Abstract

Maximizing one's blood plasma concentrations of retinol and beta-carotene may become recognized as an important cancer-prevention measure. As such, this report explores the relationship between many personal characteristics (such as alcohol consumption and smoking status) and blood plasma concentrations of these two micronutrients in subjects who had elective surgery to remove or biopsy a non-cancerous lesion. In addition, it briefly investigates the relationship between concentrations of retinol and beta-carotene in the subjects' blood. Model building to find personal characteristic determinants for retinol concentration as well as to discover a relationship between concentrations of the two in proved fruitless for this subject group. However, I did discover that beta-carotene blood plasma concentrations are higher for those who consume fiber and beta-carotene, especially in conjunction with supplemental vitamins and fat, and are lower for those who smoke and are overweight. Further study with a broader population and more personal characteristics may shed more light on patterns of retinol and beta-carotene blood plasma concentrations.

I. Introduction

The elusive nature of an effective cure for cancer has prompted interest in preventive strategies. Much research has focused on the effects of nutrients, including the antioxidants retinol and beta-carotene. Recent research suggests there may be connections between retinol and beta-carotene and lower incidence of several types of cancer. If the link is proven to be conclusive, it will be useful to know more about how one can maximize one's retinol and beta-carotene absorption, and thus minimize cancer risk.

In this study, I began to discover the nature of these micronutrients in the bloodstream and how personal characteristics affect their concentrations. Specifically, I will address two questions:

1. What blood plasma levels of these two micronutrients are usual, and are they related?
2. What personal characteristics impact blood plasma retinol and beta-carotene levels?

To begin, I will examine the distributions and applicable transformations of each of the studied personal characteristics and discuss questionable data points (in the Data section). Then, in Analysis and Discussion, I will address each of the questions in turn, including model exploration and selection. Finally, I will summarize my conclusions in the last section.

II. Data

The data used in this analysis were obtained from Dr. Jane Doe, a research physician in the oncology department, who worked with several colleagues over a three-year period to collect general personal characteristics of 315 subjects who had elective surgery to biopsy or remove a non-cancerous lesion of the lung, colon, breast, skin, ovary or uterus.

The personal characteristics and blood plasma levels studied are:

1. age

2. sex
3. smoking status (abbreviated smokstat)
4. quetelet index (weight divided by squared height)– values above 27 in females and above 28 in males indicate obesity
5. vitamin use (abbreviated vituse)
6. number of calories consumed per day
7. grams of fat consumed per day
8. grams of fiber consumed per day
9. number of alcoholic drinks consumed per week
10. milligrams of cholesterol consumed per day
11. micrograms of dietary beta-carotene consumed per day (abbreviated betadiet)
12. micrograms of dietary retinol consumed per day (abbreviated retdiet)
13. nanograms of beta-carotene per milliliter of blood plasma (abbreviated betaplasma)
14. nanograms of retinol per milliliter of blood plasma (abbreviated retplasma)

A. Item Distributions and Outliers.

I looked at each of these factors individually to get an idea about how the each item was distributed for the study subjects and to identify outliers and possible miscodings. Boxplots for each item are displayed in Figure 1. From these and subsequent analysis, I have identified several unusual points, enumerated below and discussed as they relate to model building in Appendix A. These subjects have either been coded incorrectly or are somehow part of a different population. For all five identified questionable points, no correction was easily discernable so I removed the subject from the study. The subject's complete removal, rather than the removal of only the questionable data point, is valid because the study had ample subjects and thus the loss of good data for the five subjects I have removed will not significantly affect the analysis.

- **Alcohol.** The Alcohol outlier is subject 62, who is reported to have drunk 203 alcoholic beverages per week; this is either a coding error (because the subject would have to consume about two drinks for every hour he was awake), or this subject is unusual to the point of being in a different population from the rest of the subjects. Based on these arguments, I have removed this subject from the study. Replotting alcohol reveals two high outliers at 35 drinks per week–a plausible amount to consume, so I am not justified in removing these points on this basis alone.
- **Calories.** The Calories outlier is also for subject 62, who was reported to consume 6662.2 calories per day. This is high even for an obese person, but the recorded quetelet shows that this subject is below the obesity threshold by several points. Although this could be the effect of the subject's large alcohol consumption (which would account for 4300 of those calories), I believe that this subject either reported erroneous data or was miscoded in more than one item.

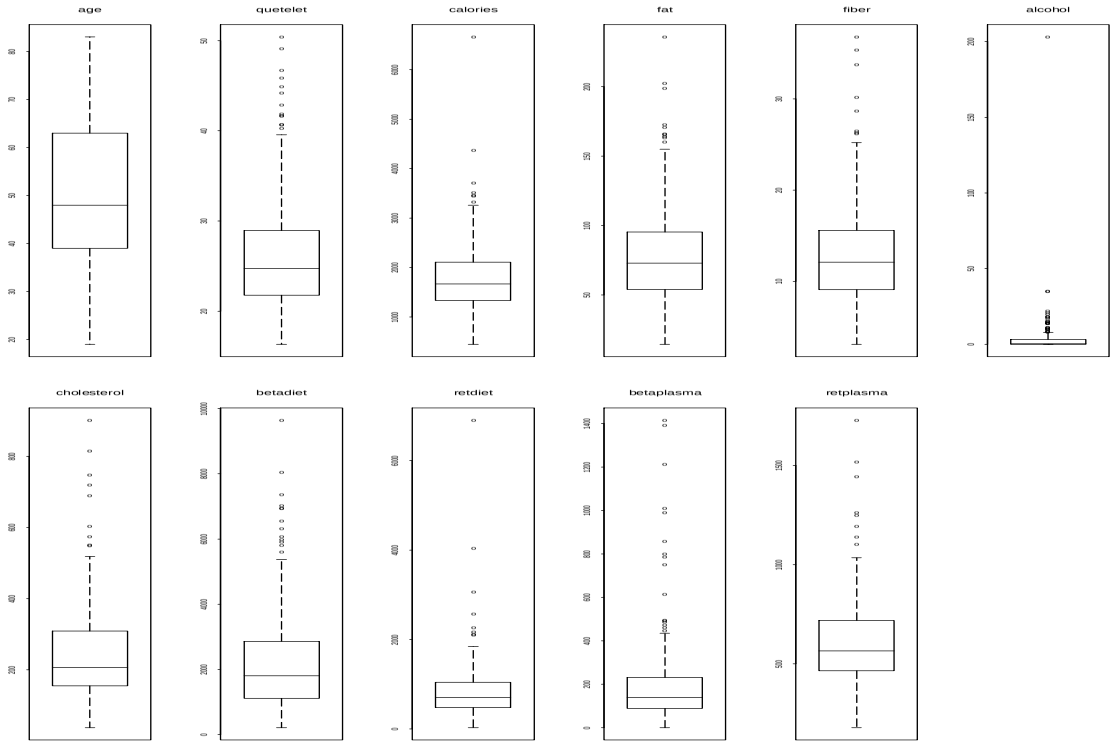

Figure 1: Boxplots of Each Variable

- **Fat.** There are three Fat outliers (subjects 42, 95 and 152). The data for each are as follows:

```
    age sex smokstat quetelet vituse calories   fat fiber alcohol
 42  66   2        2 27.49609      3   3184.8 199.0  16.8     0.2
 95  43   2        2 23.03810      1   3711.0 202.7  14.9    18.0
152  54   2        1 37.86868      1   4373.6 235.9  22.9     0.1
    cholesterol betadiet retdiet betaplasma retplasma male female
 42       362.6     2100    1083        102       838    0      1
 95       469.2     1861     783        125       592    0      1
152       814.7     2912    2104        121       416    0      1
```

  It is reasonable to expect that women consuming 199, 202, and 235 grams of fat per day, respectively, would be obese. However, one subject's quetelet score is significantly below the obese level–number 95. Although I considered excluding her on this basis, subsequent analysis (see Appendix A) suggested she should not be removed from the analysis.

- **Retdiet.** The two Retdiet outliers are for subjects 94 and 171, with 4041 and 6901 mcg per day of retinol in their diet, respectively. Looking at the spread of the rest of the data, as well as the recommended daily allowance for retinol consumption (a maximum of 1300 mcg per day for the highest consumption group[1]), these Retdiet should be somewhere between 100 and 1500 mcg. Since I can not correct these data points, I will remove them.

- **Betaplasma.** I have also found an outlier for Betaplasma, subject 257, who has no beta-carotene in her blood plasma. Given the wide variety of foods that contain beta-carotene,[2] this seems unlikely unless the subject has some medical condition that inhibits beta-carotene absorption and also places the subject in a separate population; I will remove this point as well.

- **Age.** I have found one 19-year-old subject and four subjects in their 80s. These do not seem unusual participants in the study, so I will not remove them.

Boxplots for the corrected data are shown in Figure 2.

It is obvious from Figure 2 that few of the continuous variables are approximately normally distributed. Because data points that are far from the average are unduly influential in regression analysis, I used the following transformations to centralize the data (as is displayed in Figure 3):

- log quetelet
- log calories
- log fat
- log fiber
- quarter root alcohol[3]

[1] www.ivillagehealth.com/library/onemed/content/0,7064,241012_248741,00.html
[2] www.ivillagehealth.com/library/onemed/content/0,7064,241012_248515,00.html
[3] A log transformation, which may have been more appropriate, could not be used because of the large number of subjects who abstain from drinking. An alternative to my quarter root approach is to treat alcohol as a categorical variable; I saw no plausible medical mechanism that might justify partitioning, and so opted for a reasonable continuous transformation.

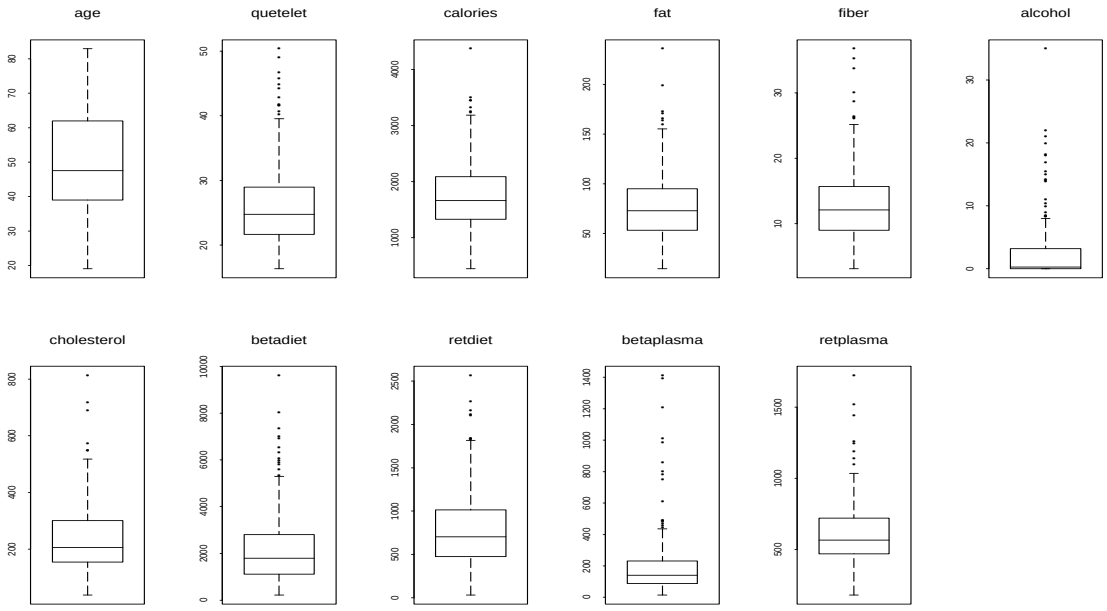

Figure 2: Boxplots of Each Variable with Questionable Points Removed

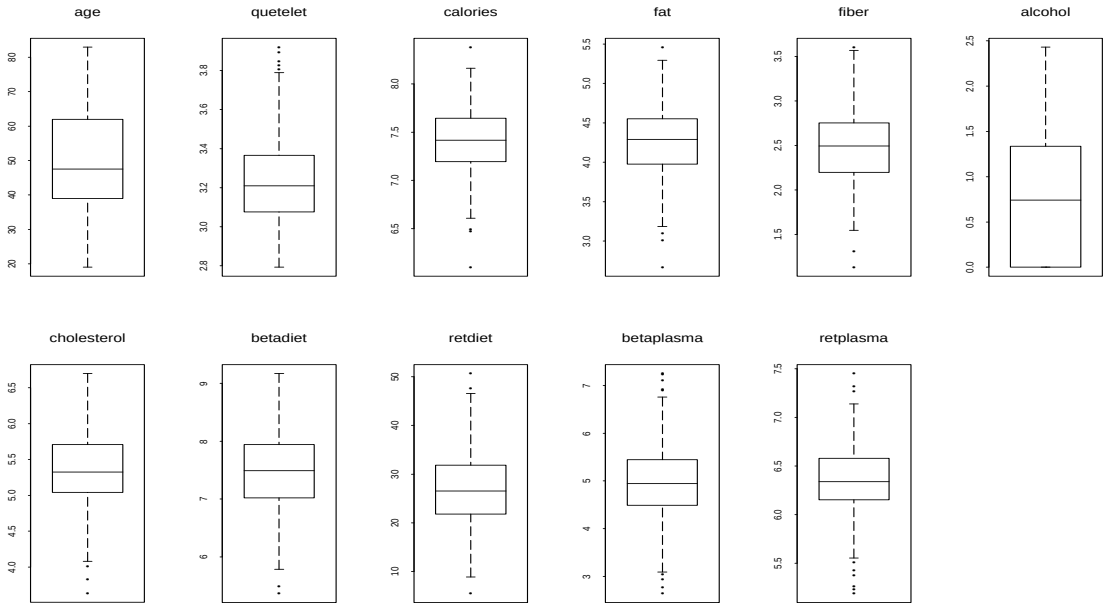

Figure 3: Boxplots of Each Transformed Variable with Questionable Points Removed

- log cholesterol
- log betadiet
- square root retdiet
- log betaplasma
- log retplasma

