_QUETELET <- anova(fit)[6,2]/(SSTO-sum(anova(fit)[1:5,2]))
round(Rsq_Y_QUETELET*100,2) #8.97
#this means adding quetelet into the model containing VITUSE+SMOKSTAT
+BETADIET+RETDIET,
#marginal reduction in SSE in proportional is 8.97%.
avPlots(fit,~QUETELET, ylim=c(-1,1),xlim=c(-0.2,0.2))

#check VITUSE
fit = lm(log(BETAPLASMA)~ SMOKSTAT  + CALORIES + FIBER+ QUETELET +
BETADIET+VITUSE ,data = plasma.t)
anova(fit)
SSTO <- sum(anova(fit)[,2])
Rsq_Y_VITUSE <- anova(fit)[6,2]/(SSTO-sum(anova(fit)[1:5,2]))
round(Rsq_Y_VITUSE*100,2) #3.78
avPlots(fit,~VITUSE, ylim=c(-1,1),xlim=c(-0.2,0.2))

#check CALORIES
fit = lm(log(BETAPLASMA)~VITUSE + SMOKSTAT + BETADIET + QUETELET+
FIBER + CALORIES ,data = plasma.t)
anova(fit)
SSTO <- sum(anova(fit)[,2])
Rsq_Y_CALORIES <- anova(fit)[6,2]/(SSTO-sum(anova(fit)[1:5,2]))
round(Rsq_Y_CALORIES*100,2) #2.58%
avPlots(fit,~CALORIES, ylim=c(-1,1),xlim=c(-0.2,0.2))

#check SMOKSTAT
fit = lm(log(BETAPLASMA)~VITUSE   + CALORIES + FIBER+ QUETELET +
BETADIET+ SMOKSTAT,data = plasma.t)
anova(fit)
SSTO <- sum(anova(fit)[,2])
Rsq_Y_SMOKSTAT <- anova(fit)[6,2]/(SSTO-sum(anova(fit)[1:5,2]))
round(Rsq_Y_SMOKSTAT*100,2) #2.43
avPlots(fit,~SMOKSTAT, ylim=c(-1,1),xlim=c(-0.2,0.2))
```

```
#check BETADIET
fit = lm(log(BETAPLASMA)~VITUSE + SMOKSTAT  + CALORIES + FIBER+
QUETELET + BETADIET,data = plasma.t)
anova(fit)
SSTO <- sum(anova(fit)[,2])
Rsq_Y_BETADIET <- anova(fit)[6,2]/(SSTO-sum(anova(fit)[1:5,2]))
round(Rsq_Y_BETADIET*100,2) #1.59
avPlots(fit,~BETADIET, ylim=c(-1,1),xlim=c(-0.2,0.2))

#check fiber
fit = lm(log(BETAPLASMA)~VITUSE + SMOKSTAT + BETADIET + CALORIES +
QUETELET+ FIBER ,data = plasma.t)
anova(fit)
SSTO <- sum(anova(fit)[,2])
Rsq_Y_fiber <- anova(fit)[6,2]/(SSTO-sum(anova(fit)[1:5,2]))
round(Rsq_Y_fiber*100,2) #1.3%
avPlots(fit,~FIBER, ylim=c(-1,1),xlim=c(-0.2,0.2))


par(mfrow = c(1,1))
avPlots(fit,~QUETELET, ylim=c(-1,1),xlim=c(-0.2,0.2))
avPlots(fit,~VITUSE, ylim=c(-1,1),xlim=c(-0.2,0.2)) #2
avPlots(fit,~CALORIES, ylim=c(-1,1),xlim=c(-0.2,0.2)) #1
avPlots(fit,~SMOKSTAT, ylim=c(-1,1),xlim=c(-0.2,0.2)) #2
avPlots(fit,~BETADIET, ylim=c(-1,1),xlim=c(-0.2,0.2)) #1
avPlots(fit,~FIBER, ylim=c(-1,1),xlim=c(-0.2,0.2))    #1


#########################################################
#########One Way Anova on SEX, SMOKSTAT, VITUSE
#########################################################
#SEX
par(mfrow = c(1,3))
plot(BETAPLASMA~SEX, data=plasma)
g <- lm(BETAPLASMA~SEX, data=plasma)
plot(g,which=1)
plot(g,which=2)

plot(log(BETAPLASMA)~SEX, data=plasma)
g <- lm(log(BETAPLASMA)~SEX, data=plasma)
plot(g,which=1)
plot(g,which=2)
anova(g)
summary(g)
n=dim(plasma)[1]
I = length(levels(plasma$SEX))
qt(0.975,n-I)
D = mean(log(plasma$BETAPLASMA[plasma$SEX=="FEMALE"]))-mean(log(plasma
$BETAPLASMA[plasma$SEX=="MALE"]))
```

