Determinants of Plasma Concentration of Retinol and Beta-carotene

Department of Statistics
University of California, Davis
Meng-Tien Tsai | mtts@ucdavis.edu
Wun-Syuan Wu | wswu@ucdavis.edu

I. Abstract

The relationship between personal characteristics and dietary factors and plasma concentrations of retinol and beta-carotene was studied with 11 personal variables over 315 patients by using multiple regression analysis. Quetelet appeared to negatively affect the level of beta-carotene plasma the most while age was the most influential in determining the level of retinol plasma concentration. However, the relationship between personal variables in the data and the level of retinol plasma could be subtle. Under the 5% significance level, there was not enough evidence to conclude that a linear relationship between beta-carotene plasma and plasma retinol existed.

Keywords: Multiple Regression analysis, Beta-carotene, Retinol

II. Introduction

While the relationship between increased risk of developing certain types of cancer and low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids has been widely studied, relatively few studies have observed the determinants of plasma concentrations of these micronutrients. Therefore, this study is presented to deliver an initial investigation of the relationship between personal characteristics as well as dietary factors and the levels of plasma concentration of both the micronutrients retinol and beta-carotene.

Following are the questions that are expected to be addressed with this study:

- What personal characteristics or dietary factors can be influential to the level of beta-carotene and retinol plasma concentration?
- Do all the intake of micronutrients and personal characteristics affect the level of plasma beta-carotene and plasma retinol?
- Does a transformation of variables need to take place for the relationship between predictor variables and the level of plasma to be linear?
- Are the concentrations of the two micronutrients retinol and beta-carotene related to each other?

The data used in this study contains 14 variables including 12 personal characteristics and dietary factors and 2 measurements of the levels of plasma concentrations, and the data source was CMU Stat Lib. These variables were obtained from 315 patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary, or uterus that was found to be non-cancerous.

III. Methods and Results

3.1. Exploratory Data Analysis

An exploratory data analysis was first performed to analyze and investigate the data set and summarize the main characteristics for later modeling. No missing values for each variable were detected. Summary statistics, histograms, boxplots, scatter plot matrix, and pairwise correlations were used on quantitative data, including Age, Quetelet, Number of calories consumed per day (Calories), Grams of fat consumed per day (Fat), Grams of fiber consumed per day (Fiber), Number of alcoholic drinks consumed per week (Alcohol), Cholesterol consumed (Cholesterol), Dietary beta-carotene consumed (Betadiet), Dietary retinol consumed (Retdiet), Plasma beta-carotene (Betaplasma), and Plasma retinol (Retplasma), while pie charts, frequency tables, and side-by-side boxplots were used on qualitative data, including Sex, Smokstat (Smoking status), and Vituse (Vitamin use).

According to summary statistics (Figure 1), the mean and the 3rd quartile of Alcohol were 3.279 and 3.200 respectively, with a range of 0 to 203 showing that few extreme values could be included, which was also suggested by the histogram of Alcohol (Figure 2).

By observing the histograms (Figure 3), the distribution of Quetelet, Grams of fat consumed per day, Grams of fiber consumed per day, Cholesterol consumed, Dietary beta-carotene consumed, Dietary retinol consumed, Plasma beta-carotene and Plasma retinol were all right-skewed. In particular, the histogram of age implied that its distribution was bimodal distribution. In addition, as the response variable, Plasma beta-carotene appeared to need transformation.

In Figure 4, the boxplot of Age revealed that the distribution of it was symmetric. All the boxplots supported the previous observations made by the histograms. The boxplot of alcohol and the boxplot of retdiet both showed that 1 extreme case could be included in each Alcohol and Dietary retinol consumed, which could also be observed through the scatter plot matrix (Figure 5). To clearly check the relationship between variables, another scatter plot matrix without the extreme cases mentioned above was conducted (Figure 6). Through the scatter plot matrix, Calories, and Grams of fat consumed per day were highly correlated ($r = 0.90$), and Grams of fat consumed per day and Cholesterol consumed also have a clear relationship ($r = 0.7$). Overall, no obvious non-linearity patterns and clusters were detected.

With the barplot (Figure 7) and pie chart (Figure 8) of each qualitative variable, the number of numbers and percentage of different factors in the variable was displayed. The plots showed that females were the majority(87%) in the subjects, and non-smokers accounted for half of the testers larger than the amount of those who smoked before and who were currently smoking. In addition, the number of subjects who did not often intake vitamins was less than the number of people who often did and who didn't use vitamins.

From side-by-side boxplots (Figure 9), the mean and the range of Retplasma in males were larger than females while the mean and the range of the two genders are similar when it comes to Betaplasma. Also, in a side-by-side boxplot by Smokstat, non-smoker had a larger range of the level of Betaplasma, and those who often take vitamins get a larger range of the level of Betaplasma as well.

3.2. Preliminary Model Investigation

The preliminary model investigation was delivered with the full model of the first ordered model. Since relatively few studies had investigated in relation to the determinants of plasma concentrations, a full model was being fitted here to first examine the assumptions and discover whether a possible transformation to Betaplasma and Retplasma was needed.

In linear regression, the assumptions here were linearity of the regression relation, normality of the distribution of residuals, constant variance, and independence of the residuals.

As the diagnostics plots of the full model were checked in the beginning, obvious heavy right-tailed was shown in both normal Q-Q plots (Figure 10, 11), indicating that the normality assumption was not held well. Thus, by the Box-Cox procedure (Figure 12, 13), log transformation (= 0) on Betaplasma and Retplasma were suggested. However, since the log transformation was unable to handle response variables that were equal or less than 0, 1 observation with Betaplasma as response variable needed to be removed before conducting the Box-Cox procedure. For the same reason, in the subsequent analysis, the values of Betaplasma were added by 1 before taking log transformation.

After the transformation, residual v.s. fitted value plots and normal Q-Q plots of Betaplasma (Figure 14) and Retplasma (Figure 15) showed an obvious augment on the assumptions, indicating a better hold for linearity and normality. Moreover,

histograms (Figure 16, 17) also implied distributions of both Betaplasma and Retplasma were closer to a normal distribution than previous.

Since the goal of this study was to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene, variables Betaplasma and Retplasma were respectively taken as response variables. However, with the weak correlation between them, they were not included simultaneously when fitting a model. Showed in the correlation matrix (Table 1), their correlation is 0.07157 which is approximately equal to 0. And by hypothesis testing with null hypothesis H0: r = 0, alternative hypothesis H1: r 0, and significance level alpha = 0.05, there was not enough evidence to conclude that a linear relationship between Betaplasma and Retplasma existed. As a result, two model buildings with Betaplasma and Retplasma as a response variable respectively but without each other as one of the predictors would be pulled out in the following. Both the datasets used in two modeling buildings were sampled into 80% training data and 20% validation data.

3.3. Response variable: Betaplasma

Two model selection approaches were employed here which were the best subset selection approach and stepwise selection approach. In the best subset selection (Table 2), only the first-order model was fitted since the second-order model contained 2^12 different model subsets which were too heavy for computing. The procedure was set with at most 12 variables selected.

