# Linear and Logistic Regression

*Project 2: Logistic Regression - Determinants of Low Plasma Beta-Carotene Levels*

Isabelle Frodé, Olof Bengtsson and Erik Kindt

May 2021

## Abstract

The aim of this project was to create a model that predicts the probability of having low plasma beta-carotene concentration based on personal characteristics and dietary factors. Multiple logistic regression was performed to create a classification model. The final model was correct 79 % of the times. The model suggest that the probability of having low levels decrease with a higher age and increase if an individual is smoking or have a high BMI. The dietary factors expected to decrease the probability of having low levels were fibers and dietary beta-carotene consumed per day. Not using vitamins often and high intake of calories per day were factors expected to increase the probability of having low plasma beta-carotene levels. Sex and the dietary variables fat, cholesterol, and alcohol consumed per day were not included in the final model for different reasons.

# Introduction

The aim of this project was to create a model that estimates the probability of having low levels of plasma concentration of beta-carotene, based on personal characteristics and dietary factors. Various studies suggests that there might be a relationship between low plasma concentration of beta-carotene and an increased risk of cancer [1]. The goal of this project was to forecast whether a certain diet or lifestyle can affect the probability of having low plasma beta-carotene levels. The study was based on a data set 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.

Multiple logistic regression was used to fit a model to the observed data containing the relevant variables. The model indicates if the variable increase, decrease or do not have an impact on the probability of having low levels. The model could then be used to predict the probability of low plasma beta-carotene levels for an individual.

Before we started to work on the model, a suitable cut-off value was identified to distinguish between low and high levels. A variable for low levels of plasma beta-carotene, *lowplasma*, was created. In part 1 of the project an *Age* model was created, investigating the relationship between low levels and age. In the next section, the *Age* model was extended with the other personal characteristic variables forming the *Background* model. The *Diet* model, another multiple logistic regression model created using the dietary variables was also created. In the end of part 2 a *Final* model was created taking into account both personal characteristics and dietary variables. In part 3 the four models were compared and evaluated. The best model was selected.

# Part 0

The cut-off value for high levels of beta-carotene had been reported to be 0.42 $\mu mol/l$ [1] but the data was given in the unit of $ng/ml$ meaning that the cut-off value had to be converted. This was done by finding the molar mass of beta-carotene, which according National Library of Medicine was 536.9 $g/mol$ [2]. With this value the converted cut-off value $a$ could be calculated according to

$$a = 536.9 \ \frac{g}{mol} \cdot 0.42 \ \frac{\mu mol}{l} = 225.498 \ \frac{\mu g}{l} = 225.498 \ \frac{ng}{ml} \tag{1}$$

The data was then separated by classifying data with beta-carotene $< a$ as low and other as high in a categorical variable called *plasmacat*. For future use a numeric variable was also created where high levels were represented by 0 and low levels by 1. Table 2 show the number of occurrences for high and low levels of beta-carotene.

| Category | Frequency |
|---|---|
| Low | 235 |
| High | 80 |

**Table 2:** Frequency table of the categorical variable *plasmacat*.

As table 2 suggests there are more samples with low levels of beta-carotene, namely 74.6 % of the data has low levels. This is what we can expect since high levels are thought to be an abnormality.

# Part 1

## Comparing low beta-carotene concetrations with smoking status.

To start with, the relationship between low plasma beta-carotene and smoking status was studied. A cross-tabulation was performed and the frequency of each category is presented in table 3.

Using these numbers, the probabilities and odds of having a low concentration of beta-carotene could be calculated. The results are also presented in table 3. The reference category used was the one with the most observations, i.e. "Never".

| Smoking status | High betaplasma | Low betaplasma | Low betaplasma probabilities p | Odds p/(1-p) |
|---|---|---|---|---|
| Never | 48 | 109 | 69.427 % | 2.271 |
| Former smoker | 29 | 86 | 74.483 % | 2.966 |
| Current smoker | 3 | 40 | 93.023 % | 13.333 |

**Table 3:** Frequency table regarding smoking status and low or high beta-carotene concentrations, with corresponding probabilities and odds.

Next, a logistic regression model was fitted using smoking status as explanatory variable. Th $e^{\beta}$-estimates e $\beta$-estimates andare presented in table 4 below.

| Parameter | $\beta$-estimate | CI | $e^{\beta}$-estimate | $e^{CI}$ |
|---|---|---|---|---|
| $\beta_0$ | 0.820 | [0.487, 1.168] | 2.2708 | [1.628, 3.216] |
| $\beta_{smokstat2}$ | 0.267 | [-0.207, 0.815] | 1.3059 | [0.764, 2.259] |
| $\beta_{smokstat3}$ | 1.770 | [0.695, 3.222] | 5.8716 | [2.003, 25.092] |

**Table 4:** $\beta$-estimates, $e^{\beta}$-estimates and their confidence intervals, CI, for a logistic regression model for lowplasma with smokstat as the explanatory variable.

Here it is worth noting that 0 is included in the confidence interval for $\beta_{smokstat2}$, which might mean that this is not significantly different from the reference variable. However since $\beta_{smokstat3}$ is significant the variable is kept.

The odds and probabilities $p$ were now expressed in terms of $\beta$-estimates as shown in equations 2 and 3. The results from the calculations agree with the results in table 3.

$$odds_i = e^{\beta_0 + \beta_1 x_{smokstat2} + \beta_2 x_{smokstat3}} \tag{2}$$

$$p_i = \frac{odds_i}{1 + odds_i} \tag{3}$$

The odds and probabilities of each category expressed in $\beta$-parameters are shown in equation 4 - 9.

$$odds_1 = e^{\beta_0 + \beta_{smokstat2}0 + \beta_{smokstat3}0} = 2.271 \tag{4}$$

$$odds_2 = e^{\beta_0 + \beta_{smokstat2}1 + \beta_{smokstat2}0} = 2.966 \tag{5}$$

$$odds_3 = e^{\beta_0 + \beta_{smokstat2}0 + \beta_{smokstat3}1} = 13.333 \tag{6}$$

$$p_1 = \frac{odds_1}{1 + odds_1} = 0.694 \tag{7}$$

$$p_2 = \frac{odds_2}{1 + odds_2} = 0.748 \tag{8}$$

$$p_3 = \frac{odds_3}{1 + odds_3} = 0.930 \tag{9}$$

A prediction with this model was made and the probability of having a low plasma beta-carotene concentration was examined for each smoking category. The probabilities along with the confidence intervals are presented in table 5 below.

| Smoking category | Low beta-carotene probability | Confidence interval |
|---|---|---|
| 0, Never | 0.694 | [0.618, 0.761] |
| 1, Former smoker | 0.748 | [0.661, 0.819] |
| 2, Current smoker | 0.930 | [0.805, 0.977] |

**Table 5:** Probabilities of having low beta-carotene concentrations for each smoking category, with respective confidence intervals.

To test the model a Wald test was performed. The test was chosen because it can determine which category would have a significant impact on the probability for a low level beta-carotene concentration.

