Determinants of Plasma Beta-Carotene Levels

Project 1. Linear Regression.

Authors: Caroline Olsmats, Felix Persson & Tom Richter

2022-04-11

Abstract: During the project several regression models have been constructed and tested. Among these models, a model based on the data set's continuous variables *age*, *quetelet*, *calory* and *fiber intake*, *dietary beta-carotene intake* and the categorical variables *sex* and *smoking status* was elected to be the best. The model was a logarithmic-linear model determined through stepwise selection based on Akaike information criterion (AIC), where improvements in AIC were considered equal to model improvements. Modelling the logarithm of the beta-carotene was done since it implied a stronger linear relationship between y-estimates and x-values. Also the residuals showed stronger tendencies towards being normally distributed with respect to the x-values and y-estimates. The model was elected to be the best model since its adjusted squared residuals bested the other built and refined models.

Introduction

The aim of this study is to investigate the relationship between a number of personal characteristics and dietary factors, and blood plasma concentrations of *beta-carotene* and other carotenoids. The understanding of determinants of plasma concentration of these substances is a relatively new, yet important, field. A low plasma concentration and intake of beta-carotene has been suggested to increase the risk of developing some types of cancer. The factors investigated in this study are presented in *Table 1*.

Table 1. Personal characteristics and dietary factors investigated in the study.

Factor	**Categories**	**Unit**
Age		years
Sex	Male, Female	
Smoking status	Non smoker, Former smoker, Current Smoker	
Quetelet (continuous BMI)		kg/m^2
BMI category	Underweight, Normal, Overweight, Obese	
Vitamin use	Fairly often, Not often, No	
Calorie intake		nr/day
Fat intake		g/day
Fiber intake		g/day
Alcohol intake		drinks/week
Cholesterol intake		mg/day
Dietary beta-carotene intake		µg/day
Plasma beta-carotene concentration		ng/ml

Methodology

Data was gathered from 315 patients who were undergoing surgery to biopsy or to remove a lesion of the lung, colon, breast, skin, ovary or uterus. The patient's lesions were found to be non-cancerous. The data was collected in the data file plasma.txt and analysed using RStudio.

The study is divided into three parts. In part 1, the relationship between *plasma beta-carotene concentration* and *age* is investigated. Two models, one linear and one logarithmic, are fit to the data and their residuals analysed in order to find a suitable model for the relationship. Confidence and prediction intervals are calculated for the beta-carotene concentration.

Model 1: $$pbc = \beta_0 + \beta_1 \cdot age \quad (1)$$

Model 2: $$ln(pbc) = \beta_0 + \beta_1 \cdot age \Leftrightarrow pbc = e^{\beta_0} \cdot (e^{\beta_1})^{age} \quad (2)$$

In the second part further variables are added to the model. Those are the dummy variables s*ex*, *smoking status* and *BMI category*. Appropriate reference categories are chosen in order to achieve a good precision. Confidence intervals are calculated for the coefficients and different aspects of the model are tested through global and partial F-tests and a t-test.

Model 3: $$ln(pbc) = \sum_{k=0}^{3} \beta_k x_k \quad (3)$$

Model 4: $$ln(pbc) = \sum_{k=0}^{7} \beta_k x_k \quad (4)$$

Where in equation 3:

- $\beta_0 x_0$ is the intercept, ie $x_0 = 1$, with either *Underweight* or *normal* as bmi reference variable
- $[\beta_1 x_1, \beta_2 x_2, \beta_3 x_3]$ correspond to *normal* (or underweight, depending on which is reference variable), *overweight* and *obese.* $x_i \; \forall \; i \in [1, 2, 3]$ are binary variables

And where in equation 4:

- $\beta_0 x_0$ is the intercept, ie $x_0 = 1$, with *normal* as bmi reference variable, *Female* as sex reference variable and *non-smoker* as smoking status reference variable
- $[\beta_1 x_1, \beta_2 x_2]$ correspond to *age* (x_1 continuous) and *male* (x_2 binary).
- $[\beta_3 x_3, \beta_4 x_4]$ correspond to *current smoker* and *former smoker* (x's binary).
- $[\beta_5 x_5, \beta_6 x_6, \beta_7 x_7]$ correspond to *underweight*, *overweight* and *obese.* x's are binary

In the subsequent model 5 we used *quetelet* as bmi variable as opposed to *bmicat*. Hence, our model looked as in equation 5:

Model 5: $$ln(pbc) = \sum_{k=0}^{5} \beta_k x_k \quad (5)$$

And where:

- $\beta_0 x_0$ is the intercept, ie $x_0 = 1$, with *Female* as sex reference variable and *non-smoker* as smoking status reference variable
- $\left[\beta_1 x_1,\ \beta_2 x_2\right]$ correspond to *age* (x_1 continuous) and *male* (x_2 binary).
- $\left[\beta_3 x_3,\ \beta_4 x_4\right]$ correspond to *current smoker* and *former smoker* (x's binary).
- $\beta_5 x_5$ correspond to *quetelet* (x_5 continuous)

In model 6 we included both *bmicat* and *quetelet*. Hence the model looked as in equation 6

Model 6: $$ln(pbc) = \sum_{k=0}^{8} \beta_k x_k \qquad (6)$$

Where:

- $\left[\beta_1 x_1,\ \beta_2 x_2,\ \beta_3 x_3,\ \beta_4 x_4,\ \beta_5 x_5\right]$ are as in model 5
- $\left[\beta_6 x_6,\ \beta_7 x_7,\ \beta_8 x_8\right]$ are equivalent to $\left[\beta_1 x_1,\ \beta_2 x_2,\ \beta_3 x_3\right]$ from model 3

In part 3 the dietary variables (*vitamin use, calorie intake, fat intake, fiber intake, alcohol intake, cholesterol intake* and *dietary beta-carotene intake* are added to model 4 to improve the model further. Also a final model was determined.

Model 7: $$ln(pbc) = \sum_{k=0}^{11} \beta_k x_k \qquad (7)$$

Where:

- $\left[\beta_1 x_1,\ \beta_2 x_2,\ \beta_3 x_3,\ \beta_4 x_4,\ \beta_5 x_5\right]$ are as in model 5
- $\beta_6 x_6$ correspond to *vitamin use*, x_6 continuous
- $\beta_7 x_7$ correspond to *calorie intake*, x_7 continuous
- $\beta_8 x_8$ correspond to *fat intake*, x_8 continuous
- $\beta_9 x_9$ correspond to *fiber intake*, x_9 continuous
- $\beta_{10} x_{10}$ correspond to *alcohol intake*, x_{10} continuous
- $\beta_{11} x_{11}$ correspond to *cholesterol intake*, x_{11} continuous

The remaining models explored in this project are reductions of the above ones and will be explained more in detail later on.

Since the models aren't nested and have a different number of parameters, the adjusted coefficient of determination R^2_{adj} is used to compare the models, together with the measurements AIC and BIC.

One of the tasks was to express the relative difference in beta-carotene concentration under both the model with *BMI category* as a variable (i.e. discrete categories for bmi) and *quetelet* (i.e. continuous bmi). These relative differences are expressed in the equations 8 and 9 below.

$$\frac{Y_2 - Y_1}{Y_1} = e^{\beta_{obese}} - 1 \quad (8)$$

$$\frac{Y_2 - Y_1}{Y_1} = e^{\Delta_{quetelet}\beta_{quetelet}} - 1 \quad (9)$$

Results & analysis

Part 1. Modelling plasma beta-carotene and age

Models

One of the subjects reported a concentration of 0.00 ng beta-carotene per ml blood. The result seems unlikely and the data point is therefore removed from the dataset in order to enable logarithmic modelling.

In *Figure 1*, the plasma beta-carotene concentration is plotted against *age* together with the two models. *Table 2* shows the beta estimates and confidence intervals for model 1.

