Linear and Logistic Regression

Project 1: Linear Regression - Determinants of Plasma Beta-Carotene Levels

Isabelle Frodé, Olof Bengtsson and Erik Kindt

April 2021

Abstract

The aim of this project was to create a model that describes how the plasma concentration of beta-carotene depends on personal characteristics and dietary factors. The final model estimates the logarithm of the plasma beta-carotene levels and it can explain 24% of the variability of the log-transformed values. The reason for the log-transformation was to transform a skewed variable into a more normalized dataset with a linear relationship. The final model suggest that the plasma beta-carotene levels depends on all personal characteristics studied i.e. age, sex, smoking status, and BMI. Higher age is associated with an increase in beta-carotene levels. Higher BMI and smoking is associated with a decrease. The model further suggest that men generally have lower levels than women. The dietary factor expected to increase the levels of beta-carotene is fiber. Not using vitamins often and high intake of calories per day are expected to decrease the plasma beta-carotene levels. The variables fat, cholesterol, alcohol and dietary beta-carotene consumed per day was not included in the final model for different reasons.

Introduction

The aim of this project was to create a model that describes how the plasma concentration of beta-carotene of an individual depends on personal characteristics and dietary factors. Various studies suggests that there might be a link between low plasma concentration of beta-carotene and an increased risk of cancer. The goal of this project was to investigate whether a certain diet or lifestyle can affect the plasma beta-carotene levels. The study was based on a dataset containing 315 observations of patients' plasma beta-carotene levels on 13 variables. The variables are described below in table 1.

age	Age (years)
sex	Sex (1 = Male, 2 = Female)
smokstat	Smoking status (1 = Never, 2 = Former, 3 = Current Smoker)
quetelet	Quetelet (weight/height2kg/m2) a.k.a. BMI
bmicat	BMI category (1 = Underweight, 2 = Normal, 3 = Overweight, 4 = Obese)
vituse	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
calories	Number of calories consumed per day
fat	Grams of fat consumed per day
fiber	Grams of fiber consumed per day
alcohol	Number of alcoholic drinks consumed per week
cholesterol	Cholesterol consumed (mg per day)
betadiet	Dietary beta-carotene consumed (g per day)
betaplasma	Plasma beta-carotene (ng/ml)

Table 1: Variable description of the different variables presented.

Linear regression was used to fit a model to the observed data containing the relevant variables. The model indicates if the variable increase, decrease or does not have an impact on the plasma beta-carotene levels. The model could also be used to predict plasma beta-carotene levels of an individual.

Part 1 of the project investigated the relationship between plasma beta-carotene levels and age, i.e. the *Age* model was formulated. By performing residual analysis it was decided whether a linear or log-transformed linear model were to be used. In part 2, the *Age* model was extended with the other personal characteristic variables. Statistical tests was performed to determine if categorical or continuous variables were to be used and the *Background* model could then be formulated. In Part 3 a *Dietary* model was created using backward elimination. Outliers and potentially influential observation was analyzed. By using a stepwise procedure the *Background* and *Dietary* model could then be combined into a new model using first the *AIC*- and then the *BIC* criterion. All models were then evaluated and the best one chosen as the final model.

Part 1

Examining a linear and a log-transformed linear model

The first step was to determine if a linear or logarithmic model was to be used to describe the relationship between the levels of plasma beta-carotene and a varying age. The first step was to set up the models describing the linear relationship according to equation 1 and 2 below. One of the observations, no 257, had a plasma beta-carotene level equalling to zero and this would cause problems in the logarithmic case, hence it was omitted.

$$plasma = \beta_0 + \beta_{age} age \tag{1}$$

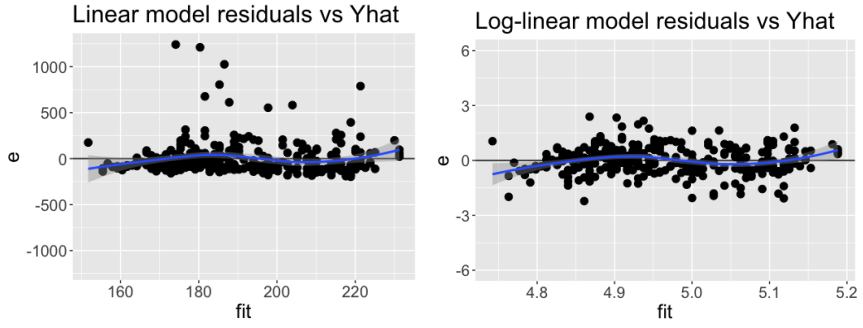
$$log(plasma) = \beta_0 + \beta_{age} age \tag{2}$$

When the models were set predictions were made and the residuals examined. The parameter estimates and confidence intervals are shown in table 2. One thing worth noting is that in the linear case, the idea of $\beta_1 = 0$ cannot be discarded at confidence level 95 %. This means that there is a chance that β_1 doesn't have any affect on the linear model.

Model	Parameter	Estimate	Confidence interval, 95 %
Linear	β_0	128.142	[55.495, 200.790]
	β_1	1.243	[-0.148, 2.633]
Log-linear	β_0	4.610	[4.315, 4.906]
	β_1	0.007	[0.001, 0.013]

Table 2: Parameter estimates and confidence intervals, linear and log-linear model

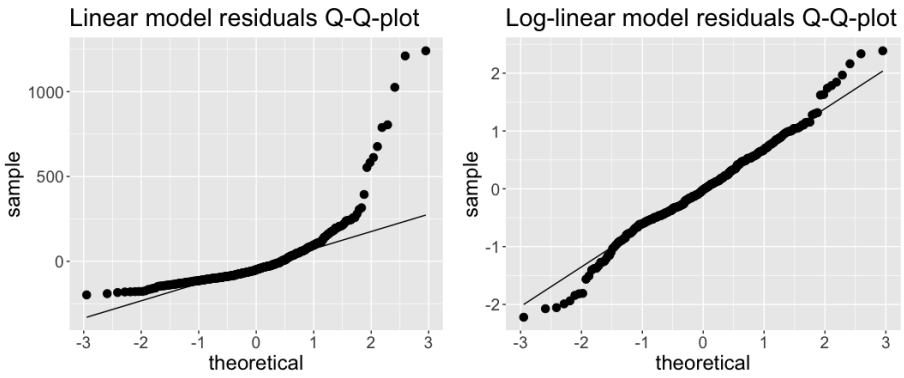
Then a basic residual analysis was conducted. The residuals for the two models were plotted against the predicted values Y-hat, shown in figure 1.