Two null hypotheses for the test could be formulated $H_0^{(1)} : \beta_{smokstat2} = 0$ and $H_0^{(2)} : \beta_{smokstat3} = 0$ with corresponding alternative hypothesis $H_1^{(1)} : \beta_{smokstat2} \neq 0$ and $H_1^{(2)} : \beta_{smokstat2} \neq 0$. The test values are presented in table 6.

| Null hypothesis | Z-value | Distribution | P-value | Conclusion |
|---|---|---|---|---|
| $\beta_{smokstat2} = 0$ | 0.97 | $Z \sim N(0,1)$ | $3.33 \cdot 10^{-1}$ | cannot reject $H_0^{(1)}$ |
| $\beta_{smokstat3} = 0$ | 2.84 | $Z \sim N(0,1)$ | $4.50 \cdot 10^{-3}$ | reject $H_0^{(2)}$ |

**Table 6:** Test statistic, distribution of the test statistic when $H_0$ is true, P-value and conclusion for each null hypothesis using the Wald test.

As shown in table 6, the $H_0^{(1)}$ hypothesis could not be rejected at a significance level of 95 %. The reasoning behind this is that the Z-value of the hypothesis

was not larger than the 95 %-quantile $\lambda_{0.025} = 1.96$. The P-value was also larger than 0.05. However for the second null hypothesis, $\beta_{smokstat3} = 0$, the Z-value of the hypothesis was larger than $\lambda_{0.025} = 1.96$, which implies that the null hypothesis could be rejected in favour of $H_1^{(2)}$. The results tell us that a current smoker has an increased risk of having low levels of plasma beta-carotene compared to a non-smoker. It also tell us that it is uncertain whether being a former smoker will increase the probability, which seems reasonable.

## Low plasma as a function of age

The next task was to examine the probability of having low beta-carotene concentration with increasing age. According to figure 1, a higher age decreases the chances of having low levels of beta-carotene. It means that the findings agrees with the findings in project 1, where higher age indicated higher levels of plasma beta-carotene.

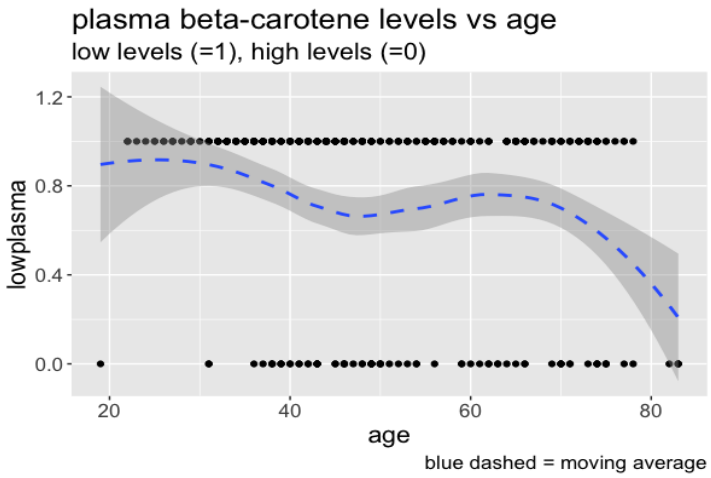



**Figure 1:** Low plasma beta-carotene levels against age with a moving average. The moving average represents the probability of low beta levels.

Then a logistic regression model for lowplasma as a function of age was fitted, referred to as the *Age* model. The parameter estimates and confidence interval for the model are presented in table 7 below.

| Parameter | $\beta$-estimate | CI | $e^{\beta}$-estimate | $e^{CI}$ |
|---|---|---|---|---|
| $\beta_0$ | 2.551 | [1.603, 3.550] | 12.825 | [4.967, 34.822] |
| $\beta_{age}$ | -0.028 | [-0.047 -0.011] | 0.972 | [0.955, 0.989] |

**Table 7:** $\beta$-estimates, $e^{\beta}$-estimates and the confidence intervals, CI, for the *Age* model.

The Wald test was performed once again with the null hypothesis $H_0$: $\beta_{age} = 0$, and $H_1$: $\beta_{age} \neq 0$. The test was chosen since it investigates if a variable has a significant effect on the probability, meaning that it tests if a variable changes the log-odds. The test was used to examine the significance of the variable *age*. The test results are presented in table 8 below.

| Hypothesis | Z-value | Distribution | P-value | Conclusion |
|---|---|---|---|---|
| $H_0 : \beta_{age} = 0$ | -3.17 | $Z \sim N(0,1)$ | $1.537 \cdot 10^{-3}$ | reject $H_0$ |

**Table 8:** Test statistic, distribution of the test statistic when $H_0$ is true, P-value and conclusion for each null hypothesis using the Wald test on the *Age* model.

Using the same reasoning as before, since the absolute value of the Z-value is larger than the 95 %-quantile $\lambda_{0.025} = 1.96$, $H_0$ could be rejected at significance level 95 %. The results tells us that age has a significant effect on the probability of having low levels of plasma beta-carotene, which was expected with respect to the results presented in figure 1.

Then the change in odds when the age increased by one year was examined. This was examined using the ratio, $r$ as shown in equation 10 below.

$$r = \frac{e^{\beta_0 + \beta_{age}(age+1)}}{e^{\beta_0 + \beta_{age}(age)}} = e^{\beta_{age}} = 0.972 \tag{10}$$

As seen in equation 10, the odds of having low beta-carotene levels decreases with approximately 3 % every year. This was also done for an age difference of ten, calculated as in equation 10 and the results are presented in table 9 below.

| Change in year | Estimate | Confidence interval |
|---|---|---|
| 1 | 2.82 % | [1.10, 4.54] % |
| 10 | 24.8 % | [10.5, 37.2] % |

**Table 9:** Estimated decrease in probability of having low beta-carotene levels.

As seen in table 9, the odds of having low beta-carotene levels significantly drops when the age increases with 10 year. The predicted probabilities are plotted in figure 2.

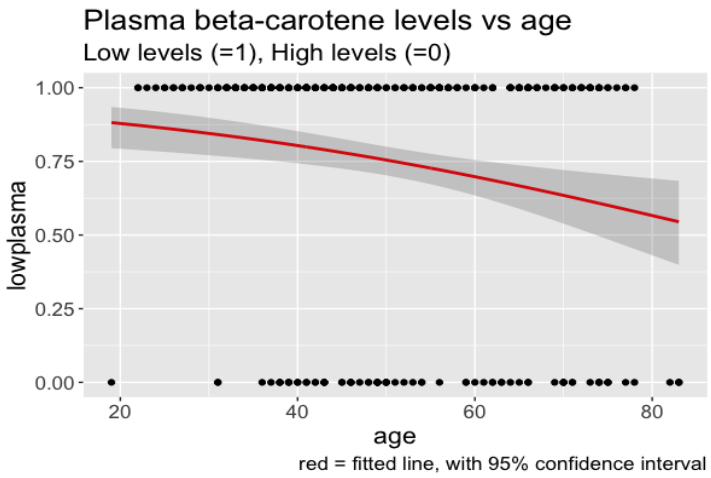



**Figure 2:** Predicted probability of having low levels of beta-carotene as a function of age with 95 %-interval using the *Age* model.