After calculating several criteria: adjusted determinant coefficients(Ra2), Mallow's Cp criterion, AIC, and BIC, two suggested models were presented (Table 3). The first one (Best Subset1) suggested by adjusted r squared(Ra2 = 0.2330), Mallow's Cp (Cp = 7.2871), and AIC (AIC = -212.0463) was the fitted model containing 8 predictors: Age, Sex, Smokstat, Quetelet, Vituse, Fat, Fiber, Betadiet with Ra2 =0.2327. Since SmokstatNever, VituseNotoften, VituseOften were chosen in the model, qualitative variables Smokstat and Vituse were included in it as one of the predictors. The second model (Best Subset2) was suggested by BIC (BIC = -183.803). It contained 5 predictors: Smokstat, Quetelet, Vituse, Fat, Betadiet.

In the stepwise selection approach, which was more computationally effective, the first-order model and second-order model were both taken into consideration by using bidirectional elimination. When fitting the first-order model, the best model (Stepwise Model1) here included 7 predictors, Quetelet, Vituse, Betadiet, Fat, Smokstat, Sex, and Age, with Ra2 =0.2302 (Figure 18). When fitting the model with interactions, the

best model (Stepwise Model2) was obtained which included 17 predictors, Cholesterol, Fiber, Quetelet, Vituse, Betadiet, Age, Retdiet, Smokstat, Vituse*Betadiet, Cholesterol*Retdiet, Age*Retdiet, Vituse*Smokstat, Retdiet*Smokstat, Cholesterol*Quetelet, Betadiet*Age, Fiber*Smokstat, and Cholesterol*Fiber, with Ra2 =0.2376 (Figure 19).

With the aim of examining the validity of the above four models, the testing data was utilized to examine internal model validation and the validation data was utilized for external validation. For internal validation, Mallow's Cp and Pressp and SSE were obtained to conceive whether Cp was approximately equal to p and whether Pressp and SSE were close. As for external validation, percentage changes in coefficient and standard error of the four models were obtained and compared. Additionally, MSPE would also be taken into consideration for the final model.

With the model validation of Best Subset1, Best Subset2, Stepwise Model1, Stepwise Model2, the Best Subset 1 was selected as the best model (Figure 20). In Table 3 and Table 4, Best Subset 1 was chosen by AIC (AIC = -210.9984). It had a second smaller SSE (SSE = 99.9662), Cp (Cp = 12.972734), Pressp (Pressp = 109.5064), and MSPE (MSPE = 58248.69) and second larger adjusted r squared (Ra2 =0.2397) as well. Even though Stepwise Model2 had the smallest MSPE (MSPE = 58228.58) yet it contained a large number of predictors (p = 17) which was far from its CP = 10.9325 that could possibly indicate overfitting. MSPE between 2 models had only small differences and the Pressp (Pressp = 132.0332) of Stepwise Model2 is relatively bigger than the other three models. Overall, Stepwise Model2 appeared to be too complicated.

After fitting the whole dataset into Best Subset1, outliers and influentials detection implied 0 outlier in response variable with studentized deleted residual compared to Bonferroni threshold under alpha = 0.1 and 44 outliers in predictor variables. In Figure 21, 3 influential cases were detected by Cook's distance plot within 44 outliers which were case 35th, 39th, and 257th. However, since the percentage change on the fitted value with or without outliers was small (1.01%, 0.97%, and 0.9% respectively), no case had an unduly large influence on prediction and thus all cases may be retained.

3.4. Response variable: Retplasma

Equivalent to the previous model building with Betaplasma as the response variable, best subset selection, and stepwise selection were the model selection approaches used here. According to the best subset selection (Table 5), Model 1 containing Age, Smokstat, Fat, and Betadiet was the best in adjusted R square (Ra2 =0.0687), Model

2 with Age, Smokstat, and Fat performed the best in Mallow's Cp and AIC criteria (Cp = -2.9369; AIC = -551.9184), and Model 3 only including Age showed the smallest BIC (BIC = -539.9583) corresponding to the BIC criteria preferring a smaller model than AIC.

Bidirectional elimination was used in the second model selection approach: Stepwise model selection. However, no matter the predictor pools were from first or second-order models, the selection all suggested that the best model was the same as Model 2 with Age, Smokstat, and Fat.

With the model validation of Model 1, 2, and 3, Model 2 (Figure 22) was selected as the best model. In Table 6, 7, Model 2 overall performed better with most of the criteria, including a best Pressp (Pressp = 28.65852) that its division by the number of observations (Pressp/n = 0.1127) was also the closest to its MSE (MSE = 0.1104), highest adjusted R square (Ra2 =0.104) with validation data, relatively consistent parameter estimation in training and validation data sets, and reasonable numbers of predictors, while MSPE (MSPE = 435048.2), SSE (SSE = 27.26162) were pretty similar among 3 models.

After fitting the whole dataset into Model 2, outliers and influential cases detection implied 0 outlier in response variable with studentized deleted residual compared to Bonferroni threshold under alpha = 0.1 and 44 outliers in predictor variables and revealed 2 possible influential cases with Cook's distance plot, the 36th and 296th cases (Figure 23). However, with the calculation of the average absolute difference in the fitted values, it suggested that the percentage change on the fitted value with or without the case is very small (0.040% respectively). Therefore, no case had an unduly large influence on prediction and thus all cases may be retained.

IV. Conclusions and Discussion

According to the previous analysis, several findings were presented below:

- Quetelet appeared to affect the level of beta-carotene plasma the most. In order to decrease the risk of developing cancer with higher beta-carotene plasma, we suggested that control of a lower Quetelet is needed; Age could be influential in determining the level of retinol plasma concentration. Apart from age, fat consumed per day and smoking status would possibly affect the levels of plasma retinol. However, the influences could be subtle.

- Number of calories consumed per day, Number of alcoholic drinks consumed per week, Cholesterol consumed, Dietary retinol consumed might not affect the

level of beta-carotene. Moreover, except for fat, most of the dietary intakes and sex did not influence the level of plasma retinol.

- To obtain a better estimation of concentrations of the two micronutrients, it does need a transformation of the level of plasma for the relationship between predictor variables and the level of plasma to be linear.

- The concentrations of the two micronutrients retinol and beta-carotene did not show a significant linear relationship.

