Determinants of Plasma Retinol and Beta-Carotene Levels

36-707 Fall 2001, Project 1

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
e-mail *@stat.cmu.edu*

Abstract

The associations of beta-carotene and retinol plasma concentration with eleven personal variables were studied in 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 using multiple regression analysis. Retinol levels were positively related to Age and number of alcoholic drinks consumed per week, with a multiple correlation coefficient of 0.3. Beta-carotene were positively related to vitamin use, smoking status and gram of fiber consumed per day. On the other hand Cholesterol consumed per day and Quetelet index were negatively related with Beta-carotene levels and a multiple correlation coefficient of 0.47. Even though these findings confirm the importance of some of predictors cited in the literature and identified vitamin use and Queletet index as new predictors of micronutrient levels, a better understanding of this relationship require further study. Finally, plasma concentrations of retinol and beta-carotene were not correlated.

1 Introduction

Beta-carotene and retinol are among the most widely studied compounds in various populations, for both human plasma (or serum) concentrations and dietary intake [1] [2, 3, 4, 5]. This situation dues to their inverse relationship with the development of several diseases, *e.g.*, cancer, cardiovascular disease, and cataracts [6, 7, 8].

A few studies have suggested that drinking and smoking habits, dietary intake, sex and age influence plasma concentration of carotenoids, and to a lesser extent, the concentrations of retinol [2, 3, 4]. The present study was designed to establish which other personal characteristics lead to lower levels of plasma concentration of the micronutrients retinol and beta-carotene and how concentrations of the two micronutrients are related to each other [9].

The report is presented giving first a brief description of data, second, the methods and criterions that have been used to analysis the data, next, the results that were found and finally, a discussion and recommendations for further analyses.

2 Description of data

A cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects (N = 315) were 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 [10].

3 Methods

An exploratory data analysis was carried out for all the variables registered in the study in order to detect some characteristics of the variables or subjects that later helped to address the study of the relations we are interested on.

Multiple regression analysis was used to study the relation between plasma beta-carotene and age, sex, smoking status, use of vitamins, calories, fat, fiber, cholesterol and dietary beta-carotene consumed daily, alcohol consumed weekly and Quetelet index [2]. The relation between the same independent variables except for dietary beta-carotene consumed daily, that was replaced by dietary retinol consumed daily, and the Plasma retinol levels were also studied using multiple regression analysis. Incremental sum of squares were used to assess the significance of independent variables and to simplify the model.

The statistical package Splus using the functions *lm* and *anova* were used for the multiple regression analysis, and all the statistical tests were performed with a level of significance of $\alpha = 0.05$ unless otherwise stated.

[1] For this project, I suppose that the paper *Determinants of plasma levels of beta-carotene and retinol* [1] did not exist, in order to make important this related study

[2] weight (kg)/height$^2(m)$

4 Results and Discussion

The population considered in this report consisted of 273 women 42 men (13 per cent) and (87 per cent). Their age are between 19 and 83 years with a median of 48 years. When the study was done, 65 per cent were taking vitamins. The mean Quetelet index, used as a measure of obesity or relative weight, was 26.16, with a range of 16.33 to 50.40. Fourteen per cent were currently smokers and 37 per cent were former smokers. One subject in this study presents high consumed of alcohol, at this is not clear if this true or it was a mistake in the entry of data, for this reason it was eliminate of the dataset for the regression analyses.

The histogram of the plasma beta-carotene levels was skewed positively but was closer to symmetric on the log scale (Figure 1). Vitamin use, grams of fiber consumed per day, smoking status, Quetelet index and Cholesterol were statistically associated with plasma beta-carotene levels ($R^2 = 0.23$). Of these variables, low levels of beta-carotene were associated with high values of Quetelet and Cholesterol.

In the case of retinol, the histogram was fairly skewed positively but in a log scale it was fairly symmetric (Figure 1). The only variables that were statistically associated were Age and number of alcoholic drinks consumed per day ($R^2 = 0.09$).

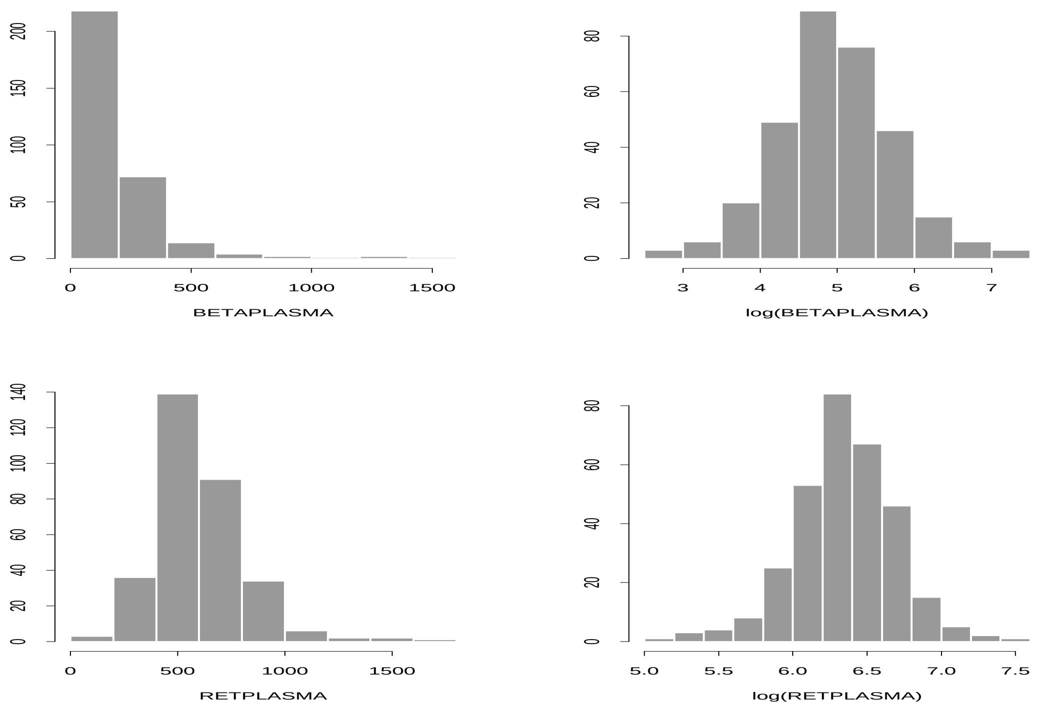

Figure 1: Histogram for the variables BETAPLASMA and RETINOL in the original and log scale

There was no evidence of a statistically significant interaction of any of the predictors with sex, current smoking status and vitamin use for any of the plasma micronutrients. This finding differs to the findings of Stryker et.al.[4].

Finally, There were not association between plasma retinol and beta-caretene levels ($r = 0.07$, p-value = 0.2169)

It is important to note that in most of the personal characteristics, some information was inconsistent, that is, the variables took some values that hard to believe (commonly over-reported), for instance, a subject with dietary retinol consumed over 6000 when the maximum allowed before the retinol becomes a toxin is 3000. Therefore more detail checking of these data should be done before trying to do further analysis. Another inportant feature of the plasma concentrations of the micronutrients is that they varied widely from subject to subject making more difficult to explain their variability, it can be seen with the low values of the R^2 for the regressions, 23 and 9 per cent, respectively. As possible solution we may try to include more subjects in the study and a more detailed checking of the information, before publishing these data and analysis in a good medical journal.

5 Conclusions

We may summarize the findings as follow:

- The personal characteristics that lead to lower levels of plasma beta-carotene are Quetelet index and cholesterol, therefore subjects with high Quetelet index and Cholesterol have more probability to develop several diseases, *e.g.*, cancer, cardiovascular disease, and cataracts [6, 7, 8]. Thus, we can recommend to the general public to take care of their consume of cholesterol and to control their obesity. It is important also that they control (reduce) the consume of alcohol.
- Dietary intake and sex did not influence plasma concentration of beta-carotene and retinol, contrary to former studies [6, 7, 8].
- There were not a significant linear association between plasma retinol and beta-caretene levels.

6 Technical Appendices

The analysis was done in two steps:

- Exploratory data analysis.
- Multiple Regression Analysis.

The statistical package Splus using the functions *lm* and *anova* were used for the multiple regression analysis. All the statistical tests were performed with a level of significance of $\alpha = 0.05$ unless otherwise stated, and the Bonferroni correction was used for joint confidence regions.

1. Exploratory Data Analysis (EDA).

   The data file contains 315 observations on 14 variables. Variable names:

   AGE: Age (years).
   SEX: Sex (1=Male, 2=Female).

SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker).
QUETELET: Quetelet (weight/height^2).
VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No).
CALORIES: Number of calories consumed per day.
FAT: Grams of fat consumed per day.
FIBER: Grams of fiber consumed per day.
ALCOHOL: Number of alcoholic drinks consumed per week.
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

The EDA is carried out using boxplots, histograms and scatterplots for the continuous variables. A different approach is used for the categorical variables, in this case it is important to study if there are differences in the amount of plasma beta-carotene or retinol, between the different categories of the variables Sex, Smoking status and Vitamin use. For this reason, Boxplots for each of these two continuous variables by each category of the variables Sex, Smoking status and Vitamin use have been used.