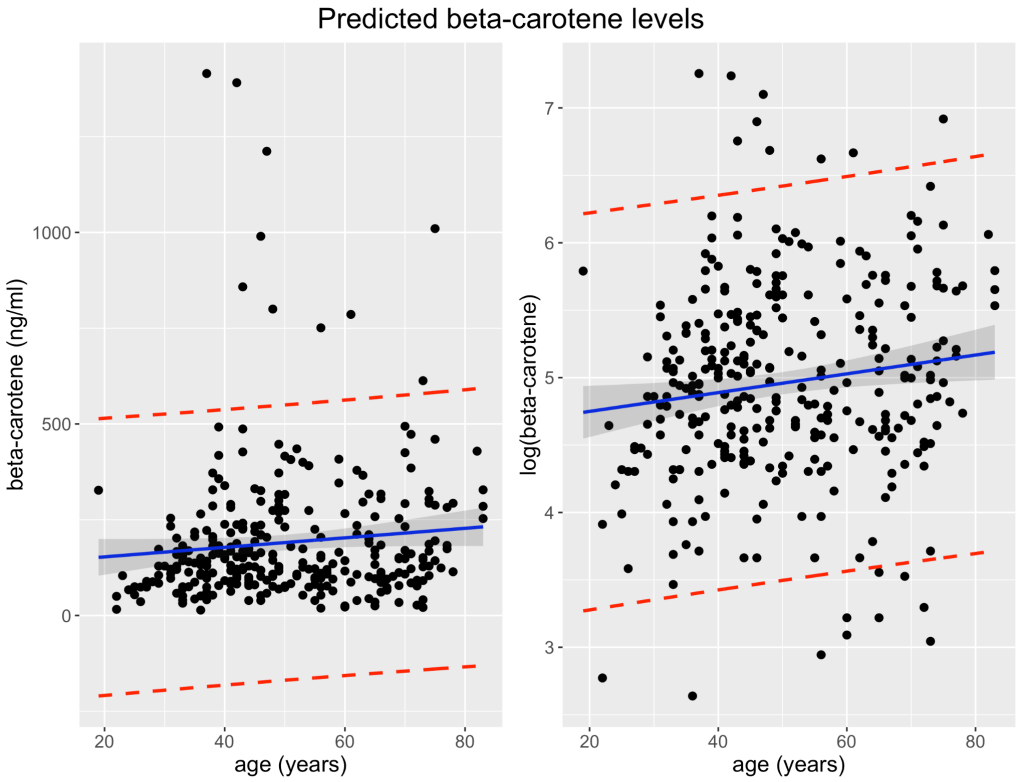

Figure 1. The plot to the left shows plasma beta-carotene concentration against *age,* stemming from a linear model adjusted to the data. The plot to the right is the corresponding values for a log-lin model. 95% confidence and prediction intervals are included for both models.

Table 2. Beta estimates and confidence intervals for model 1.

Constant	**Estimate**	**Lower boundary**	**Upper boundary**
β_0	151.7525	103.9054584	199.599522
β_1	1.2427	-0.1477424	2.633078

Table 3. Beta estimates and confidence intervals for model 2.

Constant	**Estimate**	**Lower boundary**	**Upper boundary**
β_0	4.742361371	4.547702833	4.93701991
β_1	0.006965795	0.001309118	0.01262247
e^{β_0}	114.70474	94.41527	139.354342
e^{β_1}	1.00699	1.00131	1.012702

Residuals

Residuals are defined as the difference between the observed y-values and the corresponding predicted values. If they are randomly distributed around zero, when plotting their values against the predicted values or the x-values, it is usually implied that the underlying data lacks a linear relationship. In *Figure 2*, the residuals are plotted against *age* for the two models.

Residual plots

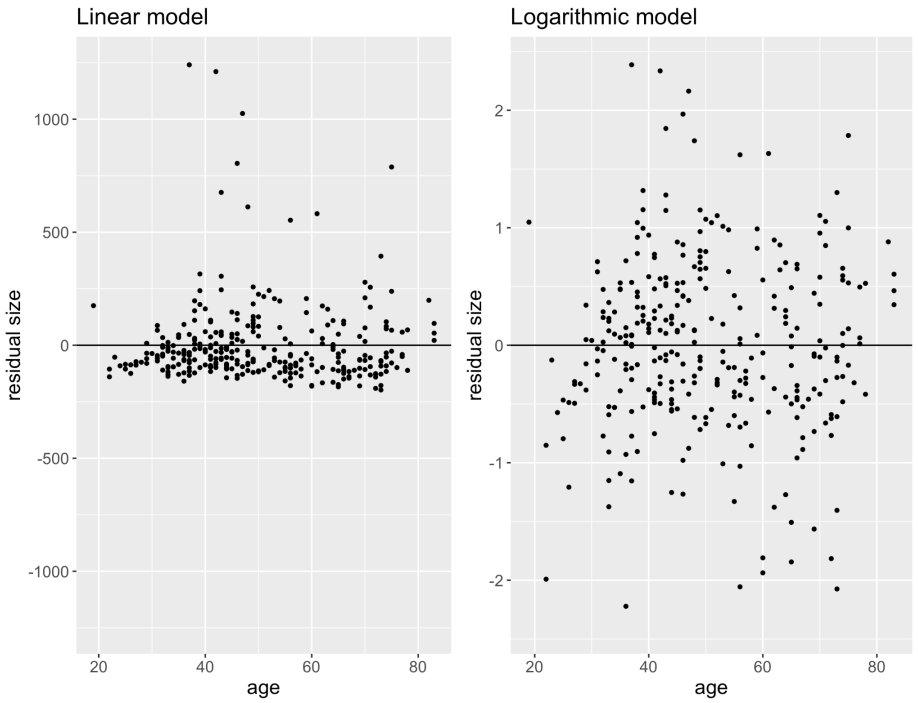

Figure 2. The residuals related to model 1 and model 2 plotted against the x-variable *age.*

QQ-plots

QQ-plots are used in order to determine the normality of the residuals and hence if there are an underlying linear relationship in the data. *Figure 3* contain QQ-plots for the two models presented so far. The QQ-plots show how well the residuals follow a normal distribution by plotting the residual quantiles on the y-axis and the quantiles of a normal distribution on the x-axis.

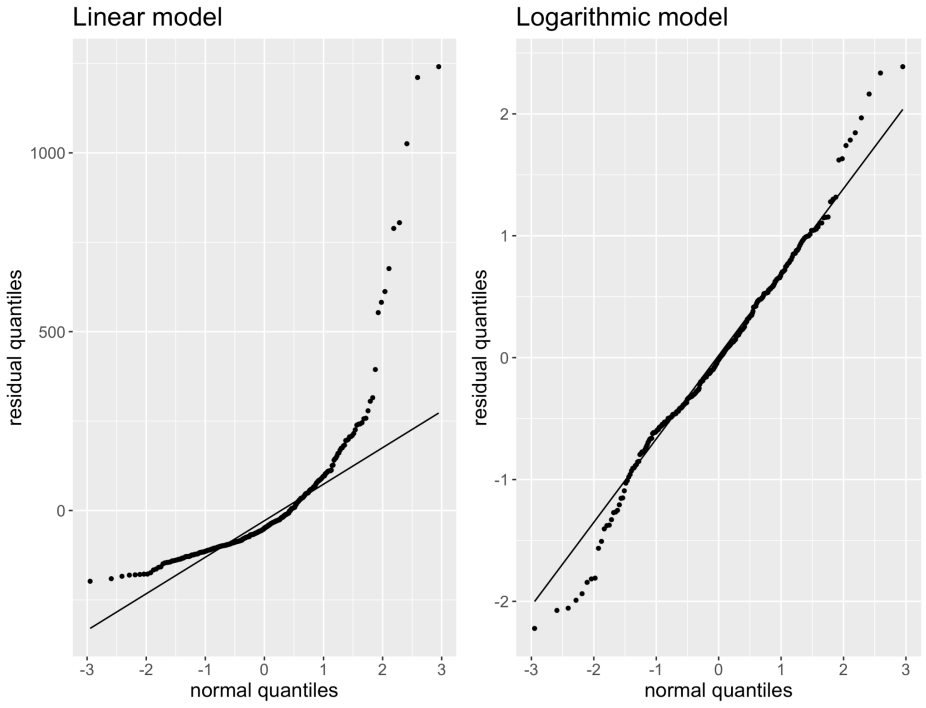

Figure 3. Normal QQ-plot of the residuals related to both models

Histograms

In order to further evaluate the normality of the residuals, histograms were generated for the two models, which are presented in *Figure 4*.