V. Appendices

Appendix 1: Figures and tables (if any)

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

Figure 1: Summary Statistics

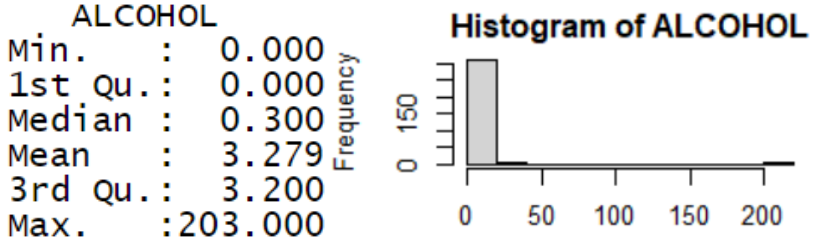

Figure 2 : Alcohol : Summary statistics and Histogram

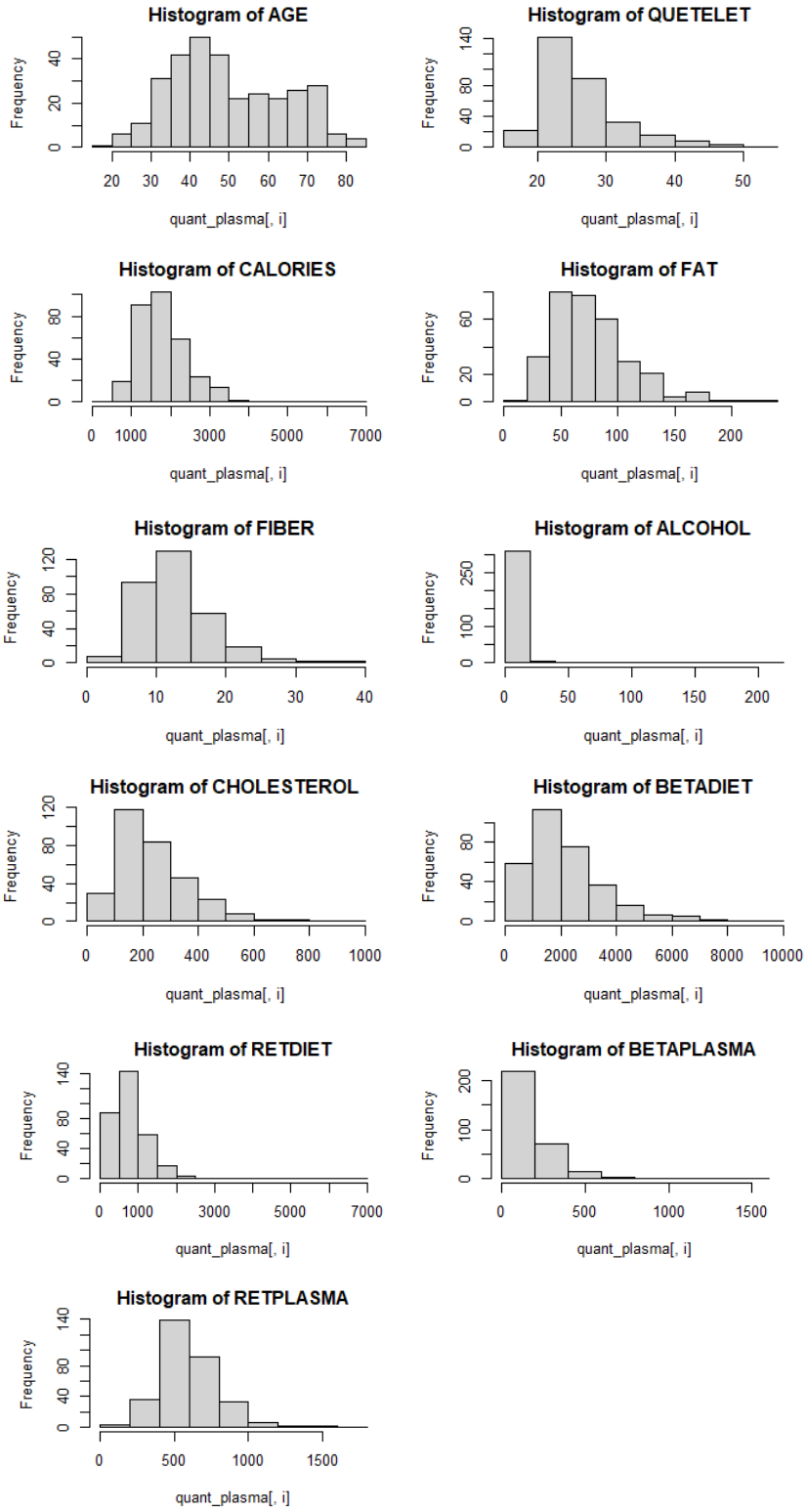

Figure 3 : Histograms of each quantitative variables

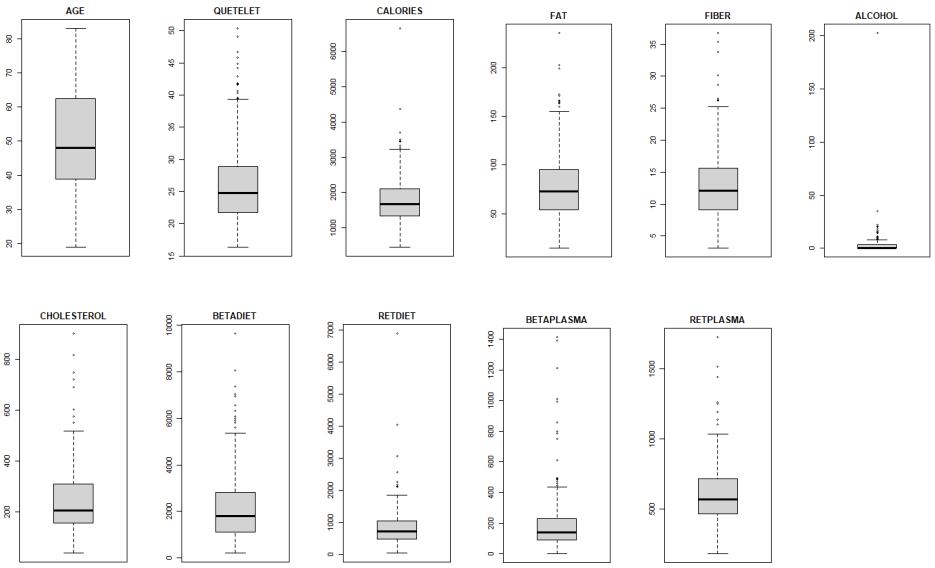

Figure 4 : Boxplots of each quantitative variables

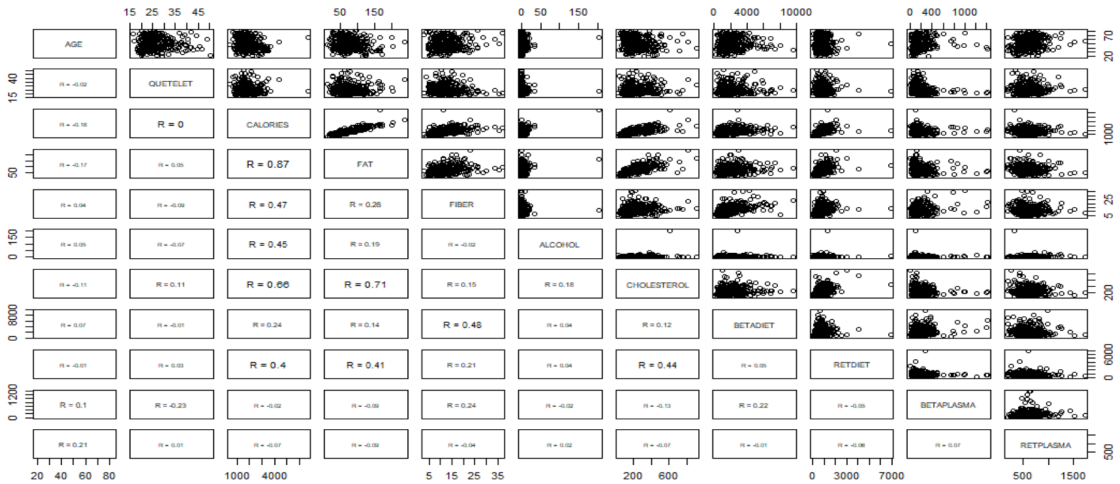

Figure 5 : Scatter plot matrix of correlation between each quantitative variables