The yearly change for a 30 and a 70 year old was then examined, if the change are the same, the relationship could be described as equation 11.

$$\Delta_{31-30} = \Delta_{71-70} \tag{11}$$

The yearly changes were calculated according to equations 12 and 13 below. In equation 12, $t_1 = 31$ and $t_0 = 30$. In equation 13, $t_1 = 71$ and $t_0 = 70$.

$$\Delta_{31-30} = e^{\beta_0+\beta_1 t1} - e^{\beta_0+\beta_1 t0} = e^{2.551-0.028\cdot 31} - e^{4.610-0.028\cdot 30} = -0.378\ \% \tag{12}$$

$$\Delta_{71-70} = e^{\beta_0+\beta_1 t_1} - e^{\beta_0+\beta_1 t_0} = e^{2.551-0.028\cdot 71} - e^{2.551-0.028\cdot 70} = -0.665\ \% \tag{13}$$

It's clear that equation 11 does not hold and that the change in beta-carotene gets more negative with an increasing age. This explains the increasing negative slope in figure 2.

## Leverage for the *Age* model

The leverage was plotted for the *Age* model and for a linear model, also using age as covariate. The plot is shown below in figure 3.

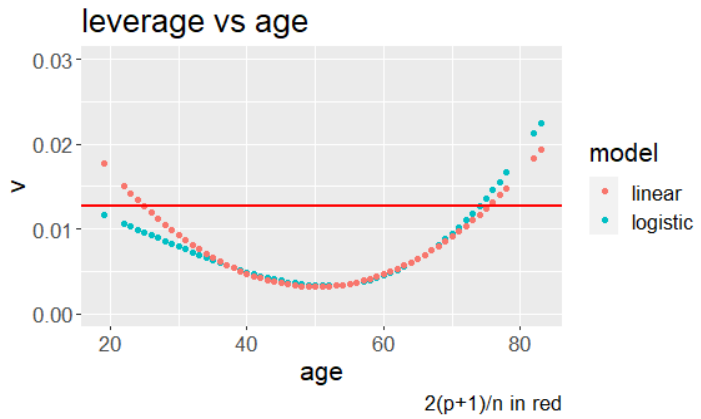



**Figure 3:** The leverage for the logistic and linear age models.

The data points with the highest leverage for the logistic *Age* model were then identified. The identified observations are marked in the figure 4. The plot indicated that the logistic model perform better for lower age than the linear model, as the leverage is lower. In terms of leverage, the models perform similar for ages 40-70. For higher age the linear model seem to perform a little bit better with respect to leverage.

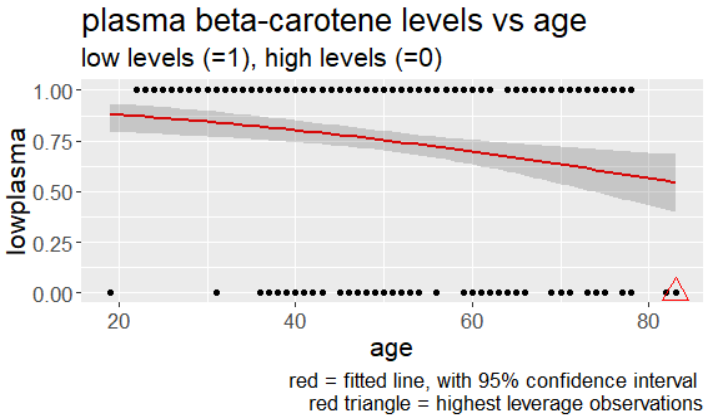



**Figure 4:** Predicted probability of having low levels of beta-carotene as a function of age with the data points with highest leverage marked with a triangle.

## Standardized deviance residuals for the *Age* model

The standardized deviance residuals were then determined for the *Age* model and plotted against age. The results are shown in figure 5. The high and low beta-carotene levels were divided by color and the high leverage data points were highlighted.

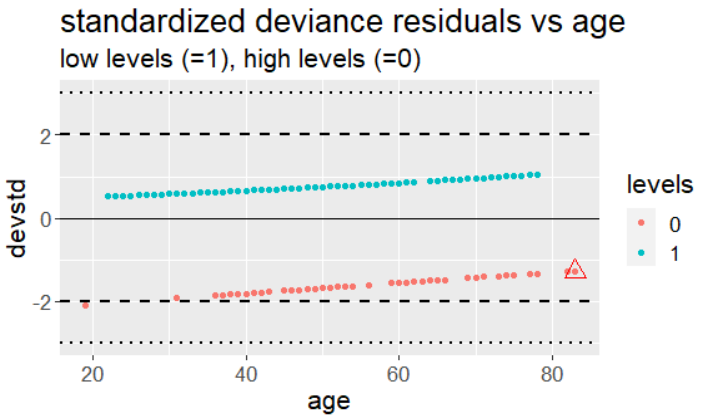



**Figure 5:** Standard deviance residuals of the age model as a function of age.

Figure 5 show that the observations with the highest leverage had the smallest residuals out of all high concentration observations. The high leverage observations were well within the limit for where they would affect the model negatively.

The high levels are always 0, low levels are always 1 of the observed values, and the probability ranges between 0 and 1. This means that if the probability was 0.5 the absolute value of the residuals would be the same for the high and low concentration groups. Since the probability according to the *Age* model is always $> 0.5$ the high level observations will always have larger residuals, which is what we see in figure 5. The slope of the residuals against age is the same for high and low levels.

The observation with the largest residual was then localized and highlighted, see figure 6.

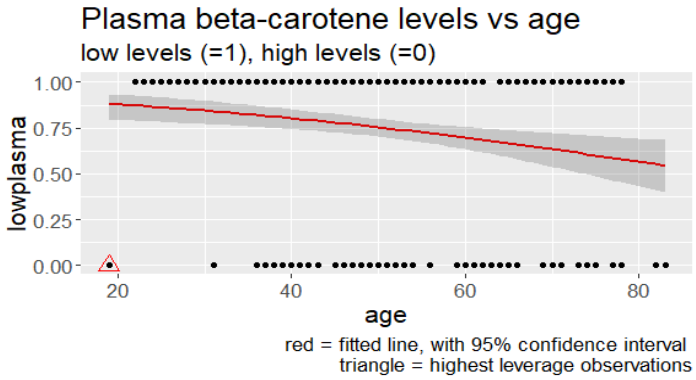



**Figure 6:** Predicted probability of having low levels as a function of age. Data point with largest standardized deviance residual marked with a triangle.

The observation with the largest residual was the observation that had the largest offset with respect to the estimated model. Figure 2 convinced us that the residual must be the largest for the observation with high plasma beta-carotene level and smallest age, since the probability for low level was the highest there. Figure 6 confirms that.

## Cook's distance for the *Age* model

The Cook's distance was calculated for each point and the results are depicted in figure 7 below.