B. Pairwise Item Correlation

I used a graphical display of pair-wise relationships between the variables (Figure 4) to begin to explore which variables will be influential in answering the first two questions (what characteristics impact retinol and beta-carotene plasma levels). Unfortunately, the plot shows very few strong relationships between variable pairs. In fact, only calories, fat and cholesterol have a very clear relationship. Other possible correlations that could be explored further in a subsequent study to clarify relationships among personal characteristics include:

- age and betaplasma
- sex and quetelet
- sex and cholesterol
- smokstat and quetelet
- quetelet and alcohol
- quetelet and betaplasma
- calories and fiber
- calories and alcohol
- calories and betadiet
- calories and retdiet
- calories and retplasma
- fat and fiber
- fat and retdiet
- fiber and cholesterol
- fiber and betadiet
- fiber and retdiet
- fiber and betaplasma
- cholesterol and betadiet
- cholesterol and retdiet
- betadiet and retdiet

III. Analysis, Results and Discussion

A. Retanol and Beta-Carotene Blood Plasma Concentrations

1. Methods

In this section, I will explore the first question I posed: What blood plasma levels of retinol and beta-carotene are usual, and are these concentrations related? For the first part, simple distribution examination will provide an overview of what levels are most common in the sample population. Correlations between the two micronutrients and a simple regression model will be used to address the second part of the question.

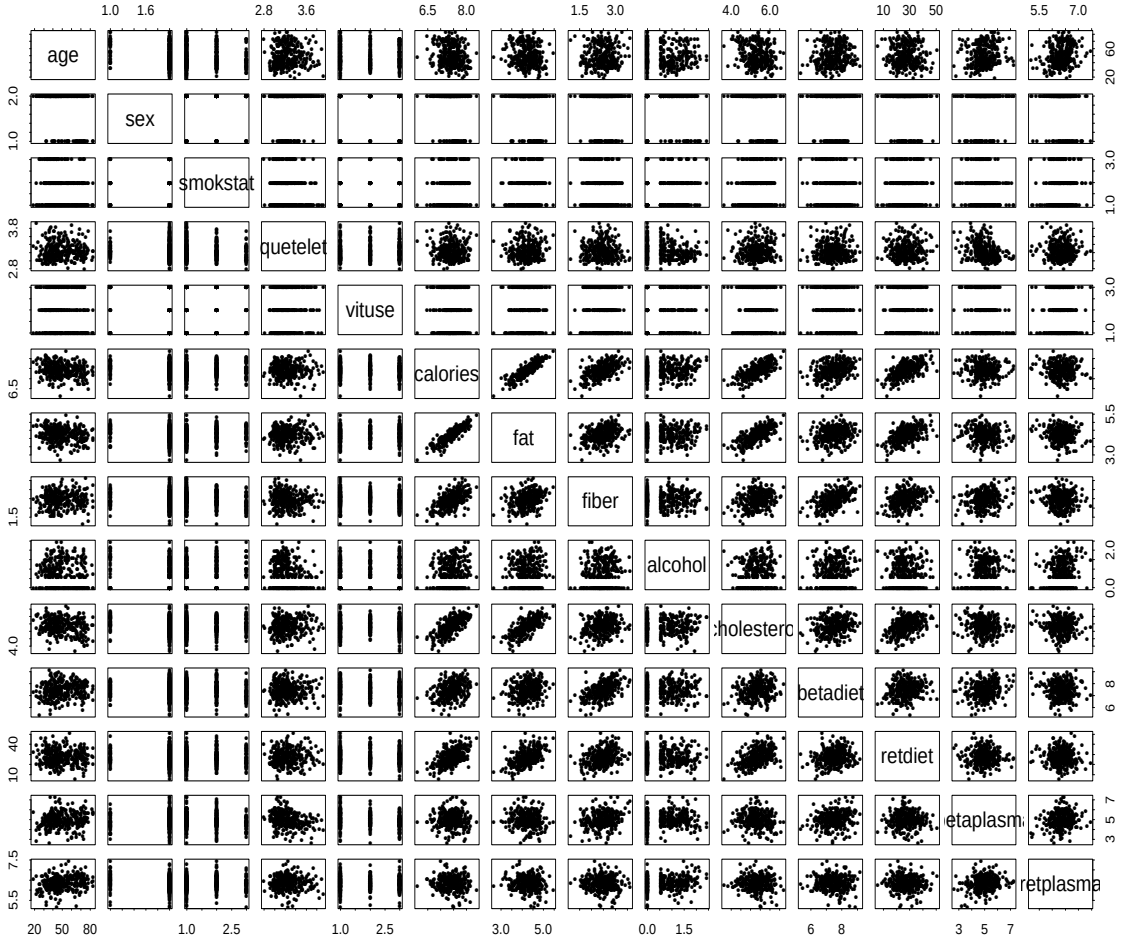

Figure 4: Pairwise relationships for Transformed Variable with Questionable Points Removed

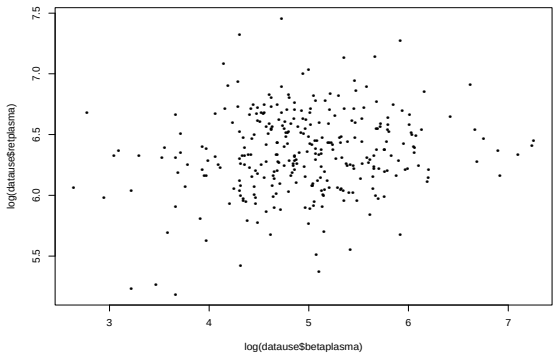

Figure 5: log Retplasma vs. log Betaplasma

2. Usual Retinol and Beta-Carotene Concentrations

The distributions of the blood plasma concentrations of these two micronutrients are displayed in Figure 2, and their summary statistics are:

	Retinol	Beta-Carotene	Beta-Carotene without 257
Mean	602.8	189.9	191.1
Median	566.0	140.0	141.0
Minimum	179.0	0.0	14.0
First Quartile	470.2	90.0	89.5
Third Quartile	720.5	230.0	230.5
Maximum	1727.0	1450.0	1450.0

The numerical analysis provides a good guideline of what an average member of this population's blood plasma micronutrient levels would be. However, one must be cautious when using this information. Because previous studies have linked retinol and beta-carotene levels to cancer prevention, those with even benign lesions may have different concentrations of these micronutrients in their blood than the general public, and so these results have limited generalizability.

3. The Relationship Between Retinol and Beta-Carotene

A simple graphical look at the plasma levels of these micronutrients does not suggest much of a relationship between the two (see Figure 5).

I fit a simple regression of log(retplasma) on log(betaplasma) that resulted in the model:

$$\log(\text{Retplasma}) = 5.99 + 0.0724 \log(\text{Betaplasma})$$

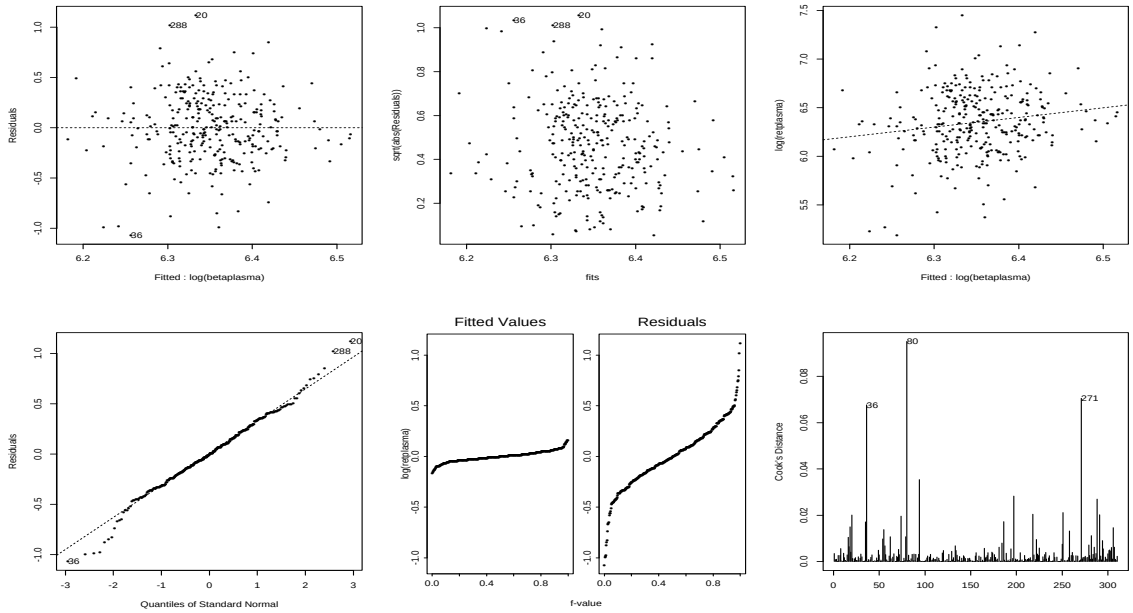

Figure 6: Regression of log Retplasma on log Betaplasma

Or alternatively,

$$\text{Retplasma} = 399.41 \times \text{Betaplasma}^{0.0724}$$

Not surprisingly, while the graphical analysis (Figure 6) confirms a valid model and the F-test suggests the slope of the model is larger than zero, the R-squared value undermines these conclusions because it indicates this model only accounts for 2.6% of the variation in the data. So, although the regression suggests a positive relationship between retinol and beta-carotene blood plasma concentrations (a hypothesis that makes sense because beta-carotene is converted to retinol by the body), the model does not account for enough variation in the data to make this conclusion.

B. Personal Characteristics' Impact on Blood Plasma Concentrations of Retinol and Beta-Carotene

1. Methods

I used several methods to build valid models that explain if and how retinol and beta-carotene are related to the studied personal characteristics. The premise on which I built my models is that variables most

correlated to retplasma and betaplasma are most likely to have a non-zero slope in the best model. Thus, I began model building by selecting those variables with the largest correlation magnitude. I performed simple regressions on these individual variables to verify the transformations discussed in the Data Section above. Then, I began with the additive model including all selected variables and added and subtracted variables and interactions (comparing each change to the previous "best model" until I found model that could not be improved, as defined by two possible elimination criteria:

1. If the models being compared were nested (i.e., the larger model was simply the smaller model with added terms), I used a F-test to compare the magnitude of the two models' variations from the observed data. In these tests, I rejected the smaller model for p-values[4] greater than 0.05. That is, if the F-statistic's p-value was less than 0.05, the larger model was considered best model of the two, and vice versa.

2. If the models being compared were not nested (i.e., each model had terms not contained in the other), I compared the validity of the models in predicting new values by splitting the data into two randomly chosen pieces, fitting the two models on the first half, then predicting the dependent variable for the unmodeled half. Then, the model for which the average squared difference between the predicted and observed data was smaller was considered the best model of the two.