(a) Boxplots

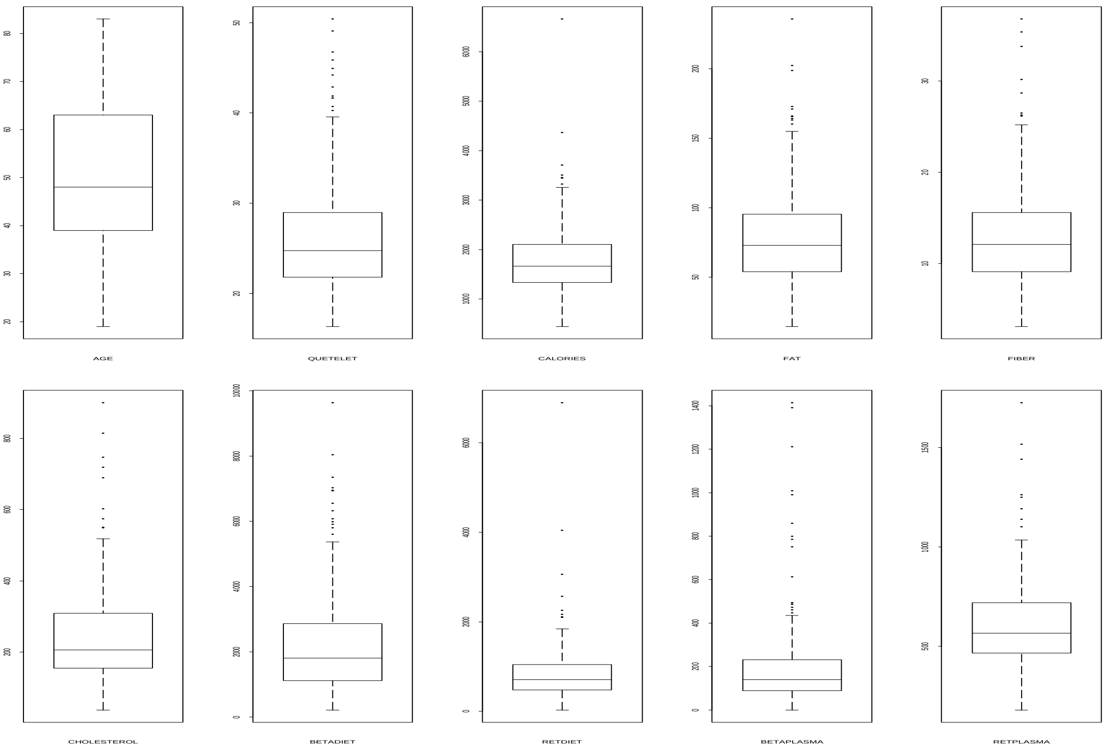

Figure 2: Boxtplot for the continouos variables

The boxplot of the variables *Smoking Status, Queletet index, Calories, Fat, Fiber, Cholesterol, Dietary beta-carotene consumed, Diatary retinol consumed, Plasma beta-carotene* and *Plasma Retinol* in Figure 2 clearly reveal the skewness of the distributions: The lower whisker is shorter than the upper whisker; and there are several outside observations at the upper end of the distributions, but not at the lower end. The boxplot of Age reveals that the distribution of *Age* is fairly symmetric. Figure 3 shows the boxplot for *Alcohol* without the subject 62 (since this subject, that is an outlier, hides other possible outliers), it reveals that the distribution for *Alcohol* is positively skewed and there are several outliers.

Based on the boxplots and since skewing in the dependent or independent variables may suggest a transformation that improves the regression model, in the second part of this analysis some transformations should be studied.

Figure 3: Boxtplot for the *Alcohol*

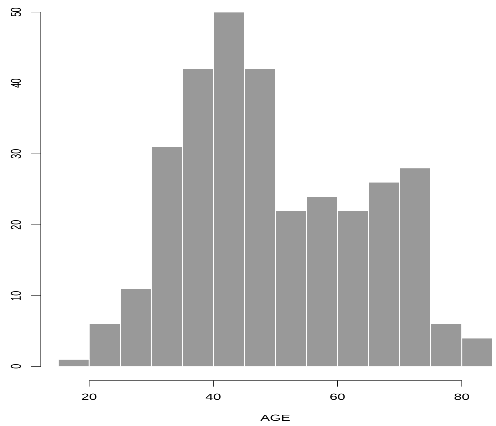

Figure 4: Histogram for *Age*

(b) Histograms

For most of the variables the histogram shows more less the same information that the boxplot except for the variable *Age*. As can be seen in Figure 4, there is an apparent bimodality in distribution of *Age*, situation that is not reveal by the boxplot.

(c) Scatterplot Matrix

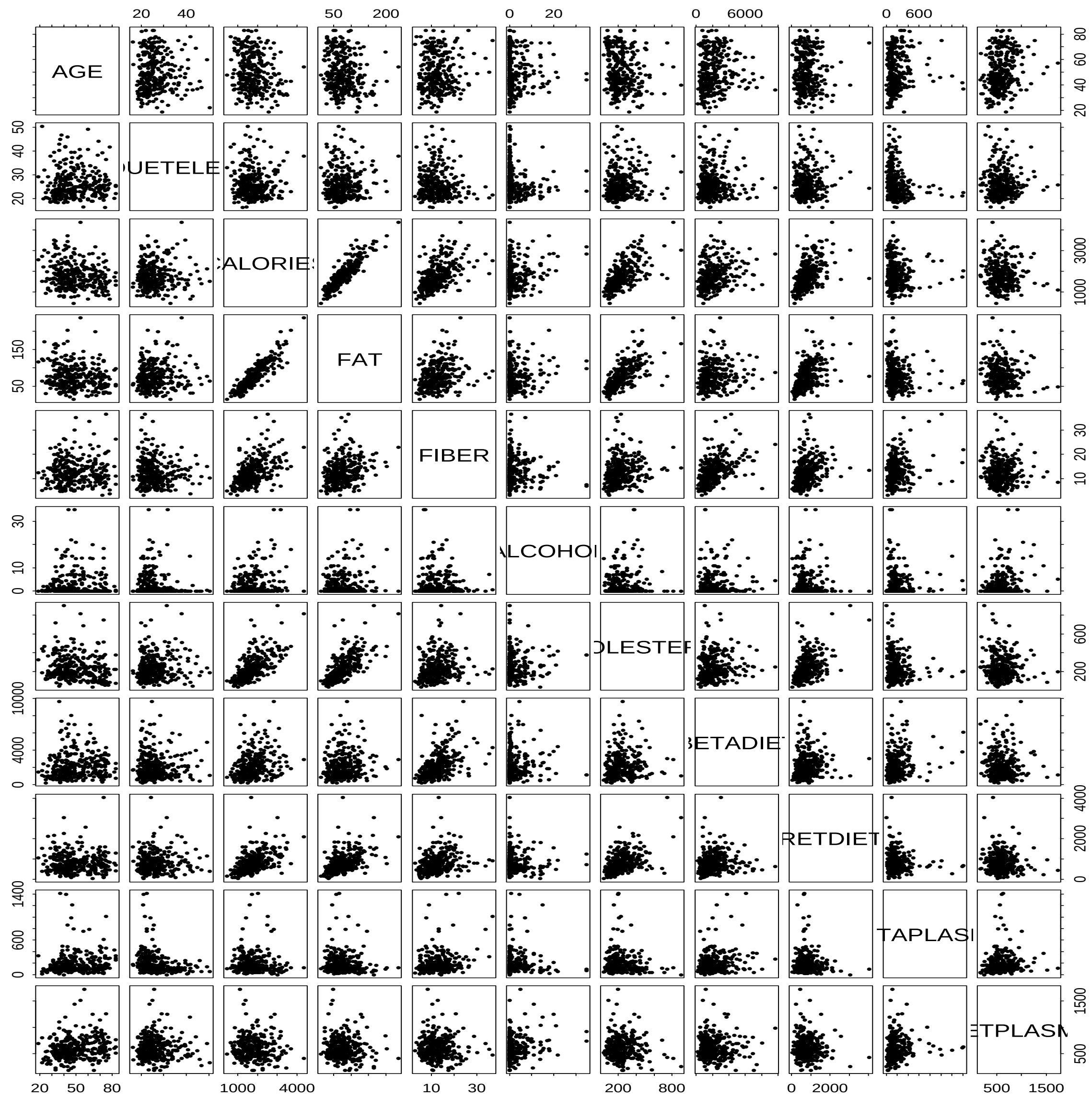

Figure 5: Scatterplot Matrix

The Scatterplot Matrix for *Smoking Status, Queletet index, Calories, Fat, Fiber, Cholesterol, Dietary beta-carotene consumed, Diatary retinol consumed, Plasma beta-carotene* and *Plasma*

Retinol is shown in Figure 5, in this case I did not include the observations 62 and 171 since they are "extreme" outliers for the variables *Alcohol* and *Retdiet* respectively, and therefore they may hide some relations. As reveal in this graph:

- there is a clear increasing linear relationship between *Calories* and *Fat*.
- there is a less pronounced increasing, but still fairly clear, relationship between *Cholesterol* and the variables *Calories* and *Fat* separately
- there is a lesser pronounced increasing relationship between *Dietary retinol consumed* and the variables *Cholesterol* and *Fat* separately; It is also true between *Dietary beta-carotene consumed* and *Fiber*.
- For the rest of pairs of variables, It seems that there are some weaker relationships and at this step of our study may not be clear enough. Nevertheless, later some transformations or some further study of the outliers should be made to find out some relationships that may be not clear right now.