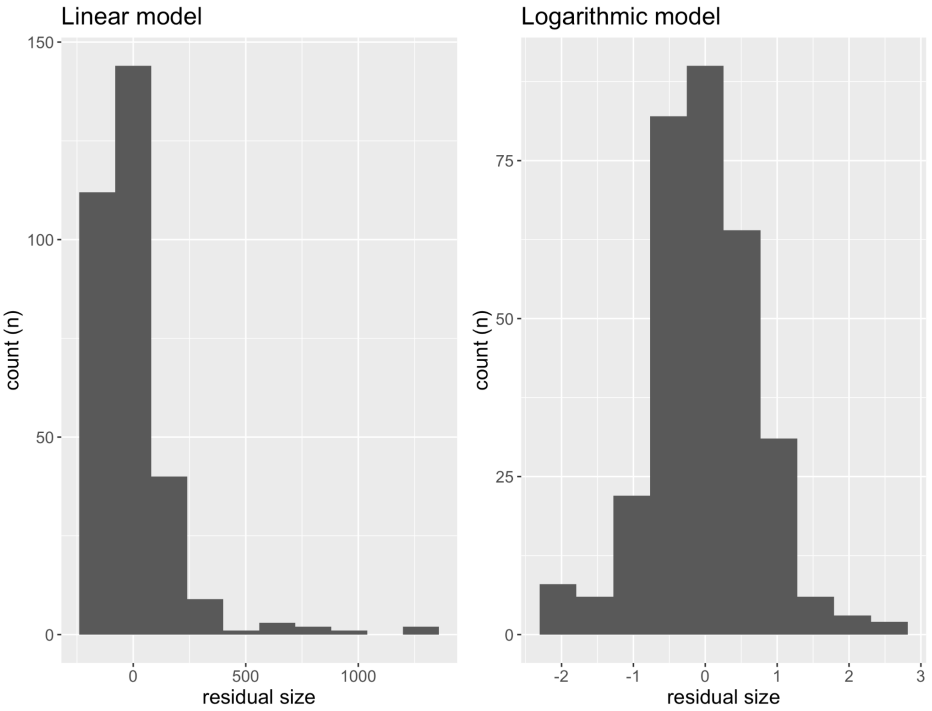

Figure 4. Histogram of the residuals belonging to the two models.

The effect on predicted values when changing the x-variable

For the logarithmic model, the predicted amount of plasma beta-carotene is multiplied by e^{β_1} = 1.00699 ng/ml as the age increases by one year, according to equation 10. A 95% confidence interval for the average increase is (1.00131, 1.012702). Hence, the average plasma beta carotene concentration increases with between 0.31% and 1.27% per year, with a certainty of 95%. (Since the estimates are logarithmic and the x-variable is linear, an additive change in age gives a relative change of beta-carotene concentration.) Therefore, the absolute increase is larger among older people who have a higher average beta-plasma concentration than younger people. Furthermore, the predictions and corresponding confidence interval widths for the ages 25, 26, 75 and 76 are presented in *Table 4*.

$$pbc_{new} = e^{\beta_0} \cdot (e^{\beta_1})^{age_{new}} = e^{\beta_0} \cdot (e^{\beta_1})^{age+1} = e^{\beta_0} \cdot (e^{\beta_1})^{age} \cdot e^{\beta_1} = pbc \cdot e^{\beta_1} \quad (10)$$

Table 4. Predicted average and prediction interval width for 25, 26 75 and 76 years old people

Age	**Fit**	**Prediction interval width**
25	119.6004	39.53499

26	120.4364 (= 119.6004 + 0.836)	38.62251
75	169.431	55.40826
76	170.6154 (= 169.431 + 1.1844)	57.49162

The results show that the confidence interval is wider for a 75 year old person than for a 25 year old person. Since the predicted value for an old person is larger than for a young person, the confidence interval for the model will also be expected to be wider after exponentiating it.

Analysis

The large number of data points above the prediction interval and none below in *Figure 1* indicates that model 1 is not a good representation of the dataset. *Figure 4*, the histogram of model 1 is significantly skewed, indicating that the residuals are not normally distributed, which is confirmed by *Figure 3*. Model 2 seems overall to be a good representation of the data, showing no systematic pattern, a constant variance and a residual distribution close to normal distributed. Through the rest of the project, the relationship between plasma beta-carotene levels and age are therefore considered logarithmic-linear.

Part 2. Plasma Beta-carotene and the background variables

In part two, the model is expanded to include all of the relevant background variables. The background variables, presented in *Table 1,* consist of both continuous and categorical variables. The categorical variables, their potential values and frequencies are presented in the frequency table below (*Table 5)*.

Table 5. Frequency table for the categorical variables *sex*, *smoking status* and *BMI category*

Variable	**Categories**	**Frequency**
Sex	Male	42
Sex	*Female*	272
Smoking status	*Non Smoker*	156
Smoking status	*Former Smoker*	115
Smoking status	*Current Smoker*	43
BMI category	*Underweight (BMI < 18.5)*	4
BMI category	*Normal (18.5 < BMI < 25)*	160
BMI category	*Overweight (25 < BMI < 30)*	89
BMI category	*Obesity (BMI > 30)*	61

Modelling based solely on BMI category

Modelling a logarithmic-linear relationship based solely on the categorical variable *BMI category* is referred to as model 3, and has a cost of three degrees of freedom as the variable has four categories. Estimated coefficients, with *underweight* used as a reference category, are presented in *Table 6*.

Table 6. Betas and corresponding standard errors for a logarithmic model solely based on *BMI category*. *Underweight* comprise the reference category

Coefficient	**β-value**	**Standard error**
β_0 *(Intercept)*	5.3602	0.3615
β_1 *(Normal weight)*	-0.2324	0.3660
β_2 *(Overweight)*	-0.4870	0.3695
β_3 *(Obese)*	-0.7421	0.3732

In order to reduce the standard errors, the reference category was changed to the value occurring the most, namely *normal weight*. When releveling the model with respect to the *normal weight* category the following parameters were obtained, presented in *Table 7*.

Table 7. Betas and corresponding standard errors for a logarithmic model solely based on *BMI category. Normal weight* comprise the reference category.

Coefficient	**β-value**	**Standard error**
β_0 *(Intercept)*	5.12774	0.05716
β_1 *(Normal weight)*	0.23244	0.36598
β_2 *(Overweight)*	-0.25451	0.09560
β_3 *(Obese)*	-0.50968	0.10879

Further on, the categories *sex* and *smoke status* were releveled too. The new reference categories were the most frequent categories *female* and *non smoker*. The releveling was once again performed to increase the precision of β_0 by choosing the largest categories in the dataset.

Modelling based on *age*, *sex*, *smoke status* and *BMI category - The full model*

In the next step, all three categorical variables *sex*, *smoke status* and *BMI category* were fitted in a log-lin model referred to as model 4 . The coefficients obtained are presented in *Table 8*. The table also include the predicted values and confidence intervals for e^{β}, which show the actual changes in beta plasma concentration (not the logarithmed value).

Table 8. β-values, exponentiated β-values and corresponding confidence interval for the model based on *age*, *sex*, *smoke status* and *BMI category*

Coefficient	**β-value**	**e^{β}-value**	**Lower bound e^{β}-value**	**Upper bound e^{β}-value**
β_0 *(Intercept)*	5.037492663	154.0831919	123.5356642	192.1844203
β_1 *(age)*	0.007462448	1.0074904	1.0018332	1.0131795
β_2 *(sexMale)*	-0.339105835	0.7124070	0.5603093	0.9057922
β_3 *(smokstatFormer)*	-0.108180556	0.8974655	0.7561136	1.0652425
β_4 *(smokstatCurrent)*	-0.449326978	0.6380574	0.4992583	0.8154443
β_5 *(bmicatUnderweight)*	0.306684718	1.3589125	0.6759210	2.7320397
β_6 *(bmicatOverweight)*	-0.217344152	0.8046530	0.6706075	0.9654925
β_7 *(bmicatObese)*	-0.547742627	0.5782537	0.4700572	0.7113545

Inspection of parameters β_5, β_6 and β_7 (and the corresponding e^{β_5}, e^{β_6} and e^{β_7}) in *Table 8* reveals that the average beta-carotene levels decrease with an increasing bmi (which is consistent with the results when using *quetelet* instead of *BMI category,* which we will see on pages 10-12).

Testing various aspects of the full model and it's reduction

In order to test whether all variables make a significant contribution to the model, several nested models were tested out and put in comparison. The results are shown in *Table 9*.