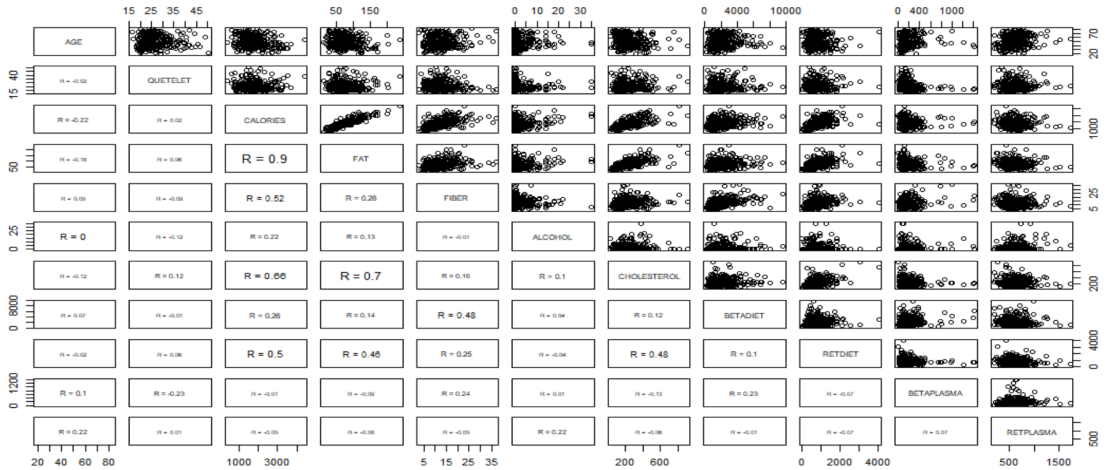

Figure 6 : Scatter plot matrix of correlation between each quantitative variables (Remove 2 extreme value)