(d) One-Way ANOVA's using Boxplots

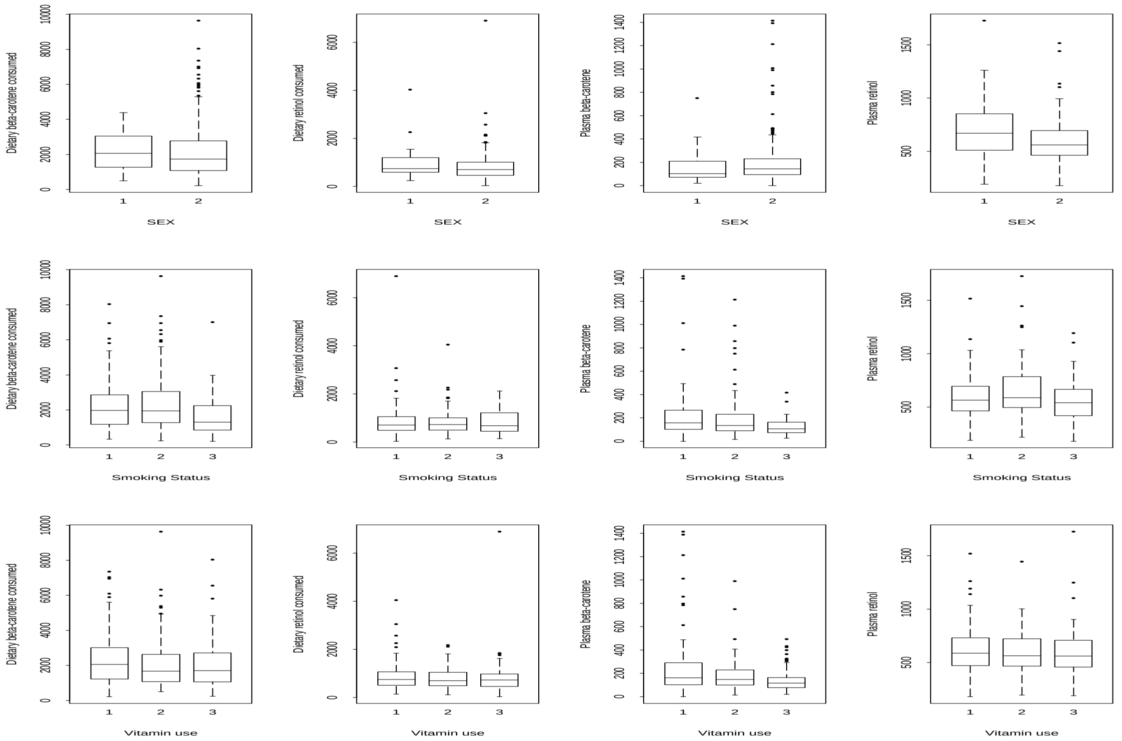

Figure 6: Boxplot categorical

All the boxplots in Figure 6 show that there is not a clear separation **between** the medians of *Plasma beta-carotene* or *Plasma retinol* for each of the categories of the variables *Sex*, *Smoking* *status*, and *Vitamin use*. This situation may be not give us too much hope in order to answer the questions of this study, but as it has been notice in the literature [4, 5, 1], interactions of this variables may be helpful to explain some relationship between our variables of interest, therefore in our next stage of analysis interactions of this variables could still be used.

2. Multiple Regression Analysis

We want to fit two models, one for plama Beta-carotene and one for plasma retinol, but before to fit them the next two steps were done:

- ⋄ Creation of indicator variables for each of the categories of the three categorical variables: *Sex*, *Smoking status*, and *Vitamin use*. This was done because contemplating a single predictor variable suggests that these treatments are ordered and that the effect of changing from one to another is the same for all of them, and these assumptions (ordered categories and equal spacing) are not justifiable.

```
> P.S1<-ifelse(SEX==1,1,0)
> P.SM1<-ifelse(SMOKSTAT==1,1,0)
> P.SM2<-ifelse(SMOKSTAT==2,1,0)
> P.V1<-ifelse(VITUSE==1,1,0)
> P.V2<-ifelse(VITUSE==2,1,0)
```

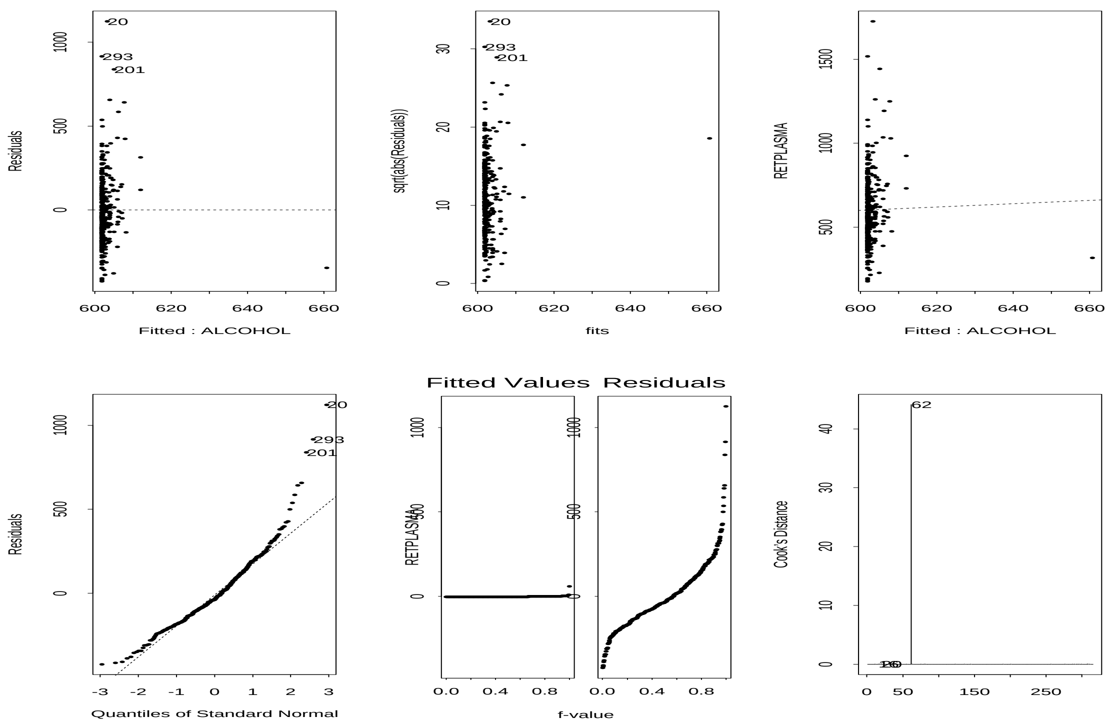

Figure 7: Standard diagnostic plots for regressing *RETPLASMA* on *Alcohol*

◇ As we saw in the EDA, the variable *Alcohol* has an "extrem" outlier that covers all possible relationships between alcohol and the rest of the variables, therefore this observation is eliminated in all further analyses. The necessity of this elimination is also clear with the analysis of residuals (Figure 7) for any regression on *Alcohol*, for instance lm(RETPLASMA ∼ ALCOHOL), since this observation is clearly a high leverage point.

Dependable variable: BETAPLASMA

- As mentioned in the EDA, the high asymmetry of the distribution of *Plasma beta-carotene* shows the need of transform the data prior to the analyses. This fact is confirmed regressing BETAPLASMA on BETADIET (or any other of the continuous variables) and noticing the asymmetry of the residuals on the original scale (QQ Normal plot of the residuals in the Figure 8)[3].