Table 9. Various tests performed on the full model versus various reduced models. The table contains the type of test, null hypothesis, the test statistic's value, it's distribution, a corresponding p-value and whether rejection of H_0 takes place

Type of test	**Null hypothesis (H_0)**	**Test statistic**	**Distribution of test statistic**	**P-value**	**Rejection of H_0***
Global F-test	$\beta_1 = 0$ $\beta_2 = 0$ … $\beta_n = 0$	8.187	$F_{7,306}$	3.786e-09	Yes
Partial F-test	$\beta_1 = 0$	8.433	$F_{6,306}$	1.763e-08	Yes

	$\beta_2 = 0$ … $\beta_n = 0$				
Global F-test	$\beta_{age} = 0$	6.800496	$F_{1,306}$	0.00956	Yes
Global F-test	$\beta_{sex} = 0$	7.719798	$F_{6,306}$	0.005799	Yes
Global F-test	$\beta_{smokstat} = 0$	6.499011	$F_{2,306}$	0.001721	Yes
Global F-test	$\beta_{bmicat} = 0$	9.837328	$F_{3,306}$	3.266e-06	Yes
T-test	$\beta_{underweight} = 0$	0.8641311	t_{306}	0.3881927	No

** The tests were all performed on a significance level* $\alpha = 0.05$*, all tests are two sided.*
*** The sub-index reg.age refers to a reduced full model in respect to the variable age*

Figure 5 shows the model's prediction on the dataset together with confidence intervals, prediction intervals with the data points overlayed. It shows clearly where we have outliers. The predictions are split into *BMI categories* as well as *smoking status*

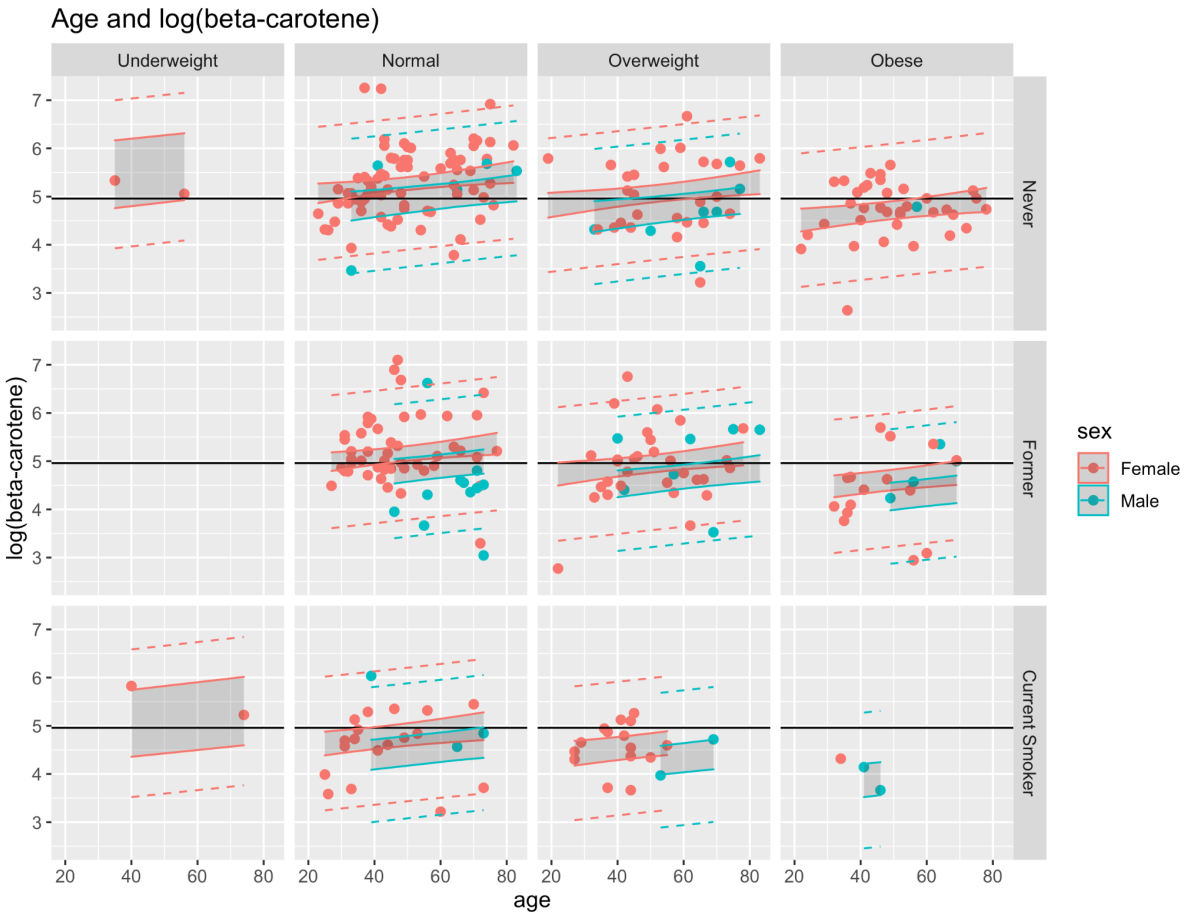

Figure 5. The model's prediction on the dataset together with confidence intervals, prediction intervals as well as the underlying data points. The predictions are split into *BMI categories* as well as *smoking status*

Replacing the categorical BMI category with continuous variable quetelet

The next decision is whether to include either the categorical variable *BMI category* or the continuous variable *quetelet* in the model or both. *Table 10* shows the estimated values of the coefficients and confidence intervals when using only the continuous variable *quetelet*. *Table 11* shows how the coefficients changed when switching from the categorical variable BMI category to the continuous variable *quetelet. Table 12* and *13* show the estimates and confidence intervals for average beta-carotene levels for males and females for the two different models. Because the vast majority of participants in the study were female one can see that the confidence intervals are wider for men than for women under both models and weight categories/quetelets. *Table 14* shows the estimated values of the coefficients and confidence intervals for the different options.

Table 10. Parameters with the continuous bmi variable *quetelet*

Coefficient	**β-value**	**e^{β}-value**	**Lower bound e^{β}-value**	**Upper bound e^{β}-value**
β_0 (Intercept)	5.84624796	345.9339845	230.6548793	518.8284854
β_1 *(Age)*	0.00744712	1.0074749	1.0018797	1.0131014
β_2 *(Sex)*	-0.34358297	0.7092246	0.5590113	0.8998022
β_3 *(SmokstatFormer)*	-0.11489066	0.8914636	0.7527599	1.0557249
β_4 *(SmokstatCurrent)*	-0.45142411	0.6367207	0.5009615	0.8092704
β_5 *(Quetelet)*	-0.03706947	0.9636092	0.9513219	0.9760552

Table 11. Changes in the other parameter values when using *quetelet* (continuous) as opposed to *BMI category* (discrete) as additional parameters

	β_0 *(Intercept)*	β_1 *I(age - minage)*	β_2 *sexMale*	β_3 *smokstatFormer*	β_4 *smokstatCurrent*
Δ	-8.088e-01	1.532e-05	4.477e-03	6.710e-03	2.098e-03

Table 12. Estimates and confidence intervals for average (median) beta-carotene level for *males* and *females* under the two different models with *quetelet* = 22 or *bmicat* = *normal weight*

	Male, model 4	**Female, model 4**	**Male, model 5**	**Female, model 5**
Estimate	4.263985	4.603091	4.250607	4.59419
Lower	3.911853	4.347106	3.908486	4.349421

Upper	4.616116	4.859075	4.592727	4.838959

Table 13. Estimates and confidence intervals for average (median) beta-carotene level for *males* and *females* under the two different models with *quetelet* = 33 or *bmicat* = *overweight*

	Male, model 4	**Female, model 4**	**Male, model 5**	**Female, model 5**
Estimate	3.842843	4.186425	4.046641	4.385746
Lower	3.490758	3.923847	3.689129	4.115041
Upper	4.194927	4.449004	4.404152	4.656452

Table 14. Coefficients values when including both *BMI category* and *quetelet* in the model

Coefficient	**β-value**	**e^{β}-value**	**Lower bound e^{β}-value**	**Upper bound e^{β}-value**
β_0 (Intercept)	5.624342542	277.0900496	139.4417024	550.6164531
β_1 (Age)	0.007429845	1.0074575	1.0018201	1.0131267
β_2 *(Sex)*	-0.33908680	0.7124206	0.5607892	0.9050516
β_3 *(SmokstatFormer)*	-0.11174205	0.8942749	0.7538413	1.0608700
β_4 *(SmokstatCurrent)*	-0.45455325	0.6347315	0.4970465	0.8105561
β_5 *(Quetelet)*	-0.02652380	0.9738249	0.9456134	1.0028780
β_6 *(BmicatUnderweight)*	0.18600549	1.2044289	0.5929404	2.4465343
β_7 *(BmicatOverweight)*	-0.08136042	0.9218614	0.7280752	1.1672261
β_8 *(BmicatObese)*	-0.17433040	0.8400193	0.5289667	1.3339828

Analysis

By looking at *Table 7* and *8* we see that the standard errors for the two models based solely on *BMI category* (using either *underweight* or *normal weight* as reference categories) are significantly lower when using *normal weight*, and hence gives a more reliable estimation. This is due to the fact that *normal weight* has by far the highest frequency. Therefore we used *normal weight* as the bmicat reference variable in all models thereafter.

