Lowering the risk of cancer everyday: a study on the relationships between personal facts and plasma concentrations of risk related micronutrients

Project1 - Applied Regression Analysis

Abstract

Previous studies suggest that low plasma concentrations of retinol, beta-carotene, or other carotenoids may be associated with high risk of developing certain types of cancer. We investigate now the determinants of these plasma concentrations by developing an observational study on 315 patients who underwent a surgical procedure but were found to be sound. We show that dietary intakes of fiber, vitamins and alcohol are related to high levels of beta carotene, while smoke, high body mass, cholesterol and fat intake are related to lower levels of beta carotene. We also show that plasma retinol is raised by alcohol use, and lowered by smoke and fat intake; though the study presents lack of fit problems. Some suggestions for future studies are made.

1 Introduction

Previous studies suggest that low plasma concentrations of retinol, beta-carotene, or other carotenoids may be associated with high risk of developing certain types of cancer. But few other studies have investigated the determinants of plasma concentrations of this risk lowering micronutrients.

On the other hand, there are many popular beliefs that certain personal facts and habits can influence the risk of developing a cancer. Many of these ideas have been proved to be true, like the fact that smoking is related to lung cancer; but many other haven't yet been investigated at all.

So we are now interested in investigating what personal characteristics, dietary factors and lifestyles lead to higher or lower levels of retinol and betacarotene, thus lowering or increasing the risk of developing a cancer.

In this way, we would suggest a way to connect personal facts and risk of cancer. We will develop a statistical regression analysis to do this.

Section 2 will present the data set and do some Exploratory Data Analysis on each variable. Section 3 will present some simple regression models, to investigate the relationships between plasma level of micronutrients and each variable. Sections 4, 5 and 6 will present a model for the prediction of plasma levels of, respectively: plasma retinol, beta carotene, and both of them together. Section 7 performs a discussion of the analysis, conclusions and recommendations.

2 The Data

The data used in this analysis was collected surveying 315 subjects who had an elective surgical procedure, to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. Fourteen variables were recorded, among which there are the plasma levels of beta-carotene and retinol (both in ng/ml), the dietary intake of beta carotene and retinol (mcg per day); grams of fat, cholesterol, calories and fiber consumed per day; the Quetelet index (the weight divided by the squared height), smoking status, use of vitamins, sex and age of the patient. See technical appendix A for the head of the datafile.

2.1 Plasma Retinol

The observed mean of plasma retinol was 602.8 ng/ml, with a standard deviation of 208.9 ng/ml. We can see from the boxplot in the first graph in Figure 1 that this variable has some high outliers (1249, 1262, 1443, 1517, 1727) and is slightly skewed. Since this is one of the response variables, we want it to be normally distributed. From the last two graphs in Figure 1 we can see that a log transformation of the variable works well (i.e., there is no evidence of non-normality in the log-transformation).

2.2 Plasma Beta-Carotene

The levels of beta carotene in plasma go from a minimum of 0 to a maximum of 1415; with a sample mean of 189.9 ng/ml and a standard deviation of 183.0 ng/ml. So plasma beta carotene has a stronger variability, in the sample, than plasma retinol. It is surprising that there can be no beta carotene in the blood, so it is interesting to look at the observation which achieves the minimum:

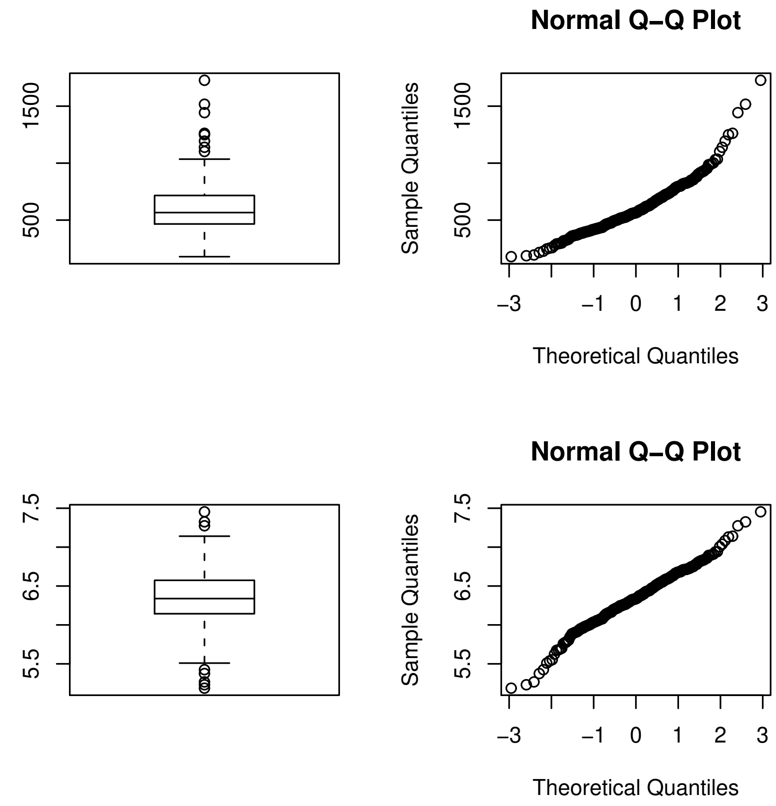

Figure 1: Plasma Retinol

```
    age sex smoke     quet vit    cal   fat fiber alc colest betad retd beta
257  40   2     1 31.24219   1 3014.9 165.7  14.4   0  900.7  1028 3061    0
    retpl
257   254
```

So a 40 years old woman, obese (quet> 27), has a 0 level of plasma beta-carotene together with a fairly low 254 level of plasma retinol. Nevertheless, it seems impossible that there could be no plasma beta-carotene, and some biological investigation should be done to see whether this is possible or not. This is the other response variable, and outlyingness in it can be a problem[1]. So we will omit this observation when using betaplasma as a response and, usually, maintain it when it is a predictor.

From Figure 2 we can see a quantile normal plot for the variable, without observation 257 in the model. I chose to transform this variable with a power of .2, and the result can be seen in the lower plots. In the technical appendix B there is some perspective on the motivations of this transformation.

2.3 Age, Sex, Dietary variables

The age of the subjects goes from 19 to 83 years, and there are only 42 males over 315 observations. This shouldn't be a problem, while I think will be interesting in the future to study the levels of beta-carotene and retinol in kids and teenagers, as old studies prove that kids are more likely to develop certain types of cancer.

122 persons used vitamins "often", 82 used vitamins "sometimes" and 111 never used vitamins.

In Figure 3 we can see a boxplot for the Quetelet index. There are some high outliers, and a long right tail. Women with an index over 27, and men over 28 are considered obese. We can see that 111 (32% of the observations), were found to be obese.

The calories intake has an observed mean of 1796 and a standard deviation of 680. There is an high outlier, consuming 6662 calories per day. We can expect this outlier to be informative in the analysis. 101 persons consume more than the 2000 threshold of recommended calories intake per day. It is interesting to notice that only 31 of this 101 are obese (in the sense determined by the Quetelet index).

There is nothing particularly interesting to notice about fat, cholesterol, beta-carotene and fiber intake: there are some high outliers and a light

[1] Cfr handout on outliers, http://www.stat.cmu.edu/˜brian/707/, more details in technical appendix B

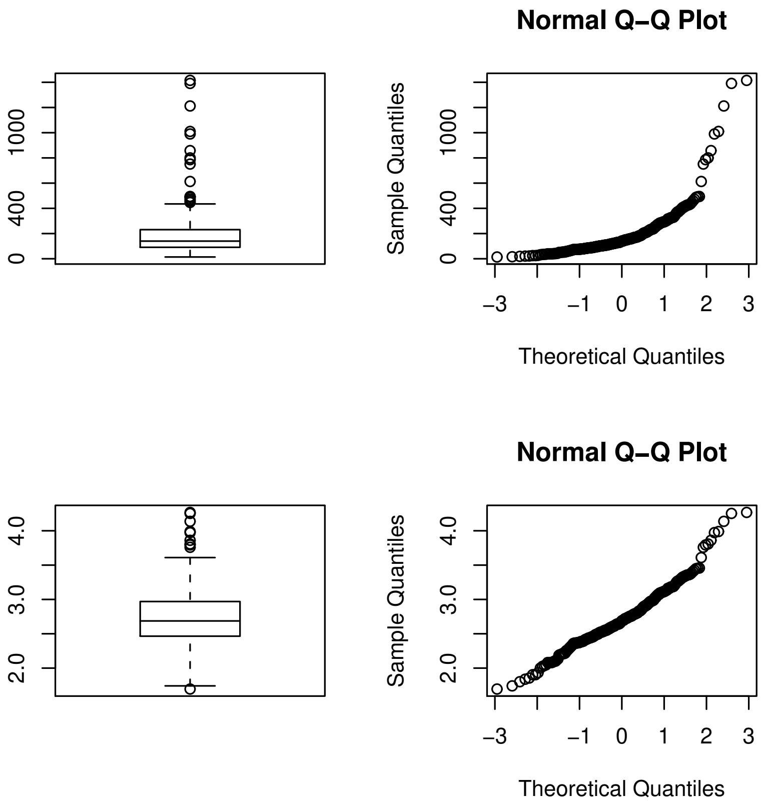

Figure 2: Beta-Carotene

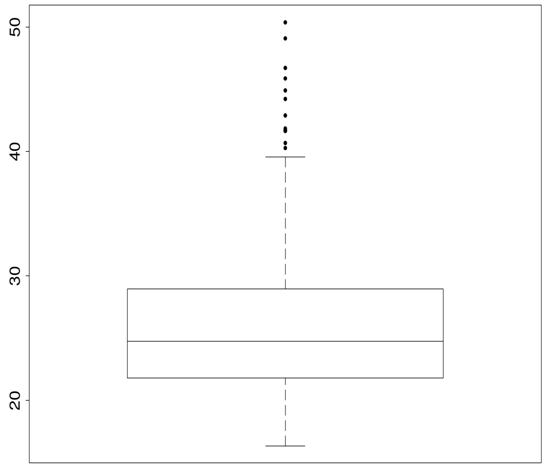

Figure 3: Quetelet index

positive skewness.

The intake of retinol is more skewed, as can be seen in the first two graphs in Figure 4. The very high outlier is 6901, while the minimum is 30. I decided to use a log transformation of this variable in the analysis, and the reasons are explained in the technical appendix B.