(a) Linear model **(b)** Log-linear model

Figure 1: Residuals plotted against the predicted Y-hat

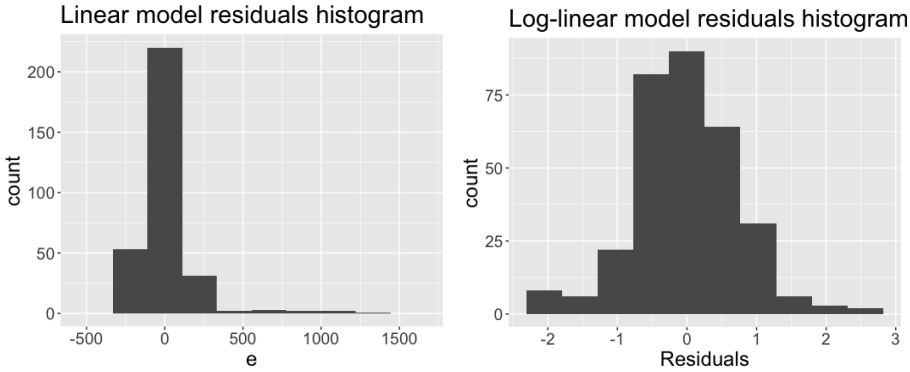
It's clear that the residual variance increases with Y-hat for the linear model. Especially note the scale in the figures above, the values are much larger for the linear case. The Q-Q plots are shown in figures 2 below.



(a) Linear model **(b)** Log-linear model

Figure 2: Q-Q plot of residuals

In the Q-Q plots it's clear that the residuals for the linear case is not normally distributed around zero, however for the log-linear case they are. In the linear case there exists non-linear trend while for the logarithmic the trend is linear.



(a) Linear model **(b)** Log-linear model

Figure 3: Histogram of the residuals

The histogram plots in figure 3 show that the linear model residuals are left skewed and the log-linear model show are more normally distributed around zero. Using these results the model chosen was the logarithmic one.

Change rate regarding increasing age

In figure 4 the estimated relationship between plasma beta-carotene concentration and age is shown, with confidence and prediction intervals on an 95 % level.

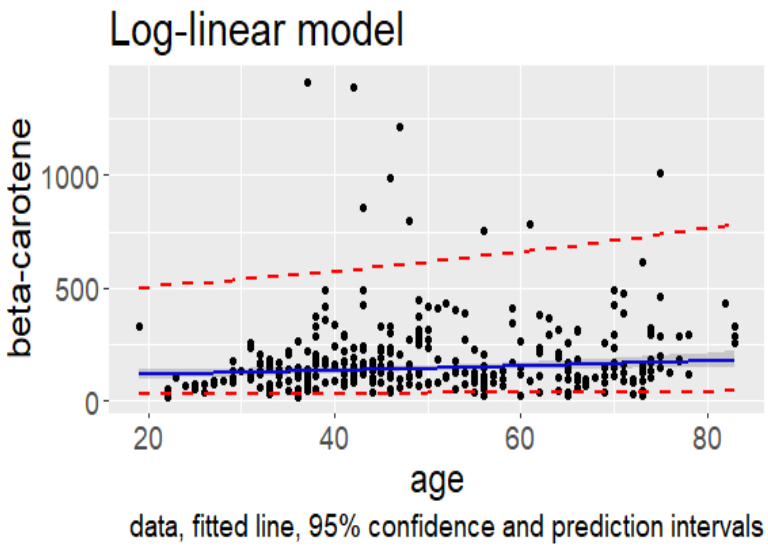


Figure 4: Plasma beta-carotene concentration plotted against age with fitted line from log-linear model (blue), confidence interval (dark grey) and prediction intervals (dotted red).

To see what happens to the plasma beta-carotene levels when the year is increased by one, the slope of the model was examined. This is the parameter β_1 from table 2. This gives that the average plasma beta-carotene level is increased every year with $e^{\beta_1} = e^{0.007} = 1.007$ ng/ml. This is the change rate and the confidence intervals are displayed in table 3 below.

Change rate	Estimate ng/ml	Confidence interval
e^{β_1}	1.007	[1.001, 1.013]

Table 3: The change rate with a 95 % confidence interval.

Next part to be examined was that if a yearly change for 30 year-olds and 70 year-olds results in the same difference in plasma beta-carotene levels. If they are the same this can be described as equation 3.

$$\Delta_{31-30} = \Delta_{71-70} \tag{3}$$

This was calculated for both cases according to equations 4 and 5 below. In equation 4, $t_1 = 31$ and $t_0 = 30$. In equation 5, $t_1 = 71$ and $t_0 = 70$.

$$\Delta_{31-30} = e^{\beta_0+\beta_1 t_1} - e^{\beta_0+\beta_1 t_0} = e^{4.610+0.001\cdot 31} - e^{4.610+0.001\cdot 30} = 0.866 \quad (4)$$

$$\Delta_{71-70} = e^{\beta_0+\beta_1 t_1} - e^{\beta_0+\beta_1 t_0} = e^{4.610+0.001\cdot 71} - e^{4.610+0.001\cdot 70} = 1.144 \quad (5)$$

It's shown that equation 3 doesn't hold and this is expected since it's a logarithmic model. If the model would have been linear equation 3 would hold, but because of the logarithmic behaviour of the model $\Delta_{31-30} \neq \Delta_{71-70}$.

The concentration of plasma beta-carotene for a 30- and a 70-year old.

Using the model and examining the concentration of plasma beta-carotene concentration for different age, one can conclude that the prediction interval was wider for a 70 year-old than a 30 year-old, as shown in table 4. This would be reasonable since the 95 % prediction interval is expected to contain 95 % of the predictions. The data has a wider spread of plasma beta-carotene levels for people with higher age, as shown in figure 4, thus the prediction interval is larger.

Age	Prediction Interval	Interval width	$e^{\text{Prediction interval}}$	$e^{\text{Interval width}}$
30	[3.35, 6.29]	2.93	[28.50, 539.2]	510.7
70	[3.63, 6.56]	2.93	[37.71, 706.3]	668.6

Table 4: Prediction interval for a 30- and a 70-year old

Part 2

Extended model

In this part an extended model where plasma beta-carotene depended on personal characteristics, i.e. background variables sex, BMI, smoking status and age. Sex and smoking was categorical variables and BMI had both categorical and continuous representations whereas age was continuous as before.