Looking at *Table 9*, we can conclude that the model using all of *age, sex, smoking status* and *BMI category* is significantly better than the null hypotheses (using only an intercept, using only age, reduced models taking one out of *age, sex, smoking status* and *BMI category*). However, the

underweight subcategory to *BMI category* is essentially redundant in predicting beta-carotene levels, as shown by the T-test in the final line of *Table 9*.

Looking at *Figure 5*, we can see that there are no underweight males in our data set (and only 4 females). This is certainly a drawback in our models. There are only 2 underweight former smokers. The question becomes whether we should rely on our model for estimates with these variables, and the answer is naturally no since the variance is high and one can be certain that the data set does not present the population at large.

When using *quetelet* instead of *BMI category*, the coefficients β_1 to β_4 remain almost unchanged, while the intercept β_0 changes significantly. This is due to the fact that choice of reference category affects the intercept when using the categorical variable while the continuous variable *quetelet* has no such impact on the intercept.

Adding both the categorical variable *BMI category* and the continuous variable *quetelet* to the model does not add any relevant information, since the variables measure the same thing and ence are strongly correlated. However, it reduces the model's degrees of freedom, leading to a less accurate estimation with larger uncertainty. This phenomena is investigated further in Part 3.

Part 3. Model validation and selection

Table 15. Model comparison between model containing continuous quetelet or BMI category

Model	R^2	**adjusted** R^2	**Degrees of freedom**	**AIC**	**BIC**
$Model_4$	0.01846886	0.01532293	9	671.5858	705.3303
$Model_5$	0.15774521	0.13847794	7	665.1774	691.4232

Further on the $Model_5$ will be referred to as the background model.

The continuous variables in the dataset

When looking into the continuous variables in the dataset the following relationships were discovered.

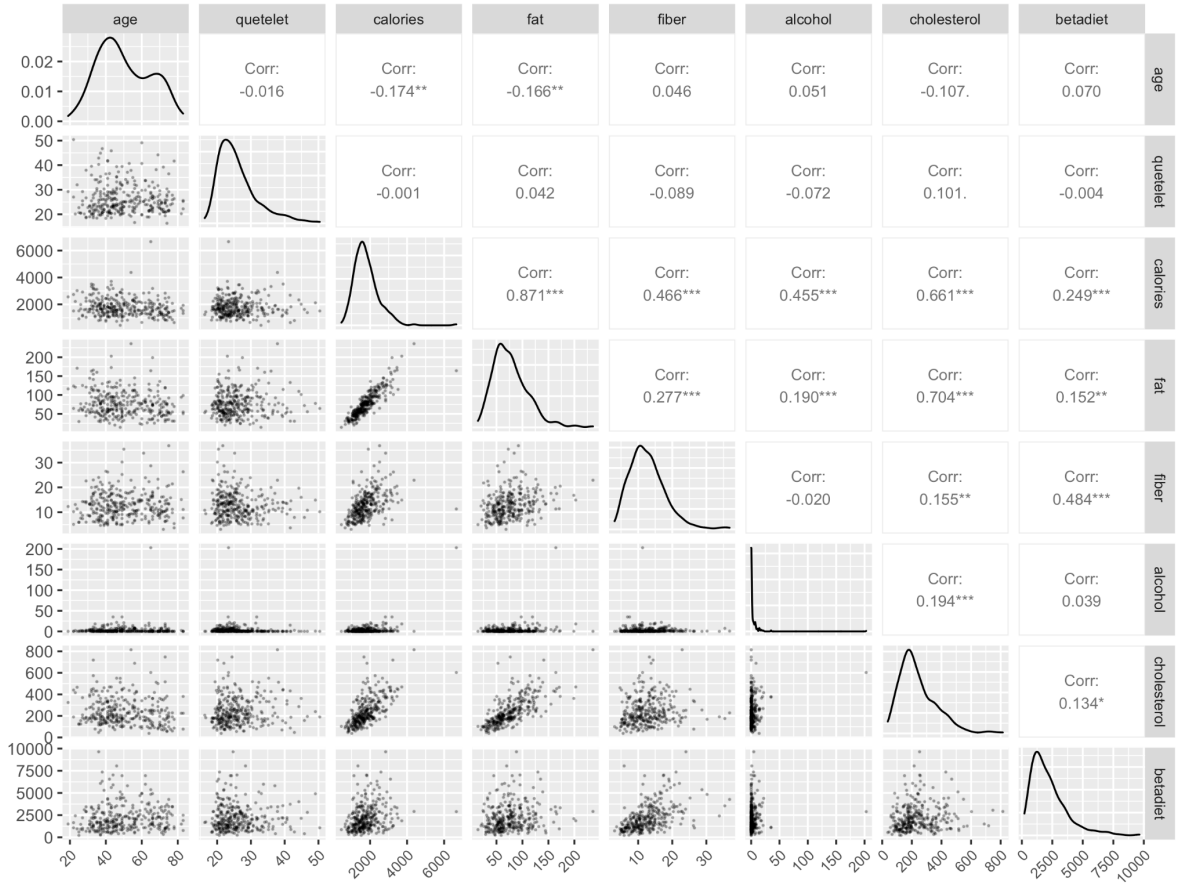

Figure 6. All continuous data categories in the dataset plotted against each other together with the corresponding correlation

Fat/Calories and Fat/Cholesterol have positive correlations greater than 0.7. In the plots there are clear linear patterns between the variables, see *Figure 7*.

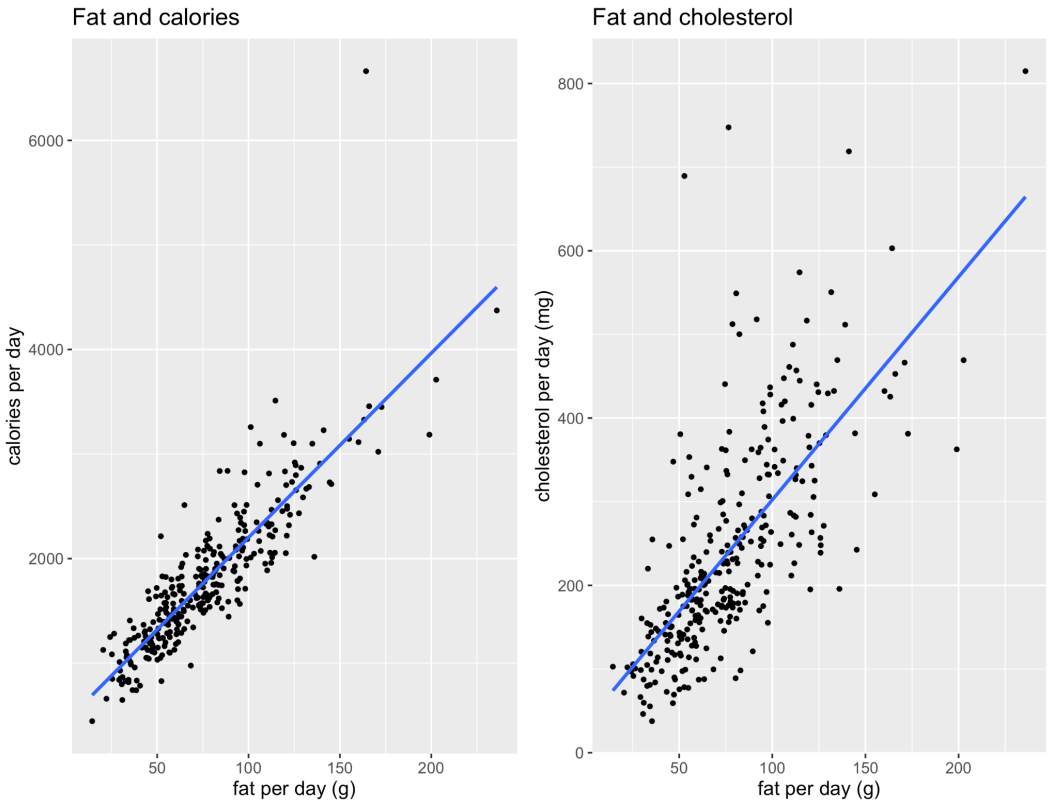

Figure 7. A closer look on the *calories* vs. *fat* per day and *cholesterol* vs. *fat* per day

A new model based on dietary variables

In this part, a new model is to be developed based on dietary variables. These are *vitamine usage*, *calories*, *fat*, *fiber*, *alcohol*, *cholesterol* and *dietary beta-carotene concentration*. In regards to the categorical variable *vitamin usage*, *Table 13* presents the occurring values within the category together with the corresponding frequency.