Once I identified the best model, I validated it similarly to the method used in the second elimination criterion. I split the data into two randomly chosen halves and used the first half to fit the final model. Then, I predicted dependent variable values for the subjects in the second half of the data, and calculated the average squared difference between the predicted and observed data, and the standard error of this difference. A valid model has small (ideally zero) average squared difference.

2. Retinol Concentrations

Pairwise Relationships. I used the correlation of the measured blood plasma retinol levels with each of the gathered personal characteristics (all transformed and without questionable data points) to select variables that are most likely to have a significant relationship with blood plasma retinol levels. I chose:

- alcohol (correlation = 0.167)
- age (correlation = 0.227)
- sex (correlation = -0.153)
- fat (correlation = -0.088)
- fiber (correlation = -0.056)
- calories (correlation = -0.050)

The plots of Retplasma verses the selected variables are shown in Figure 7. Each of these correlations is quite weak, a characteristic reflected in the diagnostics for the simple regressions of retplasma on the individual

[4]The probability that a randomly chosen difference is less likely than the calculated difference in magnitude.

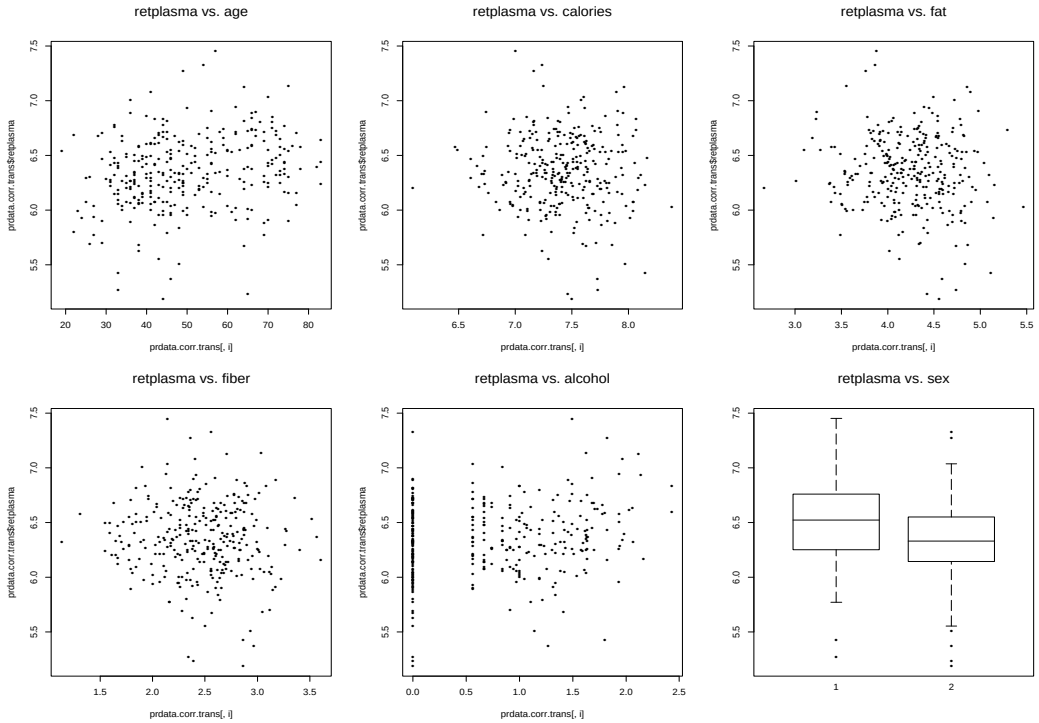

Figure 7: Plots of Transformed Variables

variables. While the graphical evidence for these transformed simple regressions show that the models and transformations discussed in the Data Section are valid (residuals without any discernable pattern, see Figure 8, and distributed within an error envelope of a normal distribution, see Figure 9), and the F-statistics indicate that the slopes of these simple relationships are not zero,[5] the R-squared values (which varied from 0.3 to 5.2%) imply that the models do not account for very much of the variation in the data.

Model Selection. The expanded models do not show much more promise than the pairwise regressions. The simple model:

$$\log \text{Retplasma} \sim \sqrt[4]{\text{Alcohol}} + \text{Age} + \text{Sex} + \log(\text{Fat}) + \log(\text{Fiber}) + \log(\text{Calories})$$

[5] The models compared here are a horizontal line and the simple one-variable model.

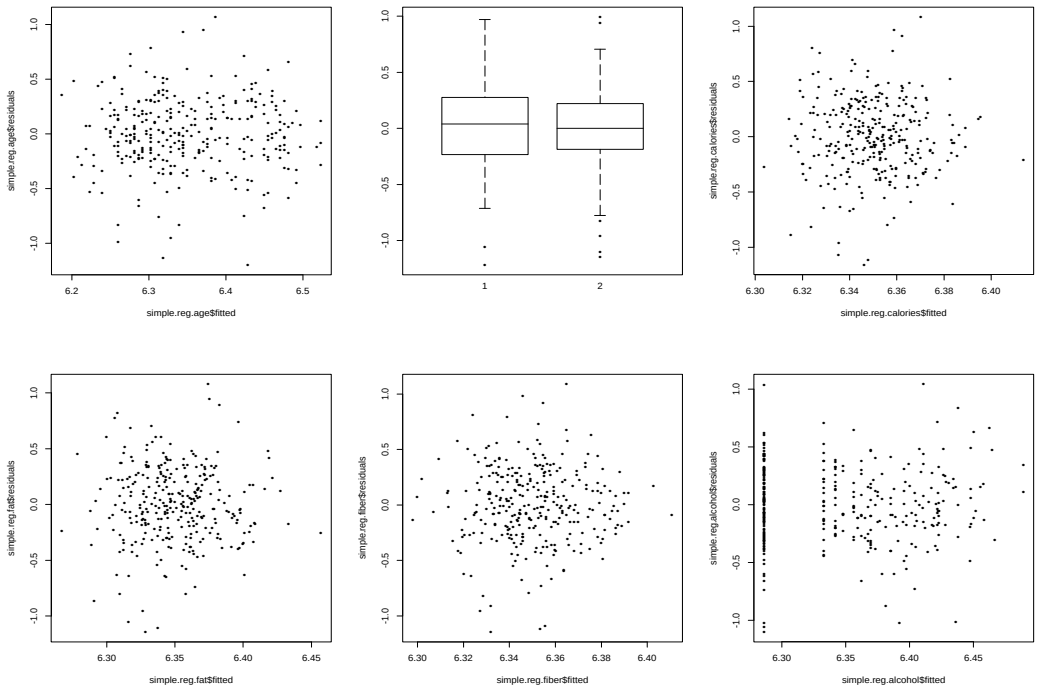

Figure 8: Residuals of Regressions of Retplasma on Age, Sex, Calories, Alcohol, Fiber, and Fat (Clockwise From Upper Left)

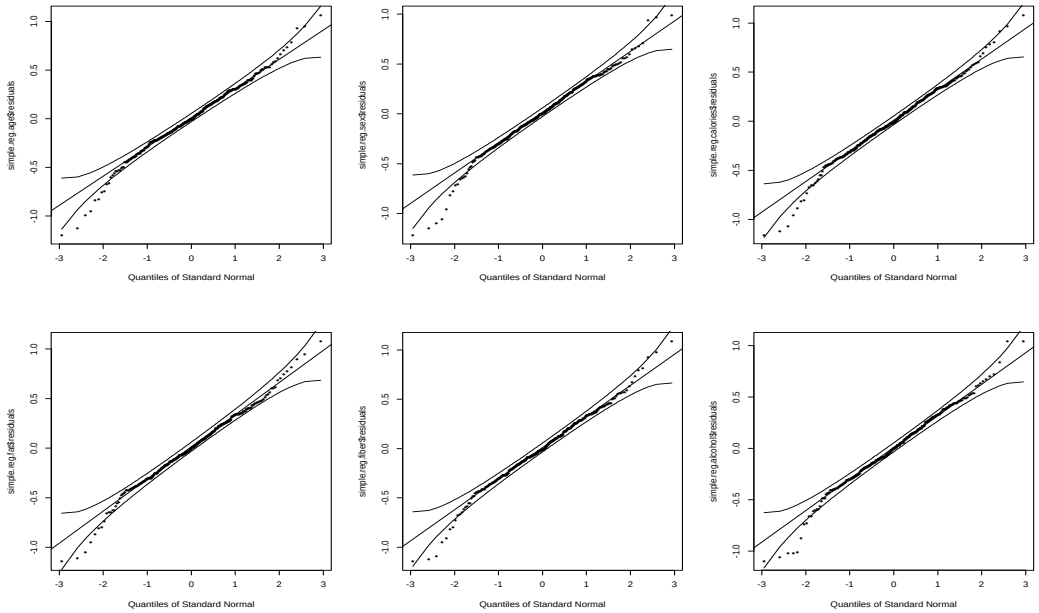

Figure 9: Residuals of Regressions of Retplasma on Age, Sex, Calories, Alcohol, Fiber, and Fat (Clockwise From Upper Left)

has an R-squared value of only 12%, but the F-statistic's p-value (0.000) shows that the model is better than a horizontal line.

This model can not be improved by the single addition of any of the other variables (according to F-tests with p-values between 0.38 and 0.98) or intuitively chosen interaction terms (retdiet and fat, and retdiet and vituse had p-values of 0.21 and 0.57, respectively).[6] Additionally, sex may be removed from the model without causing a statistically significant change with 95% confidence to leave the best model:[7]

$$\begin{aligned}\log(\text{Retplasma}) &= 4.1037 + 0.0046\text{Age} + 0.5033\log(\text{Calories}) - \\ &\quad 0.34400\log(\text{Fat}) - 0.1513\log(\text{Fiber}) + 0.0996\sqrt[4]{\text{Alcohol}}\end{aligned}$$

The graphical summaries of the regression using the above model are shown in Figure 10 and the numerical summary is:

```
Call: lm(formula = retplasma ~ age + calories + fat + fiber + alcohol, data
 = data1)
Residuals:
     Min      1Q   Median     3Q  Max
 -0.9407 -0.1839 -0.01988 0.1874 1.03

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  4.1037  1.1539     3.5565  0.0005
        age  0.0046  0.0018     2.5591  0.0115
   calories  0.5033  0.2698     1.8655  0.0641
        fat -0.3400  0.1778    -1.9126  0.0577
      fiber -0.1513  0.0987    -1.5335  0.1273
    alcohol  0.0996  0.0397     2.5085  0.0132

Residual standard error: 0.32 on 150 degrees of freedom
Multiple R-Squared: 0.1127
F-statistic: 3.809 on 5 and 150 degrees of freedom, the p-value is 0.002815

Correlation of Coefficients:
         (Intercept)     age calories     fat   fiber
     age -0.2931
calories -0.9688      0.2044
     fat  0.8356     -0.1587 -0.9348
   fiber  0.7192     -0.1605 -0.7774   0.6538
 alcohol  0.2765     -0.0576 -0.2977   0.2684  0.1943
```

[6] These interaction terms were chosen because retinol is a fat-soluble nutrient, and absorption of nutrients is often dependent on reactions with other nutrients that could be supplied by supplemental vitamins.

[7] Appendix B contains a summary of the models I tested (using the principles discussed in Methods above) and their appropriate rejection criteria.