2.4 Alcohol Use

We can see from Figure 5 that the average number of drinks consumed per week has a very long right tail. There are 111 observations with a 0 average, and 50% of the observations consume less than .3 drinks per week, while the three higher values observed were 35,35 and an amazing 203. I guess it is possible that an alcoholic could consume 203/7=29 drinks per day. This is the observation:

```
   age sex smoke     quet vit    cal   fat fiber alc colest betad retd beta
62  65   1     3 23.37617   3 6662.2 164.3  11.3 203    603  2893 1364   96
   retpl
62   317
```

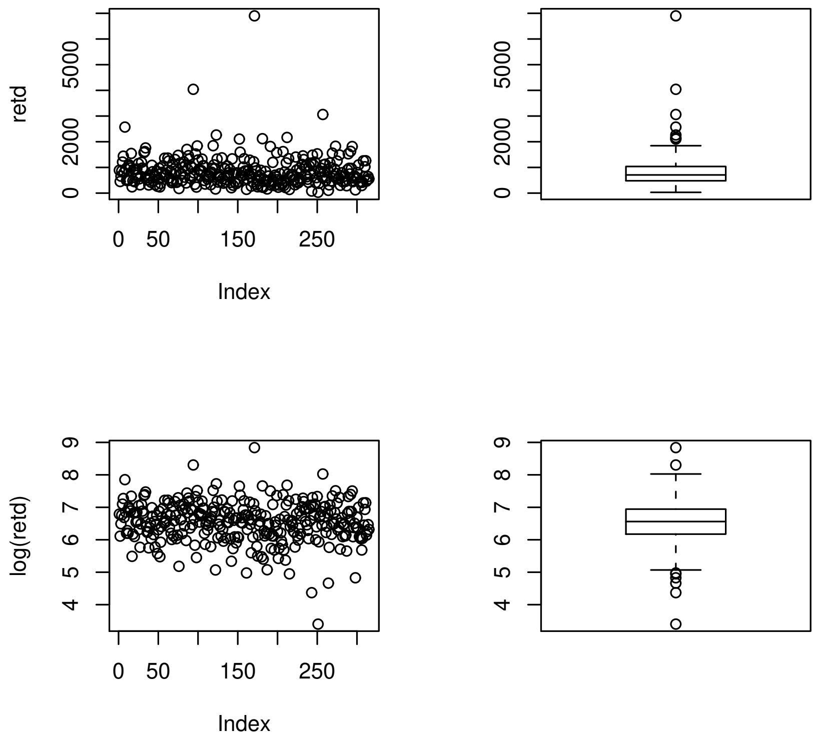

Figure 4: Dietary retinol

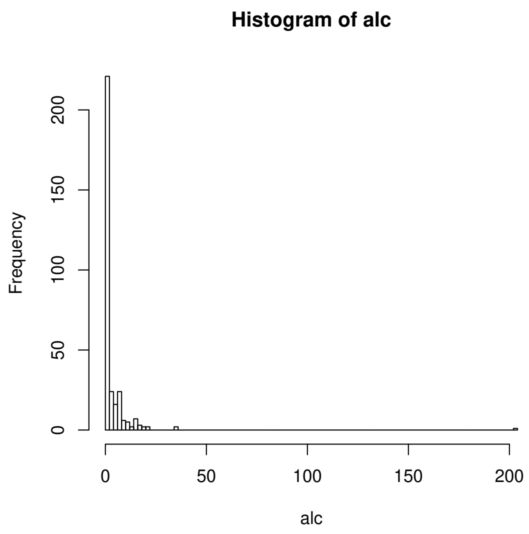

Figure 5: Alcohol

This man is a current smoker, non obese but consuming more than six thousand calories per day; and has a low level of plasma retinol and a very low level of plasma beta-carotene. We can make the hypothesis that a consistent part of the high number of calories consumed by this man comes from alcohol; so, in a sense, the value of “cal” is a validation for the value of “alc”, and vice-versa. This is the same outlier observed for the ”calories” variable.

The behavior of this variable is pretty reasonable: there are a lot of no alcohol users, many low drinkers and some heavy drinkers. We need to work with a suitable transformation, as this skewness can interfere with the analysis. The problem is that there are too many zero-values, which are not affected by usual transformations. So I think that a good idea can be transforming this quantitative variable into a qualitative one. I decided to divide the variable into 5 levels

1. Non drinkers (0 drinks per week)

2. Very low drinkers (drinking less than 1 drink per week)

3. Moderate drinkers (between 1 and 11 drinks per week)

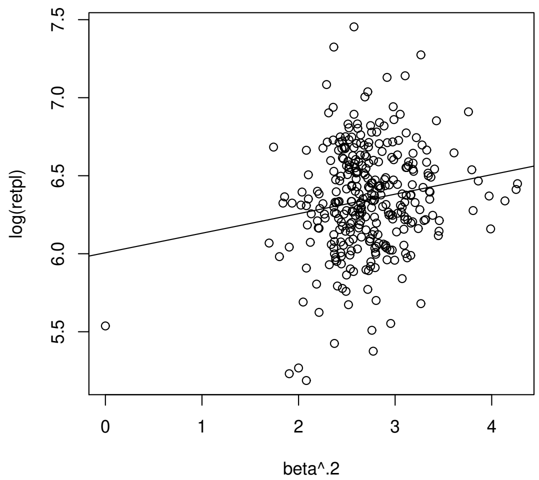

Figure 6: Plasma beta and retinol

4. Hard drinkers (more than 11 drinks per week)

This categorization could be made in many different ways, and it can be objected that the third class has got a really big range. However this should be a good way to divide this variable, and in technical appendix B there are more details and justifications for this categorization.

3 Simple Analyses

3.1 Between the micronutrients

Let's begin studying the relationship between plasma retinol and beta-carotene. They are very low positively correlated (.07).

Figure 6 shows a scatterplot between transformed plasmas. We can note the "stand alone" point in the lower left corner, with a plasma beta of 0. The line is fitted without this outlier. There seems to be a strongly significant increasing relationship ($p < 0.007$) between the two micronutrients, though the fit is not very good (Multiple R-Squared: 0.05). However, high levels of one micronutrient will raise the levels of the other one; and so what is effective on one of the two, indirectly, should be important also for the other.

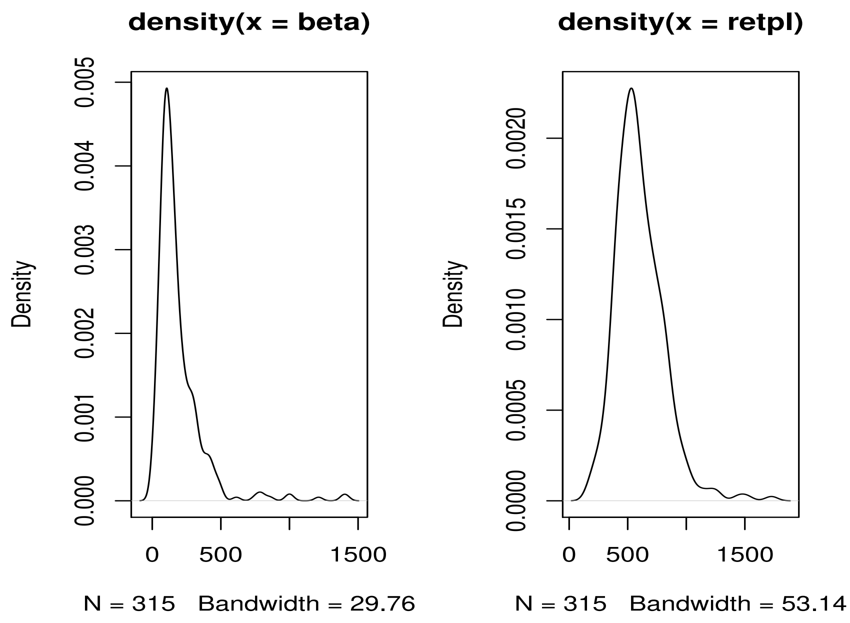

Figure 7: Micronutrients density

Figure 7 shows a plot of the estimated density of the two micronutrients. They show approximately the same behavior for the two variables. As for betaplasma, we can see a sharp mode close to 0 and a long right tail, with other small modes. The same can be said for plasma retinol, though its density is smoother and it achieves a lower mode (this is because of the differences in variability). This similar behavior suggests that, maybe, micronutrients are usually distributed in this way; and we can safely accept the value of the outliers as "possible". That is, we can always expect to see some individuals with unusual high levels of one or both of the two micronutrients. Let's end this analysis pointing out that it is not necessary that somebody with unusual high levels of plasma retinol should have high levels of the other micronutrient (i.e., no outlier value for the two variables is achieved by the same observation). More details are given in the technical appendix C.

3.2 Micronutrients and Dietary Intakes

It seems pretty obvious that high dietary intakes of one micronutrient will account for high levels of it. This is true for plasma beta-carotene, which has a particularly strong relationship with its dietary intake ($p < 0.0002$); but surprisingly it is not true for plasma retinol ($p > 0.24$). In technical appendix C, I put some considerations on the residuals of this regression fits. It is amazing that dietary intake of retinol won't affect the plasma retinol level. We can expect that dietary intakes, in general, will be ineffective on plasma retinol, which may be influenced by other kind of variables. It may be that levels of plasma retinol are auto-regulated by the organism, till certain conditions don't come to change the "regular" levels. This could be an explanation, also, for the lower variability with respect to plasma beta-carotene. We will say more about this later.

Table 1 shows approximated p-values and coefficient estimates for simple regressions of each dietary variable on plasma beta[2]. Fat intake per day seems to mildly lower the levels of beta carotene, while cholesterol intake has a milder lowering effect but is more strongly significative. There seems to be no relationship with the total amount of calories consumed per day, while the fiber intake seems to have a very strong effect on raising the plasma beta-carotene levels. The R-Squared for this regression is .06, being slightly high for the standards in this data set (no R-Squared has been higher than 3% till now, in the simple regressions). It is interesting however to notice that, even if we can conclude that high intakes of fiber will end up with an higher plasma beta-carotene, the fiber intake is strongly correlated with beta-carotene intakes (.48), so our conclusions are a little bit too optimistic. If we study the relationship between betaplasma and fiber intake net of beta intake, we find a parameter estimate of only .0147 (see technical appendix C for more details on conditional coefficients).

On the other hand, plasma retinol hasn't got any significative relationships with fiber, calories, and cholesterol intakes. There is only a not strongly significative relationship with fat ($p = .03$).

Table 2 shows the correlation structure between the dietary variables. As one could expect, all the variables are strongly positively correlated; and we will need to take into account this multicollinearity problem when fitting a multiple regression, avoiding direct interpretation of the parameters.

[2]Important: each variable has been regressed on plasma beta separately. I put a zero parameter estimate where the slope was found to be non significative.

	Parameter	p-value
Fat	-.0015	.036
Cholesterol	-.0005	.006
Fiber	.02	0
Calories	0	.344

Table 1: Dietary variables on beta

	Fat	Cholesterol	Calories	Fiber
Fat	1	.71	.87	.28
Cholesterol		1	.66	.15
Calories			1	.46
Fiber				1

Table 2: Correlation of dietary variables

3.3 Lifestyle and Personal Characteristics Effects

The distinction between food and lifestyle variables is done only for better understanding of the study, and it doesn't have any statistical consequences.

I decided to consider the Quetelet index as a personal characteristic; on the grounds that, at a cursory glance, it isn't significatively related with any "dietary" variable. The one variable significatively related with the Quetelet index is "smoke". This is reasonable, as it is well known that smoking can lower the weight.

I decided moreover to consider the number of drinks consumed per week and vitamin use as "lifestyle" indicators, even if there is a very strong relationship between alcohol and calories intake (the regression between the two ends with a p-value very close to 0). I did so because it is reasonable to expect that no alcohol consumption and/or vitamin use are indicators of an healthy lifestyle, and they can go together with other healthy behaviors (i.e., with gym exercise).

3.3.1 Vitamin Use

The first graph in Figure 8 suggests that daily use of vitamins (vit= 1) can sky-rocket the plasma levels of beta carotene in certain individuals: the outliers of the first boxplot achieve strongly higher values of other categories' outliers. In an ANOVA, we find that the vitamin intake has a very strong

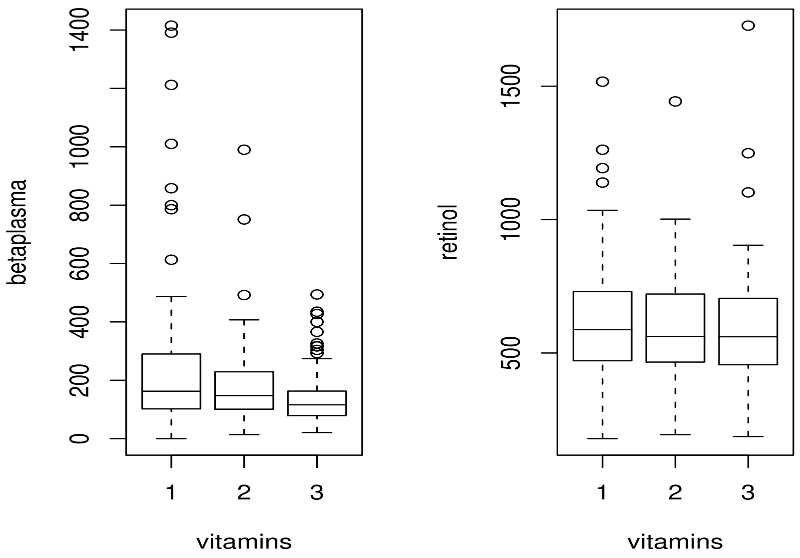

Figure 8: Vitamin Intake

effect on beta-carotene presence ($p < 0.0004$). The mean of betaplasma in the group of vitamin users is 241, 186 for the occasional users, and only 137 in the group of non users. If we confront the last two groups, we'll find that the difference between them is significative ($p = 0.003$), so even occasional use of vitamins can make plasma levels of beta-carotene higher than levels in no vitamin users. See technical appendix C for some descriptions on the dummy variables used and for contrasts issues.