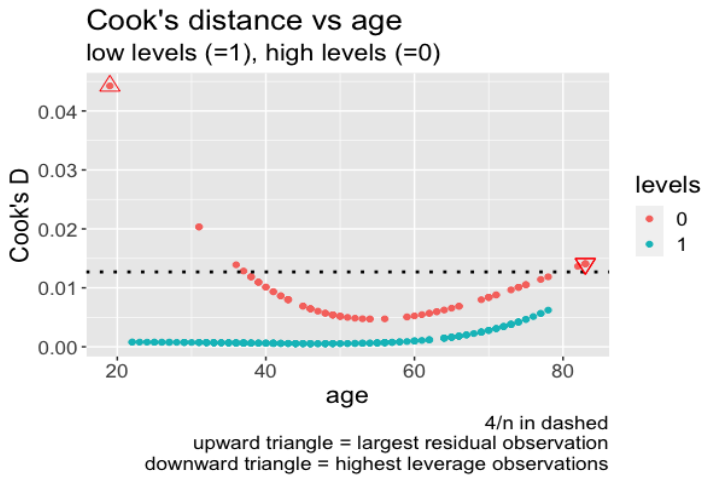



**Figure 7:** Cook's distance for all observations. Highest leverage and largest residual observations highlighted.

Figure 7 makes it clear that the data point with the largest residual also had the largest Cook's distance. This observation which had the lowest age but still high levels of beta-carotene concentration had the distinguishable largest influence. Examining the moving average shown in figure 1, it is clear that the largest residual observation had a large influence. If the observation was not there the moving average would continue towards 1.0 as age $\rightarrow$ 20. Instead, the moving average goes down. Figure 2 also show that the slope of the predicted probability was less steep as the age $\rightarrow$ 20 than it was for the age categories in the middle. It is therefore not unexpected that this observation would have the largest influence. Figure 2 further shows that the high leverage observations causes the largest confidence interval for the predicted probability. These observations had the second largest influence.

# Part 2

## Forward selection using AIC

Part 2 of the project concerned multiple logistic regression and selecting a model. Firstly, a null model was selected with only an intercept and the Akaike information criterion, or AIC, was used to determine which variables to be used in the model. The largest model allowed was one containing all background variables, *age*, *sex*, *smokstat*, and *quetelet*.

The order in which the variables were added to the model was: quetlet, smokstat and age. The variable sex was not used in this model, which is referred to as the *Background* model. The AIC changed with every added variable according to table 10 below.

| Variable | AIC | AIC decrease |
|---|---|---|
| null model | 358.99 | 0 |
| quetelet | 347.79 | - 11.20 |
| smokstat | 336.92 | - 10.87 |
| age | 330.69 | - 6.23 |

**Table 10:** The AIC forward step selection.

The $\beta$- and $e^{\beta}$-estimates with confidence intervals are presented in table 11.

| Parameter | $\beta$-estimate | CI | $e^{\beta}$-estimate | $e^{CI}$ |
|---|---|---|---|---|
| $\beta_0$ | -0.571 | [-2.316, 1.123] | 0.565 | [0.099, 3.075] |
| $\beta_{quetelet}$ | 0.108 | [0.053, 0.171] | 1.114 | [1.054, 1.186] |
| $\beta_{smokstat2}$ | 0.297 | [-0.260, 0.865] | 1.346 | [0.771, 2.375] |
| $\beta_{smokstat3}$ | 1.867 | [0.753, 3.343] | 6.470 | [2.124, 28.306] |
| $\beta_{age}$ | -0.027 | [-0.045 -0.008] | 0.974 | [0.956, 0.991] |

**Table 11:** $\beta$- and $e^{\beta}$-estimates together with their confidence intervals, CI, for the *Background*-model.

Because of the difficulties of plotting two continuous variables, *quetelet* and *age*, the variable *age* was divided into three discrete classes. These classes were $age \leq 40$, $40 < age \leq 55$ and $age \geq 55$. Plots of observed beta-carotene levels against age and BMI with probability of low beta-carotene are shown in figures 8 and 9.

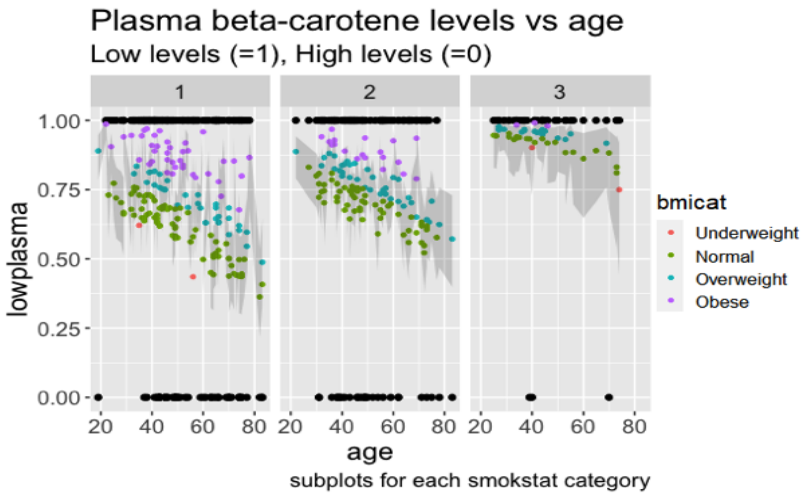



**Figure 8:** Observed beta-carotene and predicted probability of low levels with confidence intervals plotted against age, for each smokstat category. Each BMI-category is marked with different coloured markers.

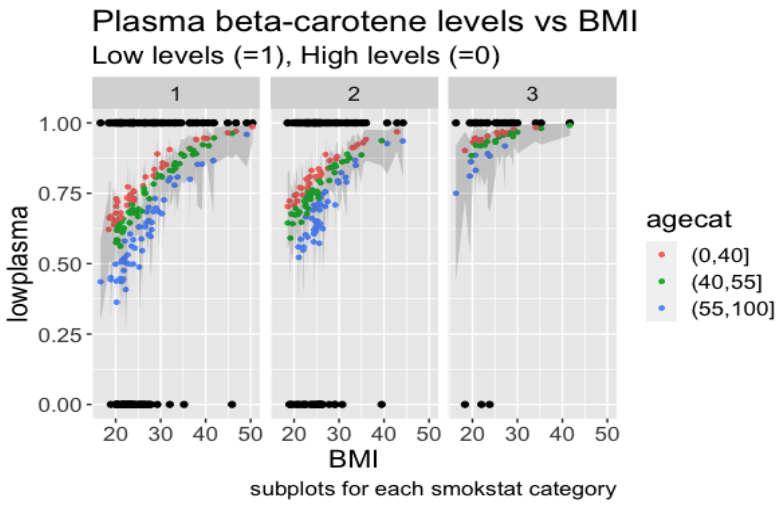



**Figure 9:** Observed beta-carotene and predicted probability with confidence intervals plotted against BMI, for each smokstat category. Each age-category is marked with different coloured markers.

Figures 8 and 9 show that the probabilities of having low beta-carotene levels decreases with an increase in age, increases with an increase in BMI and increases with an "increase" in smoking. This correlates well to table 11, since the slope is negative for $\beta_{age}$ and positive for $\beta_{quetelet}$, $\beta_{smokstat2}$ and $\beta_{smokstat3}$.

## Leverage, residuals and Cook's distance for the *Background* model

The leverage, the standardized deviance residuals and Cook's distance were then calculated for the *Background* model whereupon they were plotted. Below are the plots of the leverage as a function of both age and BMI with the number of covariates $p = 5$ and the number of observations $n = 315$. The reference lines are depicted at $2(p+1)/n = 2 \cdot 6/315$.