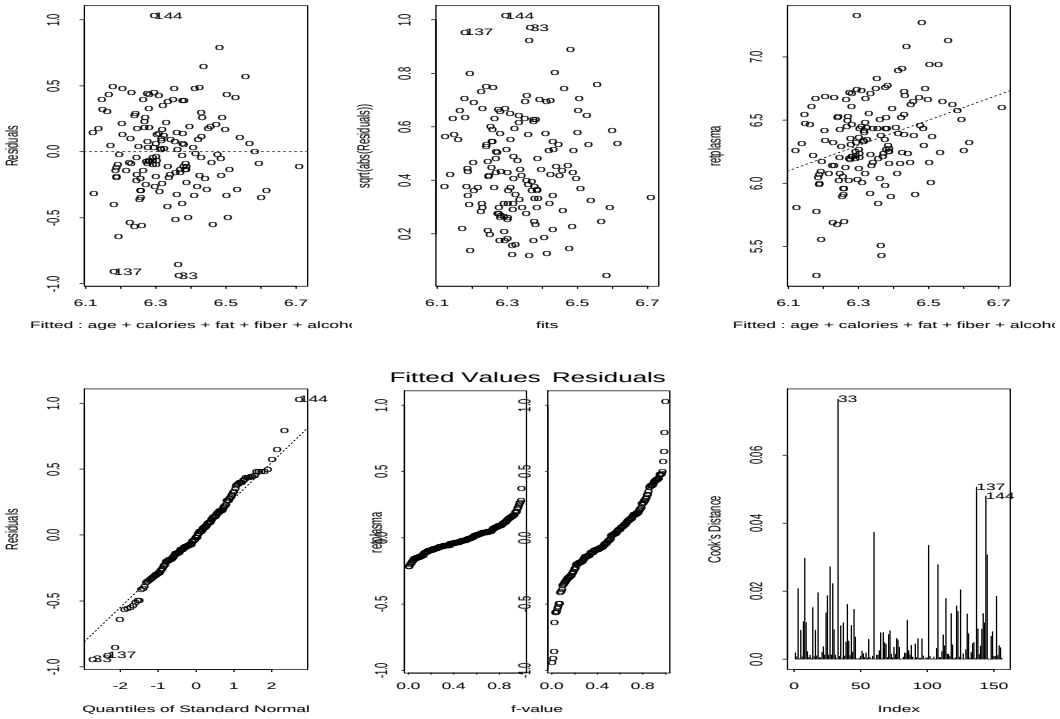

Figure 10: Regression of Retplasma on Age + Quarter Root Alcohol + Calories + Fat + Fiber

The verification process revealed that a prediction using this model has squared differences from the observed data with a mean of only 0.0044 and a standard error of 0.0029. Thus, the mean squared difference could easily be zero since its error is large compared to its distance from zero.

Model Discussion The best model can be written:

$$\begin{aligned} \text{Retplasma} &= 60.56 \times \exp(0.0046 \cdot \text{Age}) \times \text{Calories}^{0.5033} \times \\ &\quad \text{Fat}^{-0.34400} \times \text{Fiber}^{-0.1513} \times \exp(0.0996 \sqrt[4]{\text{Alcohol}}) \end{aligned}$$

The model suggests that increases in blood plasma concentrations are dominated by the baseline value (60.56), but are also related to increases in age, caloric intake and alcohol consumption, and are related to decreases in fat and fiber consumption. However, the R-squared is small, indicating that the model accounts for only 11% of the variation in the data. Although observational studies such as this one rarely allow large R-squared values, 11% is too small to consider the model informative. This model was only affected by a single removed subject (62); the behavior of models for retplasma when this point is included is discussed in Appendix B.

3. Beta-Carotene Concentrations

Pairwise Relationships I have used the correlation of each variable with betaplasma to identify those characteristics that are most likely to be related to beta-carotene concentrations. Good candidates and their correlation values are:

- quetelet (-0.285)
- vituse (-0.255)
- fiber (0.216)
- betadiet (0.194)
- smokstat (-0.186)
- age (0.142)
- sex (0.128)
- fat (-0.109)
- cholesterol (-0.109)

The plots of Betaplasma verses the above variables are shown in Figure 11. Each of these correlations is quite weak, a characteristic reflected in the diagnostics for the simple regressions of retplasma on the individual variables. While the graphical evidence for these transformed simple regressions show that the models are valid (normally-distributed residuals, see Figures 12 and 13), and the F-statistics indicate that the slopes of

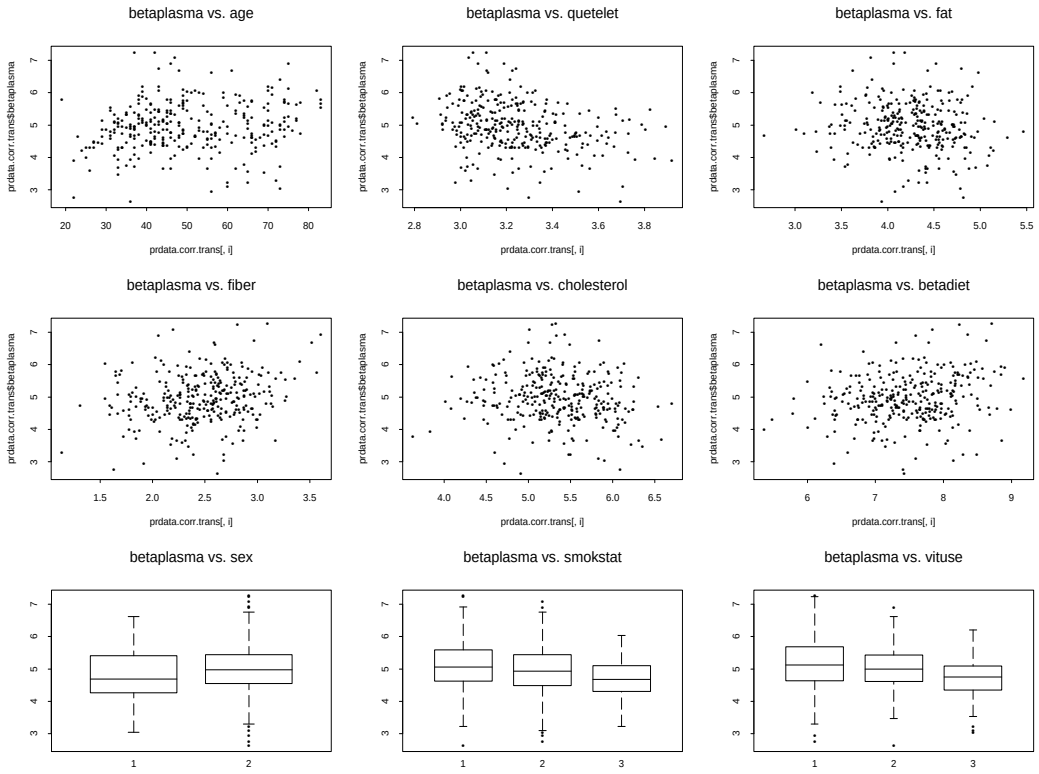

Figure 11: Plots of Betaplasma vs. Variables (all Transformed)

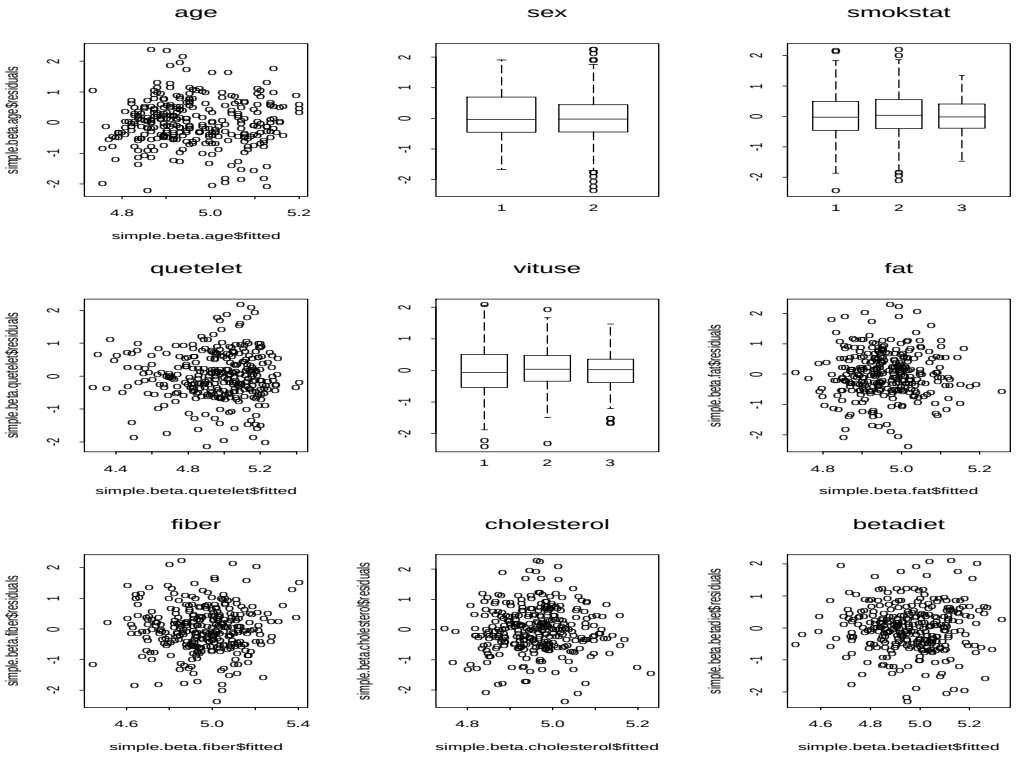

Figure 12: Residuals of Regressions of Betaplasma on Selected Variables

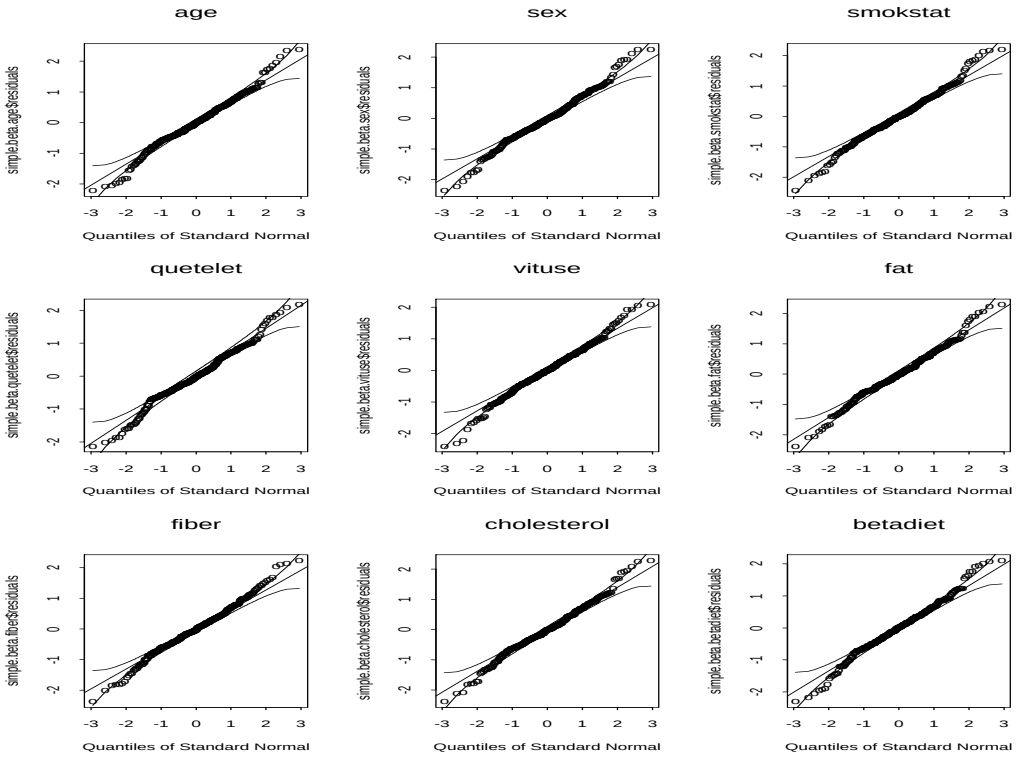

Figure 13: Residuals of Regressions of Betaplasma on Selected Variables

these simple relationships are not zero, the R-squared values (which varied from 1.2 to 8.1%) imply that the models do not account for very much of the variation in the data.

Model Selection. The expanded models are improved. The simple model:

$$\begin{aligned}\log(\text{Betaplasma}) \sim\ & \text{Age} + \text{Sex} + \text{Smokstat} + \log(\text{Quetelet}) + \text{Vituse} + \\ & \log(\text{Fat}) + \log(\text{Fiber}) + \log(\text{Cholesterol} + \log(\text{Betadiet})\end{aligned} \tag{2}$$

has an R-squared value of 23%, and the F-statistic's p-value (0.000) shows that the model is much better than a horizontal line.