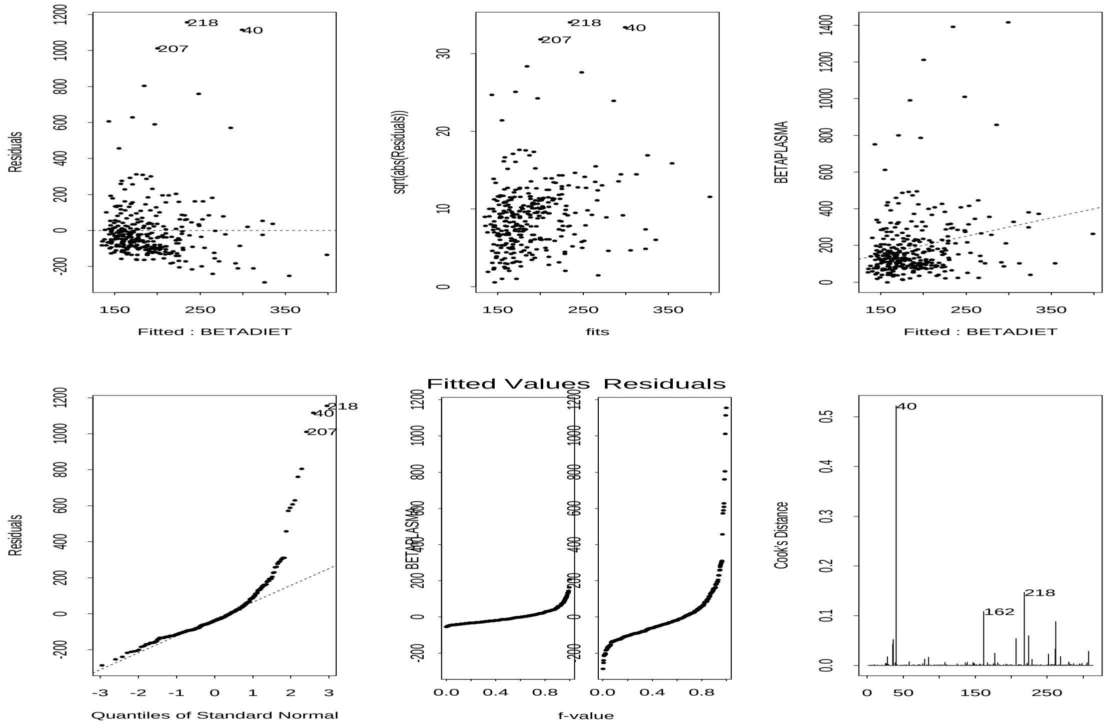

Figure 8: Standard diagnostic plots for regressing *BETAPLASMA* on *BETADIET*

The selected transformation for *Plasma beta-carotene* is the log transformation since the distribution of this variable is positive skew, and this transformation make the distribution more symmetric, as can be seen in Figure 9.

- To start the multiple regression analysis, matrix of correlations for the continuous variables is computed:

[3]Note how the points lie above the comparison line in both tails of the distribution.

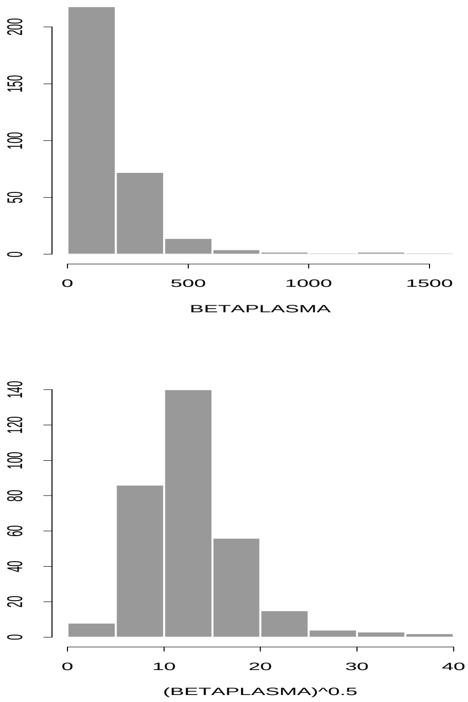

Figure 9: Histogram of *BETAPLASMA* and three transformations

	AGE	QUETELET	CALO	FAT	FIBER	ALCOHOL	CHOLEs	BETDI	RETDI	BETPL	RETPL
AGE	1.00	-0.02	-0.22	-0.18	0.05	0.00	-0.12	0.07	-0.01	0.10	0.22
QUETELET	-0.02	1.00	0.02	0.05	-0.09	-0.12	0.12	-0.01	0.03	-0.23	0.01
CALORIES	-0.22	0.02	1.00	0.90	0.52	0.22	0.66	0.25	0.42	-0.01	-0.05
FAT	-0.18	0.05	0.90	1.00	0.28	0.13	0.70	0.14	0.41	-0.09	-0.08
FIBER	0.05	-0.09	0.52	0.28	1.00	-0.01	0.16	0.48	0.22	0.24	-0.05
ALCOHOL	0.00	-0.12	0.22	0.13	-0.01	1.00	0.10	0.04	0.00	0.01	0.22
CHOLESTEROL	-0.12	0.12	0.66	0.70	0.16	0.10	1.00	0.11	0.44	-0.13	-0.06
BETADIET	0.07	-0.01	0.25	0.14	0.48	0.04	0.11	1.00	0.05	0.23	-0.01
RETDIET	-0.01	0.03	0.42	0.41	0.22	0.00	0.44	0.05	1.00	-0.04	-0.06
BETAPLASMA	0.10	-0.23	-0.01	-0.09	0.24	0.01	-0.13	0.23	-0.04	1.00	0.07
RETPLASMA	0.22	0.01	-0.05	-0.08	-0.05	0.22	-0.06	-0.01	-0.06	0.07	1.00

The two last rows of the matrix of correlations contains the simple correlation of each of the two dependent variables with each of the independent variables. The four variables *QUETELET*, *FIBER*, *CHOLESTEROL*, and *BETADIET* have the highest correlations with *BETAPLASMA*, and the two variables *AGE* and *ALCOHOL* have the highest correlations with *RETPLASMA*. The impact of these correlations on the regression results is noted as the analysis goes forward. On the other hand, there appears to be almost no correlation between *BETAPLASMA* and *RETPLASMA*. To test if there is not correlation between *BETAPLASMA* and *RETPLASMA*, the confidence interval of the correlation were found:

```
> tanh(atanh(cor(BETAPLASMA,RETPLASMA))+c(-1,1)*qnorm(0.975)/sqrt(length(BETAPLASMA)))
[1] -0.04089959  0.17836351
```

Since the confidence interval contains zero, we conclude that there is not relation between these to variables.
Another important thing is that since the variables *FAT* and *CALORIES* are highly correlated, therefore we are aware of possible collinearity problems if both variables are included at the same time in a model.

- Two possible approaches for starting the regression analysis are: use the full model and eliminate subsets until we get a good model or use the model founded in the preliminary analysis done by this unit [4][9] and add more variables until get a good model. In this study, the analysis starts using the full model.
  It is necessary to mention here that since I do not have a previous knowledge of the problem and the literature found does not provide it neither, in the next process of determining an appropriate model, there is not an "logical" ording of the variables according to their relative importance that allows to simplify this process.
  The results of the multiple regression analysis using: all the independent variables, and the dependent variable *log(BETAPLASMA +1)*[5] are summarized in the next Splus output:

```
Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)    5.0881    0.2881   17.6612   0.0000
        AGE    0.0056    0.0031    1.7840   0.0754
   QUETELET   -0.0318    0.0069   -4.6358   0.0000
   CALORIES   -0.0001    0.0002   -0.4219   0.6734
        FAT    0.0003    0.0034    0.0920   0.9268
      FIBER    0.0253    0.0118    2.1490   0.0324
    ALCOHOL    0.0039    0.0093    0.4173   0.6768
CHOLESTEROL   -0.0011    0.0004   -2.4296   0.0157
   BETADIET    0.0001    0.0000    2.0522   0.0410
       P.S1   -0.1630    0.1336   -1.2201   0.2234
      P.SM1    0.2471    0.1288    1.9181   0.0560
      P.SM2    0.1883    0.1309    1.4384   0.1514
       P.V1    0.2545    0.0970    2.6234   0.0091
       P.V2    0.2767    0.1056    2.6213   0.0092

Residual standard error: 0.7013 on 300 degrees of freedom
Multiple R-Squared: 0.2477
F-statistic: 7.597 on 13 and 300 degrees of freedom, the p-value is 5.476e-013
```

  The test of the composite hypothesis that all 13 regression coefficients are zero is highly significant. The coefficient of determination is just .2477, thus 25 % of the variability of the log-transformated *BETAPLASMA* can be associated with the variation of these 13 independent variables.
  The t test of the partial regression coefficients $H_0 : \beta_j = 0$ (using the bonferroni correction) would seem to suggest that ten of the 13 independent variables are unimportant and could be dropped from the model, however removing one variable from the model will cause the regression coefficients to change, therefore a simultaneous test was used to eliminate subsets of variables. The first subset to drop, is the set of variables with a really high p-value: *CALORIES*, *FAT* and *ALCOHOL*. The Splus output[6] for this hypothesis is:

[4] Unit in the research hospital where I am working at.
[5] the transformation $\log(Y+1)$ is used because one subject has *BETAPLASMA* equals zero
[6] It was a little bit modified, since it was too long.

```
Analysis of Variance Table

Response: log(BETAPLASMA +1)
        Terms  Resid. Df     RSS  Test Df Sum of Sq   F Value     Pr(F)
1 Small Model        303 147.7937
2  FULL MODEL        300 147.5476        3 0.2460758 0.1667772 0.9187273
```

Since the p-value is greater than $\alpha = 0.05$, we can drop these tree variables from the model. The results for the multiple regression analysis without the independent variables *CALORIES*, *FAT* and *ALCOHOL* are summarized in the following Splus output:

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   5.0436   0.2679    18.8274   0.0000
        AGE   0.0061   0.0030     2.0624   0.0400
   QUETELET  -0.0322   0.0067    -4.7658   0.0000
      FIBER   0.0208   0.0087     2.3997   0.0170
CHOLESTEROL  -0.0013   0.0003    -3.8423   0.0001
   BETADIET   0.0001   0.0000     2.1040   0.0362
       P.S1  -0.1590   0.1301    -1.2217   0.2228
      P.SM1   0.2515   0.1280     1.9650   0.0503
      P.SM2   0.1926   0.1299     1.4821   0.1393
       P.V1   0.2455   0.0946     2.5941   0.0099
       P.V2   0.2722   0.1042     2.6109   0.0095

Residual standard error: 0.6984 on 303 degrees of freedom
Multiple R-Squared: 0.2464
F-statistic: 9.907 on 10 and 303 degrees of freedom, the p-value is 2.287e-014
```

The coefficient of determination in this case is .2464, almost the same that including all the available variables. Once again, there are some variables that could be dropped from the model, therefore the last procedure was done again with the variables: *P.S1* , *P.SM1* [7], and *P.SM2* [8]:

```
Analysis of Variance Table

Response: log(BETAPLASMA + 1)

        Terms Resid. Df      RSS              Test Df Sum of Sq  F Value     Pr(F)
1 Small Model       306 150.6049
2   Big Model       303 147.7937 +P.S1+P.SM1+P.SM2  3  2.811183 1.921121 0.1261315
```

This case is similar to the last case and we can drop the variables *P.S1*, *PSM.1* and *P.SM2*. And the summary for the small model is

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   5.1651   0.2620    19.7109   0.0000
```

[7] Note that P.SM2 does not have a high p-value, but since we are dropping P.SM1, P.SM2 should also be dropped

[8] In the output, the model without the variables we are dropping is called "Small Model"

AGE	0.0057	0.0028	2.0599	0.0402
QUETELET	-0.0300	0.0067	-4.4746	0.0000
FIBER	0.0229	0.0086	2.6553	0.0083
CHOLESTEROL	-0.0014	0.0003	-4.6055	0.0000
BETADIET	0.0001	0.0000	2.2167	0.0274
P.V1	0.2880	0.0932	3.0893	0.0022
P.V2	0.3035	0.1035	2.9315	0.0036

```
Residual standard error: 0.7016 on 306 degrees of freedom
Multiple R-Squared: 0.2321
F-statistic: 13.21 on 7 and 306 degrees of freedom, the p-value is 6.994e-015
```

The results in the last Splus output would seem to suggest that the variables *AGE* and *BETADIET* could be dropped from the model, but the F test in this case indicates us that these two variables may remain in the model, as can be seen in the next summary. With this situation an appropriate model would be:

$$Log(BETAPLASMA_i + 1) = 5.1651 + .0057 \cdot AGE - .0300 \cdot QUETELET + .0229 \cdot FIBER$$
$$- .0014 \cdot CHOLESTEROL + .0001 \cdot BETADIET + .2880 \cdot P.V1 + .3035 \cdot P.V2 + \epsilon_i$$

```
Analysis of Variance Table

Response: log(BETAPLASMA + 1)

        Terms Resid. Df      RSS          Test Df Sum of Sq  F Value       Pr(F)
1 Small Model       308 155.3900
2   Big Model       306 150.6049 +AGE+BETADIET  2  4.785158 4.861259 0.008348876


Analysis of Variance Table

Response: log(BETAPLASMA + 1)

        Terms Resid. Df      RSS Test Df Sum of Sq  F Value      Pr(F)
1 Small Model      307  152.6933
2   Big Model      306  150.6049 +AGE  1  2.088477 4.243382 0.04024901
```

The model we have obtained so far seems to be appropiate according with the preliminar analysis [9], but since the coefficient of determination is small, some other approaches are studied.

- In the EDA, it was mentioned that interactions between some of the variables could be studied. Based on some previous studies [4, 5], all main effects and first-order interactions of the predictors with *Sex*, *smoking status*, and *Vitamin use* were included in the initial model. The multivariate regression analysis was done as before with the only difference that the **principle of marginality**, that is, whenever an interaction is fitted, all of the lower-order interactions and simple terms should also be fitted [11] was used. The results founded with this model were equal to the results already found, therefore there is not gain with this process.
- Even though the symmetry property is not necessary for the predictors, asymmetry may suggest a transformation that improves the regression model, therefore the log transformation was used with the variable *BETADIET*. It is important to mention that transformations for other

variables were done but, they did not present an improvement. Carrying out the multivariate regression analysis using the extra-sum-of-squares principle, a similar model was obtained, the only difference is: that *Age* could be dropped from the model. The summary of the regression and the last F test to decide if the Age should be dropped from the model are shown in the Splus output below. It is worth mentioning that the p-value is near the cut-off point 0.05.

```
Analysis of Variance Table

Response: log(BETAPLASMA + 1)

                                                             Terms Resid. Df      RSS
1       log(BETADIET) + FIBER + v1 + v2 + QUETELET + CHOLESTEROL       307 152.6403
2 AGE + log(BETADIET) + FIBER + v1 + v2 + QUETELET + CHOLESTEROL       306 150.8405

  Test Df Sum of Sq  F Value      Pr(F)
1
2 +AGE  1  1.799749 3.651028 0.05696893

Coefficients:
                Value Std. Error t value Pr(>|t|)
  (Intercept)  4.4077  0.5017     8.7853  0.0000
log(BETADIET)  0.1651  0.0702     2.3517  0.0193
        FIBER  0.0227  0.0088     2.5857  0.0102
         P.V1  0.2873  0.0937     3.0664  0.0024
         P.V2  0.2736  0.1030     2.6572  0.0083
     QUETELET -0.0301  0.0067    -4.4716  0.0000
  CHOLESTEROL -0.0015  0.0003    -4.9538  0.0000

Residual standard error: 0.7051 on 307 degrees of freedom
Multiple R-Squared: 0.2217
F-statistic: 14.57 on 6 and 307 degrees of freedom, the p-value is 1.21e-014
```

The model

$$Log(BETAPLASMA_i + 1) = 4.4077 + .1651 \cdot log(BETADIET) + .0227 \cdot FIBER + .2873 \cdot P.V1$$
$$+ .2736 \cdot P.V2 - .0301 \cdot QUETELET - .0015 \cdot CHOLESTEROL + \epsilon_i$$

has concordance with the preliminary analysis that have been found, then it is our "best" candidate for final model [9].

- Residual analysis. Figure 10 shows the standard diagnostic plots for the model

$$Log(BETAPLASMA_i + 1) = 4.4077 + .1651 \cdot log(BETADIET) + .0227 \cdot FIBER + .2873 \cdot P.V1$$
$$+ .2736 \cdot P.V2 - .0301 \cdot QUETELET - .0015 \cdot CHOLESTEROL + \epsilon_i$$

As can be seen in figure 10, the variance seems to be constant, there is not tendency in the residuals, except for one point the residuals seem to follow a normal distribution, and the

[9] Many others models were fitted without success, that is, without a better R^2.