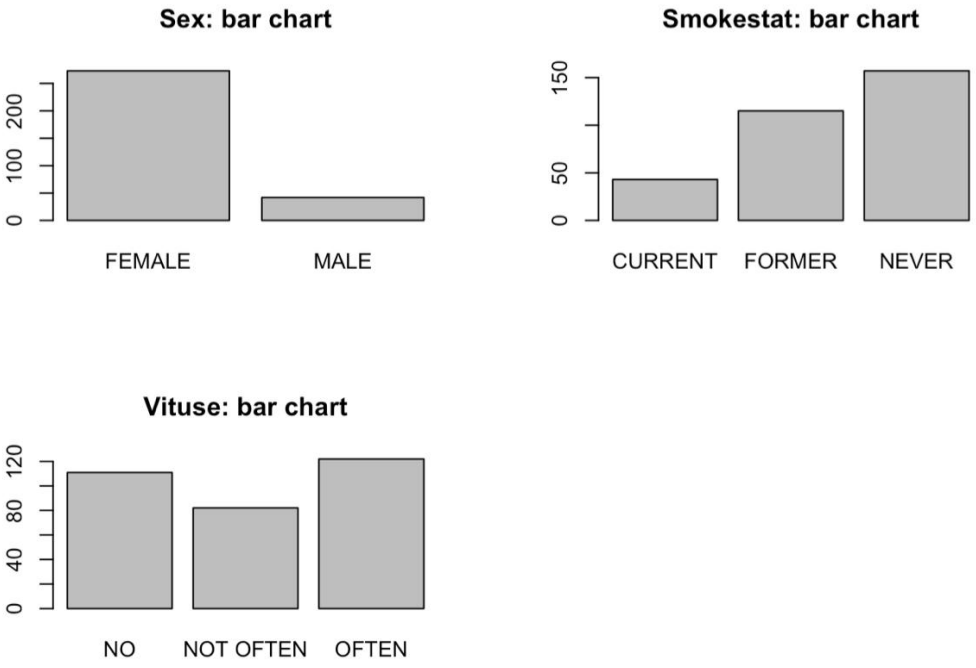

Figure 7 : Barplot of each qualitative variables

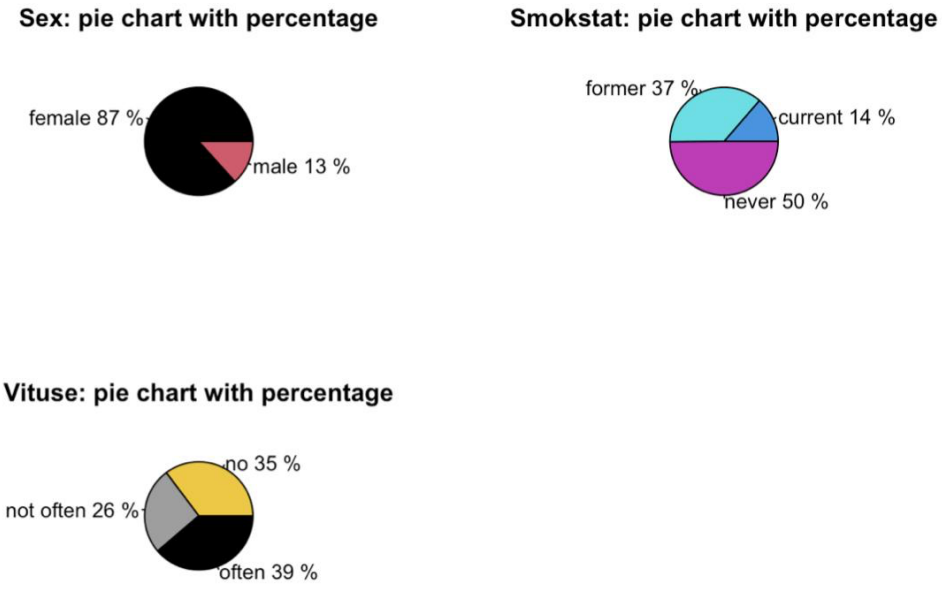

Figure 8 : Pie chart of each qualitative variables

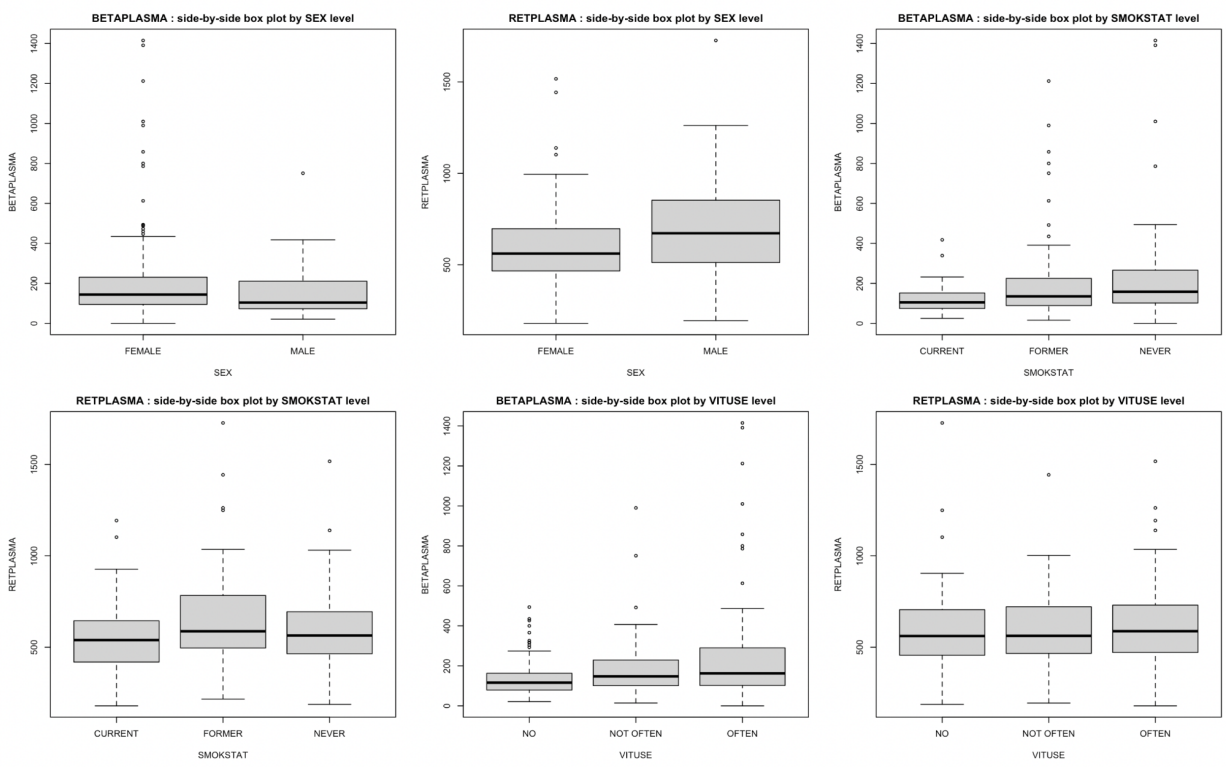

Figure 9 : Side-by-side boxplots of each qualitative variables with *Betaplasma* and *Retplasma*