Table 16. Frequency table for *vitamin usage*.

Category	**Frequency**
1 - Yes, fairly often	121
2 - Yes, not often	82
3 - No	111

The reference category chosen for *vitamin usage* is category 3 - *no*, although category 1 - *fairly often* has the highest frequency. However, it seems reasonable to choose *no* as the reference category and investigate the impact from adding vitamins to the diet.

Dietary model - leverages

In order to investigate the leverages in regards to the dietary model, leverage was plotted against *age* as well as *alcohol intake*. The plots, presented in *Figure 8*, indicate that there is an outlier with particularly high leverage and particularly high intake of alcohol.

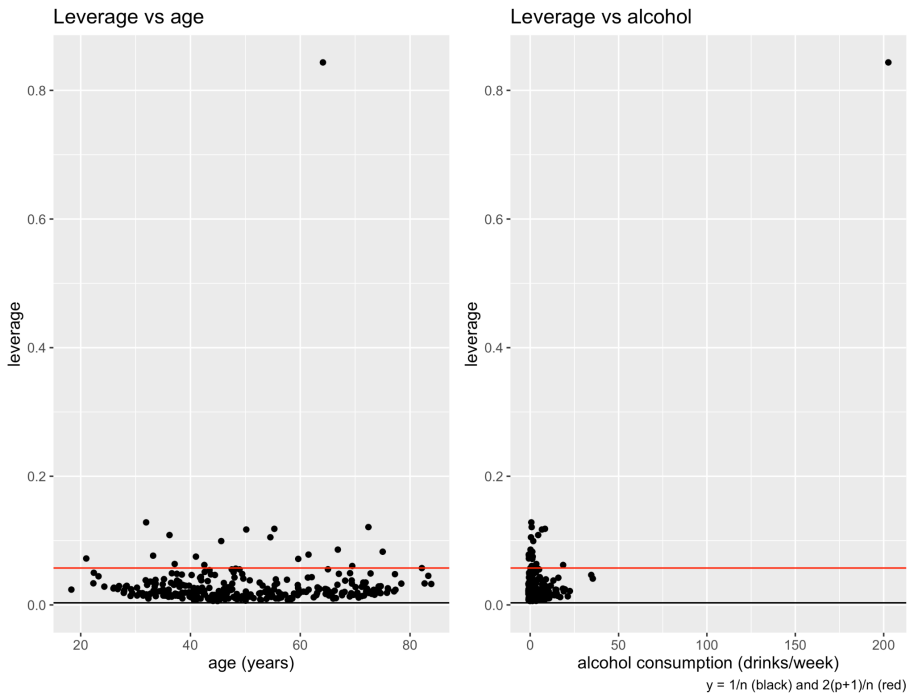

Figure 8. Leverage plotted against *age* and *alcohol intake*

To determine whether the outlier is an outlier in other categories too, *alcohol intake* is plotted against various continuous variables in the dataset in *Figure 9*. *Figure 10* displays a histogram of the alcohol intake, showing the extremely high intake of the outlier.

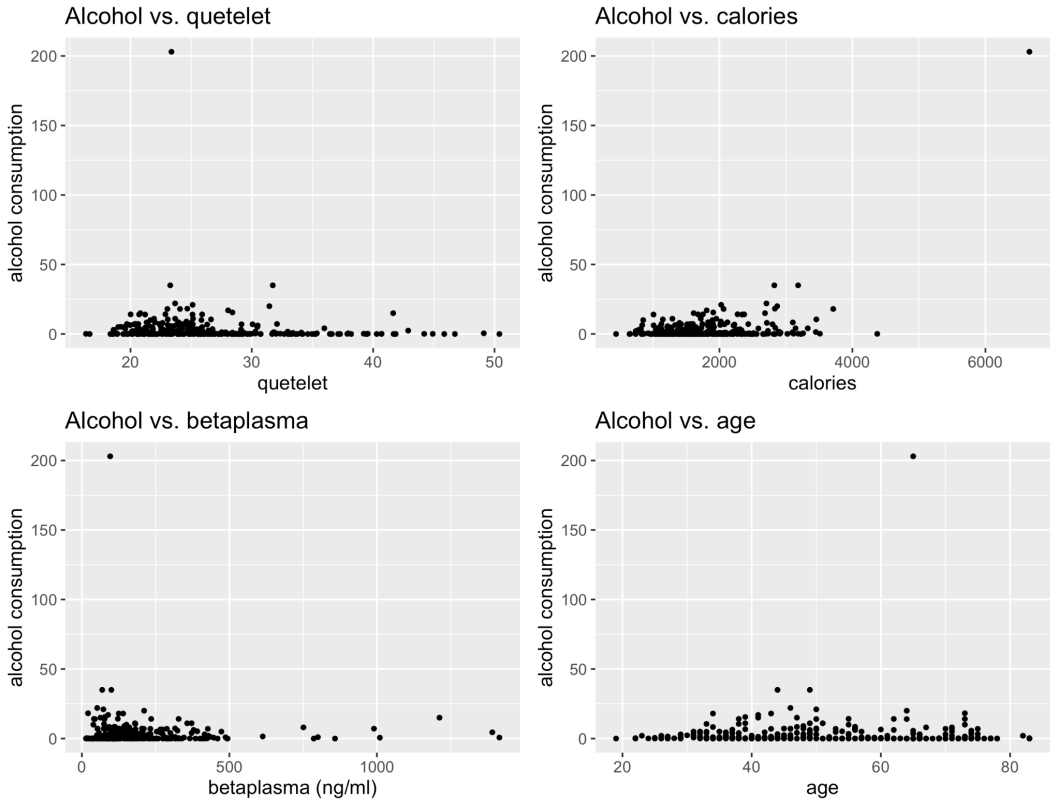

Figure 9. Alcohol intake vs various x-variables

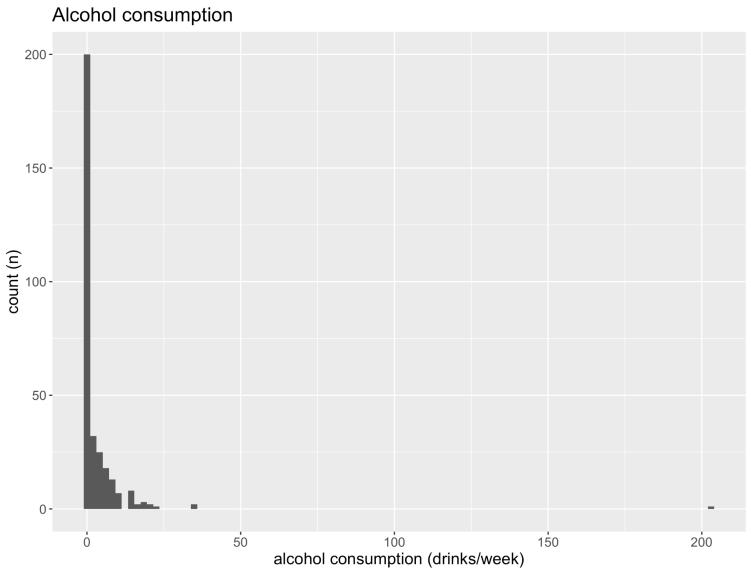

Figure 10. A histogram over the alcohol intake.

It can be seen that the alcohol intake generally seems to be quite small. Doing a category count, see *attached code*, it can be seen that there are 110 people in the sample that do not consume alcohol at all. Consequently, taking the logarithm of the values would not be a good idea, due to the substantial number of non-drinkers.

One can distinguish that there is an outlier and that this person also comprises an outlier in regards to calorie intake. In *Figure 11* follows a normal-QQ-plot of the residuals as well as a plot of the studentized residuals related to the dietary model against y-estimates. Thereafter, in *Figure 12,* follows a 3D-plot of *alcohol intake* vs. leverage and residual size.