However, this model can be improved by the addition of intuitively chosen interaction terms (betadiet and fat, and betadiet and vituse had p-values of 0.040 and 0.035, respectively).[8] Additionally, cholesterol, fat, sex, smokstat, fiber and vituse may be removed from the model without causing a statistically significant change with 95% confidence. Appendix C contains a summary of the models I tested and their appropriate exclusion criteria; the process of addition and elimiation produced a model that adhered to the normality requirement for the residuals better if the betadiet term was not transformed as discussed in the Data Section (see Figure 14). Thus, the best model is:

$$\begin{aligned}\log \text{Betaplasma} = \ & 7.2726 + 0.0076\text{Age} - 0.8507\log(\text{Quetelet}) - 0.0063\log(\text{Betadiet}) : \log(\text{Fat}) + \\ & 0.0002\text{Betadiet} : (\text{Often Vitamin Use}) + 0.0001\text{Betadiet} : (\text{Some Vitamin Use}) + \\ & 0.0000\text{Betadiet} : (\text{No Vitamin Use})\end{aligned}$$

The graphical summaries of the regression using the above model are shown in Figure 15 and the numerical summary is:

```
Call: lm(formula = betaplasma ~ age + quetelet + exp(betadiet):vitoft + exp(
betadiet):vitsome + exp(betadiet):vitno + betadiet:fat, data =
data1)
Residuals:
  Min      1Q   Median     3Q   Max
 -1.9 -0.3436 -0.03894 0.4281 1.717

Coefficients:
                        Value Std. Error t value Pr(>|t|)
          (Intercept)  7.2726  0.9304     7.8166  0.0000
                  age  0.0076  0.0038     1.9789  0.0497
            quetelet -0.8507  0.2470    -3.4446  0.0007
 exp(betadiet):vitoft  0.0002  0.0001     3.4997  0.0006
exp(betadiet):vitsome  0.0001  0.0001     2.0004  0.0473
```

[8] These interaction terms were chosen because, like retinol, beta-carotene is a fat-soluble nutrient, and absorption of nutrients is often dependent on reactions with other beneficial substances that could be supplied by supplemental vitamins

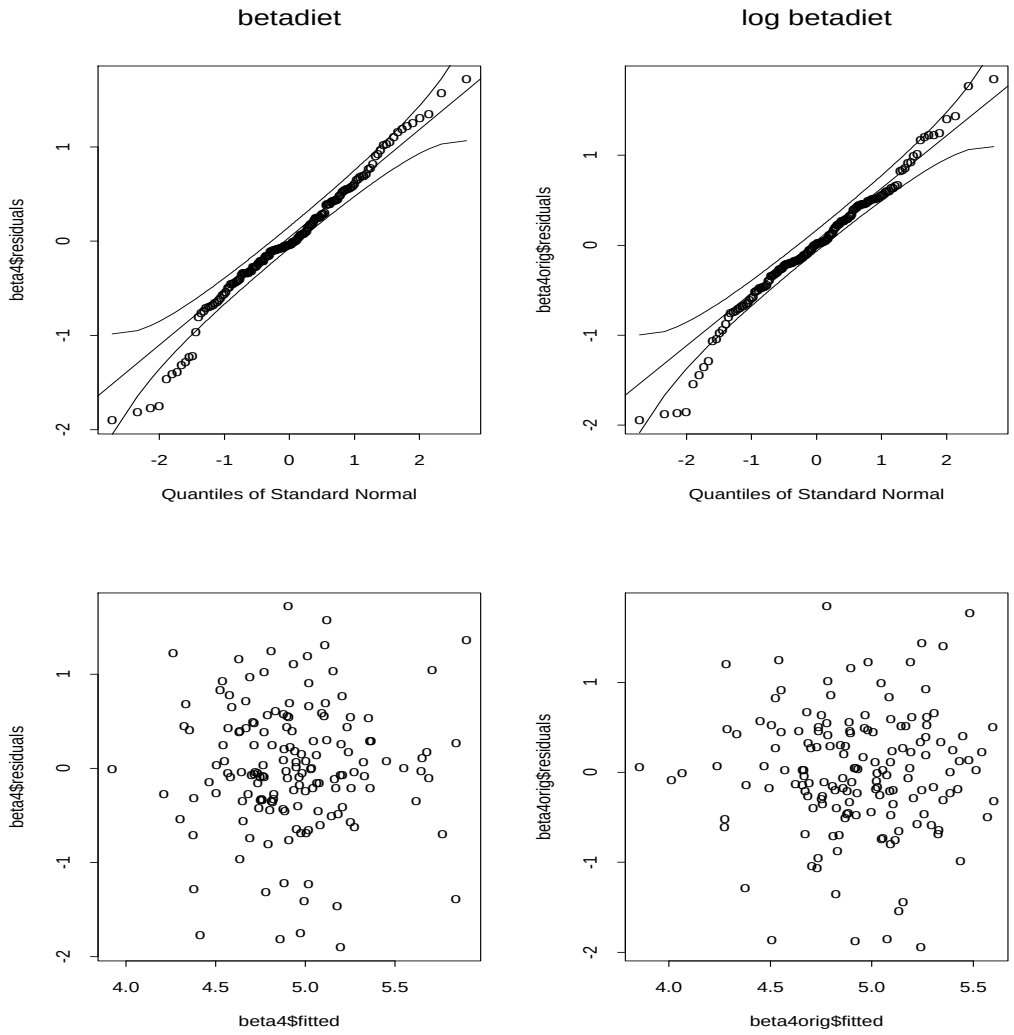

Figure 14: Comparison of Regressions with Transformed and Untransformed Betadiet

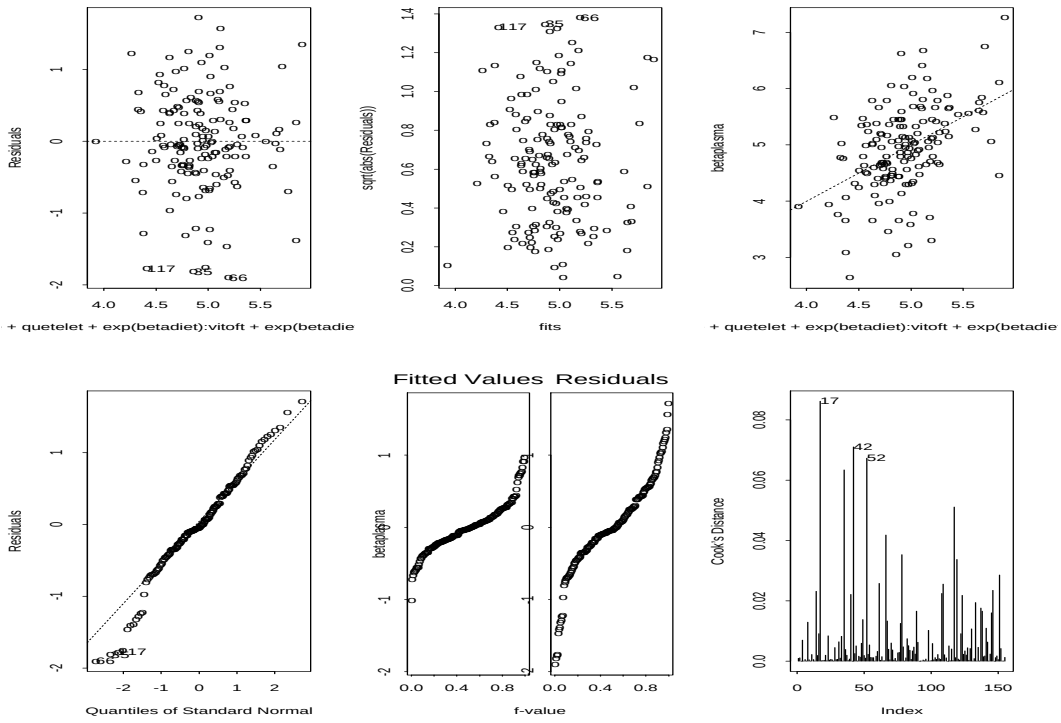

Figure 15: Regression of Betaplasma on Age, Quetelet, Betadiet:Fat, and Betadiet:Vituse

```
  exp(betadiet):vitno  0.0000  0.0001     0.0345  0.9725
         betadiet:fat -0.0063  0.0159    -0.3981  0.6911

Residual standard error: 0.6818 on 148 degrees of freedom
Multiple R-Squared: 0.211
F-statistic: 6.596 on 6 and 148 degrees of freedom, the p-value is 3.394e-006

Correlation of Coefficients:
                      (Intercept)     age quetelet
                  age -0.2426
              quetelet -0.8517      0.0182
 exp(betadiet):vitoft  0.2016     -0.0027  0.0304
exp(betadiet):vitsome  0.1558      0.0212  0.0487
  exp(betadiet):vitno  0.2609     -0.1262 -0.0085
         betadiet:fat -0.4452      0.0541 -0.0313
                      exp(betadiet):vitoft exp(betadiet):vitsome
                  age
              quetelet
 exp(betadiet):vitoft
exp(betadiet):vitsome  0.6075
  exp(betadiet):vitno  0.6544                0.5829
         betadiet:fat -0.6170               -0.5499
                      exp(betadiet):vitno
                  age
              quetelet
 exp(betadiet):vitoft
exp(betadiet):vitsome
  exp(betadiet):vitno
         betadiet:fat -0.6119
```

The verification process revealed that a prediction using this model has squared differences from the observed data with a mean of only 0.0208 and a standard error of 0.0159. Thus, the mean squared difference could easily be zero since its error is large compared to its distance from zero.

Model Discussion. The selected model can be written:

$$\begin{aligned}\{\text{Betaplasma} \quad = \quad & 1440.3 \times \exp(0.0076\text{Age}) \times \text{Quetelet}^{-0.8507} \times (\text{Betadiet} \cdot \text{Fat})^{-0.0063} \times \\ & (\text{Often Vitamin Use}) \cdot \exp(0.0002\text{Betadiet}) \times (\text{Some Vitamin Use}) \cdot \exp(0.0001\text{Betadiet})\end{aligned}$$

Like that for retinol concentration, this model is dominated by the baseline value (1440.3), but also suggests that increases in beta-carotene concentrations are related to increases in age and beta-carotene consumption in conjunction with vitamin use, and are related to decreases in body fat and fat consumption in conjunction with beta-carotene. Especially note that dietary beta-carotene is only beneficial if consumed with supplemental vitamins.

Unlike the retinol model, The R-squared for this model shows that our best model for Betaplasma accounts for 23% of the variation in the data, a small but acceptable level for an observational study such as this. Thus, this model is useful for predicting causal relationships between the involved personal characteristics and betaplasma. However, except for quetelet, the contributions made by each variable so small as to be practically insignificant.

IV. Conclusions

Analysis of the nature of retinol and betacarotene in blood plasma was inconclusive. Although I determined normal concentrations of the micronutrients for this population, these results can not be generalized to the general public because of potential differences between the subject population (which could have similar characteristics to populations with cancer, or at high risk for cancer) and the general public. Comparisons of blood plasma concentrations between the two are similarly uninformative. My analysis shows no direct link between beta-carotene and retinol levels in the blood. Explanations that the reader may have insight into or would require further exploration include:

1. The body uses these micronutrients for different purposes and so their concentrations would be expected to be unrelated.
2. The body is good at regulating beta-carotene and retinol levels in the blood stream through an imprecise buffering mechanism that keeps the micronutrients near the normal levels observed.
3. There are other variables that control retinol and beta-carotene concentrations in the blood.

In this third instance, one might expect personal characteristics to be the dominant determinants for the concentration of these nutrients in the blood stream. Unfortunately, regression analysis does not show strong evidence for this hypothesis. In the case of both retinol and beta-carotene concentrations, the dominating factors in the best models were the baseline values (i.e., the intercepts). The small amount of variability accounted for by each model (11 and 23% for retinol and beta-carotene, respectively) suggests that the studied personal characteristics are not the dominant determinants for blood concentrations of retinol and beta-carotene, although several factors do have a statistically, if not practically, significant influence.

The one clear indication from the two models is that reductions in body fat are related to increases in beta-carotene blood plasma concentrations. One possible mechanism that the reader may have insight into or may require further study is that people with less body fat can not store as much of this fat-soluble nutrient in fat, and beta-carotene is therefore stored in blood plasma at a higher concentration than for those who can use more body fat to store needed amounts.

Because the data were so uninformative on the whole, I conclude that further study under an alternative design is needed to conclusively answer any of the questions I have explored in this analysis. Improvements to the design could include:

- Broadening the study population. Both beta-carotene and retinol are being studied because of their possible reduction of cancer risk. By studying only subjects who may be at a higher risk of cancer or share characteristics of individuals with high cancer risk (i.e., those who have had non-malignant growths) the study may only involve subjects with similarly-valued personal characteristics, or with some other shared factor that has a strong influence on blood concentrations of the micronutrients.

- Broadening the personal characteristics measured. Including more habitual factors as well as concentrations of other substances in the bloodstream and body fat of the subjects could reveal a dominant determinant.

Further study along these lines could reveal important links between personal habits and characteristics and beta-carotene and retinol concentrations. While the import for cancer prevention of maximizing body concentrations of these micronutrients has yet to be proven for all members of the general population, having this knowledge will be essential should researchers come to this conclusion. I recommend broadening the study in several aspects to discover the information that may become life-saving in the future.