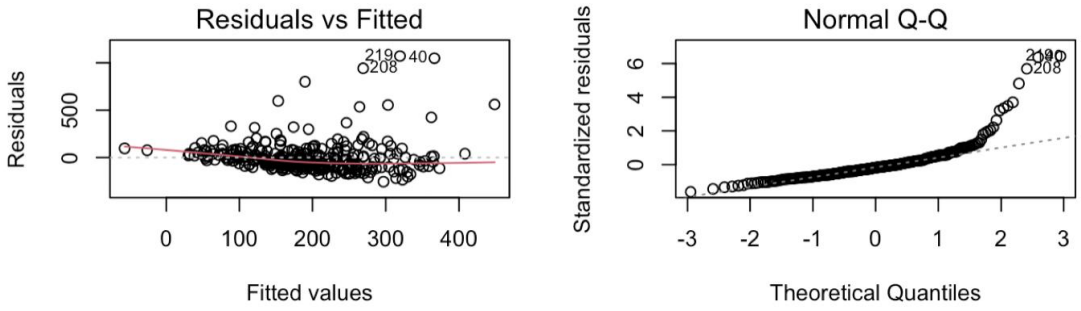

Figure 10 : Preliminary Full model of *Betaplasma*: Diagnostic plot

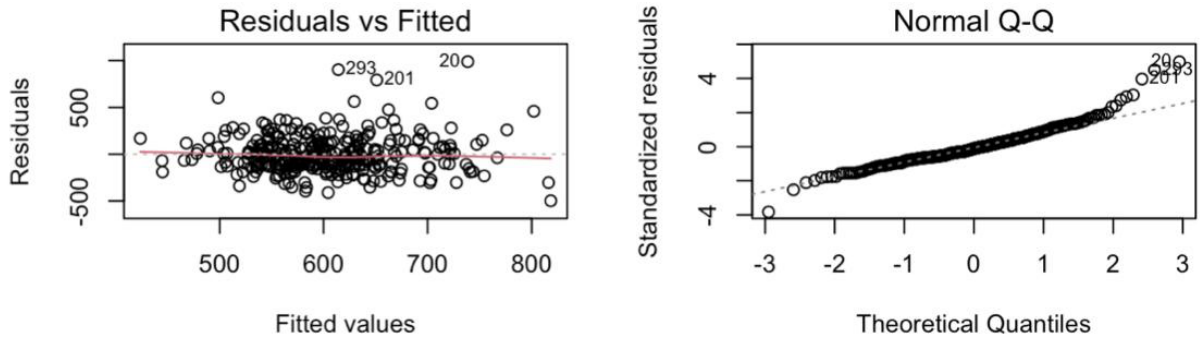

Figure 11 : Preliminary Full model of *Retplasma*: Diagnostic plot

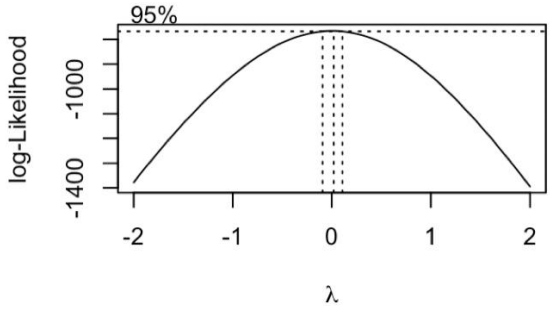

Figure 12 : *Betaplasma*: Box-Cox procedure plot

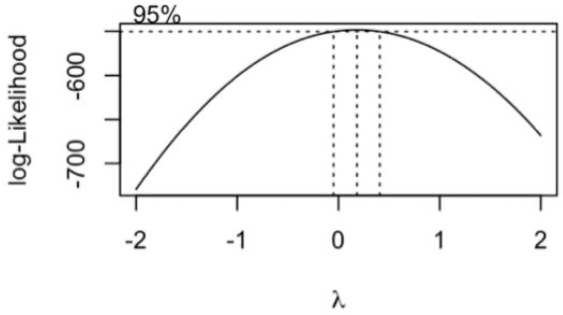

Figure 13 : *Retplasma*: Box-Cox procedure plot

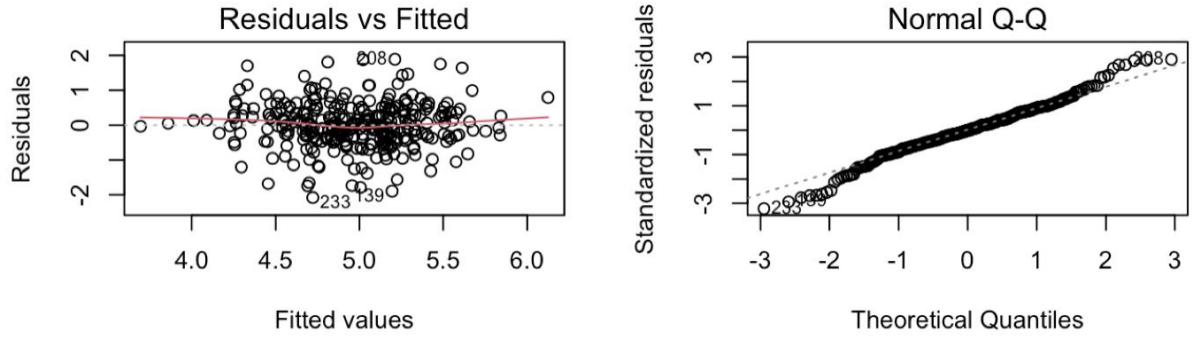

Figure 14 : Full model of log(*Betaplasma*) : Diagnostic plot

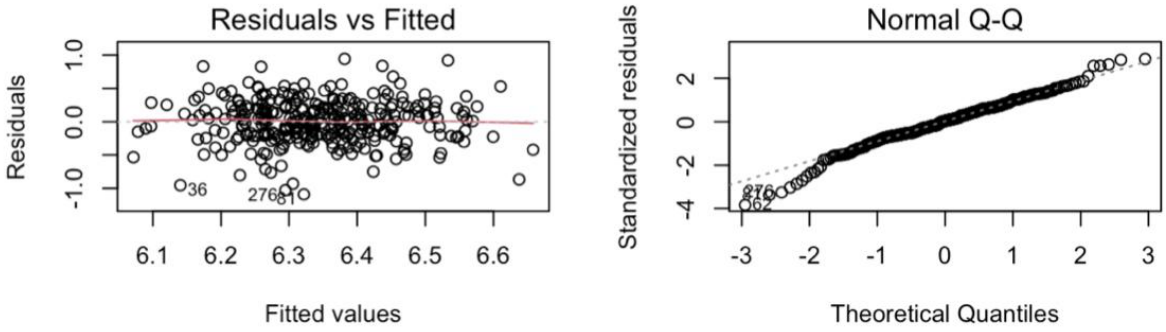

Figure 15 : Full model of log(*Retplasma*) : Diagnostic plot

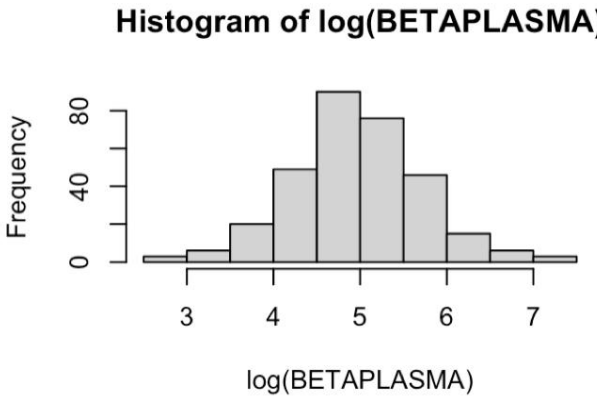

Figure 16 : Log(*Betaplasma*) : Histogram

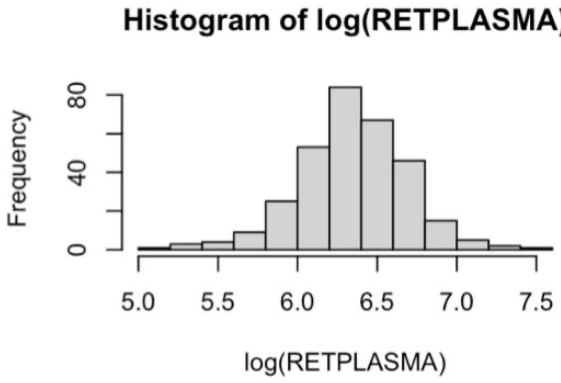

Figure 17 : Log(*Retplasma*) : Histogram

	AGE	QUETELET	CALORIES	FAT	FIBER	ALCOHOL	CHOLESTEROL	BETADIET	RETDIET	BETAPLASMA	RETPLASMA
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
AGE	1.000000000	-0.017463821	-0.176769399	-0.16947981	0.04485175	0.05158327	-0.11360597	0.071869247	-0.009610795	0.10112765	0.21167170
QUETELET	-0.017463821	1.000000000	0.003526964	0.04875033	-0.08762333	-0.07269543	0.11025724	-0.006603005	0.032055785	-0.22938737	0.01313865
CALORIES	-0.176769399	0.003526964	1.000000000	0.87184150	0.46548077	0.45146980	0.65917545	0.243376823	0.402491661	-0.02220696	-0.07332861
FAT	-0.169479813	0.048750329	0.871841504	1.00000000	0.27648356	0.18574310	0.70984794	0.143427853	0.412214814	-0.09164659	-0.09093779
FIBER	0.044851750	-0.087623326	0.465480772	0.27648356	1.00000000	-0.02011748	0.15396838	0.482643706	0.214611616	0.23595358	-0.04443071
ALCOHOL	0.051583268	-0.072695431	0.451469804	0.18574310	-0.02011748	1.00000000	0.18226398	0.039426478	0.044946841	-0.02221084	0.01713633
CHOLESTEROL	-0.113605966	0.110257245	0.659175452	0.70984794	0.15396838	0.18226398	1.00000000	0.115634801	0.443439304	-0.13030501	-0.07020134
BETADIET	0.071869247	-0.006603005	0.243376823	0.14342785	0.48264371	0.03942648	0.11563480	1.000000000	0.052866904	0.22477951	-0.01353942
RETDIET	-0.009610795	0.032055785	0.402491661	0.41221481	0.21461162	0.04494684	0.44343930	0.052866904	1.000000000	-0.04613524	-0.06280220
BETAPLASMA	0.101127650	-0.229387366	-0.022206957	-0.09164659	0.23595358	-0.02221084	-0.13030501	0.224779514	-0.046135240	1.00000000	0.07157724
RETPLASMA	0.211671702	0.013138652	-0.073328605	-0.09093779	-0.04443071	0.01713633	-0.07020134	-0.013539420	-0.062802201	0.07157724	1.00000000