As expected and suggested by the second graph in Figure 8, on the other hand, there is no relationship between plasma retinol and vitamin use. The mean for vitamin users is 613, 597 for non-often users and 595 for non-users. Even if we try to confront the group of often-users with the other two groups, we get a $p = 0.748$ and we must conclude that there is no difference between using and not using vitamins (see technical appendix C for more details).

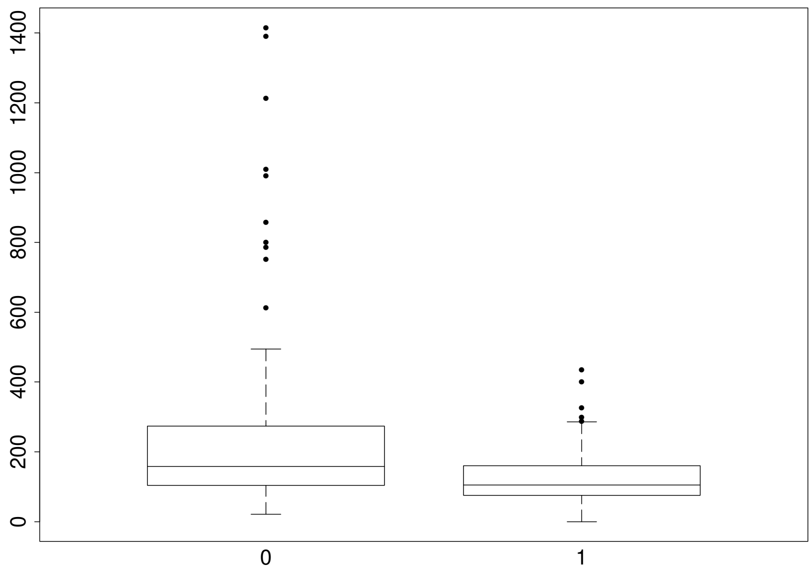

Figure 9: Obesity (0: non obese, 1: obese)

3.3.2 Body Mass Effects

The plasma beta-carotene turns out to be strongly linked with the Quetelet index, with a p value very close to 0 and an R-Squared greater than 7%; which I said in this analysis is an high value. The parameter estimate is -0.02. So it seems that higher Quetelet index brings about lower concentrations of beta-carotene. On the other hand, the Quetelet index proves ineffective on the retinol concentrations, with a p-value greater than 90%.

Another approach to the Quetelet index is given by dividing the individuals in two categories: obese and non obese. In technical appendix C there is more on this. It turns out that 101 individuals in the data set are obese, and that this cutting edge is important: the effect of obesity on the transformed betaplasma is -0.25674, with an R-Squared increased to more than 8%. Figure 9 shows higher median in the boxplot for non obese (0 label). As we could expect, obesity is not influential on plasma retinol.

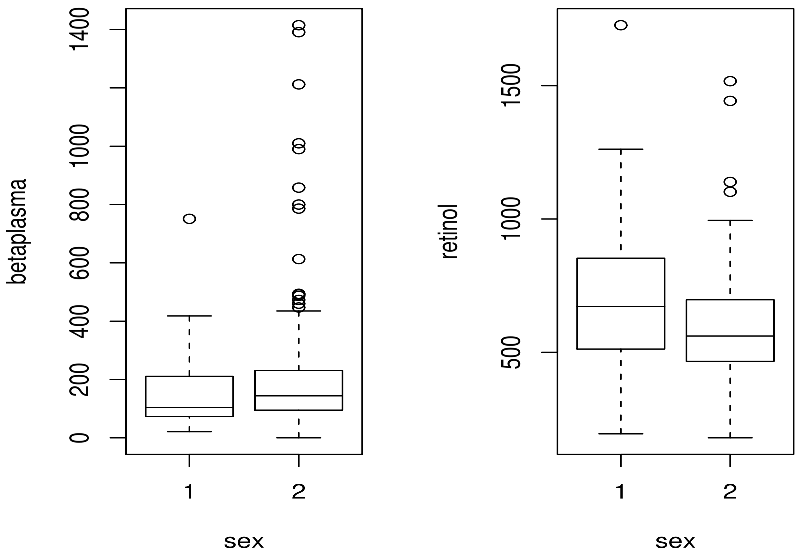

Figure 10: Sex differences

3.3.3 Gender Effects

As can be seen in Figure 10, it seems like women are more likely to have higher betaplasma than men; while men are more likely to have higher retinol. From ANOVA analyses we conclude that this difference is strongly significative in both cases ($p < .022$ for plasma retinol, see technical appendix C for some issues on the residuals of this ANOVA; and $p = 0.0228$ for beta-carotene). So males tend to have lower betaplasma and higher plasma retinol than females.

3.3.4 Age Effects

When we study the relationship between age and the micronutrients, we find strong increasing relationships, with p-value close to 0. One hypothesis is that elderly can retain more retinol and beta-carotene in the plasma. The other one is that elderly lead healthier lifestyles. Maybe more studies should be made towards the answer to this doubt; but we can investigate this also with a multiple regression, later in the report.

3.3.5 Smoke Effects

Figure 11 shows something counterintuitive: being a former smoker is better than having never smoked, in terms of plasma retinol.

Running ANOVA analyses, we can see that this difference is strongly significative ($p < .009$). The mean of plasma retinol in the group of former smokers is 644, while 583 in the group of no-smokers and 563 in the group of current smokers. If we use contrasts to confront the first two groups, we conclude that is better having smoked and given up, than having never smoked. This is counterintuitive, as I said, and it may be that individuals dropping smoking are more likely to get other healthy habits. However, it can also be that while smoking lowers the levels of retinol retained in plasma; when smoking is dropped the organism compensates retaining more retinol in the plasma. Future analyses can be performed to better disclose the effects of smoke on retinol. It is obviously not true to conclude that smoking for a period and then dropping is healthy, as smoke is connected with lots of illnesses and lung cancer. More details on this amazing behavior of retinol are given in the technical appendix C.

As for plasma beta-carotene, more intuitive results come up: the smoke has a strong effect of the levels of betaplasma ($p = .002$), though the best group now is the no-smoking one, with a mean of 206. The former smokers

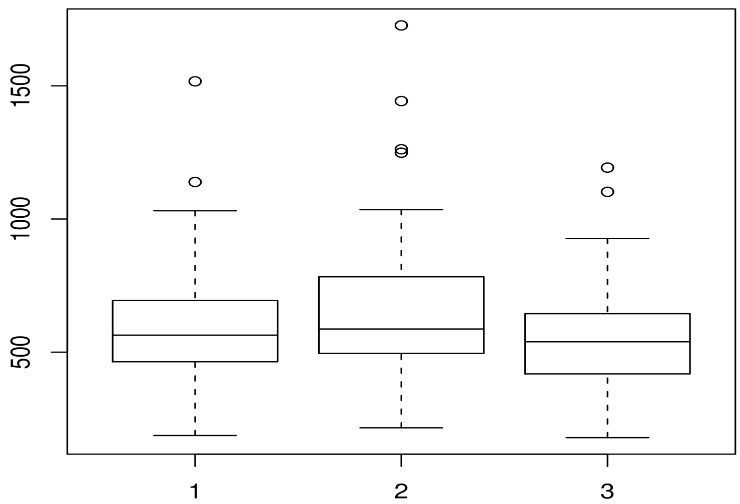

Figure 11: Smoke effects on retinol

No drinks	Less than 1	Less than 11	More than 11
572.3	598.8	619.4	716.7

Table 3: Plasma retinol means for alcohol groups

No drinks	Less than 1	Less than 11	More than 11
170.7928	190.5778	216.1053	167.1579

Table 4: Plasma beta-carotene means for alcohol groups

have, on average, a level of 193 ng/ml plasma beta-carotene and the smokers only 121.

3.3.6 Alcohol Use

As can be seen in Table 3, alcohol use raises plasma retinol concentrations. I.e., the group of "heavy drinkers" present a mean of about 716 ng/ml of plasma retinol, while non-drinkers have only 572 ng/ml. This differences are proved to be significant by ANOVA analyses, achieving a p value of less than 0.013. However, only not drinking or being hard drinkers prove to have "stand-alone" significant effects. A contrast analysis shows that the real important difference is between drinkers and non drinkers, with a $p = 0.0176$. It may be that alcoholics contain chemical elements that stimulate the production or the ability to retain plasma retinol. More details on this analysis are given in the technical appendix C, together with some residuals issues.

Table 4 shows that moderate use of alcohol can account for higher beta-plasma levels, though heavy use of it is a cause of lower plasma levels. This difference is significative ($p < 0.005$), together with the difference between the first three groups and the last one ($p \cong .0007$, average difference=25 ng/ml). The difference between the first and the last group is not significative. Influence analysis was done for this ANOVA, and it is illustrated in technical appendix C, together with the contrast analysis.

	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
smoke	2	0.8564	0.4282	4.1650	0.016422
sex	1	0.4801	0.4801	4.6703	0.031466
age	1	1.2144	1.2144	11.8127	0.000670
alcq	3	0.9991	0.3330	3.2395	0.022469
fat	1	0.6498	0.6498	6.3204	0.012450
$\sqrt[5]{\text{beta}}$	1	0.7770	0.7770	7.5578	0.006331
Residuals	305	31.3553	0.1028		

Table 5: Analysis of variance table, response=log(retpl)

4 A Model for Plasma Retinol: Multiple Regression Analysis

So the variables which have proven effective on plasma retinol are: plasma beta carotene, sex, age, fat, smoke and alcohol. The other variables had real high p-values, hence it is unreasonable to think that they would become significative in a multiple model.

I fitted some two-variables models, to see if any interaction was significative. I found a tendency to significativity ($p \cong .06$) in the interaction between sex and smoke, no interaction between age and smoke (so elderly smokers have no additional effects on plasma retinol as young ones), a tendency to significativity ($p \cong .06$) in the interaction between alcohol and smoke. However, even in the big model, the interactions between alcohol and smoke and between sex and smoke are not significative.

Table 5 shows the ANOVA table for this multiple regression. All the effects are confirmed. Unfortunately, the fit is not good, with an R-Squared of only about 14%. The transformed beta plasma has now a parameter estimate of 0.12, so, net of other significative effects, the presence of betaplasma is a strong indicator of good levels of plasma retinol. See technical appendix D for residuals issues and a validation of the model.

5 A Model for Plasma Beta: Multiple Regression Analysis

I fitted some two-variables models, to see if any interaction was significative (some details are given in the technical appendix). I found no interaction between alcohol use and smoke ($p \cong 0.16$), interaction between smoke and

	Df	Sum Sq	Mean Sq	F value	Pr($>F$)
betad	1	2.329	2.329	18.6877	2.108e-05
age	1	0.807	0.807	6.4754	0.0114470
log(retpl)	1	1.049	1.049	8.4173	0.0039965
smoke	2	1.436	0.718	5.7619	0.0035098
fat	1	0.550	0.550	4.4163	0.0364447
colest	1	0.578	0.578	4.6381	0.0320801
sex	1	0.748	0.748	5.9997	0.0148896
fiber	1	1.740	1.740	13.9598	0.0002242
vit	2	1.818	0.909	7.2929	0.0008101
quet	1	3.304	3.304	26.5086	4.802e-07
betad:vit	2	1.055	0.527	4.2311	0.0154282
smoke:vit	4	1.270	0.317	2.5472	0.0395860
Residuals	295	36.764	0.125		

Table 6: Analysis of variance table, response=$\sqrt[5]{beta}$

vitamin use ($p \cong 0.02$), smoke and fiber ($p \cong 0.008$), beta carotene intake and vitamin use ($p \cong 0.03$), vitamin and fiber ($p \cong 0.04874$). No third level interaction has been found to be significative.