Stundetized residuals

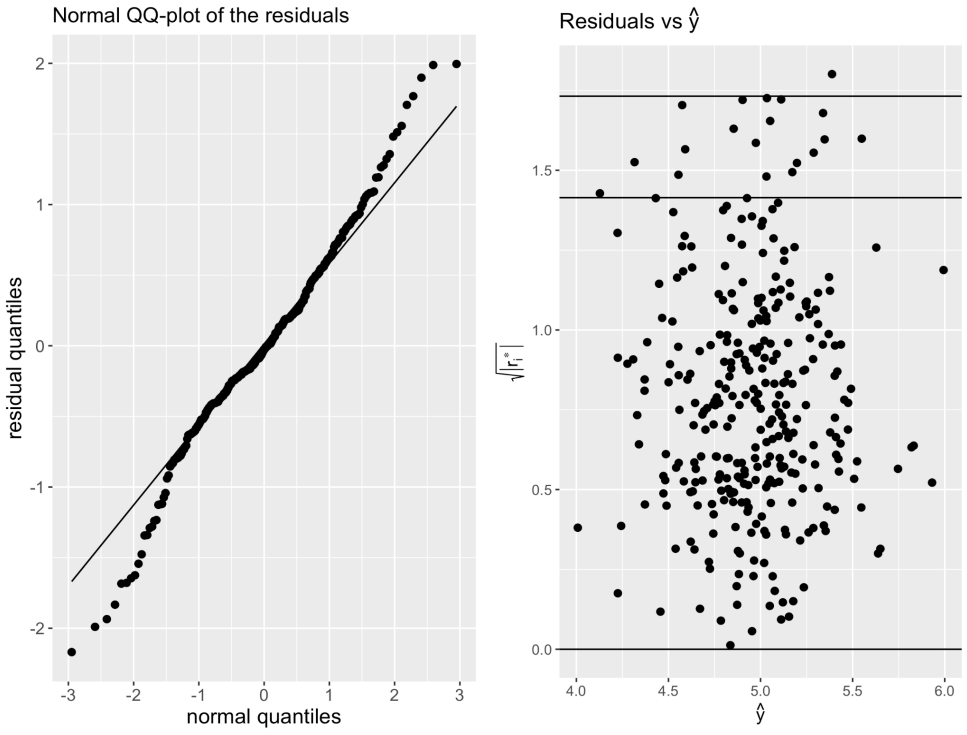

Figure 11. Normal QQ-plot of the residuals for the dietary model as well as the square root of the absolute values of the studentized residuals plotted against y-hat. The horizontal lines correspond to $\sqrt{|r_i^*|} = \sqrt{2}, \sqrt{|r_i^*|} = \sqrt{3}$.

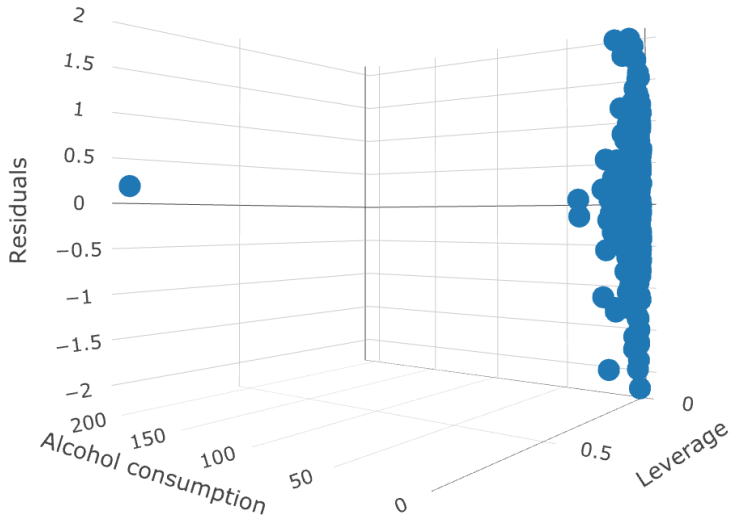

Figure 12. Residuals plotted against leverage and alcohol intake.

In order to further evaluate the outlier as well as see if any other data point has a large influence on the beta coefficients, Cook's distance was computed for all data points and plotted against the leverage in *Figure 13*.

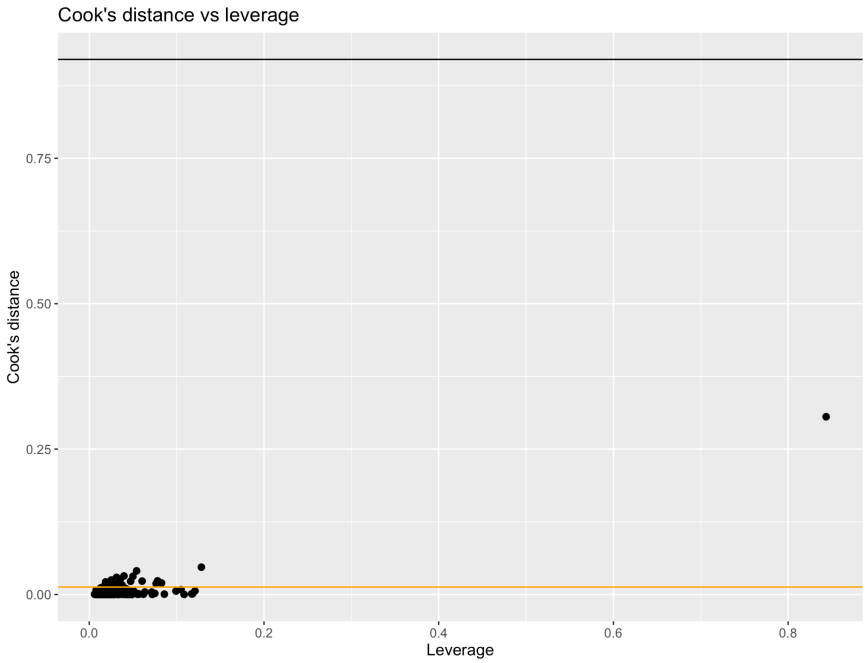

Figure 13. Cook's distance plotted against leverage. The orange horizontal line correspond to $F_{0.5,\,8,\,305}$ and the black one corresponds to $\frac{4}{n}$, where n is the number of data points.

In order to evaluate whether any categories' data points have unusually large influence on its own beta, DF-betas were computed and plotted against the corresponding x-variable. The resulting plots are shown in *Figure 14* and *15*.

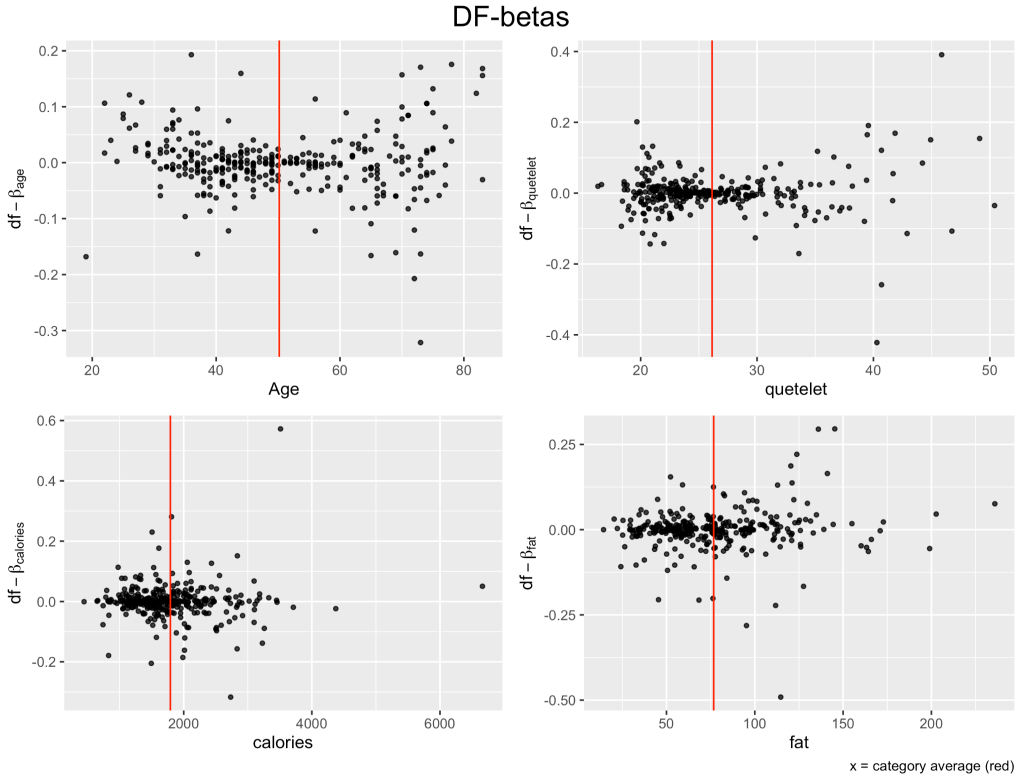

Figure 14. DF-betas plotted against their corresponding x-variable. Categories included are *age, quetelet, calories* and *fat*