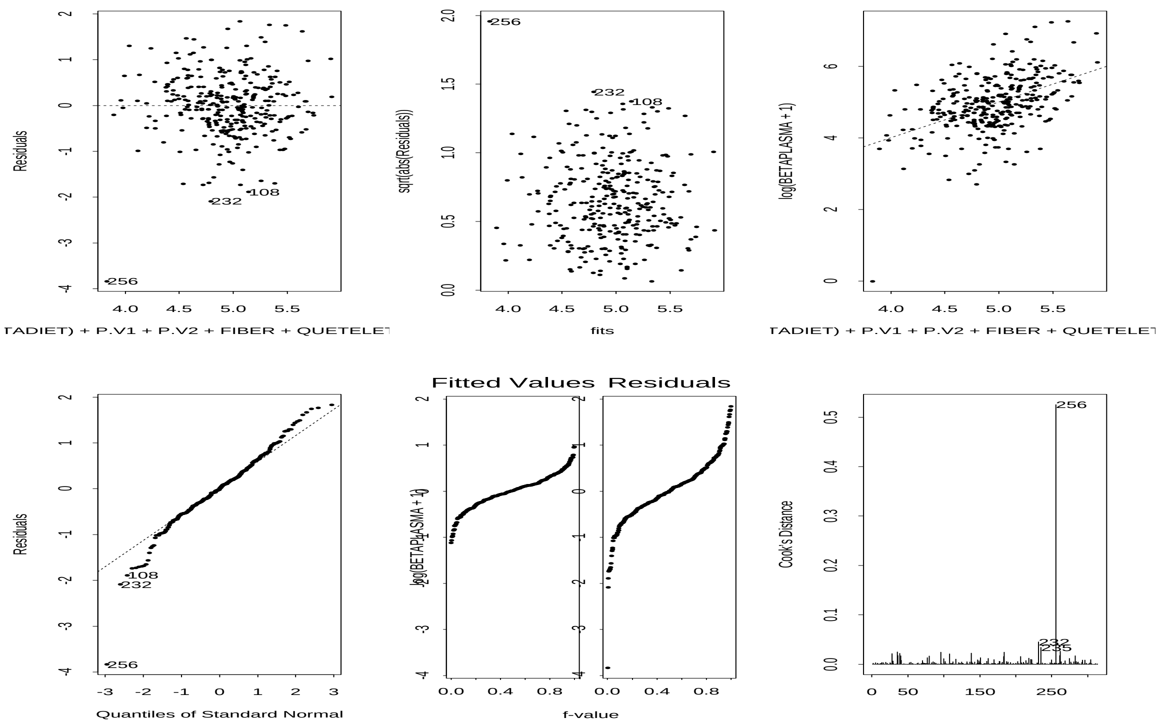

Figure 10: Standard diagnostic plots for regressing *log(BETAPLASMA+1)* on *log(BETADIET), FIBER, P.V1, P.V2, QUETELET* and *CHOLESTEROL*.

subject 256 (257 in the original dataset) is an influence point. The peculiar thing about this subject is that it has a zero level of beta-carotene. Even thought that we do not know if this situation is possible in medical terms, this observation changes completely our model, it is clear when we include the indicator variable for this observation, that this observation is an outlier, as shown in the next Splus output, therefore we should dropped it.

```
With all the variables

Coefficients:
                   Value Std. Error  t value Pr(>|t|)
(Intercept)   4.9798  0.2691    18.5050   0.0000
       P.S1  -0.2312  0.1252    -1.8459   0.0659
        AGE   0.0054  0.0029     1.8761   0.0616
   QUETELET  -0.0313  0.0064    -4.8973   0.0000
       P.V1   0.2909  0.0899     3.2365   0.0013
       P.V2   0.2660  0.0980     2.7153   0.0070
   CALORIES  -0.0001  0.0001    -1.3584   0.1753
      FIBER   0.0284  0.0098     2.8890   0.0041
    ALCOHOL   0.0041  0.0083     0.4926   0.6226
CHOLESTEROL  -0.0002  0.0004    -0.5536   0.5803
   BETADIET   0.0000  0.0000     1.6062   0.1093
      P.SM1   0.2877  0.1205     2.3875   0.0176
```

```
                     P.SM2   0.2049   0.1223     1.6753   0.0949
ifelse(BETAPLASMA == 0, 1, 0)  -4.6133   0.6990    -6.5996   0.0000

Residual standard error: 0.6554 on 300 degrees of freedom
Multiple R-Squared: 0.343
F-statistic: 12.05 on 13 and 300 degrees of freedom, the p-value is 0


With the variables in the "first final model"

Coefficients:
                                Value Std. Error t value Pr(>|t|)
                  (Intercept)  4.5250  0.4753     9.5203  0.0000
               log(BETADIET)  0.1265  0.0667     1.8953  0.0590
                         P.V1  0.3323  0.0890     3.7337  0.0002
                         P.V2  0.2739  0.0975     2.8101  0.0053
                        FIBER  0.0235  0.0083     2.8255  0.0050
                     QUETELET -0.0292  0.0064    -4.5731  0.0000
                  CHOLESTEROL -0.0010  0.0003    -3.2339  0.0014
ifelse(BETAPLASMA == 0, 1, 0) -4.2607  0.7038    -6.0534  0.0000

Residual standard error: 0.6674 on 306 degrees of freedom
Multiple R-Squared: 0.3049
F-statistic: 19.18 on 7 and 306 degrees of freedom, the p-value is 0
```

- With this situation, we should do the same process than before, that is, use all the variables in the model with interaction of second order with the categorical variables, and drop no important variables when the F-test indicates so, note that since BETAPLASMA no longer takes the value zero, we use log(BETAPLASMA) instead of log(BETAPLASMA+1) . The process of elimination of nonsignificant variables then is summarized as follows [10]:

```
MODEL WITH ALL THE VARIABLES

Coefficients:
                    Value Std. Error  t value Pr(>|t|)
      (Intercept)   4.9727   0.2718    18.2956   0.0000
        P.S1[-256]  -0.2336   0.1265    -1.8469   0.0657
         AGE[-256]   0.0054   0.0029     1.8707   0.0624
    QUETELET[-256]  -0.0317   0.0065    -4.8958   0.0000
        P.V1[-256]   0.2927   0.0908     3.2237   0.0014
        P.V2[-256]   0.2683   0.0989     2.7122   0.0071
    CALORIES[-256]  -0.0001   0.0001    -1.3618   0.1743
       FIBER[-256]   0.0287   0.0099     2.8917   0.0041
     ALCOHOL[-256]   0.0041   0.0084     0.4936   0.6219
 CHOLESTEROL[-256]  -0.0002   0.0004    -0.5498   0.5829
    BETADIET[-256]   0.0000   0.0000     1.5931   0.1122
       P.SM1[-256]   0.2905   0.1217     2.3862   0.0176
       P.SM2[-256]   0.2063   0.1235     1.6702   0.0959
```

[10] the interactions were nonsignificant but this test does not appear since the output is too messy.

```
Residual standard error: 0.6619 on 300 degrees of freedom
Multiple R-Squared: 0.2482
F-statistic: 8.252 on 12 and 300 degrees of freedom, the p-value is 1.912e-013



MODEL WITH ALL THE VARIABLES BUT ALCOHOL AND CHOLESTEROL.

Coefficients:
                  Value Std. Error  t value Pr(>|t|)
   (Intercept)   4.9907   0.2696    18.5123   0.0000
    P.S1[-256]  -0.2348   0.1222    -1.9222   0.0555
     AGE[-256]   0.0054   0.0029     1.8637   0.0633
QUETELET[-256]  -0.0325   0.0063    -5.1255   0.0000
    P.V1[-256]   0.2884   0.0897     3.2146   0.0014
    P.V2[-256]   0.2656   0.0986     2.6938   0.0075
CALORIES[-256]  -0.0002   0.0001    -2.2070   0.0281
   FIBER[-256]   0.0293   0.0094     3.1329   0.0019
BETADIET[-256]   0.0000   0.0000     1.5915   0.1126
   P.SM1[-256]   0.2930   0.1213     2.4146   0.0163
   P.SM2[-256]   0.2150   0.1227     1.7514   0.0809

Residual standard error: 0.6604 on 302 degrees of freedom
Multiple R-Squared: 0.2466
F-statistic: 9.884 on 10 and 302 degrees of freedom, the p-value is 2.509e-014


Analysis of Variance Table

Response: log(BETAPLASMA[-256])


        Terms  Res. Df      RSS                             Test Df Sum of Sq
1 Small Model      302 131.7150
2 Big Model        300 131.4391 +ALCOHOL[-256]+CHOLESTEROL[-256]  2 0.2759521

        F Value   Pr(F)
1
2   0.3149202 0.7300881



MODEL WITH ALL THE VARIABLES EXCEPT ALCOHOL, BETADIET AND CHOLESTEROL.

Coefficients:
                  Value Std. Error  t value Pr(>|t|)
   (Intercept)   4.9727   0.2700    18.4152   0.0000
    P.S1[-256]  -0.2451   0.1223    -2.0042   0.0459
     AGE[-256]   0.0057   0.0029     1.9617   0.0507
QUETELET[-256]  -0.0320   0.0064    -5.0411   0.0000
```

```
    P.V1[-256]   0.2978   0.0897     3.3192   0.0010
    P.V2[-256]   0.2677   0.0988     2.7080   0.0072
CALORIES[-256]  -0.0002   0.0001    -2.1465   0.0326
   FIBER[-256]   0.0352   0.0086     4.0759   0.0001
   P.SM1[-256]   0.2973   0.1216     2.4444   0.0151
   P.SM2[-256]   0.2285   0.1228     1.8619   0.0636

Residual standard error: 0.6621 on 303 degrees of freedom
Multiple R-Squared: 0.2403
F-statistic: 10.65 on 9 and 303 degrees of freedom, the p-value is 2.365e-014



Analysis of Variance Table

Response: log(BETAPLASMA[-256])


        Terms  Res. Df      RSS            Test Df Sum of Sq     F Value   Pr(F)
1 Small Model      303 132.8196
2 Big Model        302 131.7150 +BETADIET[-256]  1  1.104627  2.532721 0.1125539



MODEL WITH ALL THE VARIABLES EXCEPT ALCOHOL, BETADIET, AGE, SEX AND CHOLESTEROL.


Coefficients:
                  Value Std. Error  t value Pr(>|t|)
   (Intercept)   5.2662   0.2240    23.5057   0.0000
    P.V1[-256]   0.3198   0.0896     3.5702   0.0004
    P.V2[-256]   0.2754   0.0975     2.8241   0.0051
   FIBER[-256]   0.0386   0.0085     4.5103   0.0000
   P.SM1[-256]   0.3332   0.1213     2.7463   0.0064
   P.SM2[-256]   0.2472   0.1225     2.0172   0.0445
QUETELET[-256]  -0.0322   0.0064    -5.0289   0.0000
CALORIES[-256]  -0.0002   0.0001    -3.1722   0.0017

Residual standard error: 0.6663 on 305 degrees of freedom
Multiple R-Squared: 0.2254
F-statistic: 12.68 on 7 and 305 degrees of freedom, the p-value is 2.742e-014



Analysis of Variance Table

Response: log(BETAPLASMA[-256])


        Terms  Res. Df      RSS                  Test Df Sum of Sq
1 Small Model      305 135.4144
```

```
2 Big Model        303 132.8196 +P.S1[-256]+AGE[-256]  2  2.594748

        F Value   Pr(F)
1
2   2.959685 0.05333587
```