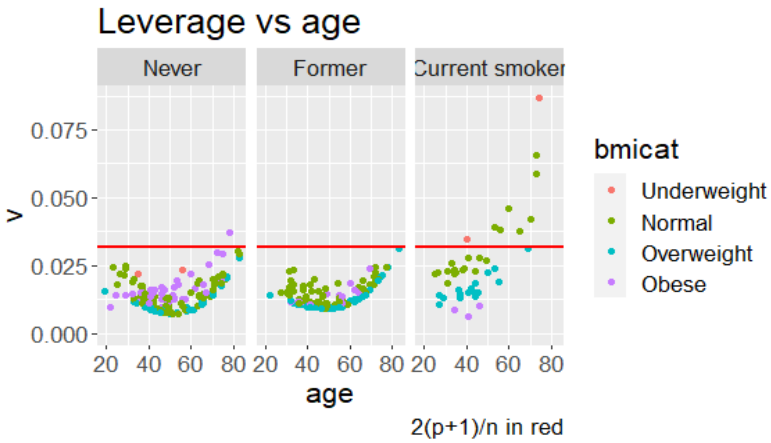



**Figure 10:** Leverage for the *Background* model plotted against age.

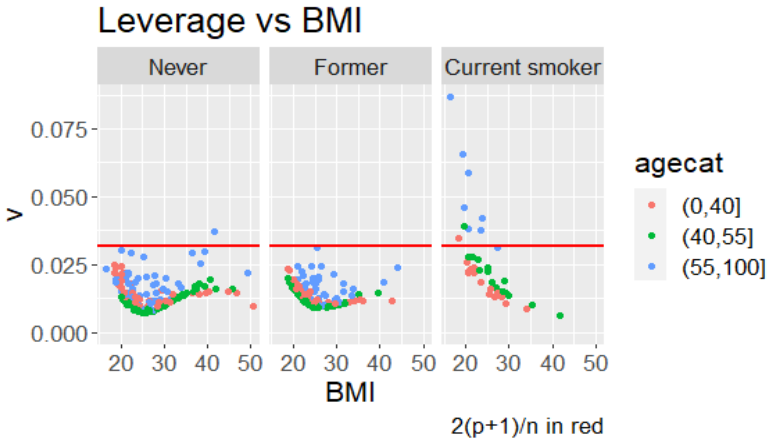



**Figure 11:** Leverage for the *Background* model plotted against BMI.

As can be seen in figure 10 and 11 observations corresponding to current smokers, underweight to normal weight and above average age are the ones who

generate the most leverage.

The standardised deviance residuals were then plotted similarly as the leverage above but with dashed lines at $\pm 2$ and $\pm 3$.

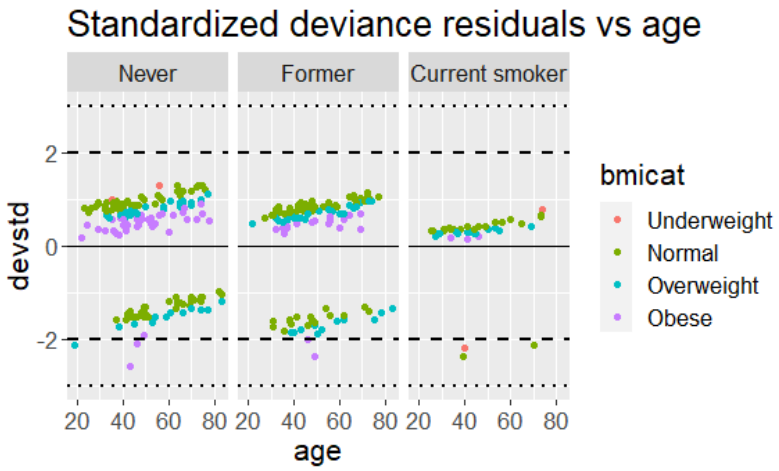



**Figure 12:** Standardised deviance residuals for the *Background* model against age.

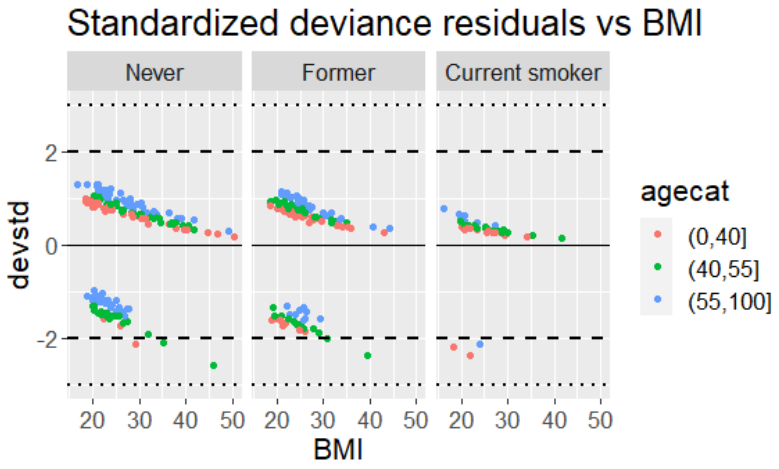



**Figure 13:** Standardised deviance residuals for the *Background* model against BMI.

The plots in figure 12 and 13 above show that the data resulting in large residuals mainly comes from observations corresponding to normal weight to overweight, average to above average age and non-smokers. This is quite the difference comparing to the characteristics that led to high leverage.

The calculated Cook's distance for the background was also plotted in the same manner as for both leverage and the standardized deviance residuals. In this plot the data points with large residuals were highlighted and a reference line was plotted at $4/n = 4/315$.

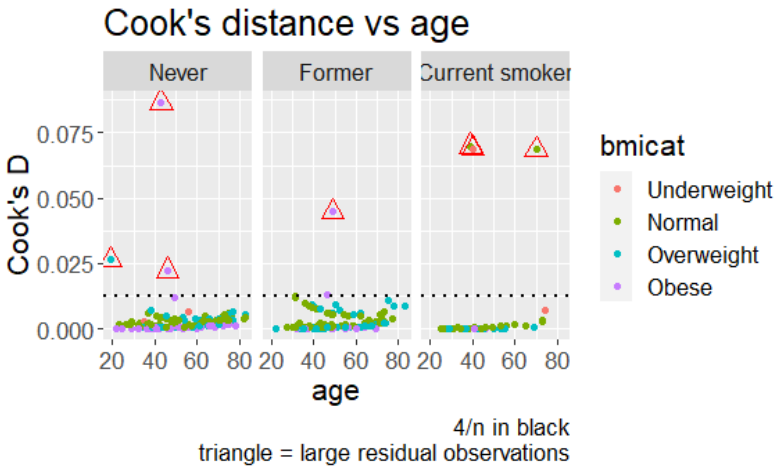



**Figure 14:** Cook's distance for the *Background* model plotted against age.

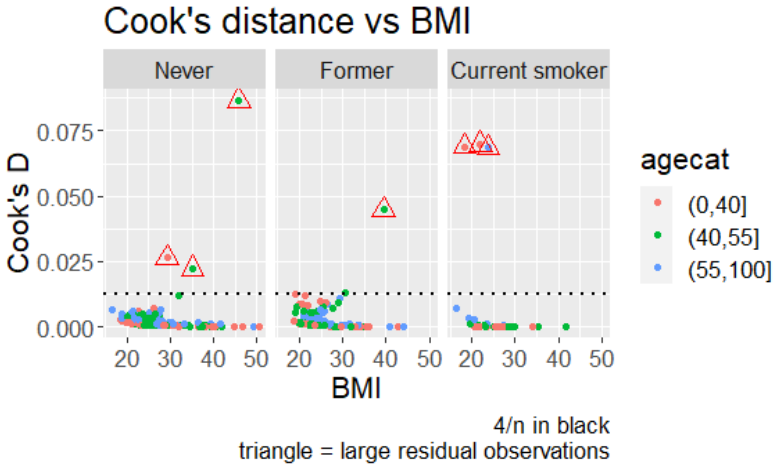



**Figure 15:** Cook's distance for the *Background* model plotted against BMI.

The two plots in figure 14 and 15 show that the data points resulting in large residuals also lead to large Cook's distance, meaning they have a large influence on the model.