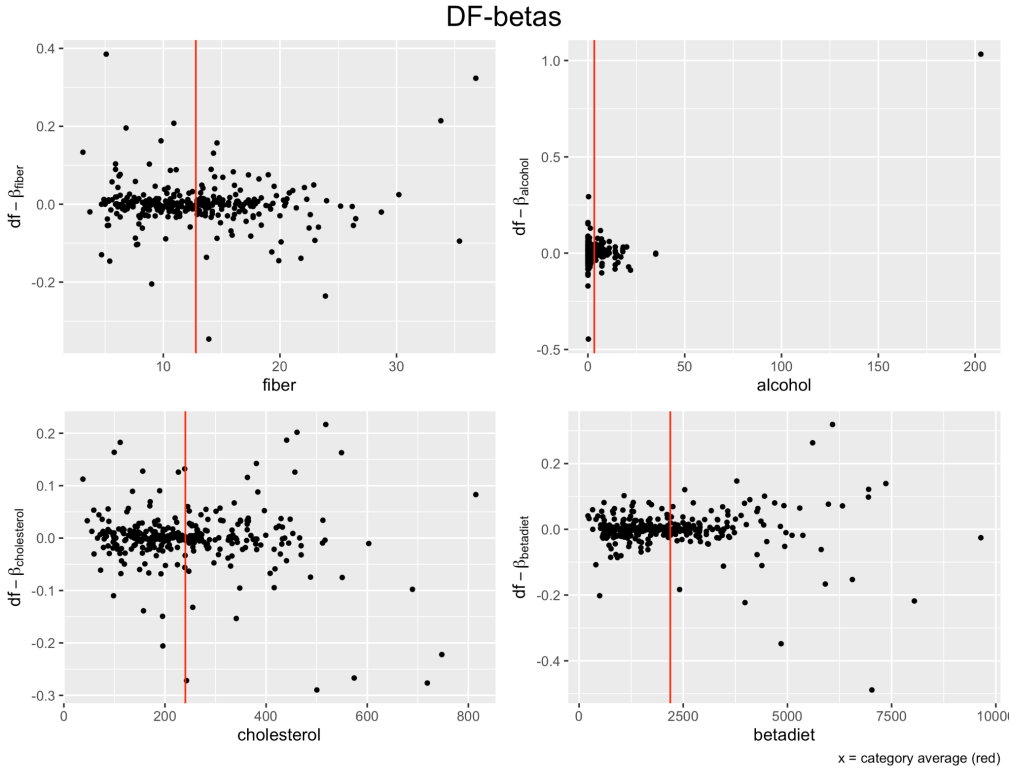

Figure 15. DF-betas plotted against their corresponding x-variable. Categories included are *fiber, alcohol, cholesterol* and *betadiet*

Backwards elimination

In order to improve the model containing the variables *age, quetelet, calories, fat, fiber, alcohol intake, cholesterol levels* and *dietary beta-carotene intake*, backwards elimination has been used. The stepping has been computed through R's built-in function *step* and the results are presented in *Table 17*.

Table 17. The Dietary model backwards elimination

Coefficient	**β-value**	e^{β}**-value**	**Lower bound** e^{β}**-value**	**Upper bound** e^{β}**-value**
β_0 *(Intercept)*	5.472456	238.0441998	150.6529284	376.1297018
β_1 *(Age)*	4.073043e-03	1.0040813	0.9987080	1.0094837
β_3 *(Quetelet)*	-3.189031e-02	0.9686128	0.9563650	0.9810175
β_3 *(Calories)*	-2.097821e-04	0.9997902	0.9996599	0.9999206
β_4 *(Fiber)*	3.451280e-02	1.0351153	1.0166025	1.0539652
β_5 *(Betadiet)*	5.876173e-05	1.0000588	0.9999996	1.0001179

Earlier a strong correlation between *fat intake* and *calories* as well as *fat* intake and *cholesterol* was detected, it can now be concluded that both these categories have been dropped.

Step AIC and BIC models

A full model, containing the union of x-variables in the dietary model and the background model. The difference between the two models is the stepping criterion either being based on AIC-values or BIC. The obtained values for beta, exponentiated betas and confidence intervals are presented below.

Table 18. The AIC model

Coefficient	**β-value**	e^{β}**-value**	**Lower bound** e^{β}**-value**	**Upper bound** e^{β}**-value**
β_0 *(Intercept)*	5.5563284264	258.8706268	160.9404769	416.3899765
β_1 *(Age)*	0.0057423124	1.0057588	1.0000909	1.0114589
β_3 *(Quetelet)*	-0.0343041386	0.9662776	0.9541716	0.9785372

β_3 *(Calories)*	-0.0001236991	0.9998763	0.9997407	1.0000119
β_4 *(Fiber)*	0.0275424008	1.0279252	1.0094849	1.0467024
β_5 *(Betadiet)*	0.0000527351	1.0000527	0.9999944	1.0001111
β_6 *(sexMale)*	-0.2979957613	0.7423045	0.5830549	0.9450499
β_7 (smokstatFormer)	-0.1060998858	0.8993348	0.7620177	1.0613967
β_{58} (smokstatCurrent)	-0.3368355380	0.7140263	0.5616209	0.9077894

Table 19. The BIC model

Coefficient	**β-value**	**e^{β}-value**	**Lower bound e^{β}-value**	**Upper bound e^{β}-value**
β_0 *(Intercept)*	5.6256456556	277.4513647	182.3564696	422.1361599
β_3 *(Quetelet)*	-0.0313659660	0.9691208	0.9567912	0.9816093
β_3 *(Calories)*	-0.0002279794	0.9997720	0.9996441	0.9999000
β_4 *(Fiber)*	0.0440066856	1.0449893	1.0280812	1.0621756

Determining the final model

A lot of the variability stems from the background variables, and not so much from the age variable itself. Among the various models, the best model according to $adj.\ R^2$ is the Step AIC model, which comprises the final model.

Table 20. R^2 as well as adjusted R^2 for the Age model, Background model, Dietary model, Step AIC model and Step BIC model.

Measure	**Age model**	**Background model**	**Dietary model**	**Step AIC model**	**Step BIC model**
R^2	0.01846886	0.1577452	0.1757349	0.2149411	0.1584833
$adj.\ R^2$	0.01532293	0.1384779	0.1623539	0.1943494	0.1503396

Analysis

The analysis in Part 2 reflects that the continuous BMI model have several advantages in regards to the discrete *BMI category* model. This is also strengthened by the results presented in *Table 15*, where it once again can be concluded, in regards to adjusted R-squared, AIC and BIC, that the

model based on the continuous BMI variable, *quetelet*, is better. Hence this model was elected to comprise the *Background model.*

In regards to the correlation between variables in the dataset, *Figure 7* illustrates the correlation between *fat intake* and *cholesterol* as well as *fat intake* and *calories*. The strong correlation will, if included, undermine the model, in the sense that its addition reduces the amount of degrees of freedom and does not make the model better at predicting. This is also strengthened by the results in *Table 17*, where one can see that backwards elimination of the Dietary model resulted in storing only one of the correlated variables.

Studying the leverages related to the Dietary model in *Figure 8*, one can see that the leverage related to the person consuming large amounts of alcohol is significantly large. It can also be concluded from *Figure 13,* illustrating Cook's distance, that this data point also has a large influence on all betas. If fitting a model for real world predictions, the models might benefit from removing this datapoint from the data set. In regards to the df-betas it can be concluded that the same data point heavily affects the slope of the beta for *alcohol intake.*

When performing AIC backwards elimination, starting with the *Dietary model* as a foundation, the final model contained the variables *age, quetelet, calories, fiber* and *betadiet*. It can then be seen that the influence of a subject's alcohol intake does not play a role anymore. Interesting future research would be to dig into whether alcohol would be included if the extreme data point was excluded.

There are some substantial differences between the model obtained through AIC stepwise selection and BIC stepwise selection, even though both of them started out as the same model. The former contains more variables of various types, whereas the latter only contains continuous variables. The *Step AIC model* outperforms the *Step BIC model* in regards to both R squared and adjusted R squared. However, the *Step BIC model* seems to outperform the background model in regards to both R squared and adjusted R squared, while containing fewer variables. This model can be of relevance if the user wants to predict the beta-carotene levels without collecting as much data as for the *Background model*. Furthermore, when predicting on a lot of datapoints, it can also be of interest for the user to not have to evaluate too many terms for each prediction, which lies in line with the *Step BIC model.* All models outperform the *Age model*.