Credits and References

Many thanks to Brian for his patient assistance in the data analysis for and writing of this report.

Rawlings, John O., Sastry G. Bantula and David A. Dickey, *Applied Regression Analysis: A Research Tool,* second edition. Springer: New York, 1989.

www.berkeleywellness.com

www.ivillagehealth.com

Appendix A
Outliers' Influence in Selected Models

Beta-carotene Models

Subject 257 was considered an outlier because she had no beta-carotene in her blood plasma. Thus, including this point in the regression requires a change in transformation because the log of zero is infinite. I will substitute the quarter root transformation for betaplasma, as pictured in Figure 16.

The new regression is graphed in Figure 17 and summarized here:

```
Call: lm(formula = retplasma \~ betaplasma, data = prdata.bet.trans)
Residuals:
    Min      1Q    Median     3Q   Max
 -1.073 -0.2137 -0.007456 0.2287 1.128

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  6.0457  0.0972    62.2205  0.0000
 betaplasma  0.0861  0.0272     3.1669  0.0017

Residual standard error: 0.3354 on 309 degrees of freedom
Multiple R-Squared: 0.03144
F-statistic: 10.03 on 1 and 309 degrees of freedom, the p-value is 0.001694

Correlation of Coefficients:
           (Intercept)
betaplasma -0.9807
```

Adding an indicator variable for the suspected outlier (subject 257) gives the following modified regression coefficient summary:

```
                              Value Std. Error  t value Pr(>|t|)
                (Intercept)  6.0923   0.1013    60.1591   0.0000
prdata.bet.trans$betaplasma  0.0733   0.0283     2.5919   0.0100
                      dummy -0.5550   0.3496    -1.5873   0.1135
```

This analysis suggests that the subject may not in fact be an outlier. However, the model is uninformative with or without subject 257 (R-squared values of 0.031 and 0.026, respectively), and so this subject's inclusion or exclusion is inconsequential. Also, because the log tranformation of betaplasma is preferable and the datapoint is inconsequential to the analysis, I chose to exclude it for ease of analysis.

Retinol Models

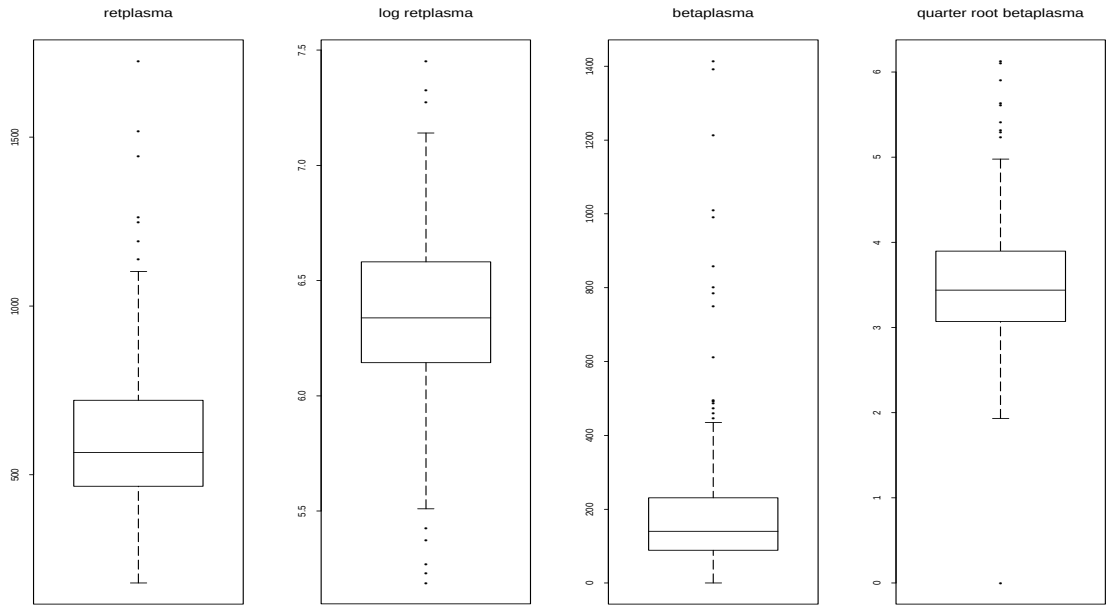

Figure 16: Retplasma on Age and Betaplasma Alternative Transformations

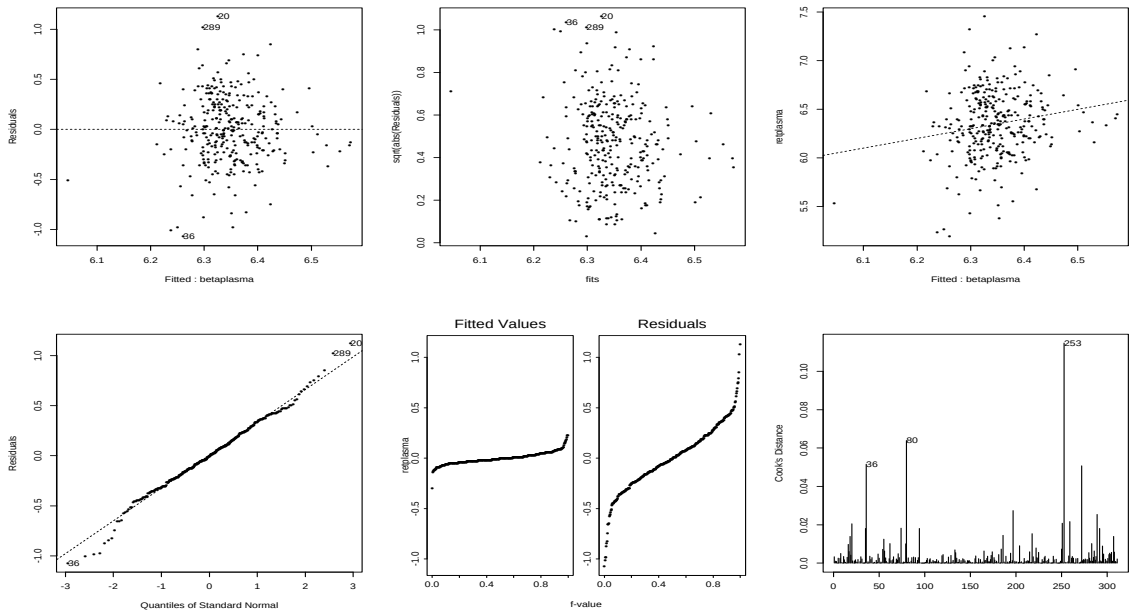

Figure 17: Regression of log Retplasma on quarter root Betaplasma

The model that included the six most correlated variables would have only been affected by possible outliers in alcohol and calories (both of which came from subject number 62, who was excluded) and fat (subject number 95, who was not excluded).

Boxplots of the data with number 62 included (Figure 18) show that the extra points do not warrant new transformations.

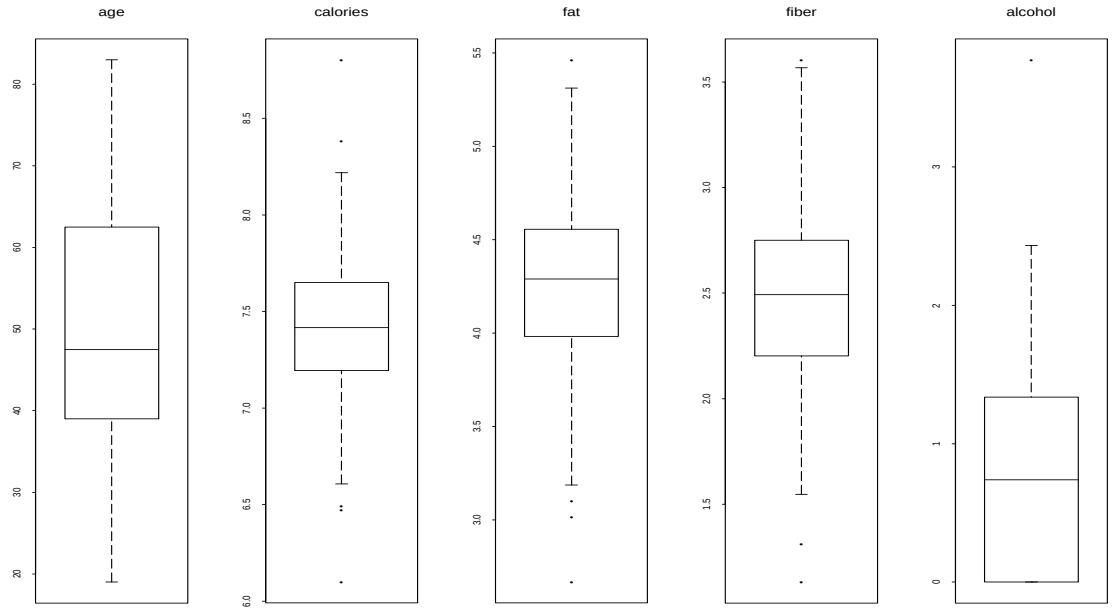

Figure 18: Boxplots of Transformed Variables Including Questionable Data Points

Regressing Retplasma on the six correlated variables (including Subject 62) gives the summary:

```
Call: lm(formula = retplasma \~ age + sex + calories + fat + fiber + alcohol, data =
prdata.ret.trans)
Residuals:
    Min      1Q   Median     3Q   Max
 -1.115 -0.1958 0.001233 0.1999 0.996

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  5.7467  0.7110     8.0821  0.0000
```

```
        age  0.0051  0.0014     3.6121  0.0004
        sex -0.0676  0.0600    -1.1275  0.2604
   calories  0.1879  0.1534     1.2243  0.2218
        fat -0.1734  0.1064    -1.6298  0.1042
      fiber -0.0964  0.0622    -1.5517  0.1218
    alcohol  0.0666  0.0279     2.3900  0.0175

Residual standard error: 0.3257 on 305 degrees of freedom
Multiple R-Squared: 0.09067
F-statistic: 5.069 on 6 and 305 degrees of freedom, the p-value is 5.63e-05

Correlation of Coefficients:
         (Intercept)     age     sex calories     fat   fiber
     age -0.3987
     sex -0.3551      0.3450
calories -0.9214      0.2165  0.1096
     fat  0.7036     -0.0891  0.0082 -0.8999
   fiber  0.5615     -0.2197 -0.1019 -0.6556   0.4857
 alcohol  0.1633      0.0738  0.1424 -0.2366   0.2098  0.1034
```

Plots of this regression are in Figure 19.

However, a comparison between this model and the one that excludes the outlying points are practically the same, because they are equally unuseful. That is, the models in both cases only account for about 9% of the variation in the data.

Dummy variables for the two outlying subjects in this model give:

```
             Value Std. Error t value Pr(>|t|)
(Intercept)  4.8095  0.7462     6.4456  0.0000
        age  0.0059  0.0014     4.2090  0.0000
        sex -0.0742  0.0591    -1.2557  0.2102
   calories  0.3946  0.1611     2.4486  0.0149
        fat -0.2843  0.1090    -2.6089  0.0095
      fiber -0.1629  0.0638    -2.5550  0.0111
    alcohol  0.0824  0.0279     2.9552  0.0034
    dummy62 -1.2987  0.3584    -3.6241  0.0003
    dummy95  0.0065  0.3260     0.0200  0.9840
```

Because the coefficient for subject 62 is significantly different from zero (with a p-value of 0.0003), its outlier status for this regression is confirmed, and my conclusion that this subject is either in a different population or has been miscoded is justified. Alternately, the coefficient for subject 95 is not significantly different than zero and so classifying her as an outlier is not justified on this basis.