When deciding on which category was to be used as a reference for each variable, the frequency tables 5-7 below was considered.

Sex	No of observations
Male	42
Female	272

Table 5: Frequency table over sex.

Smoking status	No of observations
Never	156
Former	115
Current smoker	43

Table 6: Frequency table over smoking status.

BMI category	No of observations
Underweight	4
Normal	160
Overweight	89
Obese	61

Table 7: Frequency table over BMI category.

As tables 5-7 suggests there was a high number of observations for females, people who have never smoked and a normal BMI. Since the reference is usually chosen as the largest category, these are probably the ones which are best to use.

As a first model plasma beta-carotete was fitted to only depend on the categorical BMI variable with *underweight* as reference category. The β-estimates and standard errors obtained from this model are presented below in tabular 8.

Reference: Underweight	Estimate	Standard error
β_0	5.360	0.3615
β_{normal}	-0.232	0.3660
$\beta_{overweight}$	-0.487	0.3695
β_{obese}	-0.742	0.3732

Table 8: β-estimates and standard errors for model where plasma beta-carotene depended on the categorical variable of BMI, using *underweight* as reference category.

A model was then fitted by using the largest category, *normal weight* as reference. The β-estimates and standard errors obtained from this new model are presented in tabular 9 below.

Reference: Normal weight	Estimate	Standard error
β_0	5.128	0.05716
$\beta_{underweight}$	0.232	0.36598
$\beta_{overweight}$	-0.255	0.09560
β_{obese}	-0.510	0.10879

Table 9: β-estimates and standard errors for model where plasma beta-carotene depended on the categorical variable of BMI, using *normal weight* as reference category.

Table 8 and table 9 show that the standard errors are lower for parameters β_0, $\beta_{overweight}$, β_{obese} using the normal weight category as reference. The only parameter with a high standard error with normal as reference is the $\beta_{underweight}$ parameter. Smaller standard errors means that the model fit the observed data better which implicates that the normal weight as reference is the best choice, as expected.

The difference in standard error for the two choices of reference category can be explained by table **??**. There are only 4 observations in the *Underweight* category which corresponds to 1.3 % of all observations, whilst the normal category corresponds to 51 % of all observations. The reference category is used as reference to the other categories, hence it is important that it is representative.

Expanding the model

The two categorical variables *smoking* and *sex* were then added to the model together with age, which was used in the first part. Just as for *BMI* the reference categories for the new variables were chosen as their largest category, which was *never smoked* and *female*. The new model then had the form given in equation 6. For simplicity the variables describing smoking status were changed according to *Former smoker* $\rightarrow$ *smokstat2* and *Current smoker* $\rightarrow$ *smokstat3* as in table 1. This is also going to be the case for all categorical variables introduced later.

$$\begin{aligned} log(\text{beta-carotene}) = \beta_0 &+ \beta_{age} age + \beta_{underweight} x_{underweight} \\ &+ \beta_{overweight} x_{overweight} + \beta_{obese} x_{obese} \\ &+ \beta_{smokstat2} x_{smokstat2} + \beta_{smokstat3} x_{smokstat3} \\ &+ \beta_{male} x_{male} \end{aligned} \tag{6}$$

Here $x_{category} = 1$ if the data was a part of that category and 0 for all other categories. The β-estimates and corresponding confidence intervals obtained from this model are given below in table 10

Parameter	β-estimate	Confidence interval	e^{β}-estimate	Confidence interval
β_0	4.896	[4.583, 5.208]	133.7	[97.8, 182.8]
β_{age}	0.00746	[0.00183, 0.0131]	1.007	[1.0018, 1.013]
$\beta_{underweight}$	0.307	[-0.392, 1.0050]	1.359	[0.676, 2.732]
$\beta_{overweight}$	-0.217	[-0.400, -0.0351]	0.805	[0.671, 0.965]
β_{obese}	-0.548	[-0.755, -0.341]	0.578	[0.470, 0.711]
β_{male}	-0.339	[-0.579, -0.0989]	0.712	[0.560, 0.906]
$\beta_{smokstat2}$	-0.108	[-0.280, 0.0632]	0.897	[0.756, 1.065]
$\beta_{smokstat3}$	-0.449	[-0.695, -0.204]	0.638	[0.499, 0.815]

Table 10: β-estimates, e^{β}-estimates and their confidence intervals for model where plasma beta-carotene depended on age and the categorical variables of BMI, sex and smoking.

By looking as table 10 it is clear that all parameters are not significant, where these parameters are the ones corresponding to being underweight and former smoker. If the confidence interval of a variable contains 0, it should be considered insignificant. Since these are parameters corresponding to categorical variables this probably means that the variable itself is significant, but that the plasma beta-carotene level does not vary between said the reference category and said categories.

Testing the model

Tests were then performed to see whether the additional variables had made a significant improvement in the model and were all done for $\alpha = 0.05$. First a global F-test was conducted to see if this new model was better than a model with only a constant term. Next was a partial F-test to compare the new model to the model using only age as a variable.

Further tests were then done for every individual category to see if they had a significant impact on the model. For the binary variable sex and the continuous variable age t-test were done to examine the impact of one single parameter, whereas for the remaining two variables partial F-tests had to be conducted to measure the importance of several parameters. The last test was to see whether the category *underweight* for BMI was significantly different from the reference category, which was done with a t-test.

The null hypotheses and results from all test above are shown below in table 11.

H_0	Type	Statistic	Distr.	P-value	Concl.
$\beta_i = 0,\ i \neq 0$	Global F	$F = 8.187$	$F(7, 306)$	$3.786 \cdot 10^{-9}$	Reject H_0
$\beta_i = 0,\ i \neq 0, age$	Partial F	$F = 8.4334$	$F(6, 306)$	$1.763 \cdot 10^{-8}$	Reject H_0
$\beta_{age} = 0$	t-test	$t = 2.608$	$t(306)$	0.00956	Reject H_0
$\beta_{male} = 0$	t-test	$t = 2.778$	$t(306)$	0.00580	Reject H_0
$\beta_{smokstat2} = \beta_{smokstat3} = 0$	Partial F	$F = 6.499$	$F(2, 306)$	0.001721	Reject H_0
$\beta_{underweight} = \beta_{overweight} = \beta_{obese} = 0$	Partial F	$F = 9.8373$	$F(3, 306)$	$2.66 \cdot 10^{-6}$	Reject H_0
$\beta_{underweight} = 0$	t-test	$t = 0.864$	$t(306)$	0.388193	Keep H_0

Table 11: Null hypothesis H_0, test type, test statistic, the distribution of the test statistic under H_0, P-value and conclusion for all the test conducted above.