In the multiple model, neither the alcohol use or the interaction between fiber and vitamins is any more significative. The first fact is due to the strange behavior of plasma beta-carotene levels with alcohol usage.

Table 6 shows the ANOVA table for this multiple regression. All the other effects are confirmed. The fit is reasonably good, with an R-Squared of about 32%. The multicollinearity between the dietary variables is not a problem, since we are looking at a model and not trying to interpret the single parameters (which have been given for simple regression models). Influence analysis show that observations number 36 and 39 are strongly influential. Table 7 show the values of the model variables for this two observations. Observation 36 has a lower beta plasma than expected ($residual \cong -0.6$), and observation 39 has a slightly higher plasma beta carotene than expected ($residual \cong 1.14$).

More details, together with other residuals issues and a validation of the model are given in the technical appendix D.

	age	sex	smoke	quet	vit	fat	fiber	colest	betad	beta	retpl
36	44	2	3	25.87867	1	95.3	17.5	253.1	7026	39	179
39	39	1	3	21.99912	1	109.1	4.7	461.1	998	418	665

Table 7: Influential observations

6 Tentative of a Micronutrients Model

Though the two micronutrients have proven connected with different variables, we may want to say something about them together. One way to study the effects of dietary and lifestyle variables on both micronutrients is by summing the two transformed variables. More details on this and on the following analysis are given in technical appendix D.

I used both forward inclusion and backward elimination variable selection methods. Both of them suggest the same model, whose covariates are:

1. Age
2. Quetelet Index
3. Vitamin Use
4. Cholesterol
5. Beta Intake
6. Alcohol Use.

It is interesting to notice that Sex is not in this model (it had opposite effect on the two micronutrients, so maybe its effect is "canceled"). The idea that sex is not effective on the sum of the micronutrients is confirmed by the first graph in Figure 12. The same idea apply to Smoke, though in the second graph of Figure 12 there is evidence of difference between non-current smokers (categories 1,2) and current smokers (category 3). The fact that smoking is not good for both plasma micro-nutrients has been confirmed in the separate studies. Fat is not in this model, maybe because of its strong collinearity with Cholesterol, which turns out to be the most effective in lowering micronutrients plasma levels. Age is in the model, so we should conclude that it has a direct link with raising micro-nutrients plasma concentrations.

This model achieves an R-Squared of only 20%, so more research is needed to explain plasma levels of this two micronutrients.

Figure 12: Boxplots for ineffective variables

7 Discussion & Conclusion

Part of the discussion and conclusion has been done in the previuos sections, however in this section we will try to summarize our results and give some recommendations.

The analysis showed that plasma beta-carotene is strongly influenced by dietary variables, that is:

- Fiber, alcohol, vitamin and beta carotene intake tend to raise beta concentrations, with a positive interaction between betacarotene and vitamin intake.
- Fat and cholesterol tend to lower betacarotene concentrations.

Moreover, we can conclude that smoking lowers plasma concentrations of this micronutrient; and the same is for body mass (Quetelet index), especially when people get obese. Females tend to have higher concentrations than men, and elderly tend to have higher concentrations than younger adults.

Plasma retinol is less influenced by these variables than beta carotene, though we proved some relationships, that is:

- Beta carotene plasma levels and alcohol use tend to raise plasma retinol
- Fat intake and smoke tend to lower it.

Moreover, we can conclude that elderly tend to have higher concentrations than younger adults and men tend to have higher concentrations than women.

Plasma concentrations of both micronutrients are predicted in this way:

- Vitamin, beta-carotene and alcohol intake tend to raise plasma levels.
- Cholesterol and body mass tend to lower plasma levels.

Moreover, we can conclude that elder adults tend to have higher concentrations of micronutrients in the plasma.

The analysis showed significative relationships between micronutrients concentrations and this personal facts, because we detected low p values for many variables. However, much more must be done.

First of all, lack of fit has been detected in all models: no variable does a good job as a predictor. The models can't be used for a satisfactory prediction of the response variables.

Lack of fit can be due to need of better variable selection: there should be other variables that can be used as better predictors; though some, used in this study, are promising covariates for future models.

This problem can be due also to need for a better study design: all of the observations are patients who had an elective surgical procedure, who were found not to have developed a cancer. It would be interesting to survey patients who have developed a cancer; and also patients not undergone a surgical procedure, that is, patients who never were suspected of developing a cancer. This is also a problem in generalizing our results: if we would like to claim that our conclusions hold for the entire adult population, we should develop a study on a sample of the entire adult population. However, there is no evidence of bias in the sample (that is, there is no evidence of the fact that the results should apply only to former patients of elective surgical procedures). So I think we can safely generalize our results, while developing studies on different samples.

Lack of fit problem is bigger for plasma retinol, whose model achieves a slightly low R squared (14%). As I suggested, it may be that plasma retinol is auto-regulated in some sense by the organism. This fact would be a way to explain its lower variability and its mild relations with personal facts variables. However, we can‘t make this conclusion with avaliable informations, and other studies should be designed to investigate this fact.

Other studies should also investigate the relationship between plasma retinol and smoke, because we came to counter-intuitive results in this study. It has been detected a significative difference between current and former smokers, and former smokers had higher levels of plasma retinol. We suggested that this can be due to a reaction of the organism to dropping smoking, or to a relation between dropping smoking and switching to other healthy lifestyles. However, we can't make any conclusion to explain this peculiar behavior of plasma retinol concentrations; so other studies can be done both to validate this conclusion and to try to explain its reason.

I would like to suggest that biological studies need more scientific measures for the variables. I.e., it may be better to use the grams of alcohol in the blood rather than the average number of drinks, grams of vitamins consumed rather than frequency of use, and so on.

We concluded, using this data, that there are "natural" outliers with high levels of plasma micronutrients. Recall the density plots in Figure 7, page 10. It is a challenge to understand why certain individuals have got so high levels, but this is another issue we couldn't address here. A good conclusion we can make is that outlier concentrations of plasma micronutrients are healthy high (right tail of density estimates).

While there are many questions left, we can make some important recommendations. Overall, it seems like healthier lifestyles lead to healthier (higher) levels of plasma micronutrients. In fact, we can recommend to maintain high intakes of fiber, use vitamins, drop smoking. We can also recommend to avoid heavy intakes of cholesterol and fat; and to control body mass in a healthy low level. Moreover, we have seen that use of alcohol tend to raise plasma concentrations, so we can recommend moderate use of alcohol[3].

A Appendix: Head of the Dataset

This datafile contains 315 observations on 14 variables.

Variable Names in order from left to right:

```
AGE: Age (years)
SEX: Sex (1=Male, 2=Female).
SMOKE: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUET: Quetelet index (weight/(height^2)); values above 27 kg/m^2
          (female) or 28 kg/m^2 (male) indicate obesity
VIT: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
CAL: Number of calories consumed per day.
FAT: Grams of fat consumed per day.
FIBER: Grams of fiber consumed per day.
ALC: Number of alcoholic drinks consumed per week.
COLEST: Cholesterol consumed (mg per day).
BETAD: Dietary beta-carotene consumed (mcg per day).
RETD: Dietary retinol consumed (mcg per day)
BETA: Plasma beta-carotene (ng/ml)
RETPL: Plasma Retinol (ng/ml)
```

	age	sex	smoke	quet	vit	cal	fat	fiber	alc	colest	betad	retd	beta
1	64	2	2	21.48380	1	1298.8	57.0	6.3	0.0	170.3	1945	890	200
2	76	2	1	23.87631	1	1032.5	50.1	15.8	0.0	75.8	2653	451	124
3	38	2	2	20.01080	2	2372.3	83.6	19.1	14.1	257.9	6321	660	328
4	40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153
5	72	2	1	20.98504	1	1952.1	82.6	16.2	0.0	170.8	2863	1209	92
6	40	2	2	27.52136	3	1366.9	56.0	9.6	1.3	154.6	1729	1439	148

[3]Let's remember that other studies suggest that moderate use of red wine, for instance, is an aid in preventing some diseases; so this conclusion is not particularly amazing.

```
7   65   2     1 22.01154   2 2213.9 52.0  28.7  0.0  255.1  5371  802  258
8   58   2     1 28.75702   1 1595.6 63.4  10.9  0.0  214.1   823 2571   64
9   35   2     1 23.07662   3 1800.5 57.8  20.3  0.6  233.6  2895  944  218
10  55   2     2 34.96995   3 1263.6 39.6  15.5  0.0  171.9  3307  493   81
   retpl
1    915
2    727
3    721
4    615
5    799
6    654
7    834
8    825
9    517
10   562
```

B Appendix: Transformations

In order to use beta and retpl as response variables, we must have no evidence against their normality. On the other hand, the covariates were transformed when the skewness was strong, fact that could inficiate the results of the regression. We will now examine each transformation separately.

B.1 Plasma Retinol

From the first row of graphs in Figure 13 it can be seen that plasma retinol has got a long right tail and some high outliers, thus providing evidence against the normality. A log-transformation or a power-transformation, with power less than 1, usually works well with this kind of problems. Confronting the square-root and the log-transformation, it seemed to me that the logarithm of plasma retinol was closer to normality.

B.2 Plasma Beta-Carotene

Plasma beta-carotene has got the same distribution problems of plasma retinol, but it is far more skewed and there is no evidence of normality (Figure 14, first row). Strong transformations were needed, and after some tries I decided to use the 0.2 power in order not to lose too much variability with the transformation. As I said, when using beta-carotene as a response variable, the low outlier "0" was dropped from the dataset. This is both

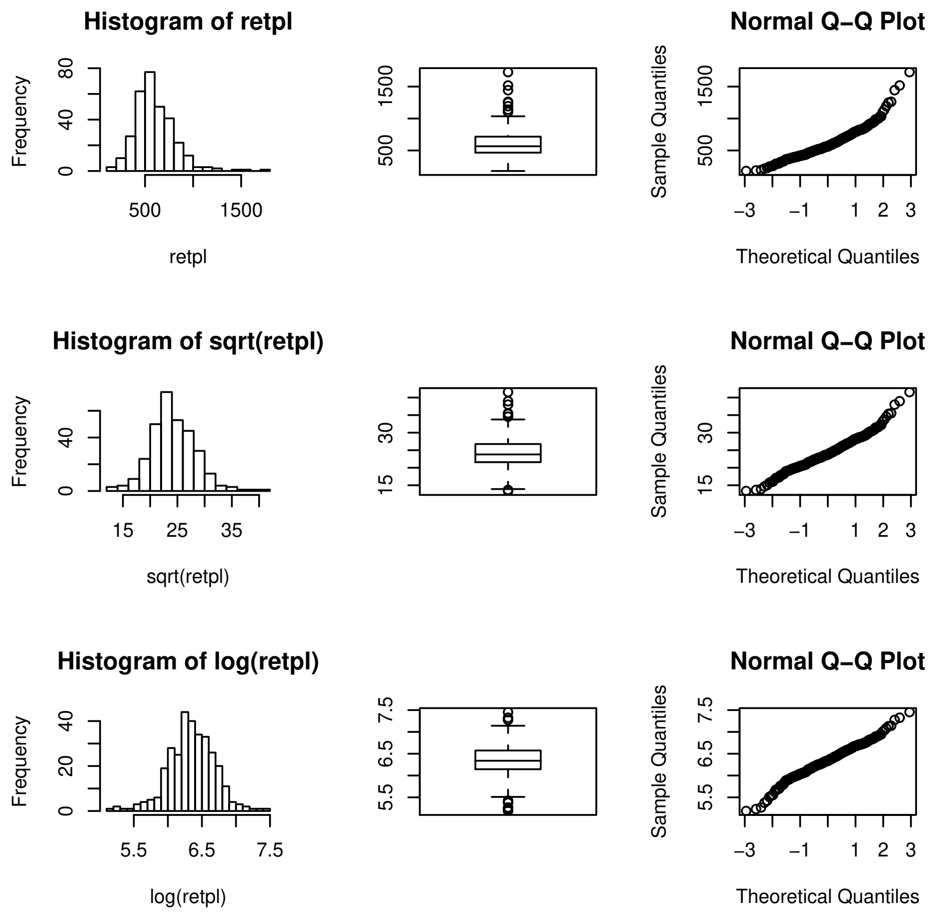

Figure 13: Transformations for retinol

because we suspected miscoding and because outlyingness in the response variable makes fit harder[4], since it is a strong evidence of non normality. The zero becomes a really strong outlier for the transformed variable (Figure 14, last row).