Finally, Figure 11 shows the standard diagnostic plots for our final model for BETAPLASMA:

$$Log(BETAPLASMA_i) = 5.2662 + .3198 \cdot P.V1 + .2754 \cdot P.V2 + 0.0386 \cdot FIBER$$
$$0.3332 \cdot P.SM1 + 0.2472 \cdot P.SM2 - .0322 \cdot QUETELET - .0002 \cdot CHOLESTEROL + \epsilon_i$$

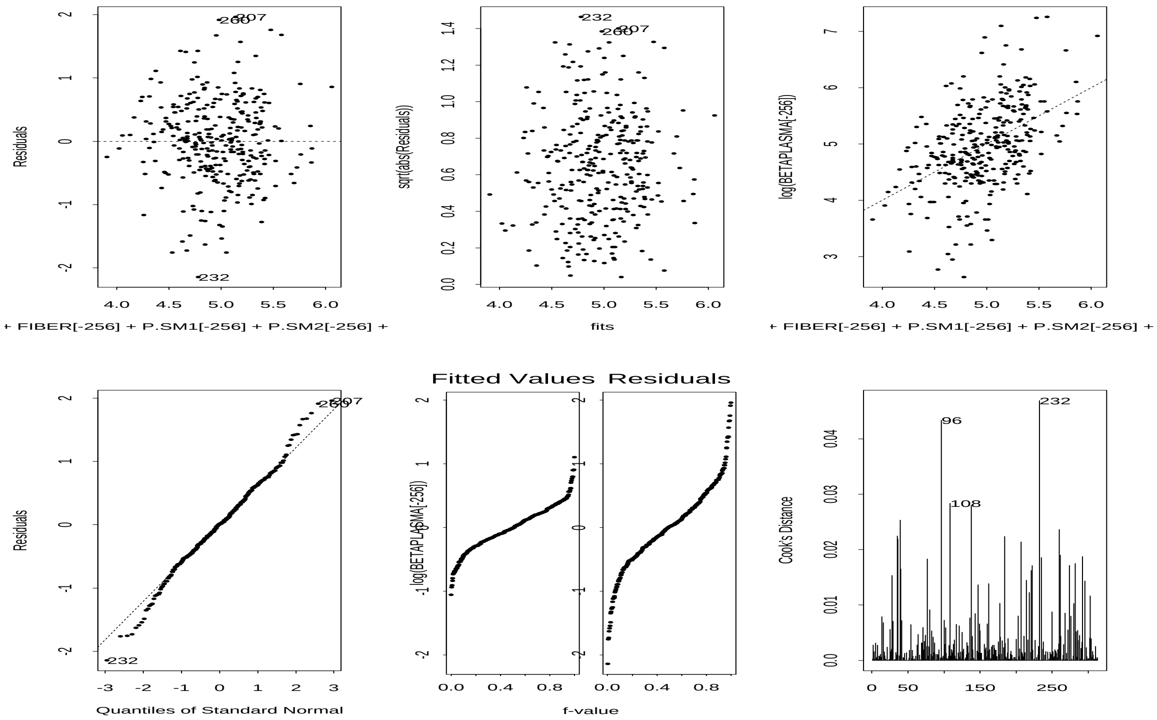

Figure 11: Standard diagnostic plots for regressing *log(BETAPLASMA)* on *P.V1, P.V2, FIBER, P.SM1, P.SM2, QUETELET* and *CALORIES*.

As is shown in Figure 11, the variance seems to be constant, there is not tendency in the residuals and they seem to follow a normal distribution. There are some observations that seem to be influential points, but in these cases there are not a strong reason to dropped, but further study of them should be done.

Dependable variable: RETPLASMA

⊙ A similar process done for the variable BETAPLASMA was done for the variable RETPLASMA, this time seems that a transformation is not necessary since the distribution is fairly symmetric. The results of the multiple regression analysis using: all the independent variables, and the dependent variable *RETPLASMA* are summarized in the next Splus output:

```
Coefficients:
                Value Std. Error  t value  Pr(>|t|)
(Intercept)  410.8705   81.1629    5.0623    0.0000
       P.S1   77.2769   37.6055    2.0549    0.0408
       P.V1   36.4480   27.2757    1.3363    0.1825
       P.V2   28.6833   29.7624    0.9637    0.3360
      P.SM1   -0.0680   36.3625   -0.0019    0.9985
      P.SM2   45.5590   36.8326    1.2369    0.2171
        AGE    2.7109    0.8794    3.0826    0.0022
   QUETELET    1.6666    1.9296    0.8637    0.3884
   CALORIES    0.0779    0.0600    1.2986    0.1951
    ALCOHOL    7.2110    2.6121    2.7606    0.0061
      FIBER   -4.1793    3.0939   -1.3508    0.1778
        FAT   -1.4478    0.9473   -1.5283    0.1275
CHOLESTEROL   -0.0900    0.1287   -0.6995    0.4848
    RETDIET   -0.0078    0.0218   -0.3576    0.7209

Residual standard error: 197.7 on 300 degrees of freedom
Multiple R-Squared: 0.1392
F-statistic: 3.733 on 13 and 300 degrees of freedom, the p-value is 0.00001543
```

The test of the composite hypothesis that all 13 regression coefficients are zero is significant. The coefficient of determination is just .1392, thus 14 % of the variability of the *RETPLASMA* can be associated with the variation of these 13 independent variables.
The t test of the partial regression coefficients $H_0 : \beta_j = 0$ (using the bonferroni correction) would seem to suggest that 11 of the 13 independent variables are unimportant and could be dropped from the model.A simultaneous test was used to eliminate subsets of variables. The first subset to drop, is the set of variables with a really high p-value: *P.SM1*, *P.SM2*, *RETDIET CHOLESTEROL*, AND *QUETELET*. The Splus output for this hypothesis is:

```
Analysis of Variance Table

Response: RETPLASMA


        Terms  Res. Df      RSS                    Test Df Sum of Sq
1 Small Model      305 11926301
2 Big Model        300 11723731 +P.SM1+P.SM2+QUETELET+CHOLESTEROL+RETDIET  5  202569.4

        F Value      Pr(F)
1
2      1.036715  0.3960879
```

Since the p-value is greater than $\alpha = 0.05$, we can drop these five variables from the model. The results for the multiple regression analysis without the independent variables dropped are summarized in the following Splus output:

```
Coefficients:
                Value Std. Error  t value  Pr(>|t|)
```

```
(Intercept)  462.8552   62.1995     7.4415    0.0000
       P.S1   79.1026   36.8478     2.1467    0.0326
       P.V1   33.5883   26.9365     1.2469    0.2134
       P.V2   28.6169   29.7109     0.9632    0.3362
        AGE    2.7189    0.8687     3.1299    0.0019
   CALORIES    0.0674    0.0588     1.1454    0.2529
    ALCOHOL    7.8467    2.5384     3.0911    0.0022
      FIBER   -3.9001    3.0054    -1.2977    0.1954
        FAT   -1.5250    0.9282    -1.6429    0.1014

Residual standard error: 197.7 on 305 degrees of freedom
Multiple R-Squared: 0.1244
F-statistic: 5.415 on 8 and 305 degrees of freedom, the p-value is 2.177e-006
```

The coefficient of determination in this case is .1244. Once again, there are some variables that could be dropped from the model, therefore the the last procedure was done again with the variables: *P.S1*, *P.V1* , *P.V2*, *CALORIES*, *FIBER*, and *FAT*.

```
Analysis of Variance Table

Response: RETPLASMA


        Terms  Res. Df      RSS                              Test Df Sum of Sq
1 Small Model      311 12314556
2 Big Model        305 11926301 +P.S1+P.V1+P.V2+CALORIES+FIBER+FAT  6  388255.4

       F Value      Pr(F)
1
2      1.654857 0.1317877
```