As seen in the table above all test showed that the null hypothesis could be rejected apart from the test where the importance of $\beta_{underweight}$ was tested. This shows that all variables are significant and makes the model better, but

that *underweight* does not seem to be significantly different from the reference category *normal weight* and might be abundant.

Investigating the model

Figure 5 show how the logarithm of the predicted plasma beta-carotene levels against age and the corresponding confidence and prediction intervals for every individual categorical variable.

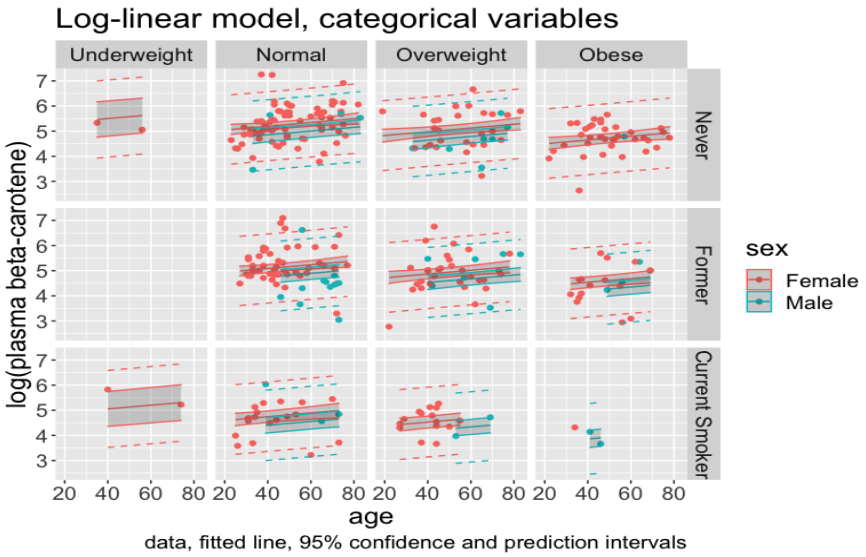


Figure 5: Logarithm of the prediction and confidence- and prediction intervals for the different categorical variables

As can be seen in figure 5 above there are very few samples for people that are underweight, and there are in particular none for former smokers that are underweight and male. The prediction and both confidence and prediction intervals for this type of person using the same model as before are shown below in table 12. Here like in figure 5 the logarithm is given.

Prediction	Confidence interval	Prediction interval
5.128	[4.386, 5.870]	[3.574, 6.682]

Table 12: Prediction and confidence and prediction intervals for the model $log(\text{betaplasma}) \sim \text{age} + \text{BMI}_{cat} + \text{smoking} + \text{sex}$

By looking at table 12 it can be seen that the intervals are similar to the other ones for an underweight person which are quite large, especially compared to the other variable and categories. Furthermore, when taking the e^{β}-estimate of these values the interval gets even bigger at [80.32, 354.4] for the confidence

interval. Since there is no data for this particular set of variables it is no surprise that the margin of error is large and because of this inaccuracy the prediction cannot be trusted.

When looking at how the logarithm of plasma beta-carotene varies as BMI increases, see figure 5 one can see that it decreases when going from underweight to obese. This is in agreement with table 10 where all statistically significant β-estimates are negative for the categorical BMI variable.

Continuous BMI

A new model was fitted using a continuous BMI variable *quetelet* instead of the categorical one, giving the model outline

$$\begin{aligned} log(\text{betaplasma}) = \beta_0 + \beta_{age} age + \beta_{quetelet} quetelet + \beta_{smokstat2} x_{smokstat2} \\ + \beta_{smokstat3} x_{smokstat3} + \beta_{male} x_{male} \end{aligned} \tag{7}$$

The estimates and intervals from this model are given in table 13 below.

Parameter	β-estimate	Confidence interval	e^{β}-estimate	Confidence interval
β_0	5.705	[5.241, 6.169]	300.3	[188.8, 477.5]
β_{age}	0.00745	[0.00188, 0.0130]	1.007	[1.002, 1.013]
$\beta_{quetelet}$	-0.0371	[-0.0499, -0.0242]	0.964	[0.951, 0.976]
β_{male}	-0.344	[-0.582, -0.106]	0.709	[0.559, 0.900]
$\beta_{smokstat2}$	-0.115	[-0.284, 0.0542]	0.891	[0.753, 1.056]
$\beta_{smokstat3}$	-0.451	[-0.691, -0.212]	0.637	[0.501, 0.809]

Table 13: β-estimates, e^{β}-estimates and their confidence intervals for model where plasma beta-carotene depended on continuous age and BMI and the categorical variables of sex and smoking.

Comparing this table to table 10 the only parameter differing significantly is β_0 which is reasonable since we have removed the categorical nature in the BMI variable so that there are no more constant plateaus for the different categories. We also see that the estimate of the parameter describing BMI is strictly negative, which further shows that plasma beta-carotene decreases when BMI increases. Using both this model and the model with categorical BMI a prediction of the average plasma beta-carotene level was made for both male and female 30-year old former smoker with a BMI of 20 (normal). The estimates and confidence intervals from this are shows below in table 14.

Sex	Model	Prediction	Confidence interval
Male	Categorical	106.9	[79.55, 143.8]
	Continuous	113.1	[84.22, 151.9]
Female	Categorical	150.1	[124.9, 180.5]
	Continuous	159.5	[132.4, 192.1]

Table 14: Estimation and confidence intervals for predicting the plasma beta-carotene level for both a male and female 30-year old former smoker with a BMI of 20.

The results above show that the prediction produced by both models are quite close to each other where the continuous is a little bit larger. This might be because the value for *quetelet* was 20 which is in the lower region of the interval of normal BMI in the categorical case. Since plasma beta-carotene is thought to decrease as BMI increases the prediction is probably going to be too low for lower range values and too high for higher range values. Furthermore the prediction intervals seem similar in range, but since models predict the logarithm of the plasma beta-carotene level the prediction is made by raising e to its power. This means that the confidence of the β-parameters really are smaller for females compared to males, which is because there are much more samples for females than males.