B.3 Retinol Intake

As can be seen in Figure 4, pag.7, the retinol intake is too strongly skewed to be useful in the analysis. I confronted usual transformations, and decided to use a log transformation confronting an index proposed in (**?**, pag.451) and for parallelism with the transformation done on plasma retinol. The skewness index used is the difference between the distances of the first and the third quartile from the median, divided by the IQR (inter-quantile range), that is $\frac{(Q_3-Me)-(Me-Q_1)}{Q_3-Q_1}$. The log transformation of retd had an index of only -0.005, while the square root an index of 0.09.

R code:

```
> (sum(quantile(log(retd),c(.25,.75)))-
+  2*median(log(retd)))/diff(quantile(log(retd),c(.25,.75)))
        75%
-0.005500357
> (sum(quantile(sqrt(retd),c(.25,.75)))-
+ 2*median(sqrt(retd)))/diff(quantile(sqrt(retd),c(.25,.75)))
      75%
0.0905506
```

B.4 Quetelet Index and Obesity

Since the Quetelet index is usually used to determine whether a subject is obese or not, I created a dummy variable for obesity and used it in the analysis. This was done giving a 1 to all males with index greater than 28 or to all females with index greater than 27; and 0 to all the other subjects. The code used was:

```
obM_ifelse(quet>=28 & sex==1,1,0)
obF_ifelse(quet>=27 & sex==2,1,0)
ob_obF+obM
```

[4] Again, cfr handout on outliers, http://www.stat.cmu.edu/~brian/707/

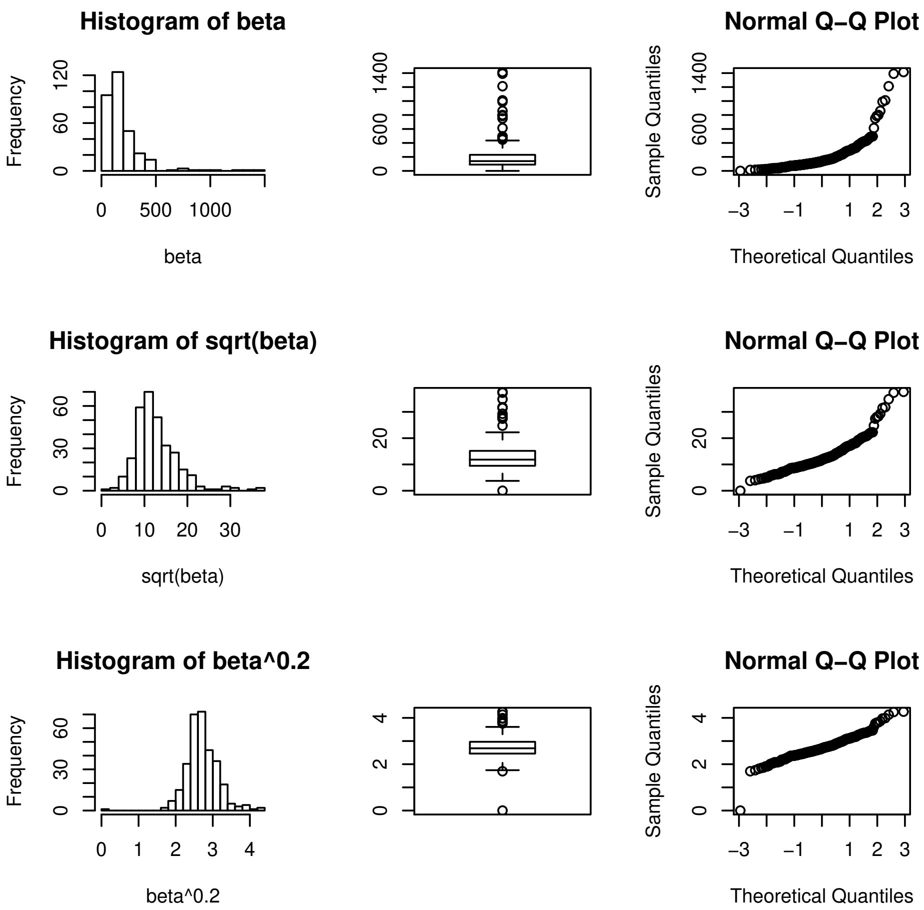

Figure 14: Transformations for beta-carotene

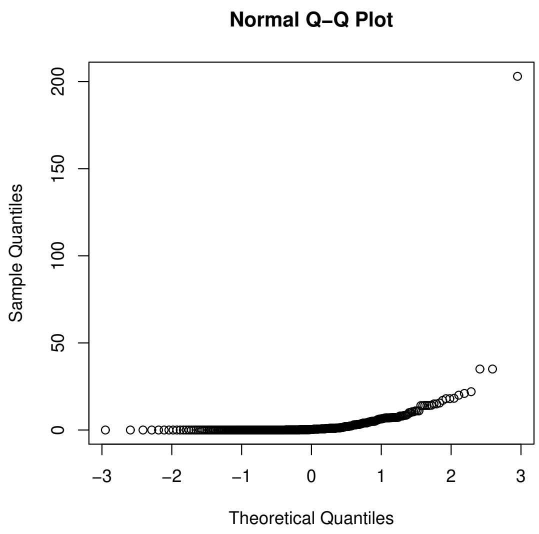

Figure 15: Quantile normal plot, alcohol

B.5 Alcohol

As can be seen in the histogram, Figure 5 on page 8, and in the quantile normal plot in Figure 15, this variable is strongly skewed.

As I noticed, there are too many 0 values (111) for any transformation to be effective, so my suggestion is to consider the average number of drinks consumed as an ordered qualitative variable, dividing it into opportune classes. There were two choices: dividing the variable into "ideal" classes, thus deciding what was to be considered as hard, medium, low drinking; or looking at the data and splitting the variable into the classes that looked natural for this dataset. I decided to use the second method, first of all because it should be better for the subsequent analysis, and then because this variable is so peculiarly distributed that no "ideal" split would have been reasonable. This variable has, moreover, a really strong outlier which have been considered separately sometimes. When dividing the variable in classes, outlyingness is no more important (robustness of the procedure: same weight is given to any drinker in the same class).

The stem and leaf is:

```
The decimal point is 1 digit(s) to the right of the |
```

```
 0 | 000000000000000000000000000000000000000000000000000000000000000000+211
 1 | 00111114444445567888
 2 | 012
 3 | 55
 4 |
 5 |
 6 |
 7 |
 8 |
 9 |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
17 |
18 |
19 |
20 | 3
```

and, without observation 62 (203 drinks per week):

```
The decimal point is at the |

 0 | 000000000000000000000000000000000000000000000000000000000000000000+137
 2 | 000000122345690001222224577
 4 | 0001112557900000267
 6 | 122455780000011222223
 8 | 0003450
10 | 0055000
12 |
14 | 001112005
16 | 0
18 | 002
20 | 00
22 | 0
24 |
```

```
26 |
28 |
30 |
32 |
34 | 00
```

So, without the highest observation, groups can be distinguished. I decided to make a group of no alcohol users, because in a biology analysis it can be useful to know whether alcohol has been introduced in the organism or not. Then I cut a group of very low drinkers, say less than 1 drink per week. The stem and leaf shows a separation between 11 and 13, so I decided to cut the third group there, and to consider the others as a fourth group of heavy drinkers. A fifth (the two 35) and a sixth (the 203) groups could be considered, and in fact I created them and used them for outliers study. I decided however to consider the division with only four groups, because there is a trade off between the number of levels and the easyness to handle. So I preferred having less groups and doing a deeper analysis to use more information given by this variable. Together with the dummy variables, a new qualitative variable `alcq` (with 4 levels) was created.

R Code:

```
> length(alc[alc<1])/314
[1] 0.6019108
> alc0_ifelse(alc==0,1,0)
> alcm1_ifelse(alc>0 & alc<=1,1,0)
> alc111_ifelse(alc>1 & alc<=11,1,0)
> alcM11_ifelse(alc>11,1,0)
> alc35_ifelse(alc==35,1,0)
> alc203_ifelse(alc==203,1,0)
> alcq_4-ifelse(alc==0,1,0)-ifelse(alc<=1,1,0)-ifelse(alc<=11,1,0)
```

C Appendix: Details on Simple Analyses

I tried to see if outlyingness in plasma levels of one micro-nutrient could be an indicator of outlyingness in the other one, but it wasn't so: no observation was an outlier for both the micronutrients, and even between the other variables there was no outlier in common (apart between alcohol and calories, as I said).

R code:

```
 any(pr[retpl>1000,1]==pr[beta>600,1])
[1] FALSE
 any(pr[retpl<300,1]==pr[beta>600,1])
[1] FALSE
any(pr[retpl>1500,1]==pr[beta<70,1])
[1] FALSE
```

C.1 Alcohol on the Two Micronutrients

Figure 16 shows the strong influence of observation 62 (alcq=4, alc=203, retpl=317) in the regression between plasma retinol and alcohol. This observation is in fact an outlier, and it can be useful to understand the effects of being an alcoholic on plasma retinol. But, in this case, we are conducting a more general analysis, so it can be more interesting to let the other observations be more influential. So I dropped observation 62 from the data set. Not surprisingly, the difference in the parameter estimate is not very sensible.

I did also some contrasts, with null hypotheses that $\mu_0 = \mu_1$, $\mu_1 = \mu_2$, $\mu_0 = \mu_3$, $\mu_1 + \mu_0 = \mu_3 + \mu_2$ and $\mu_0 = \frac{\mu_1+\mu_2+\mu_3}{3}$, where μ_i is the effect of the i-th level on plasma retinol. The third one was significative, stating that there is difference between the effects of not drinking and consuming 1 to 11 drinks per day. The fifth contrast states that there is significative difference between not drinking and drinking. The second contrast verifies the hypothesis that there is no difference between not drinking or consuming only less than one drink per day, which is not significant. So low drinking is the same of no drinking with respect to plasma retinol concentrations. The other contrasts were also found to be not significative.

Here is the R code I used (for the last contrast, for instance):

```
c1_c(3,-1,-1,-1)
temp_cbind(alc0,alcm1,alc111,alcM11)
temp1_temp %*% c1
aov(log(retpl)~temp1)
```

Coming to the relatioship between plasma beta-carotene and alcohol use, looking at the first boxplot in Figure 17, I thought that the difference could be significative because of the outliers. In particular, observation number 208 (alcq=4, alc=15, beta=1212) seemed really influential. But then I run a regression without all of the outliers, and got the same results (and, in some cases, lower p-values).