```
SE_D = summary(g)$sig*sqrt(1/length(plasma$BETAPLASMA[plasma
$SEX=="FEMALE"]) +1/length(plasma$BETAPLASMA[plasma$SEX=="MALE"]))
#gender
round(c(D-SE_D*qt(0.975,n-I),D+SE_D*qt(0.975,n-I)),3)

#SMOKSTAT Tukey b/c CI narrower than bonfer
par(mfrow = c(1,3))
plot(log(BETAPLASMA)~SMOKSTAT, data=plasma)
g <- lm(log(BETAPLASMA)~SMOKSTAT, data=plasma)
plot(g,which=1)
plot(g,which=2)
TukeyHSD(aov(g))


#VITUSE, Tukey b/c CI narrower than bonfer
par(mfrow = c(1,2))
plot(log(BETAPLASMA)~VITUSE , data=plasma)
g <- lm(log(BETAPLASMA)~VITUSE , data=plasma)
plot(g,which=1)
plot(g,which=2)
TukeyHSD(aov(g))

#########################################################
#Transform Back
#########################################################
plasma = read.table("Plasma.txt",header=TRUE)
plasma=plasma[-which.max(plasma$ALCOHOL),]
plasma=plasma[-which(plasma$BETAPLASMA==0),]
dim(plasma) #313

r = (dim(plasma))[1]
set.seed(10)
index.s <- sample(1:r, size = round(r/4), replace = FALSE)
plasma.v <- plasma[index.s,] #78
plasma.t <- plasma[-index.s,] #235
dim(plasma.t)



lm(log(BETAPLASMA)~QUETELET+VITUSE+SMOKSTAT+BETADIET+CALORIES+FIBER  ,
data=plasma.t)
#lm(log(BETAPLASMA)~QUETELET+VITUSE+SMOKSTAT+BETADIET+CALORIES
+FIBER  , data=plasma)
#best.FB.AIC
mu_quetelet = 1/(sqrt(312)*sd(plasma$QUETELET))
beta_q = mu_quetelet*best.FB.AIC$coefficients[2]
beta_v_notOften = best.FB.AIC$coefficients[3]
beta_v_Often = best.FB.AIC$coefficients[4]
beta_sm_for = best.FB.AIC$coefficients[5]
beta_sm_ner = best.FB.AIC$coefficients[6]
```

```
mu_betadiet = 1/(sqrt(312)*sd(plasma$BETADIET))
beta_betadient = mu_betadiet*best.FB.AIC$coefficients[7]
mu_calories = 1/(sqrt(312)*sd(plasma$CALORIES))
beta_calories = mu_calories *best.FB.AIC$coefficients[8]
mu_fiber = 1/(sqrt(312)*sd(plasma$FIBER))
beta_fiber = mu_fiber*best.FB.AIC$coefficients[9]

A= lm(log(BETAPLASMA)~QUETELET+VITUSE+SMOKSTAT+BETADIET+CALORIES
+FIBER  , data=plasma.t)
signif(A$coefficients,4)
signif(c(beta_q,beta_v_notOften,beta_v_Often,beta_sm_for,beta_sm_ner,b
eta_betadient,
  beta_calories,beta_fiber),4)


beta_0 = mean(log(plasma.t$BETAPLASMA))-
  beta_q*mean(plasma.t$QUETELET)-
  beta_v_Often*sum(plasma.t$VITUSE=='NOT OFTEN')/313-
  beta_v_Often*sum(plasma.t$VITUSE=='OFTEN')/313-
  beta_sm_for*sum(plasma.t$SMOKSTAT=='FORMER')/313-
  beta_sm_ner*sum(plasma.t$SMOKSTAT=='NEVER')/313-
  beta_betadient*mean(plasma.t$BETADIET)-
  beta_calories*mean(plasma.t$CALORIES)-
  beta_fiber*mean(plasma.t$FIBER)


beta_0
```

## Reference:

[1] Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology 1989;130:511-521.

[2] Mayo Clinic Staff, Diseases and conditions of lung cancer prevention, Mayo Clinic, 25 Sept 2015, http://www.mayoclinic.org/diseasesconditions/lungcancer/basics/prevention/con-20025531, 4 Dec 2016.

[3] Stryker W.S., Kaplan L.A., Stein EA, et.al. The relation of diet, cigarette smoking, and alcohol consumption to plasma beta-carotene and alpha-tocopherol levels. Am. J. Epidemiol 1998;127:283-96