The relative difference for female and male were then calculated for both models when comparing the results in table 14 to an obese person with BMI = 32. In order for the difference to be additive it was chosen to calculate the difference of the logarithms. For the model using the categorical variable of BMI the difference between the two predictions was simply β_{obese} giving the relative difference $\beta_{obese}/(32-20) = \beta_{obese}/12$. For the continuous case the difference was $32\beta_{quetelet} - 20\beta_{quetelet} = 12\beta_{quetelet}$ giving the relative difference $12\beta_{quetelet}/(32-20) = \beta_{quetelet}$. The estimates of this are shown in table 15 below.

Model	Estimate	Confidence interval
Categorical	-0.0456	[-0.0629, -0.0284]
Continuous	-0.0371	[-0.0499, -0.0242]

Table 15: Relative difference of predictions made for BMI = 32 and BMI = 20 for both models.

Table 15 shows that the relative differences are not the same, but quite close. The reason for this could again be due to the categorical approximation where the relative difference for a normal BMI of value 18.5 would decrease the estimate to -0.0406 for the categorical case. Because of this the relative differences seem reasonable and sufficiently close to each other.

Model with both categorical and continuous BMI

A model was fitted using both the categorical and continuous variables for BMI as well as the other background variables, resulting in the estimates for the BMI-parameters shown in table 16 below

Parameter	β-estimate	Confidence interval
$\beta_{quetelet}$	-0.0265	[-0.0559, 0.00287]
$\beta_{underweight}$	0.186	[-0.523, 0.895]
$\beta_{overweight}$	-0.0814	[-0.317, 0.155]
β_{obese}	-0.174	[-0.637, 0.288]

Table 16: Estimates and confidence intervals of the parameters modelling BMI in a model where the logarithm of plasma beta-carotene was fitted against sex, smoking, age and continuous and categorical BMI.

where the other parameter estimates were close to the ones they had in the other models. As table 16 suggests all parameters are statistically insignificant, which is not a surprise since they all aim to describe the same thing and therefore are highly correlated. A t-test was also made to measure the impact of *quetelet* when adding the variable to the categorical model, and unsurprisingly the test showed that the model was better with just the categorical variable.

Part 3

Continuous or discrete BMI

When determining if using a continuous or discrete BMI variable, the errors of the two different models were examined. Since the models were not nested a partial F-test could not be used. The better model was chosen according to the *principle of parsimony*, meaning that the model chosen will be one with small residuals and as few parameters as possible. To compare the two models, the R^2-values were looked at. Since the two models have different numbers of variables, the adjusted R^2-values were used, which are presented in table 17 below.

Model	Discrete BMI	Continuous BMI
Adjusted R^2	0.1385	0.1506

Table 17: The adjusted R^2 values of models using discrete and continuous BMI respectively.

The model chosen is the one with the highest R^2_{adj} value since it states how much of the variance can be described by the model, adjusted with the number of parameters considered in the model. Since the model with the continuous BMI variable has the highest R^2_{adj} value this was the model that was chosen. This model is now referred to as the *Background* model.

Correlation between variables

The next step was to consider the pairwise correlation between the variables. A new category of variables called *dietary* variables were also introduced. There were some pairs of variables that had a correlation stronger than 0.7. A strong correlation between fat and calories was found. There was also a strong correlation larger than 0.7 between fat and cholesterol. The calculated correlation is presented in table 18 and visualized in figure 6.

Variables	Correlation
$\rho_{fat,calories}$	0.8709
$\rho_{fat,cholesterol}$	0.7041

Table 18: The two pairs with a correlation larger than 0.7.

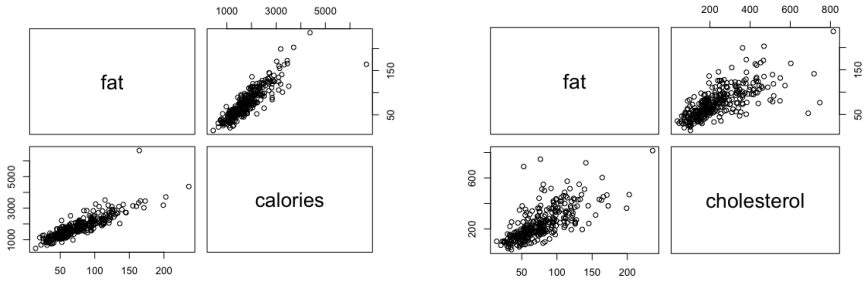


Figure 6: The correlation between fat-calories and fat-cholesterol.

One thing that was noted was that in alcohol, calories and cholesterol there was some outliers that might be problematic for the modeling. The outlier in the alcohol variable is the most prominent one, shown in figure 7.

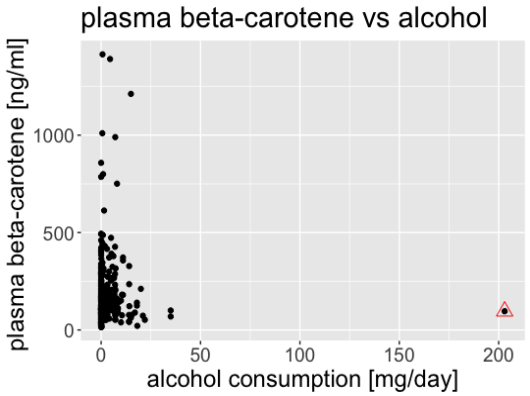


Figure 7: Plasma beta-carotene levels plotted against alcohol consumption. Outlier is marked with a upwards red triangle.

When choosing which category in vitamin use to be used as the reference category, the frequency table for the categorical variable *vitamin use (vituse)* presented in table 19 was observed.

Vitamin use	Yes, fairly often	Yes, not often	No
Observations	121	82	111

Table 19: Frequency table over the variable vitamin use.

The category with the largest number of observations, "Yes, fairly often", will be used as the reference category.

Fitting a model with all dietary variables

A model using all dietary variables was then fitted:

$$\begin{aligned} log(\text{beta-carotene}) = \beta_0 &+ \beta_{vituse2}x_{vituse2} + \beta_{vituse3}x_{vituse3} \\ &+ \beta_{calories}calories + \beta_{fat}\text{fat} + \beta_{fiber}fiber \\ &+ \beta_{alcohol}alcohol + \beta_{cholesterol}cholesterol \\ &+ \beta_{betadiet}betadiet \end{aligned} \tag{8}$$

When plotting the leverage, figure 8, it was obvious that there was a problem with one observation. This was the observation with the large alcohol consumption.