R code and output:

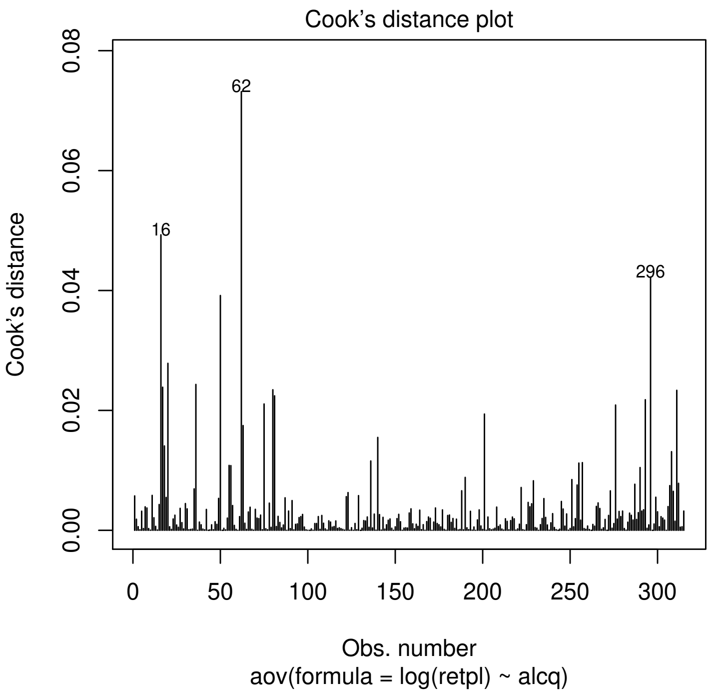

Figure 16: Cook Distances

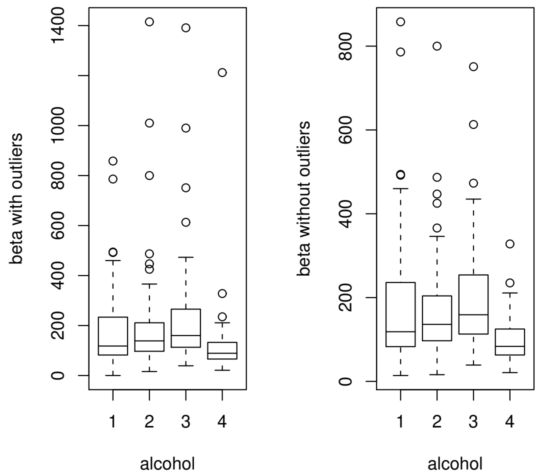

Figure 17: Boxplots for alcohol and beta

```
> a_aov(b[-257]~alcq[-257])
> summary(a)
             Df Sum Sq Mean Sq F value  Pr(>F)
alcq[-257]    3  1.720   0.573  3.4369 0.01726 *
Residuals   310 51.727   0.167
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  '
'  1

> a_aov(b[-c(257,208,40,219,262,263,3)]~alcq[-c(257,262,208,40,219,263,3)])
> summary(a)
                                         Df Sum Sq Mean Sq F value   Pr(>F)
alcq[-c(257, 262, 208, 40, 219, 263, 3)]   3  2.122   0.707 5.2317  0.001557
Residuals                               304 41.096 0.135

> a_aov(b[-c(257,208,40,219,262,263)]~alcq[-c(257,262,208,40,219,263)])
> summary(a 1 1 1 -3)
                                      Df Sum Sq Mean Sq F value   Pr(>F)
alcq[-c(257, 262, 208, 40, 219, 263)]   3  1.815   0.605 4.4327  0.004556
```

```
Residuals                              305 41.635 0.137
```

C.2 Beta Intake on Beta

This is the summary of the regression between plasma beta and beta-carotene intake.

```
Call: lm(formula = betapl^0.2 ~ betadiet)

Residuals:

    Min      1Q    Median     3Q   Max
 -2.645 -0.2455 -0.006834 0.2226 1.434

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.5803  0.0435    59.3402  0.0000
   betadiet  0.0001  0.0000     3.8175  0.0002

Residual standard error: 0.431 on 313 degrees of freedom Multiple
R-Squared: 0.04449 F-statistic: 14.57 on 1 and 313 degrees of
freedom, the p-value is 0.0001625

Correlation of Coefficients:
         (Intercept)
betadiet -0.8295
```

So the diet is not a good predictor of the presence of betaplasma, while it is obviuosly effective.

From Figure 18 we can see that there is no evidence against the normality of the residuals, but there is one strong low outlier, observation 257. This is one of the reasons why I decided to run the regressions with "beta" as response variable without this outlier.

In fact, the residuals of the model without observation 257 go from a minimum of -1 to a maximum of 1.43, and there is no evidence of not normality (Figure 19).

We can see from the following regression summary that without 257 the slope achieves a lower p-value, though being itself lower.

```
 Call: lm(formula = b[-257] ~ betad[-257])
```

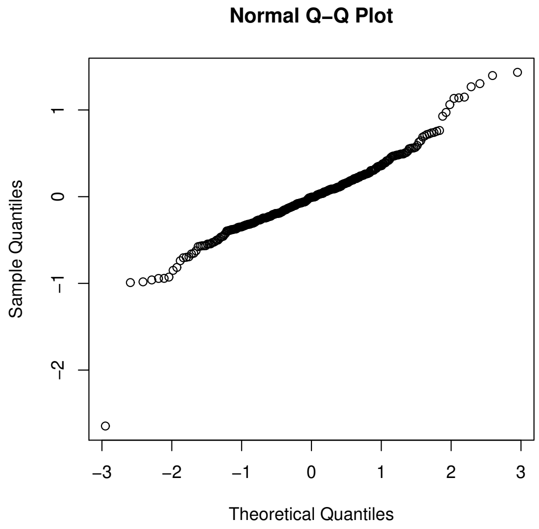

Figure 18: Residuals for beta on betadiet

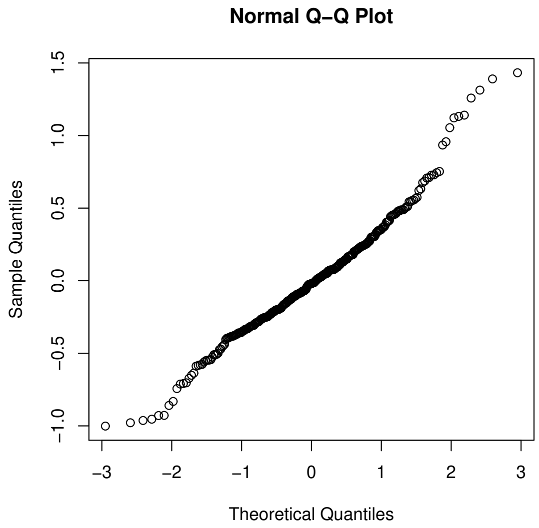

Figure 19: Residuals for beta on betadiet, without 257

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.00158 -0.25424 -0.02012  0.22174  1.43268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.599e+00  4.093e-02   63.49  < 2e-16 ***
betad[-257] 5.849e-05  1.551e-05    3.77 0.000195 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1
‘ ’  1

Residual standard error: 0.4048 on 312 degrees of freedom Multiple
R-Squared: 0.04358,    Adjusted R-squared: 0.04051 F-statistic:
14.21 on 1 and 312 degrees of freedom,     p-value: 0.000195
```

R code:

```
> a_residuals((lm(b[-257]~betad[-257])))
> a[a==min(a)]
      233
-1.001575
> a[a==max(a)]
     219
1.432675
> qqnorm(a)
> summary(a)
```

C.3 Fiber on Beta

We said that the relation between plasma beta-carotene and fiber is milder than detected in the simple regression analysis, because beta intake is strongly correlated with fiber (.48) and strongly influential on plasma beta. In fact, whenever two variables X_1 and X_2 are correlated, the parameter estimates $\hat{\beta}$ of each varies when the other is in the model, that is, $\hat{\beta}(X_1) \neq \hat{\beta}(X_1|X_2)$. If correlation is high, it can be that $\hat{\beta}(X_1)$ and $\hat{\beta}(X_1|X_2)$ have got different sign. It is important to avoid interpretation of the parameters in presence of multicollinearity, while the model is still useful for prediction. For more details, see (**?**).

C.4 Vitamins

I created two dummy variables to investigate the relationship of the micronutrients with vitamin use. This are the summaries for the regressions between plasma retinol and vitamins:

```
             Df Sum Sq Mean Sq F value Pr(>F)
vityes        1  0.012   0.012  0.1038 0.7476
Residuals   313 36.320   0.116

             Df Sum Sq Mean Sq F value Pr(>F)
vitno         1  0.011   0.011  0.0933 0.7602
Residuals   313 36.321   0.116

             Df Sum Sq Mean Sq F value Pr(>F)
vitno         1  0.011   0.011  0.0931 0.7605
vityes        1 0.004   0.004  0.0311 0.8602
Residuals   312 36.318   0.116

                Df Sum Sq Mean Sq F value Pr(>F)
as.factor(vit)   2  0.014   0.007  0.0621 0.9398
Residuals      312 36.318   0.116
```

So it is evident that there is no relationship between vitamin use and plasma retinol.

R code:

```
vityes_ifelse(vit==1,1,0)
vitno_ifelse(vit==3,1,0)
a_aov(log(retpl)~vityes)
summary(a)
etc.
```

Coming to the relationship between plasma beta and vitamin use, the coefficient estimate of often use of vitamins is 0.1815302, and this is a summary:

```
              Df Sum Sq Mean Sq F value    Pr(>F)
vityes[-257]   1  2.451   2.451  14.994 0.0001314 ***
Residuals    312 50.996   0.163
```

The means for the three groups seem distant (240, 185, 136). Some contrasts were fitted. I saw significative difference between occasional use and no vitamin use, and this is a summary (with R code):

```
> cont_c(0,1,-1)
>d1_ifelse(v==1,1,0)
>d2_ifelse(v==2,1,0) d3_ifelse(v==3,1,0)
>cc_cbind(d1,d2,d3) %*% cont
> summary.lm(aov(b[-257]~cc[-257]))

Call: aov(formula = b[-257] ~ cc[-257])

Residuals:
     Min       1Q   Median       3Q      Max
-1.12810 -0.26695 -0.03865  0.24043  1.53241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73487    0.02319  117.93  < 2e-16 ***
   cc[-257] 0.08845    0.02958    2.99  0.00301 **
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  '
'  1

Residual standard error: 0.4081 on 312 degrees of freedom
Multiple R-Squared: 0.02786,    Adjusted R-squared: 0.02474
F-statistic: 8.941 on 1 and 312 degrees of freedom,
p-value: 0.003011
```

Other R code:

```
lapply(split(beta,vit),mean)
$"1" [1] 240.959
$"2" [1] 185.6585
$"3" [1] 136.8919
```

C.5 Obesity on Beta

This is the R code I used to count obese individuals:

```
> obM_ifelse(quet>=28 & sex==1,1,0)
> obF_ifelse(quet>=27 & sex==2,1,0)
> ob_obF+obM
> sum(ob)
[1] 101
```

And this is a summary of the regression:

```
Call: lm(formula = b[-257] ~ ob[-257])

Residuals:
     Min       1Q   Median       3Q      Max
-0.97005 -0.24708 -0.03092  0.24169  1.45882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80846    0.02708 103.719  < 2e-16 ***
   ob[-257] -0.25674    0.04798  -5.351 1.70e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 0.3961 on 312 degrees of freedom
Multiple R-Squared: 0.08405,    Adjusted R-squared: 0.08112
F-statistic: 28.63 on 1 and 312 degrees of freedom,
p-value: 1.698e-007
```

C.6 Sex on Plasma Retinol

Figure 20 shows a little evidence of not normality of the residuals of this regression, but since n is very big we can conclude that normality holds anyway. The observations out of the spans are too few to have convincing evidence against the normality hypothesis.

C.7 Smoke on Plasma Retinol

I coded the dummy variables in this way:

```
FormSmok_ifelse(smoke==2,1,0)
NeverSmok_ifelse(smoke==1,1,0)
```

And this is a summary of the regression:

```
Call: lm(formula = log(retpl) ~ FormSmok)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11968 -0.20518  0.01087  0.23252  1.04339
```

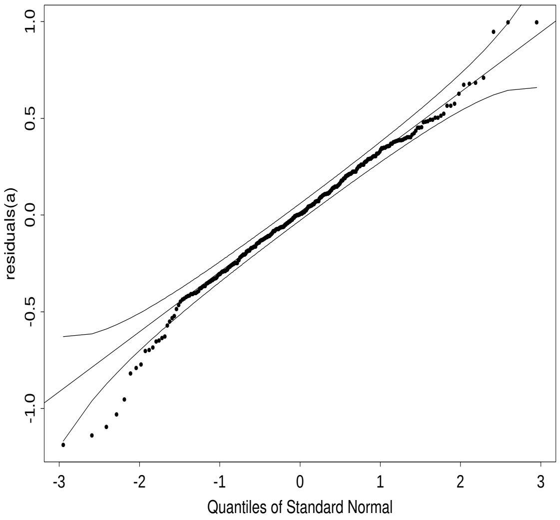

Figure 20: QQnorm residuals of Sex on plasma retinol

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.30707    0.02383 264.675  < 2e-16 ***
    FormSmok 0.10368    0.03944   2.629  0.00899 **
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1
' '  1

Residual standard error: 0.337 on 313 degrees of freedom
Multiple R-Squared: 0.0216,     Adjusted R-squared: 0.01848
F-statistic: 6.911 on 1 and 313 degrees of freedom,
p-value: 0.00899
```

So being a former smoker is a predictor of high levels of plasma retinol and, obviusly, smoking is a predictor of low levels of plasma retinol.