Table 1: Quantitative Variables: Correlation Matrix

	(Intercept)	AGE	SEXMALE	SMOKSTATFORMER	SMOKSTATNEVER	QUETELET	VITUSENOTOFTEN	VITUSEOFTEN	CALORIES	FAT	FIBER	ALCOHOL	CHOLESTEROL	BETADIET	RETDIET	SSEp	R2	Ra2	Cp	AIC	BIC
null	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	135.68773	0	0	73.568179	-154.0064	-150.477
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	125.38785	0.07590872	0.07221235	51.006533	-171.9004	-164.8415
2	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	118.43776	0.12712991	0.1201189	36.432985	-184.2705	-173.6822
3	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	112.92261	0.16777581	0.15770859	25.281264	-194.287	-180.1693
4	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	109.57139	0.19247383	0.17939648	19.289774	-199.8789	-182.2317
5	1	0	0	0	0	1	1	1	0	1	0	0	0	1	0	106.70709	0.21358333	0.19759925	14.459411	-204.554	-183.3775
6	1	0	0	0	1	1	1	1	0	1	0	0	0	1	0	104.2151	0.23194898	0.21313957	10.516869	-208.509	-183.803
7	1	0	1	0	1	1	1	1	0	1	0	0	0	1	0	102.77856	0.24253608	0.22080556	9.09122	-210.0068	-181.7714
8	1	1	1	0	1	1	1	1	0	1	0	0	0	1	0	101.21785	0.25403827	0.22947986	7.369479	-211.8628	-180.0979
9	1	1	1	0	1	1	1	1	0	1	1	0	0	1	0	100.34463	0.26047383	0.23297079	7.287136	-212.0463	-176.752
10	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	99.9662	0.26326281	0.2326928	8.384711	-210.9984	-172.1747
11	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	99.70535	0.26518522	0.23150621	9.762679	-209.6569	-167.3037
12	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	99.45994	0.26699386	0.23019021	11.17746	-208.2779	-162.3953

Table 2 : First Order Model of *Betaplasma* : Best Subset Model Selection

Ra2 <int>	Cp <int>	AIC <int>	BIC <int>
10	10	10	7

Table 3 : Best Subset Model Selection Result : Row number

```
Call:
lm(formula = log(BETAPLASMA + 1) ~ QUETELET + VITUSE +
BETADIET +
    FAT + SMOKSTAT + SEX + AGE, data = beta_plasma_t)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95934 -0.34445  0.03701  0.39190  1.76567

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.254e+00  2.675e-01  19.639  < 2e-16 ***
QUETELET        -3.391e-02  6.720e-03  -5.046 8.85e-07 ***
VITUSENOT OFTEN  2.542e-01  1.072e-01   2.372 0.018497 *
VITUSEOFTEN      2.866e-01  9.682e-02   2.960 0.003382 **
BETADIET         1.085e-04  2.855e-05   3.802 0.000182 ***
FAT             -2.656e-03  1.260e-03  -2.107 0.036150 *
SMOKSTATFORMER   1.449e-01  1.300e-01   1.114 0.266318
SMOKSTATNEVER    2.805e-01  1.265e-01   2.217 0.027533 *
SEXMALE         -3.072e-01  1.290e-01  -2.381 0.018028 *
AGE              5.266e-03  3.031e-03   1.737 0.083572 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6451 on 242 degrees of freedom
Multiple R-squared:  0.2578,    Adjusted R-squared:  0.2302
F-statistic: 9.342 on 9 and 242 DF,  p-value: 3.615e-12
```

Figure 18: Stepwise Model 1 of *Betaplasma* : Summary

```
##
## Call:
## lm(formula = log(BETAPLASMA + 1) ~ CHOLESTEROL + FIBER + QUETELET +
##     VITUSE + BETADIET + AGE + RETDIET + SMOKSTAT + VITUSE:BETADIET +
##     CHOLESTEROL:RETDIET + AGE:RETDIET + VITUSE:SMOKSTAT + RETDIET:SMOKSTAT +
##     CHOLESTEROL:QUETELET + BETADIET:AGE + FIBER:SMOKSTAT + CHOLESTEROL:FIBER,
##     data = beta_plasma_t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97244 -0.33469  0.01207  0.36536  1.66268
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.133e+00  5.947e-01  10.313  < 2e-16 ***
## CHOLESTEROL                    -2.140e-03  1.496e-03  -1.431  0.15395
## FIBER                          -2.775e-02  2.923e-02  -0.949  0.34340
## QUETELET                       -4.421e-02  1.466e-02  -3.016  0.00286 **
## VITUSENOT OFTEN                -6.311e-02  2.944e-01  -0.214  0.83045
## VITUSEOFTEN                    -5.191e-01  2.979e-01  -1.743  0.08277 .
## BETADIET                        1.576e-04  1.351e-04   1.167  0.24460
## AGE                             2.126e-03  6.866e-03   0.310  0.75716
## RETDIET                        -2.040e-04  3.569e-04  -0.572  0.56811
## SMOKSTATFORMER                 -3.972e-01  3.918e-01  -1.014  0.31169
## SMOKSTATNEVER                  -3.612e-01  3.663e-01  -0.986  0.32520
## VITUSENOT OFTEN:BETADIET        2.226e-05  8.075e-05   0.276  0.78303
## VITUSEOFTEN:BETADIET            1.525e-04  7.083e-05   2.153  0.03236 *
## CHOLESTEROL:RETDIET            -1.248e-07  4.243e-07  -0.294  0.76888
## AGE:RETDIET                     8.134e-06  6.120e-06   1.329  0.18517
## VITUSENOT OFTEN:SMOKSTATFORMER  6.574e-01  3.312e-01   1.985  0.04836 *
## VITUSEOFTEN:SMOKSTATFORMER      7.090e-01  3.190e-01   2.222  0.02725 *
## VITUSENOT OFTEN:SMOKSTATNEVER   1.457e-01  3.269e-01   0.446  0.65626
## VITUSEOFTEN:SMOKSTATNEVER       4.893e-01  3.062e-01   1.598  0.11144
## RETDIET:SMOKSTATFORMER         -1.812e-04  2.971e-04  -0.610  0.54265
## RETDIET:SMOKSTATNEVER          -1.026e-04  2.672e-04  -0.384  0.70122
## CHOLESTEROL:QUETELET            6.264e-05  5.109e-05   1.226  0.22142
## BETADIET:AGE                   -2.800e-06  2.390e-06  -1.171  0.24272
## FIBER:SMOKSTATFORMER            3.364e-02  2.828e-02   1.189  0.23551
## FIBER:SMOKSTATNEVER             5.296e-02  2.668e-02   1.985  0.04837 *
## CHOLESTEROL:FIBER              -3.074e-05  6.912e-05  -0.445  0.65695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.642 on 226 degrees of freedom
## Multiple R-squared:  0.3136, Adjusted R-squared:  0.2376
## F-statistic: 4.129 on 25 and 226 DF,  p-value: 3.895e-09
```

Figure 19: Stepwise Model 2 of *Betaplasma* : Summary

	sse <dbl>	mse <dbl>	p <dbl>	cp <dbl>	press <dbl>
best subset 1	99.96620	0.4147975	9	12.972734	109.5064
best subset 2	103.49433	0.4241571	6	15.689211	110.4410
stepwise mode1	100.70131	0.4161211	7	9.511715	109.3472
stepwise model2	93.14226	0.4121339	17	10.932479	132.0332

Table 3 : *Betaplasma* : Internal Validation

	MSPE <dbl>	Press/n <dbl>	MSE <dbl>
Best Subset1	58248.69	0.4345493	0.4147975
Best Subset2	58256.81	0.4382581	0.4241571
Stepwise model1	58260.92	0.4339174	0.4161211
Stepwise model2	58228.58	0.5239414	0.4121339