The data with large residuals were then marked in the plots in figure 8 and 9, see figure 16 and 17 below.

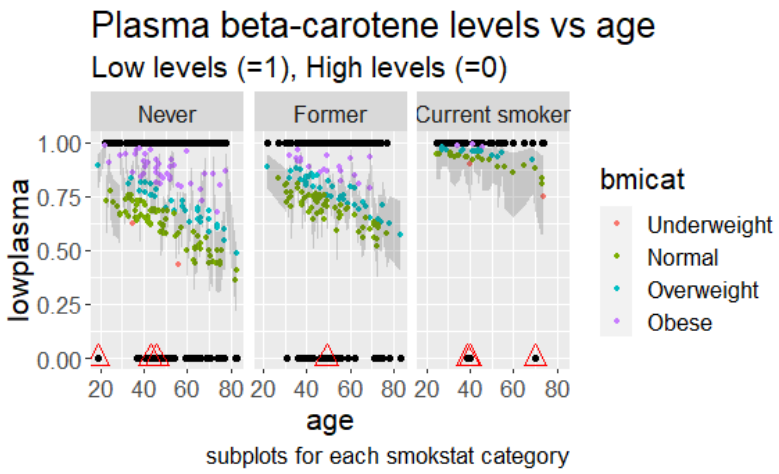



**Figure 16:** Observed beta-carotene and predicted probability with confidence intervals plotted against age, for each smokstat category. Each BMI-category is marked with different coloured markers and data with large residuals is highlighted with a triangle.

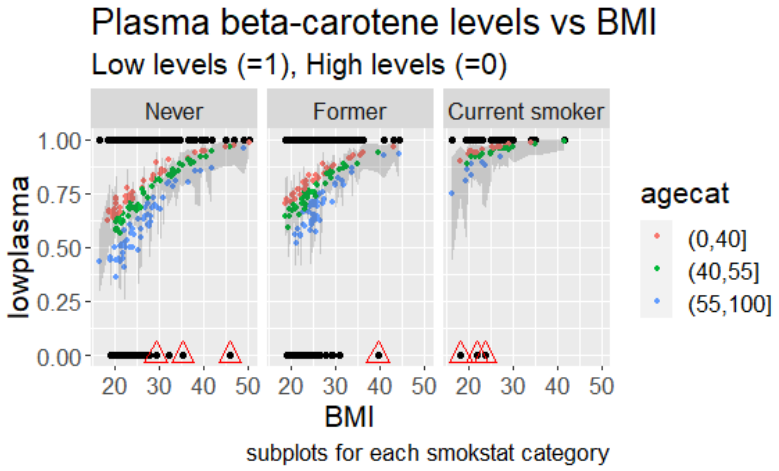



**Figure 17:** Observed beta-carotene and predicted probability with confidence intervals plotted against BMI, for each smokstat category. Each age-category is marked with different coloured markers and data with large residuals is highlighted with a triangle.

Figure 16 and 17 above shows that the data with large residuals all have high levels of beta-carotene while they according to the model would have had low levels. Removing this data and re-fitting the model could therefore be dangerous since it would cause bias. However, had the data been unreasonable it would probably be a good idea to remove the strange data and then re-fit the model, but it was concluded that this was not the case.

## Creating the *Diet* model

Next task was to build a model using the dietary variables, *vituse*, *calories*, *fat*, *fiber*, *alcohol*, *cholesterol* and *betadiet*. The model was referred to as the *Diet* model. The model was chosen using forward selection and with the AIC as the criterion. This was due to the fact that predictions with the model were going to be made and the AIC is preferable when doing predictions. Every variable that would give reasonable predictions of future observations was used in the model. The $\beta$- and $e^{\beta}$-estimates for the *Diet* model are presented in table 12.

| Variable | Estimate | CI | $e^{Estimate}$ | $e^{CI}$ |
|---|---|---|---|---|
| $\beta_0$ | 0.824 | [-0.132, 1.792] | 2.280 | [0.876, 6.000] |
| $\beta_{vituse2}$ | 0.463 | [-0.181, 1.127] | 1.590 | [0.834, 3.088] |
| $\beta_{vituse3}$ | 1.424 | [0.738, 2.166] | 4.155 | [2.091, 8.723] |
| $\beta_{diet}$ | $-2.845 \cdot 10^{-4}$ | $10^{-4}[-4.89, -8.64]$ | 0.9997 | [0.9995, 0.9999] |
| $\beta_{fiber}$ | -0.086 | [-0.153, $-0.0220$] | 0.918 | [0.858, 0.978] |
| $\beta_{calories}$ | 0.001 | $10^{-4}$[3.203, 14.76] | 0.974 | [1.000, 1.002] |

**Table 12:** $\beta$ and $e^{\beta}$-estimates for the dietary model and their corresponding confidence intervals, CI.

Now the three models *Age*, *Background* and *Diet* could be compared. This was done by examining the McFadden pseudo $R^2$, both the unadjusted $R^2_{McF}$ and adjusted $R^2_{McF,adj}$. The results with the number of covariates $p$ are presented table 13.

| Model | $R^2_{McF}$ | $R^2_{McF,adj}$ | Covariates |
|---|---|---|---|
| *Age* | 0.029 | 0.026 | 1 |
| *Background* | 0.102 | 0.091 | 4 |
| *Diet* | 0.136 | 0.122 | 5 |

**Table 13:** Table of the three different models and the corresponding $R^2_{McF}$,$R^2_{McF,adj}$ and number of covariates.

Studying the results in table 13, the model *Diet* is the preferable model, because it's the model with the highest adjusted McFadden $R^2$.

## Creating the *Final* model

A *Final* model was then created by testing a variety of different models and methods. First a model was created using stepwise selection with AIC as stop criterion, by starting with the null model and setting the maximum model as the one with all available variables. Then a model containing all variables from the final model obtained in project 1 was evaluated followed by the model that contained all background and dietary variables.