I did some contrasts, one to see if there is difference between being a former smoker or not smoking (otherwise the important is not smoking at the moment). This is a summary of the ANOVA contrast:

```
Call: aov(formula = log(retpl) ~ cc)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-1.163519 -0.198349 -0.003393  0.235611  1.058343

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.35091    0.01925 329.863   <2e-16 ***
         cc -0.04489    0.02072  -2.167    0.031 *
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1
' '  1

Residual standard error: 0.3382 on 313 degrees of freedom
Multiple R-Squared: 0.01478,    Adjusted R-squared: 0.01163
F-statistic: 4.695 on 1 and 313 degrees of freedom,
p-value: 0.03101
```

So the difference is significative.

R code:

```
cont_c(1,-1,0)
s1_ifelse(smoke==1,1,0)
s2_ifelse(smoke==2,1,0)
s3_ifelse(smoke==3,1,0)
cc_ cbind(s1,s2,s3) %*% cont
aov(log(retpl) ~ cc)
```

D Appendix: Multiple Regression Models

D.1 Plasma Retinol

This is a summary of this multiple regression:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.974581 -0.194648 -0.005489  0.213885  1.072723

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.9194880  0.1621913  36.497  < 2e-16 ***
```

```
smoke2       0.0840586  0.0407689   2.062  0.04007 *
smoke3       0.0003735  0.0571820   0.007  0.99479
sex2        -0.0939250  0.0592729  -1.585  0.11409
age          0.0038033  0.0013863   2.744  0.00644 **
alcq2        0.0973306  0.0468334   2.078  0.03852 *
alcq3        0.0605650  0.0468700   1.292  0.19727
alcq4        0.2310371  0.0827372   2.792  0.00556 **
fat         -0.0012858  0.0005720  -2.248  0.02530 *
b            0.1193555  0.0434153   2.749  0.00633 **
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'
0.05  `.'  0.1  ` '  1

Residual standard error: 0.3206 on 305 degrees of freedom
Multiple R-Squared: 0.137,      Adjusted R-squared: 0.1115
F-statistic: 5.379 on 9 and 305 degrees of freedom,
p-value: 7.651e-007
```

From the graphs in Figure 21 we can see no evidence against normality of the residuals, and that no observation is particularly influential.

This is a stem and leaf of the leverage (h_{ii}) values:

```
  The decimal point is 2 digit(s) to the left of the |

   1 | 22223334444444444555555555555555555555555566666666666666666666677777+41
   2 | 000000000000000001111111111111111111111222222222233333444444444555555566+6
   3 | 00001112222222222233333344444555555556666666666677777888899 9
   4 | 0000001111222356678
   5 | 0122233445689
   6 | 34488899
   7 | 01455669
   8 | 11
   9 | 7
  10 | 4
  11 |
  12 |
  13 |
  14 | 1
```

The large value is observation 257, obviuosly; whose cookd is 0.02082197, so even with a high leverage 257 is not heavily influential. $\bar{h} = 9/315$, and

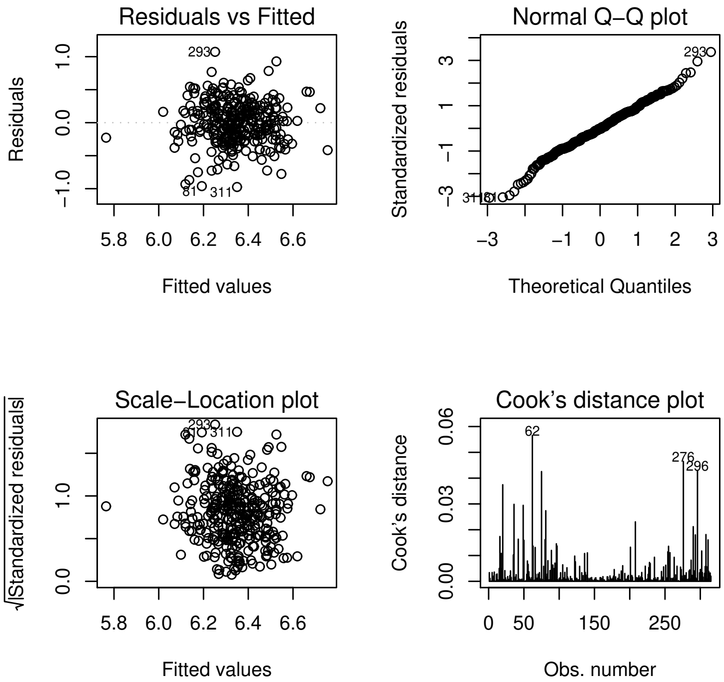

Figure 21: Residual analysis for plasma retinol

Obs	Leverage	Cookd
208	.10393	0.0185
257	.14076	0.0208
305	.09656	0.0001

Table 8: Influential observations

so the "rule of thumb" for selecting high leverage observations tells us that 3 observations are over the $3 * \bar{h}$ (0.08571429) threshold: observation 208, 257 and 305. Table 8 shows that these observations have got low cookds, so we can conclude that no observation is particularly influential in this model.

I tried to validate the model taking a random subset of the data ($n/2$ observations), fitting the model on this subset and then looking at the other observations in relation to the model. I took the predicted values on the not fitted subset. It is interesting to notice that the new model has an R-Squared of 14%, close to the real R-Squared obtained.

Figure 22 shows that the new residuals (difference between real and predicted plasma retinols) are reasonably close to normality.

R code:

```
h_lm.influence(a)$hat
stem(h)
3*sum(h)/315
cd_cooks.distance(a)
part1_sample(315,157)
a_lm(log(retpl)~smoke+sex+ age + alcq + fat
+ b, data=pr[part1,])
ft_predict.lm(a,pr[-part1,])
rs_log(retpl[-part1])-ft
```

D.2 Plasma Beta

This is a summary of the multiple regression model on betaplasma:

```
 Call: lm(formula = b ~ betad + age + log(retpl) +
smoke + fat + colest + sex + fiber + vit + quet + betad:vit +
smoke:vit)

Residuals:
```

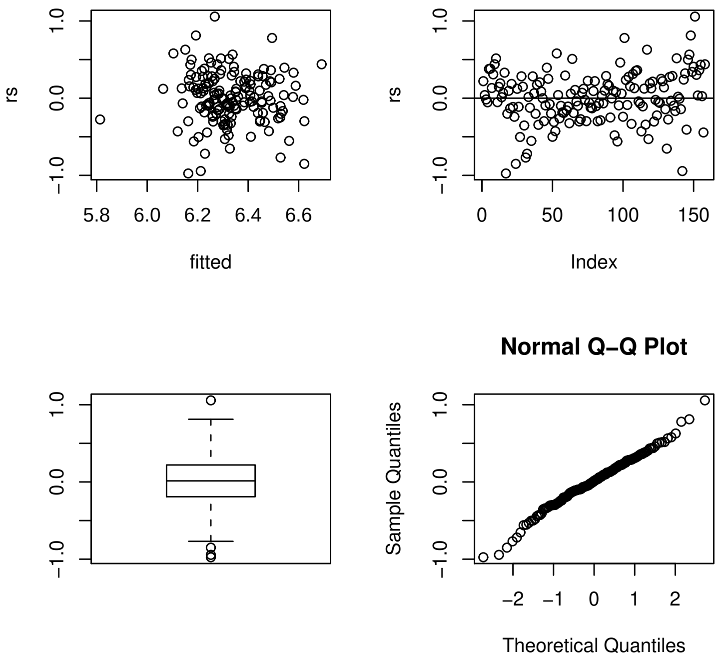

Figure 22: Validation of the model

```
     Min       1Q   Median       3Q      Max
-0.83967 -0.21046 -0.03350  0.18819  1.23010

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.597e+00  4.340e-01   3.680 0.000278 ***
betad        8.064e-05  2.292e-05   3.519 0.000502 ***
age          2.230e-03  1.560e-03   1.430 0.153780
log(retpl)   1.991e-01  6.258e-02   3.181 0.001624 **
smoke2      -4.991e-02  7.093e-02  -0.704 0.482171
smoke3      -3.519e-01  1.235e-01  -2.849 0.004697 **
fat         -4.785e-04  8.926e-04  -0.536 0.592332
colest      -2.824e-04  2.301e-04  -1.227 0.220725
sex2         9.981e-02  6.704e-02   1.489 0.137631
fiber        1.138e-02  4.560e-03   2.496 0.013109 *
vit2         5.096e-02  1.041e-01   0.490 0.624692
vit3         6.436e-02  9.874e-02   0.652 0.515026
quet        -1.678e-02  3.418e-03  -4.909 1.52e-06 ***
betad:vit2  -6.917e-05  3.394e-05  -2.038 0.042428 *
betad:vit3  -8.696e-05  3.373e-05  -2.578 0.010416 *
smoke2:vit2  1.485e-01  1.112e-01   1.335 0.182782
smoke3:vit2  2.673e-01  1.716e-01   1.557 0.120477
smoke2:vit3 -1.386e-01  1.021e-01  -1.357 0.175759
smoke3:vit3  2.249e-01  1.546e-01   1.455 0.146821

Residual standard error: 0.353 on 295 degrees of freedom
Multiple R-Squared: 0.3121,     Adjusted R-squared: 0.2702
F-statistic: 7.437 on 18 and 295 degrees of freedom,
p-value: 7.772e-016
```

Figure 23 shows a quantile-normal plot of the residuals, and there is no evidence of not normality.

Figure 24 shows some other diagnostics. We can see that the residuals aren't particularly high, though there are some outliers. There are some observations with high leverage, and another with high Cook's distance (observation number 39, whose Cookd is 0.1, residual 1.13 and leverage .14). $3 * \bar{h} = 0.1815287$, but I think we should conclude anyway that observation 39 has an high leverage, and is strongly influential.

This is a stem and leaf of the h_{ii}'s.