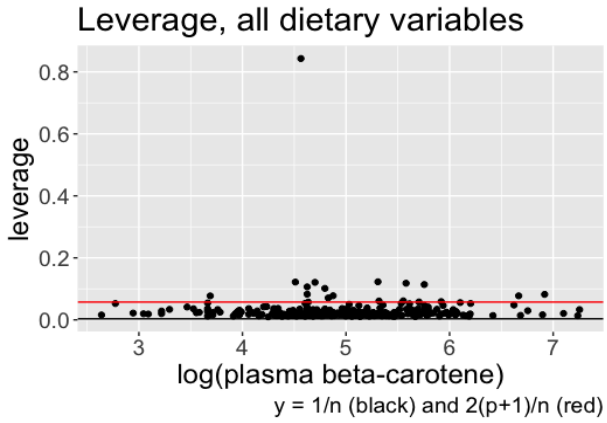


Figure 8: Leverage against log(plasma beta-carotene).

When fitting a model without the alcohol variable there is still an alarmingly high leverage for the same observation as seen in figure 8. This could mean that the same observation still has large variable values in other categories as well. The logarithm of the alcohol intake might improve the model, but this would cause problems for the 110 out of 314 observations that doesn't consume any alcohol.

When plotting alcohol against other variables it was clear that the observation with large alcohol consumption also had alarmingly large values in other categories as well. For example the same observation also had a very large value in the calories category, see figure 9, and large values in fat and cholesterol, tough it wasn't any extreme levels in the two latter.

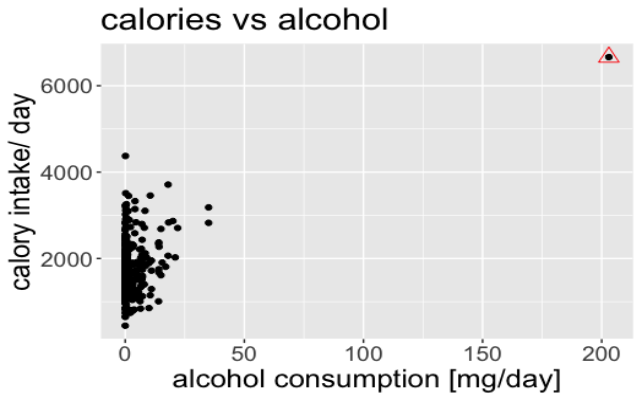


Figure 9: Calories intake plotted against alcohol consumption, observation with largest leverage is marked with a triangle

Studentized residuals

The studentized residuals was then calculated and presented in figure 10, plotted against the fitted values.

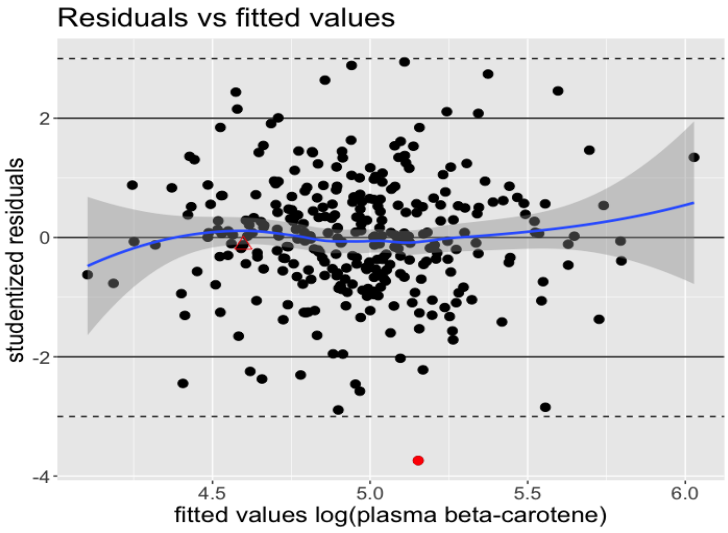
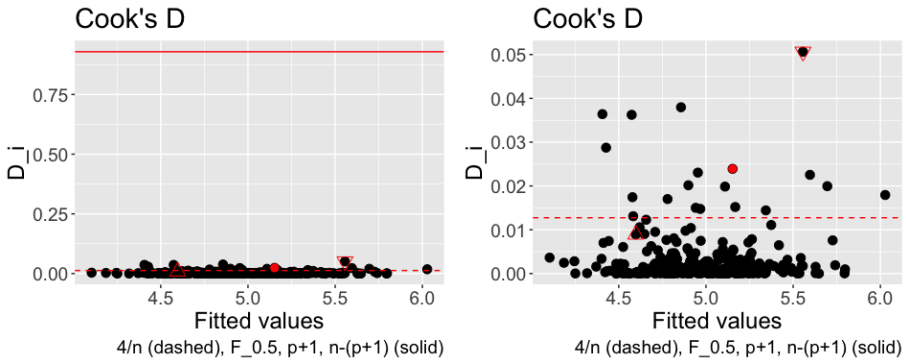


Figure 10: The studentized residuals plotted against the fitted values. The largest residual is marked as a red circle and the observation with the highest leverage with a red triangle.

Since there only is one apparent outlier in figure 10, the model used might be a good fit. In this model, the alcohol consumption was included and even though there was one extreme value, the model seems to handle it well.

Cook's distance

The next task was to inspect the Cook's distance for the full dietary model. The Cook's distance plotted against the fitted values is shown in figure 11. Marked in the figure is the 0.5-quantile of the F-distribution, visualized in the figure by a red solid line. If the Cook's distance exceeds this value it can be considered to have a large influence. If an observation has a Cooks distance smaller than the dashed line in figure 11, i.e. the number of observations divided by 4, the observation can be considered to not have a large influence.



(a) Scale $y = [0, 1]$ **(b)** Scale $y = [0, 0.05]$

Figure 11: The Cook's distance for each observation. The circle marks the observation with the largest residual, the upwards triangle marks the largest leverage and downwards triangle the largest Cook's distance. The solid line is the Cook limit.

Figure 11 show that there is not one fitted value that's close to exceeding the solid line, hence there are no value that can be considered to have an extreme influence. There are however some observations that exceeds the dashed line which would imply that those observations potentially have an influence on the model. Figure 11 show that the high leverage observation does not have an influence. The observation with the largest studentized residual may have an influence according to the plot. The observation with the largest Cook's distance is observation number 36, female overweight current smoker with high cholesterol intake per day and very low plasma beta-carotene levels.

DFBETAS was calculated for different β-parameters to determine if any of the observations with highest leverage, largest studentized residual or largest Cook's distance had an influence on the model parameters. The results are presented in table 20.