The next model was created using stepwise selection with AIC where the initial model was set as the null model ranging up to the model with all background and dietary variables. The last two models to be tested were obtained from stepwise selection going in both directions starting from the null model and the *Diet* model, ranging between the null model and model with all background and dietary variables. The model with the highest $R^2_{McF,adj}$, potentially the best model, was found to be the model containing all variables in the *Background* model and the *Diet* model. The $e^{\beta}$-estimates and their confidence intervals are given below in table 14

| Parameter | $e^{\beta}$-estimate | Confidence interval |
|---|---|---|
| $\beta_0$ | 0.729 | [0.0808, 6.330] |
| $\beta_{age}$ | 0.975 | [0.954, 0.996] |
| $\beta_{smokstat2}$ | 1.206 | [0.657, 2.233] |
| $\beta_{smokstat3}$ | 3.707 | [1.146, 16.78] |
| $\beta_{quetelet}$ | 1.103 | [1.042, 1.176] |
| $\beta_{vituse2}$ | 1.228 | [0.616, 2.476] |
| $\beta_{vituse3}$ | 3.682 | [1.809, 7.885] |
| $\beta_{calories}$ | 1.001 | [1.000, 1.0012] |
| $\beta_{fiber}$ | 0.950 | [0.885, 1.018] |
| $\beta_{betadiet}$ | 0.9997 | [0.9995, 0.9999] |

**Table 14:** $e^{\beta}$-estimates and their confidence intervals for the final model.

The pseudo $R^2$-values and the number of parameters for the *Final* model was added to table 13, as seen in table 15 below.

| Model | $R^2_{McF}$ | $R^2_{McF,adj}$ | Covariates |
|---|---|---|---|
| *Age* | 0.029 | 0.026 | 1 |
| *Background* | 0.102 | 0.091 | 4 |
| *Diet* | 0.136 | 0.122 | 5 |
| *Final* | 0.196 | 0.171 | 9 |

**Table 15:** Table of the three different models and the corresponding $R^2_{McF}$,$R^2_{McF,adj}$ and number of covariates.

# Part 3

## Confusion matrix, specificity, sensitivity, accuracy & precision

The first step in measuring the goodness-of-fit of the four models *Age*, *Background*, *Diet*, and *Final* was to calculate the confusion matrix for each model. To start with the cut-off value was set to 0.5. The confusion matrices for each of the models are presented in table 16 - **??**.

| True | Predicted ($\hat{Y}_i$) | |
|---|---|---|
| ($Y_i$) | $\hat{p}_i \leq 0.5$ | $\hat{p}_i > 0.5$ |
| $Y_i = 0$ | 0 | 80 |
| $Y_i = 1$ | 0 | 235 |

**Table 16:** *Age* model

| True | Predicted ($\hat{Y}_i$) | |
|---|---|---|
| ($Y_i$) | $\hat{p}_i \leq 0.5$ | $\hat{p}_i > 0.5$ |
| $Y_i = 0$ | 11 | 69 |
| $Y_i = 1$ | 6 | 229 |

**Table 17:** *Background* model

| True | Predicted ($\hat{Y}_i$) | |
|---|---|---|
| ($Y_i$) | $\hat{p}_i \leq 0.5$ | $\hat{p}_i > 0.5$ |
| $Y_i = 0$ | 24 | 56 |
| $Y_i = 1$ | 9 | 226 |

**Table 18:** *Diet* model

| True | Predicted ($\hat{Y}_i$) | |
|---|---|---|
| ($Y_i$) | $\hat{p}_i \leq 0.5$ | $\hat{p}_i > 0.5$ |
| $Y_i = 0$ | 49 | 31 |
| $Y_i = 1$ | 43 | 192 |

**Table 19:** *Final* model

Based on the values presented in the confusion matrices, the specificity, sensitivity, accuracy, and precision for each model could be calculated. The results are shown in table 20.

| Model | Specificity | Sensitivity | Accuracy | Precision |
|---|---|---|---|---|
| *Age* | 0.00 | 1.00 | 0.25 | 0.75 |
| *Background* | 0.14 | 0.97 | 0.76 | 0.77 |
| *Diet* | 0.30 | 0.96 | 0.79 | 0.80 |
| *Final* | 0.61 | 0.82 | 0.77 | 0.86 |

**Table 20:** Specificity, sensitivity, accuracy and precision for all models using cut-off value 0.5.

Table 20 shows that the largest spread between the models is regarding the specificity. Specificity measures the proportion of true failures that have been correctly classified as failures. The *Final* model outperforms the other models in this aspect. The *Age* model stands out because it does not manage to correctly classify any true failures. All models except the *Age* model perform similar with respect to accuracy. The *Final* model has the highest precision but also the lowest sensitivity.

## ROC-curve and AUC-value

The ROC-curves was then plotted for all four models using a threshold value of 0.5. The results are shown in figure 18.

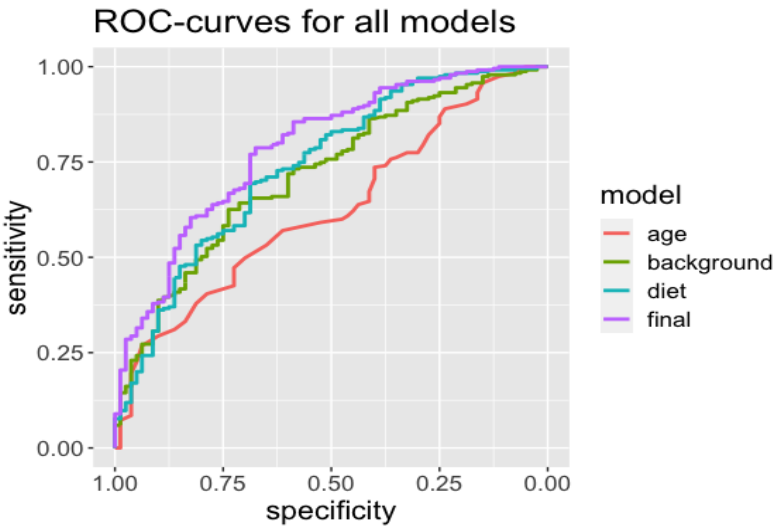



**Figure 18:** ROC-curves for all models.

The area under the curves (AUC-values) for the ROC-curves were then calculated. The values are presented in table 21 with a 95 % confidence interval.

| Model | AUC-values | Confidence interval |
|---|---|---|
| *Age* | 0.62 | [0.56, 0.69] |
| *Background* | 0.72 | [0.65, 0.78] |
| *Diet* | 0.74 | [0.67, 0.80] |
| *Final* | 0.79 | [0.73, 0.85] |

**Table 21:** AUC-vales for the ROC-curves with a 95 % confidence interval.

Pair-wise tests were then performed to compare the AUC-values for two models at a time. The test shows that there is a significant difference ($p < 0.05$) between the performance of the *Final* model and all the other models. There is also a significant difference between the performance of both the *Background* model and the *Diet* model compared to the *Age* model. There is no significant difference in performance between the *Background* and the *Diet* model. The pair-wise tests show that the *Final* model outperforms the other models with respect to having a large sensitivity and specificity when the cut-off value is 0.5.

## Optimal cut-off value

The next task was to find the optimal threshold value for $\hat{p}_i$ for each of the four models. The optimal value was the value which had high sensitivity while still keeping a high specificity. The results are shown in table 22 below.

| Model | Cut-off$_{opt}$ | Specificity | Sensitivity | Accuracy | Precision |
|---|---|---|---|---|---|
| *Age* | 0.762 | 0.613 | 0.570 | 0.581 | 0.812 |
| *Background* | 0.718 | 0.688 | 0.655 | 0.663 | 0.860 |
| *Diet* | 0.753 | 0.688 | 0.694 | 0.692 | 0.867 |
| *Final* | 0.750 | 0.700 | 0.694 | 0.695 | 0.872 |

**Table 22:** The optimal cut-off value cut-off$_{opt}$ *for each of the different models, with resulting specificity, sensitivity accuracy and precision.*

Comparing the results in table 22 to the results in table 20 with threshold value 0.5, it is clear that changing the threshold value had the largest impact on the *Age* model. The results obtained with the optimal threshold values indicates that the *Final* model performs the best in terms of highest accuracy and precision.