Table 4 : Comparison of the 4 models: MSPE, Pressp/n, MSE

$$log(BETAPLASMA + 1) = 5.171 + 0.0053 * AGE - 0.3162 * SEXMALE$$
$$+ 0.1248 * SMOKSTATFORMER + 0.2537 * SMOKSTATNEVER -$$
$$- 0.0329 * QUETELET + 0.2477 * VITUSENOTOFTEN + 0.2807 * VITUSEOFTEN$$
$$- 0.00318 * FAT + 0.0128 * FIBER + 0.00009 * BETADIET$$

$$if Male, SEXMALE = 1; if Female, SEXMALE = 0$$

$$if Former, SMOKSTATFORMER = 1; if otherwise, SMOKSTATFORMER = 0$$

$$if Never, SMOKSTATNEVER = 1; if otherwise, SMOKSTATNEVER = 0$$

$$if NOTOFTEN, VITUSENOTOFTEN = 1; if otherwise, VITUSENOTOFTEN = 0$$

$$if OFTEN, VITUSEOFTEN = 1; if otherwise, VITUSEOFTEN = 0$$

Figure 20 : Fitted regression line of the best model of *Betaplasma*

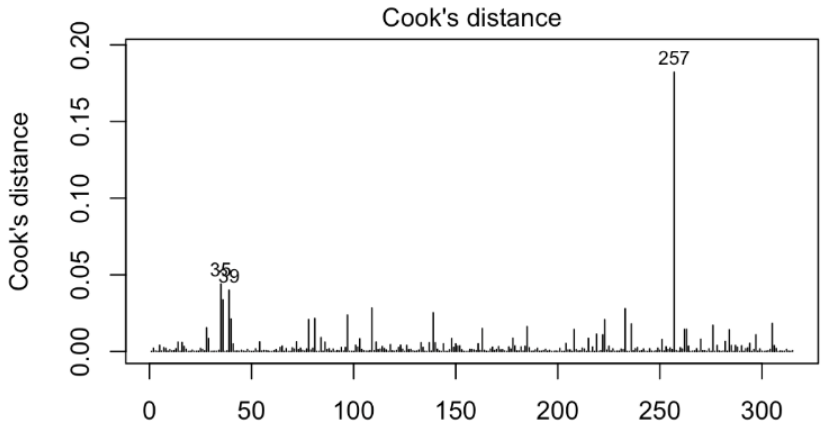

Figure 21 : Betaplasma's Final Model: Cook's Distance Plot

	(Intercept)	AGE	SEXMALE	SMOKSTATFORMER	SMOKSTATNEVER	QUETELET	VITUSENOTOFTEN	VITUSEOFTEN	CALORIES	FAT	FIBER	ALCOHOL	CHOLESTEROL	BETADIET	RETDIET	SSEp	R2	Ra2	Cp	AIC	BIC
null	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29.60318	0	0	11.2430086	-537.6699	-534.1405
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	28.29976	0.0440298	0.04020592	1.7405317	-547.0171	-539.9583
2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	27.73321	0.06316781	0.05564305	-1.25914	-550.1132	-539.5249
3	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	27.31759	0.07720749	0.06604467	-2.9269074	-551.9184	-537.8006
4	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	27.13028	0.08353489	0.06869335	-2.5798972	-551.6522	-534.0051
5	1	1	0	1	1	0	0	0	0	1	0	0	0	1	0	27.09431	0.08474999	0.06614735	-0.8973332	-549.9866	-528.81
6	1	1	0	1	0	0	0	0	1	1	1	1	0	0	0	27.05596	0.08604557	0.06366302	0.7642043	-548.3435	-523.6375
7	1	1	0	1	0	0	0	0	1	1	1	1	0	1	0	26.95609	0.08941904	0.06329582	1.8829083	-547.2754	-519.04
8	1	1	0	1	1	0	0	0	1	1	1	1	0	1	0	26.9177	0.09071602	0.06078075	3.5440815	-545.6346	-513.8697

9	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	26.89402	0.09151586	0.05772926	5.3351291	-543.8564	-508.5621
10	1	1	1	1	1	0	0	1	1	1	1	1	0	1	0	26.87563	0.09213688	0.05446621	7.1728939	-542.0287	-503.2049
11	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	26.86247	0.09258152	0.05099151	9.0567336	-540.1521	-497.799
12	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	26.85754	0.09274795	0.04719554	11.0132556	-538.1983	-492.3158

Table 5 : First Order Model of *Retplasma* : Best Subset Model Selection

	sse <dbl>	mse <dbl>	p <dbl>	cp <dbl>	press <dbl>
candidate_m1	27.09431	0.1101395	5	-2.897333	28.65852
candidate_m2	27.26162	0.1103709	4	-3.420886	28.40186
candidate_m3	28.00048	0.1124517	3	1.099474	28.67201

Table 6 : Betaplasma : Internal Validation

	MSPE <dbl>	Press/n <dbl>	MSE <dbl>
candidate_m1	435048.0	0.1137243	0.1101395
candidate_m2	435048.2	0.1127058	0.1103709
candidate_m3	435050.0	0.1137778	0.1124517

Table 7 : Comparison of the 4 models: MSPE, Pressp/n, MSE

$$log(RETPLASMA) = 6.2067 + 0.0043 * AGE + 0.0747 * SMOKSTATFORMER - 0.0451 * SMOKSTATNEVER - 0.00124 * FAT$$

$$if Former, SMOKSTATFORMER = 1; if otherwise, SMOKSTATFORMER = 0$$

$$if Never, SMOKSTATNEVER = 1; if otherwise, SMOKSTATNEVER = 0$$

Figure 22 : Fitted regression line of the best model of *Retplasam*

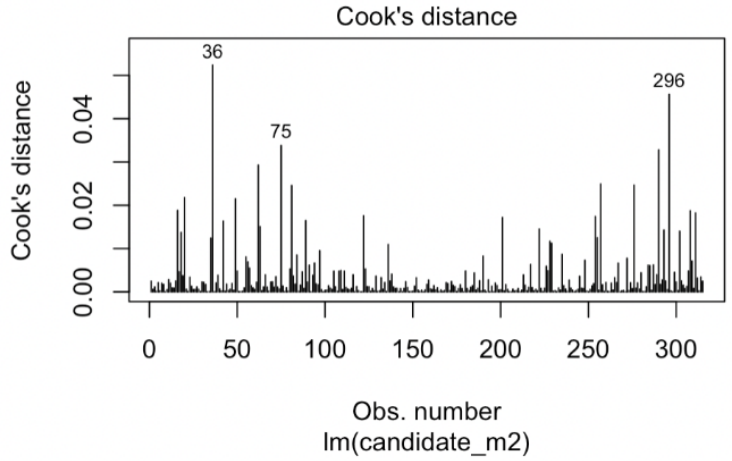

Figure 23 : Retplasma's Final Model: Cook's Distance Plot

References

- 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.