Since the p-value is greater than $\alpha = 0.05$, we can drop these six variables from the model. The results for the multiple regression analysis for the remain variables are summarized in the following Splus output:

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) 422.9405  40.7237    10.3856   0.0000
        AGE   3.1172   0.7717     4.0392   0.0001
    ALCOHOL   9.3027   2.2726     4.0933   0.0001

Residual standard error: 199 on 311 degrees of freedom
Multiple R-Squared: 0.09586
F-statistic: 16.49 on 2 and 311 degrees of freedom, the p-value is 1.566e-007
```

Therefore our best model until now for BETAPLASMA is:

$$RETPLASMA_i = 422.9405 + 3.1172 \cdot AGE + 9.3027 \cdot ALCOHOL$$

⊙ Residual analysis.

As can be seen in Figure 12, the variance does not seems to be constant, and the distribution of the residual is positively skewed. Then, we should try to transform the variable RETPLASMA after all. For this reason the transformation log(RETPLASMA) is used. The results for the multiple regression analysis are summarized in the following Splus output and the residuals plot is shown in Figure 13. It is clear that this time the distribution of the residuals is not skewed anymore, but all the other statistics remain alike.

```
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  6.0470  0.0663    91.1938  0.0000
        AGE  0.0052  0.0013     4.1700  0.0000
    ALCOHOL  0.0141  0.0037     3.8082  0.0002

Residual standard error: 0.324 on 311 degrees of freedom
Multiple R-Squared: 0.09276
F-statistic: 15.9 on 2 and 311 degrees of freedom, the p-value is 2.666e-007
```

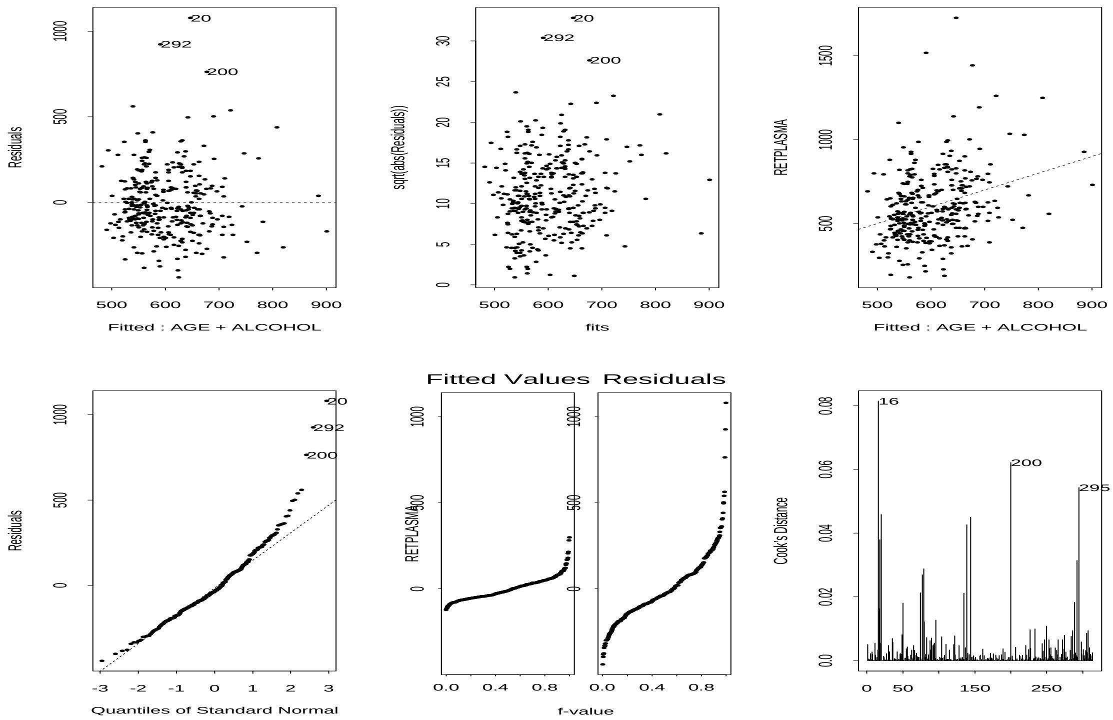

Figure 12: Standard diagnostic plots for regressing *RETPLASMA* on *AGE* and *ALCOHOL*.

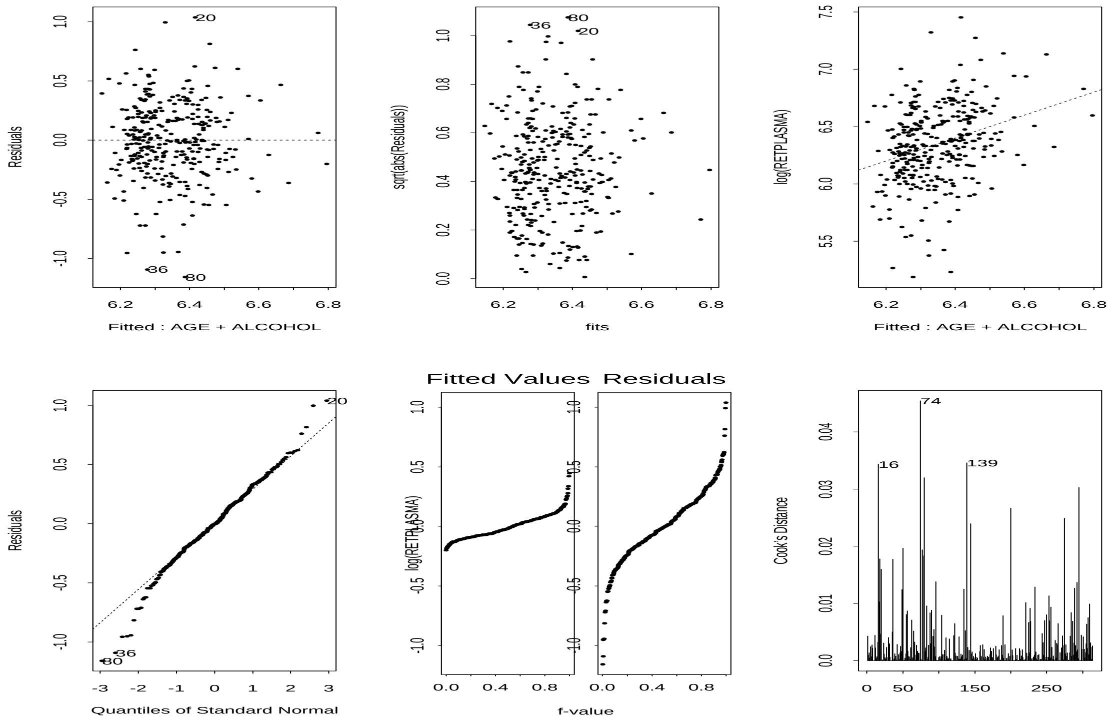

Figure 13: Standard diagnostic plots for regressing *log(RETPLASMA)* on *AGE* and *ALCOHOL*.

References

[1] Nierenberg D.W., Stukel T. A., Baron J.A., Dain B. J., Greenberg E. R., and The skin cancer prevention study group. Determinants of plasma levels of beta-carotene and retinol. Am. J. Epidemiol 1989;130:511-21.

[2] Thurnham Dl. Do higher vitamin A requirements in men explain the difference between the sexes in plasma provitamin A carotenoids and retinol?. [Abstract]Proc. Nutr. Soc. 1988;47:181.

[3] Hallfrisch J., Muller D. C., and Singh V. N. Vitamin A, E intakes and plasma concentrations of retinol, β-carotene, and α-tocopherol in men and women of the Baltimore Longitudinal Study of Aging. AM. J. Clin. Nutr. 1994;60:151-181.

[4] Stryker W.S., Kaplan L.A., Stein EA, et.al. The relation of diet, cigarrete smoking, and alcohol consumption to plasma beta-carotene and alpha-tocopherol levels. Am. J. Epidemiol 1988;127:283-96.

[5] Comstock G.W., Menkes M.S., Schober S.E., et. al. Serum levels of retinol, beta-carotene, and alpha-tocopherol in older adults. Am. J. Epidemiol 1988;127:114-23.

[6] Wald N. Retinol, beta-carotene and cancer. Cancer Surv. 1987;6:635-51.

[7] Kritchevsky D. Antioxidant vitamins in the prevention of cardiovascular disease. Nutr. Today 1992;Jan/Feb:30-3.

[8] Taylor A., Jacques P.F., and Dorey C. K. Oxidation and aging: impact on vision. Toxicol. Ind. Health 1993;9:349-71.

[9] Epidemiology Unit. Memo.

[10] http://lib.stat.cmu.edu. STATLIB.

[11] Junker B. Class notes.