## Hosmer-Lemeshow goodness of fit test

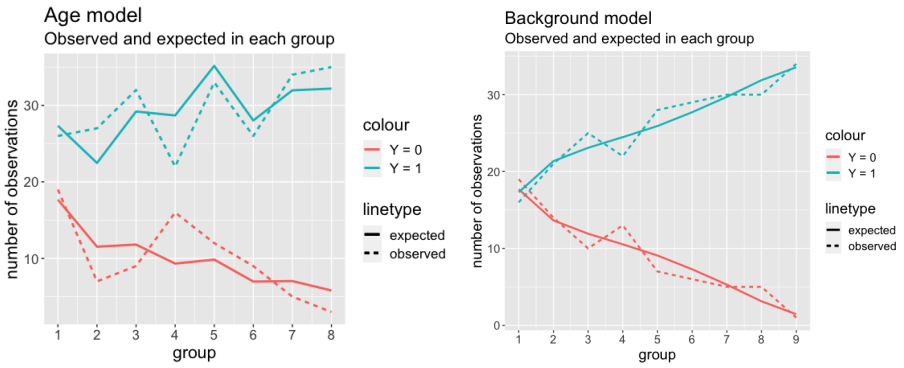
Then a Hosmer-Lemeshow goodness of fit test was performed for each model. The null hypothesis for the Hosmer-Lemeshow test is $H_0$ : "*the model gives correct probabilities*". The test is performed by estimating the probabilities of low levels, sorting them in increasing order and then dividing them into groups. Different number of groups $g > p + 1$ for each model was tested and the results changed a lot dependent on what $g$ was used. A small p-value mean that the model was a poor fit and we can reject $H_0$. For the *Age* model the p-values was for some $g$ larger than 0.05 and for other $g$ it was smaller than 0.05. The smallest expected value in each group for each model using $g = 10$ are shown in table 23 below. $g = 10$ was used for all models for comparison even though the *Final* model would require at least $g = 11$ groups to perform the Hosmer-Lemeshow goodness of fit test.

| | Smallest expected value in each group $g$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *Age* | 13.3 | 11.0 | 10.4 | 8.7 | 7.0 | 7.7 | 6.3 | 6.4 | 4.8 | 4.3 |
| *Background* | 15.6 | 12.7 | 10.9 | 9.7 | 8.9 | 7.4 | 5.8 | 4.4 | 2.5 | 1.3 |
| *Diet* | 12.1 | 14.1 | 10.9 | 8.6 | 7.3 | 6.0 | 4.8 | 3.9 | 2.8 | 1.7 |
| *Final* | 10.3 | 15.6 | 12.1 | 8.9 | 7.0 | 5.1 | 4.1 | 2.9 | 1.8 | 0.74 |

**Table 23:** Smallest expected value in each group $g$ for $g = 10$.

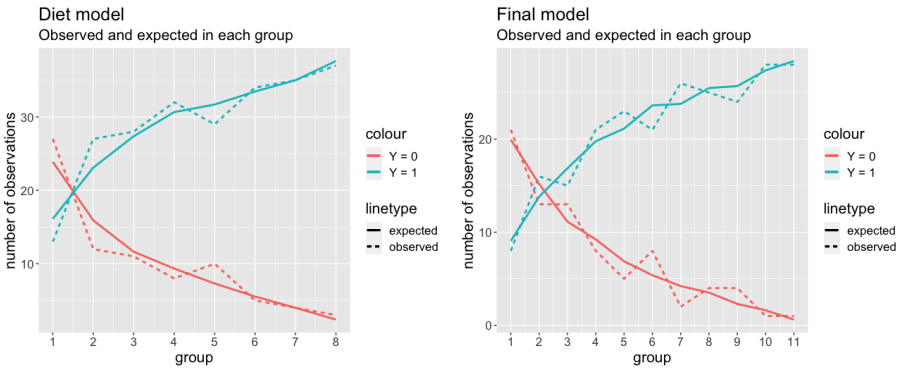
A suitable number of groups was found by testing several values and then following a majority conclusion. The expected and observed number of successes

and failures were plotted for each model. The results are presented below in figure 19 and 20.




**(a)** *Age* model **(b)** *Background* model

**Figure 19:** Expected and observed number of successes and failures.




**(a)** *Diet* model **(b)** *Final* model

**Figure 20:** Expected and observed number of successes and failures.

The Hosmer-Lemeshow goodness of fit test tell us that the *Age* model was a poor fit to the data for some number of groups $g$. The test results further tell us that $H_0$ cannot be rejected, models may be a good fit. In figure 19 and 20 it is clear that the age model had the worst fit whilst the other models seem fine. What is shown in the plots corresponds to what was found in the Hosmer-Lemeshow test results.

## Selecting the model

Taking all results into account, the *Final* model was selected as the best model to estimate the probability of having low levels of plasma beta-carotene levels.

The Hosmer-Lemeshow goodness of fit test indicates that the *Age* model was the only model that may be a poor fit to the observed data. The AUC-values to the ROC-curves also implied that the *Age* model performed significantly different (worse) than all other models, even though the accuracy and precision increased a lot when changing the threshold value. The *Age* model was therefore ruled out.

The AUC-values to the ROC-curves further implied that the *Final* model performed significantly better than both the *Background* and the *Diet* model in predicting the true outcome. The accuracy and precision achieved with the *Final* model was also higher than for the other two models using optimal threshold values. Even though the *Final* model had more parameters than the *Background* and *Diet* model it still had a higher adjusted McFadden $R^2$-value which also indicates that the *Final* model was the better choice.

# Conclusion

From the performed logistic regression data analysis it can be concluded that the best model is the *Final* model. The model contains the personal characteristic variables age, smoking status and continuous BMI. Sex was not included. The dietary factors in the model were the variables vitamin use, calories, fibers and dietary beta-carotene. The variables fat, alcohol and cholesterol were not included. The model suggests that the included variables are the factors of importance when predicting the probability of having low levels of plasma beta-carotene.

The model suggests that the risk of having low levels should decrease with a higher age and increase if an individual is smoking or have a high BMI. Not using vitamins at all seem to increase the probability of having low levels almost as much as being a current smoker according to the model. Using vitamins occasionally will still increase the probability of having low levels, but much less. A high calorie intake may increase the probability whilst eating a lot of fiber and dietary beta-carotene seem to decrease the risk of having low levels.

The AUC-value of the *Final* model was calculated to 0.79. This means that the model is correct approximately 79 % of the times which is somewhere in the middle of being an excellent classifier (100 % correct) and a bad classifier (50 % correct).

# References

[1] A. Lindgren. 2021. *Linear Logistic Regression 2021. PROJECT 2: LOGISTIC REGRESSION MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING).*

[2] National Library of Medicine. URL: `https://pubchem.ncbi.nlm.nih.gov/compound/beta-Carotene`