Beta-carotene Models

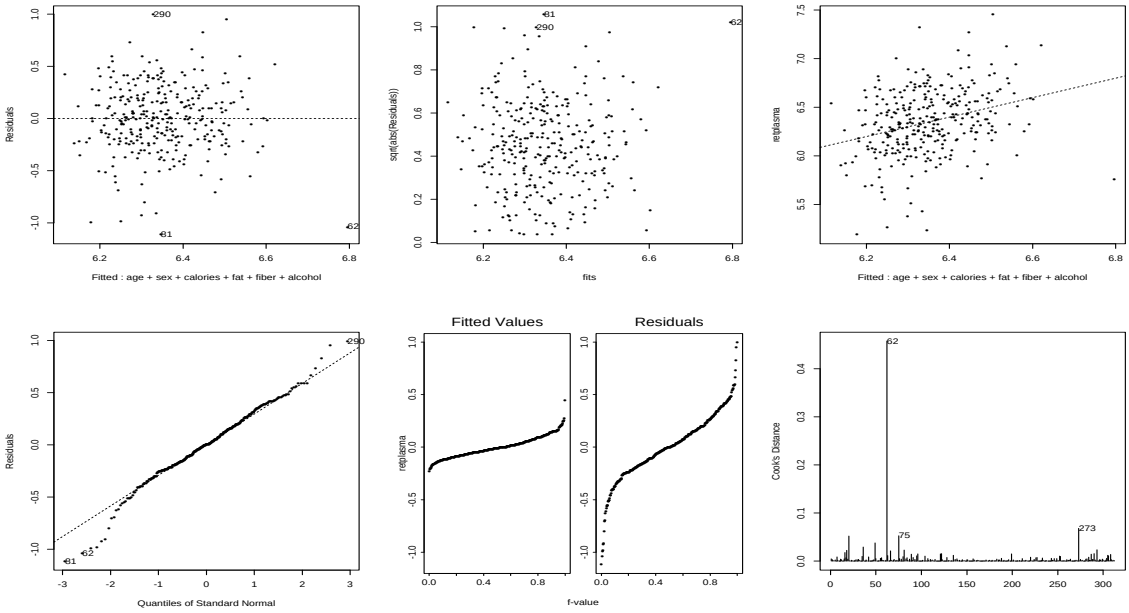

Figure 19: Regression of Retplasma on Correlated Variables Including Subjects 62 and 95

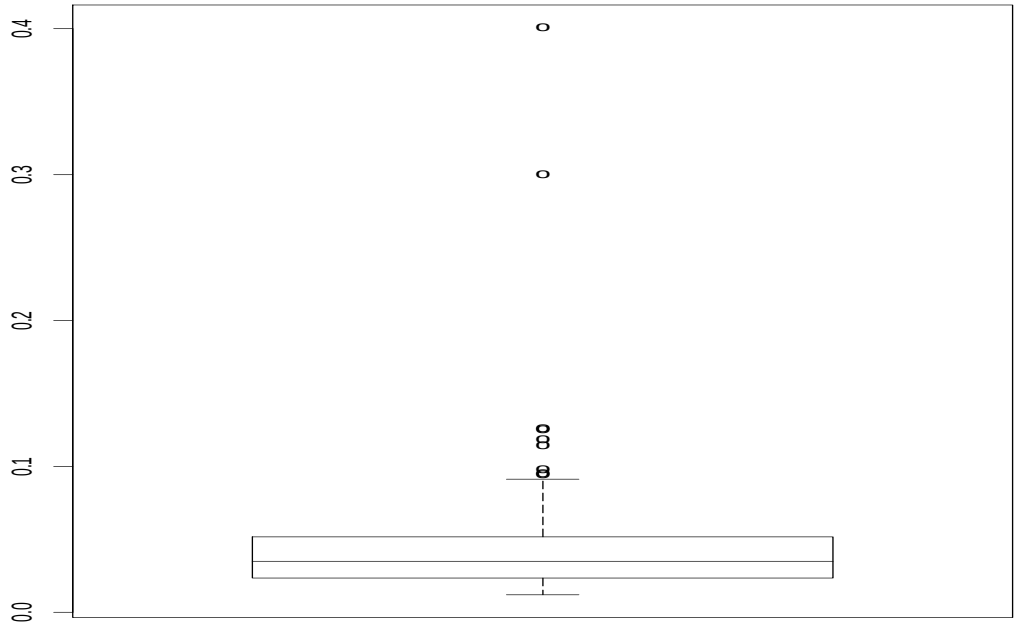

Figure 20: Regression of Retplasma on Correlated Variables Including Subjects 62 and 95

The outlier for betaplasma (subject 257) is the only one of the four excluded points that would affect the final model for Beta-carotene absorption. As discussed above, this point is inconsequential to the data analysis, which is improved by a preferred transfomation when the data point is excluded.

Another possible outlier that may have affected the model is subject 95, who seemed to have unusually high levels of fat. A diagnostic tool not used when discussing this point above is the values of the hat matrix for the regression. I have plotted these for the final model of betaplasma in Figure 20. As you can see, none of the outlying hat values are extreme, and the actual data for these points give no reason to doubt their inclusion in the model.

Appendix B

Models Tested for Regression of Retplasma on Personal Characteristics

	Small Model	Variables Added	p-value
1	retplasma~ age + sex + calories + fat + fiber + alcohol	fat:retdiet	0.897
2	retplasma~ age + sex + calories + fat + fiber + alcohol	vituse:retdiet	0.508
3	retplasma~ age + sex + calories + fat + fiber + alcohol	smokstat	0.713
4	retplasma~ age + sex + calories + fat + fiber + alcohol	quetelet	0.381
5	retplasma~ age + sex + calories + fat + fiber + alcohol	vituse	0.380
6	retplasma~ age + sex + calories + fat + fiber + alcohol	cholesterol	0.519
7	retplasma~ age + sex + calories + fat + fiber + alcohol	betadiet	0.402
8	retplasma~ age + sex + calories + fat + fiber + alcohol	retdiet	0.988
9	retplasma~ age + sex + calories + fat + fiber	alcohol	0.003
10	retplasma~ age + sex + calories + fat + alcohol	fiber	0.011
11	retplasma~ age + sex + calories + fiber + alcohol	fat	0.009
12	retplasma~ age + sex + fat + fiber + alcohol	calories	0.015
13	retplasma~ age + calories + fat + fiber + alcohol	sex	0.210
14	retplasma~ sex + calories + fat + fiber + alcohol	age	0.000

- **Single Variable Model R-squared Values and F-statistics**

```
> summary(retplas.alc)$r.squared
[1] 0.05142014
> pvalue
 0.00005759732

> summary(retplas.age)$r.squared
[1] 0.04912237
> p-value
 0.00008523392

> summary(retplas.sex)$r.squared
[1] 0.0424075
> p-value
 0.0002680169

> summary(retplas.fat)$r.squared
[1] 0.004930805
> p-value
 0.2183727

> summary(retplas.ret)$r.squared
[1] 0.001079364
> p-value
 0.5650683
```

- **Model 1**

```
                                                          Terms Resid. Df
1 age + sex + calories + fat + fiber + alcohol + fat:retdiet       302
2               age + sex + calories + fat + fiber + alcohol       303

       RSS          Test Df    Sum of Sq    F Value     Pr(F)
1 31.01421
2 31.01593 -fat:retdiet -1 -0.001713598 0.01668611 0.8973054
```

- **Model 2**

```
                                                            Terms Resid. Df
1 age + sex + calories + fat + fiber + alcohol + vituse:retdiet       302
2                  age + sex + calories + fat + fiber + alcohol       303

       RSS             Test Df   Sum of Sq   F Value     Pr(F)
1 30.97081
2 31.01593 -vituse:retdiet -1 -0.04511692 0.4399404 0.5076572
```

- **Model 3**

```
                                                    Terms Resid. Df      RSS
1 age + sex + calories + fat + fiber + alcohol + smokstat       302 31.00203
2            age + sex + calories + fat + fiber + alcohol       303 31.01593

       Test Df   Sum of Sq   F Value     Pr(F)
1
2 -smokstat -1 -0.01389923 0.1353965 0.7131588
```

- **Model 4**

```
                                                    Terms Resid. Df      RSS
1 age + sex + calories + fat + fiber + alcohol + quetelet       302 30.93722
2            age + sex + calories + fat + fiber + alcohol       303 31.01593

       Test Df   Sum of Sq   F Value     Pr(F)
1
2 -quetelet -1 -0.07870702 0.7683147 0.3814356
```

- **Model 5**

```
                                                  Terms Resid. Df      RSS
1 age + sex + calories + fat + fiber + alcohol + vituse       302 30.93671
2          age + sex + calories + fat + fiber + alcohol       303 31.01593

     Test Df   Sum of Sq   F Value     Pr(F)
1
2 -vituse -1 -0.07921409 0.7732773 0.3799043
```

- **Model 6**

```
                                                      Terms Resid. Df
1 age + sex + calories + fat + fiber + alcohol + cholesterol       302
2               age + sex + calories + fat + fiber + alcohol       303

       RSS          Test Df   Sum of Sq   F Value     Pr(F)
1 30.97335
2 31.01593 -cholesterol -1 -0.04257556 0.4151252 0.5198686
```

- **Model 7**

```
                                                    Terms Resid. Df      RSS
1 age + sex + calories + fat + fiber + alcohol + betadiet       302 30.94385
2            age + sex + calories + fat + fiber + alcohol       303 31.01593

       Test Df   Sum of Sq   F Value     Pr(F)
1
2 -betadiet -1 -0.07207359 0.7034103 0.4023035
```

- **Model 8**

```
                                                    Terms Resid. Df      RSS
1 age + sex + calories + fat + fiber + alcohol + retdiet       302 31.01590
2           age + sex + calories + fat + fiber + alcohol       303 31.01593

      Test Df     Sum of Sq    F Value     Pr(F)
1
2 -retdiet -1 -2.318079e-05 0.00022571 0.9880232
```

- **Model 9**

```
                                         Terms Resid. Df      RSS     Test Df
1 age + sex + calories + fat + fiber + alcohol       303 31.01593
2           age + sex + calories + fat + fiber       304 31.90988 -alcohol -1

   Sum of Sq  F Value       Pr(F)
1
2 -0.8939519 8.733172 0.003369907
```

- **Model 10**

```
                                         Terms Resid. Df      RSS   Test Df
1 age + sex + calories + fat + fiber + alcohol       303 31.01593
2         age + sex + calories + fat + alcohol       304 31.68413 -fiber -1

   Sum of Sq  F Value      Pr(F)
1
2 -0.6682045 6.527806 0.01110865
```

- **Model 11**

```
                                         Terms Resid. Df      RSS Test Df
1 age + sex + calories + fat + fiber + alcohol       303 31.01593
2       age + sex + calories + fiber + alcohol       304 31.71265 -fat -1

   Sum of Sq  F Value       Pr(F)
1
2 -0.6967244 6.806423 0.009533639
```

- **Model 12**

```
                                         Terms Resid. Df      RSS      Test Df
1 age + sex + calories + fat + fiber + alcohol       303 31.01593
2            age + sex + fat + fiber + alcohol       304 31.62967 -calories -1

   Sum of Sq F Value      Pr(F)
1
2 -0.6137407 5.99574 0.01490708
```

- **Model 13**

```
                                        Terms Resid. Df      RSS Test Df
1 age + sex + calories + fat + fiber + alcohol       303 31.01593
2       age + calories + fat + fiber + alcohol       304 31.17733 -sex -1

  Sum of Sq  F Value    Pr(F)
1
2 -0.161404 1.576784 0.210192
```

- **Model 14**

```
                                        Terms Resid. Df      RSS Test Df
1 age + sex + calories + fat + fiber + alcohol       303 31.01593
2       sex + calories + fat + fiber + alcohol       304 32.82933 -age -1

  Sum of Sq  F Value        Pr(F)
1
2 -1.813399 17.71541 3.384578e-05
```

Appendix C
Models Tested for Regression of Betaplasma on Personal Characteristics

	Small Model	Variables Added	p-value
1	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet	calories	0.519
2	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet	alcohol	0.104
3	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet	retdiet	0.953
4	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet	betadiet:vituse	0.040
5	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet	betadiet:fat	0.035
6	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet:fat	betadiet:vituse or betadiet	0.00086*
7	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + fiber + betadiet:fat + betadiet:vituse	cholesterol	0.096
8	betaplasma ~ age + sex + smokstat + quetelet + vituse + fat + cholesterol + betadiet:fat + betadiet:vituse	fiber	0.278
9	betaplasma ~ age + sex + smokstat + quetelet + vituse + fiber + cholesterol + betadiet:fat + betadiet:vituse	fat	0.394
10	betaplasma ~ age + sex + smokstat + quetelet + fat + fiber + cholesterol + betadiet:fat + betadiet:vituse	vituse	0.988
11	betaplasma ~ age + sex + smokstat + vituse + fat + fiber + cholesterol + betadiet:fat + betadiet:vituse	quetelet	0.001
12	betaplasma ~ age + sex + quetelet + vituse + fat + fiber + cholesterol + betadiet:fat + betadiet:vituse	smokstat	0.175
13	betaplasma ~ age + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet:fat + betadiet:vituse	sex	0.136
14	betaplasma ~ sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet:fat + betadiet:vituse	age	0.047
15	betaplasma ~ age + sex + smokstat + quetelet + fiber + betadiet:fat + betadiet:vituse	cholesterol + fiber + fat + vituse + smokstat + sex	0.102

* Difference Mean Squared Differences, not p-value

Note that this method of model selection considered all of the ancova models for the vituse variable, and selected the model that forces betadiet to have different coefficients (slopes) for each type of vitamin user, but to have the same intercept. This model says that vitamins affect the body's absorption of dietary beta-carotene, but plasma concentrations of beta-carotene are not affected by consumption of vitamins without any dietary beta-carotene intake.