Problematic observation	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Largest Cook's D	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Highest leverage	No	No	No	No	No	No	Yes	No	No
Largest residual	No	Yes	No	No	No	No	No	No	No

Table 20: Influence of largest residuals (Yes/No) for problematic observations

It is clear that the observation with the largest Cook's distance influences the model the most, having an influence on 6 out of 9 parameters. The plots of the DFBETAS plotted against the fitted values can be found in figure 12-14. Observations with DFBETAS values exceeding the dashed line can be considered to cause a change in the particular parameter, having a large influence on that parameter. Values exceeding the solid line have a very large influence. The observation with the largest Cook's distance, largest residual and highest leverage

are marked with an downward triangle, circle and upward triangle respectively in the plots 12-14.

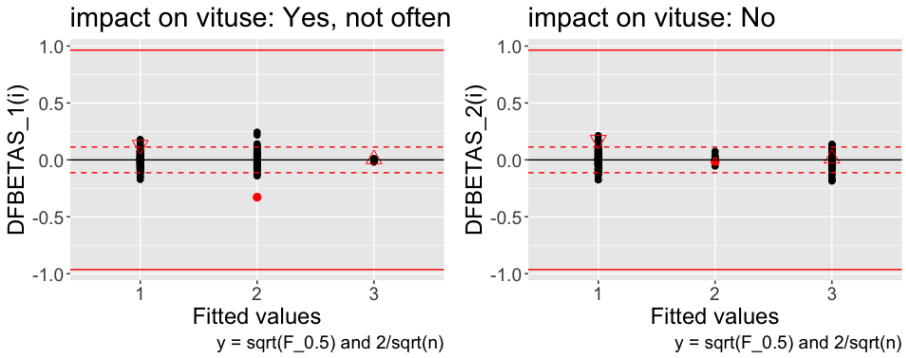


Figure 12: DFBETAS against fitted values for β_1 and β_2.

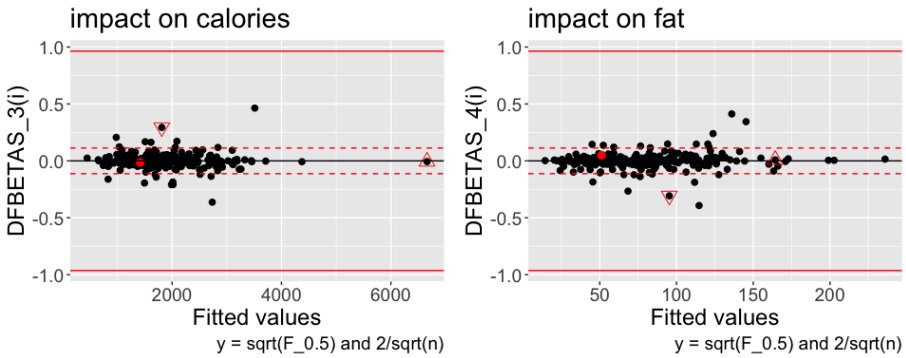


Figure 13: DFBETAS against fitted values for β_3 and β_4.

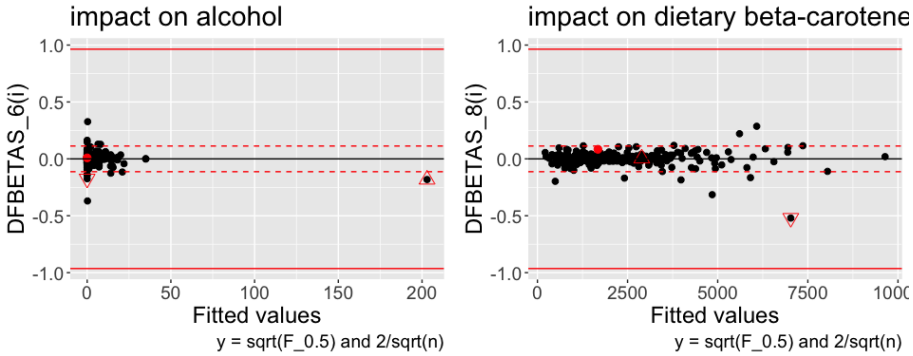


Figure 14: DFBETAS against fitted values for β_6 and β_8.

Reducing the *Dietary* model

Backward elimination was then used to determine which variables should be in the *Dietary* model. We started from a large model containing all dietary factors i.e. *vitamin use, calories, fat, fiber, cholesterol, alcohol* and *dietary beta-carotene*. In each step, the variable that would cause the largest decrease in Bayesian information criterion (BIC) was found and removed from the *Dietary* model. BIC is a criterion partly based on the likelihood function. BIC is used to choose between two or more alternative models. When using BIC, only the variables that had a significant contribution would be left in the final model, hence we get a more reliable and "true" model than we would get using the Akaike information criterion (AIC).

The final *Dietary* model obtained using BIC is:

$$\begin{aligned} log(\text{beta-carotene}) = \beta_0 + \beta_{vituse2}x_{vituse2} + \beta_{vituse3}x_{vituse3} \\ + \beta_{calories}calories + \beta_{fiber}fiber \end{aligned} \tag{9}$$

which means that *fat, alcohol, cholesterol* and *dietary beta-carotene* has been removed from the model. The β-estimates, e^{β}-estimates and their confidence intervals for *Dietary* model can be found in table 21.

Parameter	β-estimate	Confidence interval	e^{β}-estimate	Confidence interval
β_0	4.967	[4.702, 5.231]	143.6	[110.180, 187.030]
$\beta_{vituse2}$	-0.102	[0.298, 0.095]	0.903	[0.742, 1.100]
$\beta_{vituse3}$	-0.377	[-0.558, -0.196]	0.686	[0.572, 0.822]
$\beta_{calories}$	0.000	[0.000, 0.000]	1.000	[1.000, 1.000]
β_{fiber}	0.044	[0.028, 0.061]	1.045	[1.028, 1.063]

Table 21: β-estimates, e^{β}-estimates and their confidence intervals for *Dietary* model

Note that from the variable pairs found with strong correlations in table 18 there is only one variable left, i.e. *calories* which is expected. It is obvious that adding the other highly correlated variables would not bring any additional value to the model.