```
The decimal point is 2 digit(s) to the left of the |
```

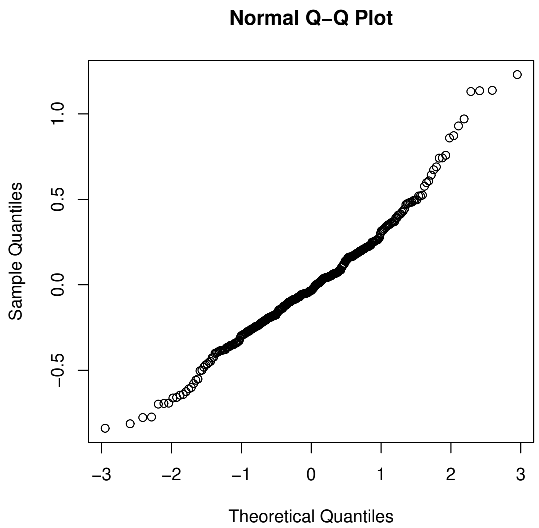

Figure 23: QQnorm, plasma beta

```
 0 | 7
 2 | 3345666777778889999900000111111111222222223333344444445555666666777+3
 4 | 000000011111111111222223333333344444445555566666667777777888889999999+26
 6 | 00011112222233334445555666667788899000122223333333334456677899 9
 8 | 0123334455778890234455556789
10 | 2455677922345667
12 | 11134497
14 | 1467
16 | 6
18 |
20 | 9
22 |
24 | 5
26 |
28 |
30 |
32 | 6
```

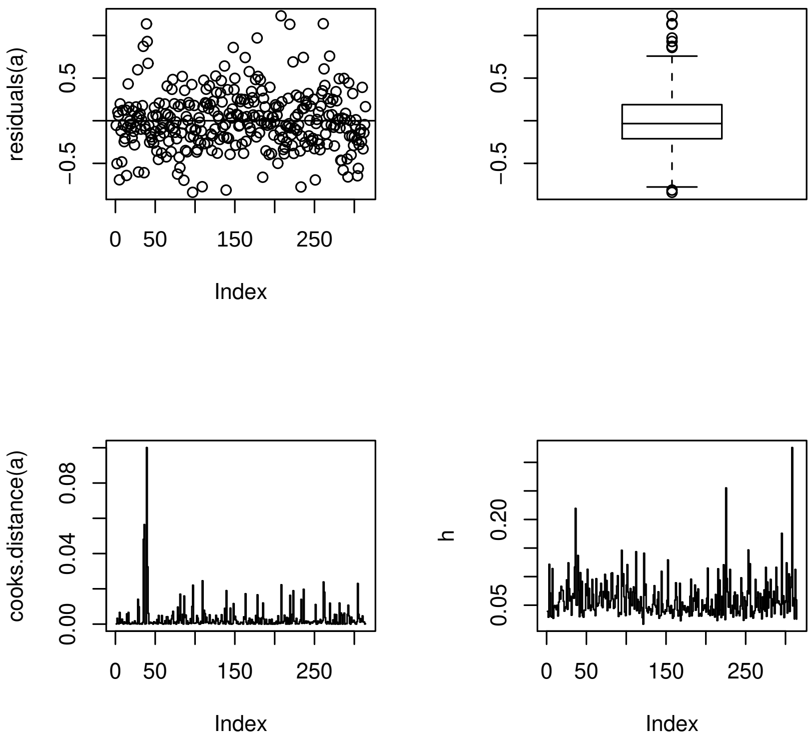

Figure 24: Residual Analysis

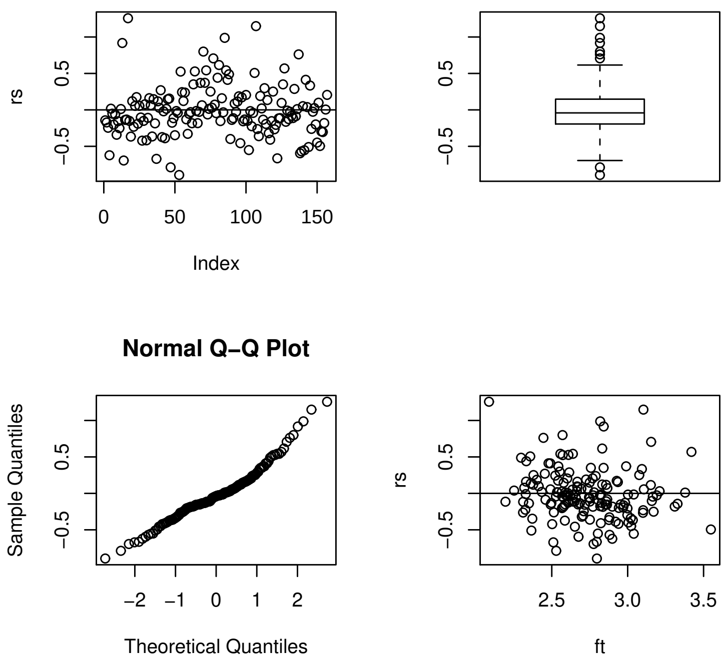

Figure 25: Validation of the model

The highest three observations, with a leverage of more than $3 * \bar{h}$, are numbers 36, 225 and 308. Their cookd is .06, .002 and .003, so we should conclude that observation 36 is strongly influential in the model. Its residual is -0.60926682. The values of the variables for this two observations are given in table 7, pag.21.

I validated this model using the same method used for plasma retinol multiple model. The graphs for this analysis are given in Figure 25. There is no evidence of problems in the model.

R code:

```
part1_sample(314,314*.5)
p1_lm(as.formula(a$call),data=pr1[part1,])
ft_predict.lm(p1,pr1[-part1,])
rs_b[-part1]-ft
```

D.3 Summed Micronutrients Model

The idea to study the variables together by summing them comes from the fact that we want to study what brings about high levels of both of them, while they have proved to react differently to the same treatment. So, a low value of the sum implies that both plasma levels are low, and of course an high level of the sum implies high levels of them both.

Moreover, we need to sum the transformed variables instead of the real ones; because the response variable in a classical regression analysis need to be normally distributed. A sum of normally distributed variables is still normally distributed, with mean the sum of the means, and variance the sum of the variances plus double the covariance. So we will end up with more variability. In this case, $var(\sqrt[5]{beta}) = 0.1938161$, $var(log(retpl)) = 0.1157071$, $cov(\sqrt[5]{beta}, log(retpl)) = 0.02830927$ and, in fact, $var(\sqrt[5]{beta} + log(retpl)) = 0.3661418$.

I used automatic variable selection methods. Summarizing, the backward elimination method starts from the complete model (all the variables in the model) and then drops one variable at a time, till the model meets certain conditions. The forward inclusion adds one variable at a time, and usually ends up with models with less variables. In this case, the result is the same. This is the R output, which uses AIC stopping method:

```
Start:  AIC= -360.44
 y ~ age + sex + smoke + quet + vit + cal + fat + fiber + colest +
    betad + retd + alcq

         Df Sum of Sq     RSS     AIC
- sex     1      0.01   90.06 -362.41
- retd    1      0.02   90.07 -362.38
- cal     1      0.02   90.07 -362.36
- fat     1      0.20   90.25 -361.75
- smoke   2      0.79   90.84 -361.69
- fiber   1      0.30   90.35 -361.39
- betad   1      0.41   90.46 -361.02
<none>                  90.05 -360.44
- colest  1      1.21   91.26 -358.24
- quet    1      1.97   92.02 -355.63
- alcq    3      3.16   93.21 -355.57
- vit     2      2.81   92.86 -354.75
- age     1      3.95   94.00 -348.93
```

```
Step:  AIC= -362.41
 y ~ age + smoke + quet + vit + cal + fat + fiber + colest + betad +
    retd + alcq

         Df Sum of Sq     RSS     AIC
- retd    1      0.02   90.08 -364.34
- cal     1      0.02   90.08 -364.32
- fat     1      0.20   90.26 -363.71
- smoke   2      0.79   90.85 -363.64
- fiber   1      0.30   90.36 -363.35
- betad   1      0.40   90.46 -363.01
<none>                  90.06 -362.41
- colest  1      1.21   91.27 -360.21
- quet    1      1.97   92.03 -357.60
- alcq    3      3.16   93.22 -357.55
- vit     2      2.82   92.88 -356.68
- age     1      4.50   94.56 -349.05

Step:  AIC= -364.34
 y ~ age + smoke + quet + vit + cal + fat + fiber + colest + betad +
    alcq

         Df Sum of Sq     RSS     AIC
- cal     1      0.02   90.10 -366.26
- fat     1      0.21   90.29 -365.62
- smoke   2      0.79   90.87 -365.60
- fiber   1      0.29   90.37 -365.32
- betad   1      0.42   90.50 -364.89
<none>                  90.08 -364.34
- colest  1      1.35   91.43 -361.65
- quet    1      1.96   92.04 -359.55
- alcq    3      3.18   93.26 -359.41
- vit     2      2.83   92.91 -358.61
- age     1      4.48   94.56 -351.05

Step:  AIC= -366.26
 y ~ age + smoke + quet + vit + fat + fiber + colest + betad +
    alcq

         Df Sum of Sq     RSS     AIC
```

```
- smoke   2      0.77   90.87 -367.58
- fat     1      0.27   90.38 -367.31
- betad   1      0.42   90.52 -366.81
- fiber   1      0.51   90.61 -366.50
<none>                  90.10 -366.26
- colest  1      1.34   91.44 -363.62
- quet    1      1.98   92.08 -361.43
- alcq    3      3.29   93.39 -360.98
- vit     2      2.87   92.97 -360.38
- age     1      4.48   94.58 -352.98

Step:  AIC= -367.58
 y ~ age + quet + vit + fat + fiber + colest + betad + alcq

         Df Sum of Sq     RSS     AIC
- fat     1      0.27   91.15 -368.63
- betad   1      0.50   91.37 -367.85
<none>                  90.87 -367.58
- fiber   1      0.63   91.50 -367.41
- colest  1      1.46   92.33 -364.57
- quet    1      1.70   92.57 -363.76
- vit     2      3.19   94.07 -360.70
- alcq    3      3.93   94.81 -360.23
- age     1      5.10   95.97 -352.39

Step:  AIC= -368.63
 y ~ age + quet + vit + fiber + colest + betad + alcq

         Df Sum of Sq     RSS     AIC
- fiber   1      0.47   91.61 -369.02
- betad   1      0.53   91.67 -368.82
<none>                  91.15 -368.63
- quet    1      1.67   92.82 -364.91
- alcq    3      3.86   95.01 -361.56
- vit     2      3.28   94.42 -361.50
- colest  1      4.70   95.85 -354.79
- age     1      5.58   96.72 -351.92

Step:  AIC= -369.02
 y ~ age + quet + vit + colest + betad + alcq
```

```
         Df Sum of Sq     RSS     AIC
<none>                  91.61 -369.02
- betad   1      1.40   93.01 -366.25
- quet    1      1.85   93.47 -364.72
- vit     2      3.43   95.04 -361.45
- alcq    3      4.04   95.65 -361.43
- colest  1      4.38   95.99 -356.32
- age     1      5.72   97.34 -351.93

Call:
lm(formula = y ~ age + quet + vit + colest + betad + alcq, data = pr)

Coefficients:
(Intercept)          age         quet         vit2         vit3       colest
  8.975e+00    9.676e-03   -1.333e-02   -1.524e-02   -2.280e-01   -9.200e-04
      betad        alcq2        alcq3        alcq4
  4.615e-05    2.243e-01    2.671e-01    2.281e-01
```

And this is a summary of the features of the model:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.62659 -0.33431  0.02758  0.32666  1.40322

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.975e+00  2.079e-01 43.170 <  2e-16 ***
        age  9.676e-03  2.216e-03  4.366 1.74e-05 ***
       quet -1.333e-02  5.368e-03 -2.484 0.013536 *
       vit2 -1.524e-02  8.015e-02 -0.190 0.849349
       vit3 -2.280e-01  7.354e-02 -3.100 0.002113 **
     colest -9.200e-04  2.410e-04 -3.818 0.000163 ***
      betad  4.615e-05  2.138e-05  2.159 0.031667 *
      alcq2  2.243e-01  7.945e-02  2.824 0.005063 **
      alcq3  2.671e-01  8.013e-02  3.333 0.000965 ***
      alcq4  2.281e-01  1.385e-01  1.647 0.100674
---
Signif. codes:  0 ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1 ‘
’  1
```

```
Residual standard error: 0.5481 on 305 degrees of freedom
Multiple R-Squared: 0.2031,    Adjusted R-squared: 0.1796
F-statistic: 8.64 on 9 and 305 degrees of freedom,
p-value: 1.496e-011
```

R code:

```
y_b+log(retpl)

lm1_lm(y~.-beta-b-retpl-alc,data=pr)

step(lm1, method="both")

lm(formula = y ~ age + quet + vit +
  + colest + betad + alcq, data=pr)
```