Alternative ancova models include:

- free slopes (as in the chosen model) as well as free intercepts; this was eliminated in the tenth model with a p-value of 0.988.
- fixed identical slopes and free intercepts; this model was eliminated in the sixth and tenth models with a difference of mean squared differences of 0.00086 and p-value of 0.988.
- no effect of vitamin use; this was eliminated in the fourth and sixth models with a p-value of 0.40 and difference of mean squared differences of 0.00086.

- **Single Variable Model R-squared Values and F-statistics**

```
> summary(simple.beta.age)$r.squared
[1] 0.02003732
                         F Value      Pr(F)
prdata.corr.trans[, i] 6.297682 0.01260253
> summary(simple.beta.sex)$r.squared
[1] 0.01641755
                         F Value      Pr(F)
prdata.corr.trans[, i] 5.141008 0.02405982
> summary(simple.beta.smokstat)$r.squared
[1] 0.0347284
                         F Value        Pr(F)
prdata.corr.trans[, i] 11.08118 0.0009779313
> summary(simple.beta.quetelet)$r.squared
[1] 0.08121964
                         F Value        Pr(F)
prdata.corr.trans[, i] 27.22702 3.330167e-07
> summary(simple.beta.vituse)$r.squared
[1] 0.06479643
                         F Value        Pr(F)
prdata.corr.trans[, i] 21.34006 5.659727e-06
> summary(simple.beta.fat)$r.squared
[1] 0.01191156
                         F Value     Pr(F)
prdata.corr.trans[, i] 3.712988 0.0549104
> summary(simple.beta.fiber)$r.squared
[1] 0.04673261
                         F Value        Pr(F)
prdata.corr.trans[, i] 15.09927 0.0001249139
> summary(simple.beta.cholesterol)$r.squared
[1] 0.01181114
                         F Value      Pr(F)
prdata.corr.trans[, i] 3.681312 0.05594964
> summary(simple.beta.betadiet)$r.squared
[1] 0.03748456
                         F Value        Pr(F)
prdata.corr.trans[, i] 11.99487 0.0006090321
```

- **Model 1**

```
         Terms
1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet +\n\tcalories
2                   age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet

  Resid. Df      RSS      Test Df  Sum of Sq   F Value     Pr(F)
```

```
1       299 133.2891
2       300 133.4752 -calories -1 -0.1860245 0.4172984 0.5187837
```

- **Model 2**

```
        Terms
1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet +\n\talcohol
2                 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet

  Resid. Df      RSS     Test Df Sum of Sq  F Value     Pr(F)
1       299 132.2982
2       300 133.4752 -alcohol -1 -1.176966 2.659998 0.1039537
```

- **Model 3**

```
        Terms
1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet +\n\tretdiet
2                 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet

  Resid. Df      RSS     Test Df    Sum of Sq     F Value     Pr(F)
1       299 133.4736
2       300 133.4752 -retdiet -1 -0.001517945 0.003400414 0.9535382
```

- **Model 4**

```
               Terms
1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet +\n\tbetadiet:vit
2                          age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + b

  Resid. Df      RSS             Test Df Sum of Sq F Value      Pr(F)
1       299 131.5951
2       300 133.4752 -betadiet:vituse -1 -1.880063 4.27173 0.03961205
```

- **Model 5**

```
             Terms
1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + betadiet +\n\tbetadiet:fat
2                       age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + beta

  Resid. Df      RSS          Test Df Sum of Sq  F Value      Pr(F)
1       299 131.5034
2       300 133.4752 -betadiet:fat -1 -1.971743 4.483162 0.03505572
```

- **Model 6**

```
> beta9fat1 <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet + betadiet:fat, data =
data1)
> beta9fat2 <- predict.lm(beta9fat1, data2, se.fit = T)
> sum(beta9fat2$se.fit^2)/length(beta9fat2$se.fit)
[1] 0.03227349
> sqrt(var(beta9fat2$se.fit^2))
[1] 0.01535
> 2 * sqrt(var(beta9fat2$se.fit^2))
[1] 0.0307
> beta9fatvit1 <- lm(betaplasma ~ age + sex + smokstat + quetelet +
vituse + fat + fiber + cholesterol + betadiet:vituse + betadiet:
fat, data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit1, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fat2$se.fit)
[1] 0.03218796
> sqrt(var(beta9fatvit2$se.fit^2))
[1] 0.0148043
> 2 * sqrt(var(beta9fatvit2$se.fit^2))
[1] 0.02960859
```

- **Model 7**

```
> beta9fatvit <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet:vituse + betadiet:fat,
data = data1)
> this <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse + fat +
fiber + betadiet:vituse + betadiet:fat, data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.002951909
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2               age + sex + smokstat + quetelet + vituse + fat + fiber + betadiet:\n\tvituse + beta
  Resid. Df      RSS         Test Df   Sum of Sq   F Value     Pr(F)
1       144 67.74719
2       145 67.84292 -cholesterol -1 -0.09572061 0.2034589 0.6526218
```

- **Model 8**

```
> beta9fatvit <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet:vituse + betadiet:fat,
data = data1)
> this <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse + fat +
cholesterol + betadiet:vituse + betadiet:fat, data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.002484642
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2         age + sex + smokstat + quetelet + vituse + fat + cholesterol + betadiet:\n\tvituse + beta
  Resid. Df      RSS   Test Df  Sum of Sq  F Value     Pr(F)
1       144 67.74719
2       145 68.30441 -fiber -1 -0.5572119 1.184381 0.2782839
```

- **Model 9**

```
> beta9fatvit <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet:vituse + betadiet:fat,
data = data1)
> this <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fiber + cholesterol + betadiet:vituse + betadiet:fat, data =
data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.003254854
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2       age + sex + smokstat + quetelet + vituse + fiber + cholesterol + \n\tbetadiet:vituse + beta
  Resid. Df      RSS Test Df  Sum of Sq   F Value     Pr(F)
1       144 67.74719
2       145 68.09093 -fat -1 -0.3437353 0.7306264 0.3941001
```

- **Model 10**

```
> beta9fatvit <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet:vituse + betadiet:fat,
data = data1)
> this <- lm(betaplasma ~ age + sex + smokstat + quetelet + fat + fiber +
cholesterol + betadiet:vituse + betadiet:fat, data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.003387361
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2          age + sex + smokstat + quetelet + fat + fiber + cholesterol + betadiet:\n\tvituse + beta
  Resid. Df      RSS    Test Df     Sum of Sq      F Value    Pr(F)
1       144 67.74719
2       145 67.74731 -vituse -1 -0.0001118554 0.0002377542 0.987719
```

- **Model 11**

```
> beta9fatvit <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet:vituse + betadiet:fat,
data = data1)
> this <- lm(betaplasma ~ age + sex + smokstat + vituse + fat + fiber +
cholesterol + betadiet:vituse + betadiet:fat, data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.0005258024
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2           age + sex + smokstat + vituse + fat + fiber + cholesterol + betadiet:\n\tvituse + beta
  Resid. Df      RSS      Test Df Sum of Sq  F Value        Pr(F)
1       144 67.74719
2       145 73.30912 -quetelet -1 -5.561926 11.82215 0.0007656273
```

- **Model 12**

```
> beta9fatvit <- lm(betaplasma ~ age + sex + smokstat + quetelet + vituse +
fat + fiber + cholesterol + betadiet:vituse + betadiet:fat,
data = data1)
> this <- lm(betaplasma ~ age + sex + quetelet + vituse + fat + fiber +
cholesterol + betadiet:vituse + betadiet:fat, data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.00312346
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2            age + sex + quetelet + vituse + fat + fiber + cholesterol + betadiet:\n\tvituse + beta
  Resid. Df      RSS      Test Df  Sum of Sq  F Value     Pr(F)
1       144 67.74719
2       145 68.62245 -smokstat -1 -0.8752577 1.860403 0.1747064
```

- **Model 13**

```
> this <- lm(betaplasma ~ age + smokstat + quetelet + vituse + fat +
fiber + cholesterol + betadiet:vituse + betadiet:fat, data =
data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.002301733
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2       age + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
  Resid. Df      RSS Test Df Sum of Sq  F Value     Pr(F)
1       144 67.74719
2       145 68.80729 -sex -1 -1.060096 2.253285 0.1355206
```

- **Model 14**

```
> this <- lm(betaplasma ~ sex + smokstat + quetelet + vituse + fat +
fiber + cholesterol + betadiet:vituse + betadiet:fat, data =
data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.002223338
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2       sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
  Resid. Df      RSS Test Df Sum of Sq  F Value      Pr(F)
1       144 67.74719
2       145 69.64027 -age -1 -1.893078 4.023831 0.04673472
```

- **Model 15**

```
> this <- lm(betaplasma ~ age + quetelet + betadiet:vituse + betadiet:fat,
data = data1)
> beta9fatvit2 <- predict.lm(beta9fatvit, data2, se.fit = T)
> this2 <- predict.lm(this, data2, se.fit = T)
> sum(beta9fatvit2$se.fit^2)/length(beta9fatvit2$se.fit) - sum(this2$
se.fit^2)/length(this2$se.fit)
[1] 0.01729368
> anova(beta9fatvit, this)
Analysis of Variance Table

Response: betaplasma


1 age + sex + smokstat + quetelet + vituse + fat + fiber + cholesterol + \n\tbetadiet:vituse + beta
2                                                                 age + quetelet + betadiet:vituse +
  Resid. Df      RSS                                        Test Df
1       144 67.74719
2       150 72.83957 -sex-smokstat-vituse-fat-fiber-cholesterol -6
  Sum of Sq  F Value     Pr(F)
1
2 -5.092372 1.804015 0.1022574
```

Appendix D
Glossary

Age: The age of a subject in years.

Alcohol: The number of alcoholic drinks a subject consumes per week.

Betadiet: The number of micrograms of beta-carotene a subject consumes per day.

Betaplasma: The number of nanograms of beta-carotene per milliliter of a subject's blood plasma.

Calories: The number of calories a subject consumes per day.

Cholesterol: The milligrams of cholesterol a subject consumes per day.

F-statistic: A statistic that is often used in regression analysis to compare two hypotheses, often referred to as the null and alternative hypotheses. In the case of simple regression, the F-statistic compares the null hypothesis that all of the independent variable coefficients (not including the intercept) are zero to the alternative hypothesis that all of the coefficients are not zero. When comparing nested models, the F-statistic compares the null hypothesis that the smaller model has less deviation from the measured dependent variable to the alternative hypothesis that the larger model has less deviation. The distribution of the F-statistic is well known if the null hypothesis is true (and is called the F distribution); therefore if the calculated F-statistic has a large probability according to the F distribution, the null hypothesis can not be rejected. On the other hand, if the calculated F-statistic has a small probability according to the F distribution, the null hypothesis can be rejected and the alternative hypothesis is accepted. See also p-value.

Fat: The grams of fat a subject consumes per day.

Fiber: The grams of fiber a subject consumes per day.

p-value: The probability that a statistic is less likely than the statistic that was calculated. The p-value is often used in hypothesis testing, since a statistic with low probability allows null hypothesis rejection. A generally accepted standard is to reject the null hypothesis when the p-value is less than 0.05. This is also known as rejecting at the 95

Quetelet: Weight divided by squared height; this is a measure of obesity, with subjects considered obese at quetelet indices higher than 28 (men) or 27 (women).

R-squared: The percentage of the variation in the data that a given model accounts for; also called the Coefficient of Determination

Regression: Fitting a line to the observed data that minimizes the differences between the observed dependent variables and those that would be predicted by the line. In this report, linear regression was used exclusively.

Retdiet: The number of micrograms of retinol a subject consumes per day.

Retplasma: The number of nanograms of retinol per milliliter of a subject's blood plasma.

Sex: The gender of a subject.

Smokstat: The smoking status of a subject (never, former, or current).

Vituse: The vitamin use for subjects (often, occasionally, or never)