Combining the *Background* and the *Dietary* model

In this section the *Background* and *Dietary* model were combined using a step-wise procedure, with the *Dietary* model as a starting point. The null model was chosen as the smallest model allowed:

$$log(\text{beta-carotene}) = \beta_0 \tag{10}$$

and the full model was a model containing all possible variables from the *Background* and *Dietary* model:

$$
\begin{aligned}
log(\text{beta-carotene}) = \beta_0 &+ \beta_{age} age + \beta_{smokstat2} x_{smokstat2} \\
&+ \beta_{smokstat3} x_{smokstat3} + \beta_{male} x_{male} \\
&+ \beta_{quetelet} quetelet + \beta_{vituse2} x_{vituse2} \\
&+ \beta_{vituse3} x_{vituse3} + \beta_{calories} calories \\
&+ \beta_{fiber} fiber
\end{aligned} \tag{11}
$$

The stepwise procedure was performed using the AIC and the BIC criterion, outputting two different models which we refer to as the *AIC* model and the *BIC* model. The β-estimates, e^{β}-estimates and their confidence intervals for the *AIC* and *BIC* model can be found in table 22 and table 23.

Parameter	β-estimate		Confidence interval	
Model	AIC	BIC	AIC	BIC
β_0	5.505	5.757	[4.971, 6.038]	[5.339, 6.174]
β_{age}	0.006		[0.001, 0.012]	
β_{male}	-0.248		[-0.488, -0.007]	
$\beta_{quetelet}$	-0.032	-0.030	[-0.045, -0.020]	[-0.043, -0.018]
$\beta_{smokstat2}$	-0.076		[-0.240, 0.088]	
$\beta_{smokstat3}$	-0.290		[-0.529, -0.051]	
$\beta_{vituse2}$	-0.029	-0.069	[-0.219, 0.161]	[-0.260, 0.121]
$\beta_{vituse3}$	-0.292	-0.349	[-0.469, -0.116]	[-0.525, -0.174]
$\beta_{calories}$	-0.0001	-0.0002	[-0.0003, 0.0000]	[-0.0003, -0.0001]
β_{fiber}	0.032	0.041	[0.016, 0.049]	[0.025, 0.057]

Table 22: β-estimates and their confidence intervals for the *AIC* and *BIC* model

Parameter	e^{β}-estimate		Confidence interval	
Model	AIC	BIC	AIC	BIC
β_0	245.843	316.324	[144.208, 419.108]	[208.340, 480.277]
β_{age}	1.006		[1.001, 1.012]	
β_{male}	0.781		[0.614, 0.993]	
$\beta_{quetelet}$	0.968	0.970	[0.956, 0.980]	[0.958, 0.983]
$\beta_{smokstat2}$	0.927		[0.787, 1.092]	
$\beta_{smokstat3}$	0.748		[0.589, 0.950]	
$\beta_{vituse2}$	0.971	0.933	[0.803, 1.174]	[0.771, 1.129]
$\beta_{vituse3}$	0.747	0.705	[0.626, 0.891]	[0.592, 0.841]
$\beta_{calories}$	0.9999	0.999	[0.9998, 1.0000]	[0.9997, 0.9999]
β_{fiber}	1.033	1.042	[1.016, 1.050]	[1.025, 1.058]

Table 23: e^{β}-estimates and their confidence intervals for the *AIC* and *BIC* model

The AIC model is presented below:

$$
\begin{aligned}
log(\text{beta-carotene}) = \beta_0 &+ \beta_{age} age + \beta_{male} x_{male} + \beta_{quetelet} quetelet \\
&+ \beta_{smokstat2} x_{smokstat2} + \beta_{smokstat3} x_{smokstat3} \\
&+ \beta_{vituse2} x_{vituse2} + \beta_{vituse3} x_{vituse3} \\
&+ \beta_{calories} calories + \beta_{fiber} fiber
\end{aligned} \tag{12}
$$

and the BIC model:

$$
\begin{aligned}
log(\text{beta-carotene}) = \beta_0 &+ \beta_{quetelet} quetelet \\
&+ \beta_{vituse2} x_{vituse2} + \beta_{vituse3} x_{vituse3} \\
&+ \beta_{calories} calories + \beta_{fiber} fiber
\end{aligned} \tag{13}
$$

Table 22 and table 23 show that all common parameters in the *BIC* and the *AIC* model are similar except from the intercept parameter β_0. In general the confidence intervals are a bit smaller for the *BIC* model, as expected. The number of parameters for the *AIC* model will however improve the accuracy of predictions.

Deciding on the final model

To decide which final model was to be used, the R^2 and the adjusted R^2_{adj} was calculated. Since the models have different numbers of variables, the R^2_{adj} was considered as the decider. The statistics for each model are presented in table 24.

Model	R^2	R^2_{adj}
age	0.0185	0.0153
background	0.1642	0.1506
dietary	0.1440	0.1330
step AIC	0.2374	0.2148
step BIC	0.2016	0.1886

Table 24: The R-statistics for the models.

Since the *step AIC* model had the highest R^2_{adj} value, this is the model considered to be the best one out of these models, hence it is chosen as the final model. The R^2 value tell us that the model can explain 24 % of the variability.

Conclusion

From the linear regression data analysis we can conclude that the best model is the *AIC* model containing the personal characteristic variables age, sex, continuous BMI and smoking status. The dietary factors in the model is the variables

vitamin use, calories and fibers. The results suggests that these are the factors that have a relevance on the plasma concentrations of beta-carotene of an individual.

The model tell us that the beta-carotene levels should increase a bit if we eat more fibers. Not using vitamins at all seem to have a large impact, causing a decrease in beta-carotene levels. Just by using vitamins occasionally, the beta-carotene levels could get much higher. If an individual have a large intake of calories per day it may cause a decrease in plasma beta-carotene levels. Looking at the confidence interval of the calorie parameter, we can see that the level of intake of calories could also lead to an increase in beta-carotene levels, or not have an impact at all.

The model suggests that men have lower levels of plasma beta-carotene than women. We can also see that the levels of beta-carotene increase with an increasing age. Having a high BMI will decrease the plasma concentration of beta-carotene. Being a former smoker will also decrease the levels compared to if an individual have never smoked according to the model. Being a former smoker is however much better than being a current smoker according to the model. Being a current smoker has almost the same effect on the plasma beta-carotene levels as not using any vitamins.

The R^2-value of the model is 24 %, which is quite low and the model is far away from perfectly describing the reality. Other factors are probably affecting the plasma concentration of beta-carotene such as genetics. Looking at the results it still seem likely that smoking, high BMI and not eating vitamins has a negative effect on the levels whilst eating fibers can have a positive effect. The model gives an idea to what extend but based on our result we cannot fully trust the model.