Project 1: Linear regression

MASM22/FMSN30/FMSN40:

Andre Edelburg, Dolev Illouz

LUNDS
UNIVERSITET

Supervisor: Peter
Department of Mathematics
Division: Mathematical Statistics
April 2021

Linear Regression

Determinants of Plasma Beta-Carotene Levels

Andre Edelburg, Dolev Illouz

Abstract

Decreased beta-carotene plasma levels have been shown to be associated with increased cancer risk. Hence, predicting an individual's beta-carotene plasma levels is of relevant clinical interest. A final model was selected by performing model validation and selection tests. In this model, the beta-carotene levels were log-transformed, as their values were exponentially distributed. Furthermore, the model considered the vitamin usage, caloric intake, fiber intake, dietary beta-carotene consumed, quetelet, smoking status, sex, and age of an individual. Both increases in age and fiber intake resulted in increased beta-carotene plasma levels. Whereas smoking of any kind, increased quetelet or caloric consumption, no vitamin usage, and being male reduced the beta plasma levels. Using this model, 24.4% of its variability can be explained.

Introduction

Beta-carotene is found in many foods and can be bought as a dietary supplement. It is the source of the orange color in pumpkins, carrots, and other vegetables. Many studies have shown that the risk of cancer increases with lower beta-carotene levels. It is less known what factors influence beta-carotene levels in human plasma. In this study, 135 patients were subject to beta-carotene screening. Their physical characteristics such as body mass index (BMI) and sex were recorded alongside their behavior, such as medication and dietary supplement usage. A total of 13 variables were considered, and their impact on beta-carotene levels was analyzed statistically.

Part 1: Plasma beta-carotene and age

a) To determine if age has a linear or log-linear relationship with beta-carotene, we plotted both of them against age. Before plotting log(beta-carotene) as a function of age, one data point had to be excluded because it has a value of 0 for beta-carotene, and log(0) is undefined. As seen in figure 1, the linear relationship seems to be more evident between age and log(beta-carotene) compared to age and beta-carotene. The residuals of the linear and the log-linear model were compared. From the residuals plot in figure 2 and 3, it is also clear that the log-transformed beta-carotene model performs better than its counterpart.

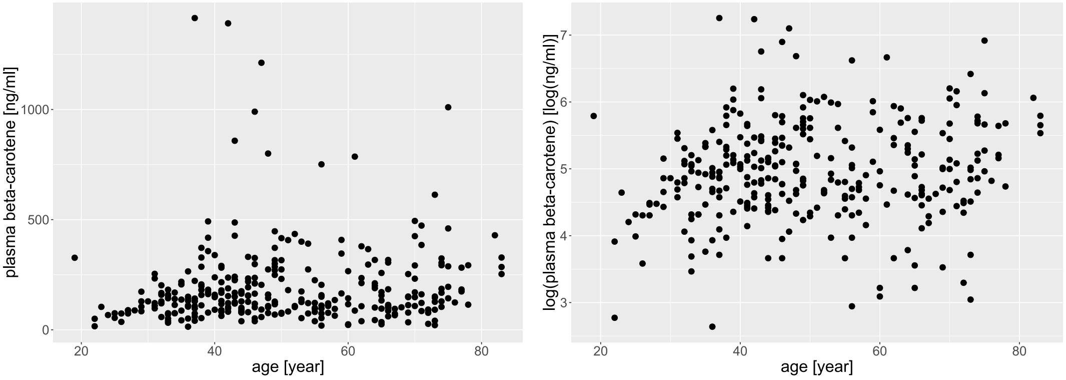


Figure 1: Plots of beta-carotene and log(beta-carotene) against age.

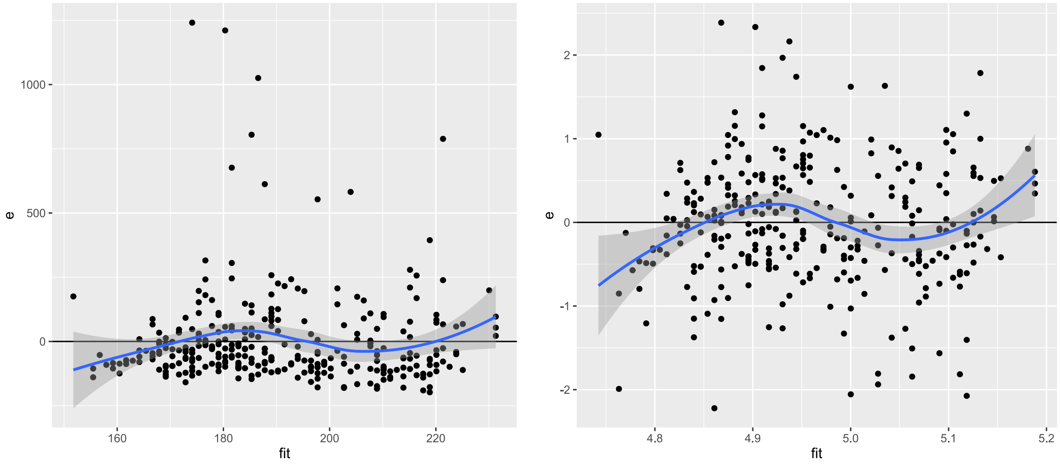


Figure 2: Residuals against fitted values for the linear and log-linear model.

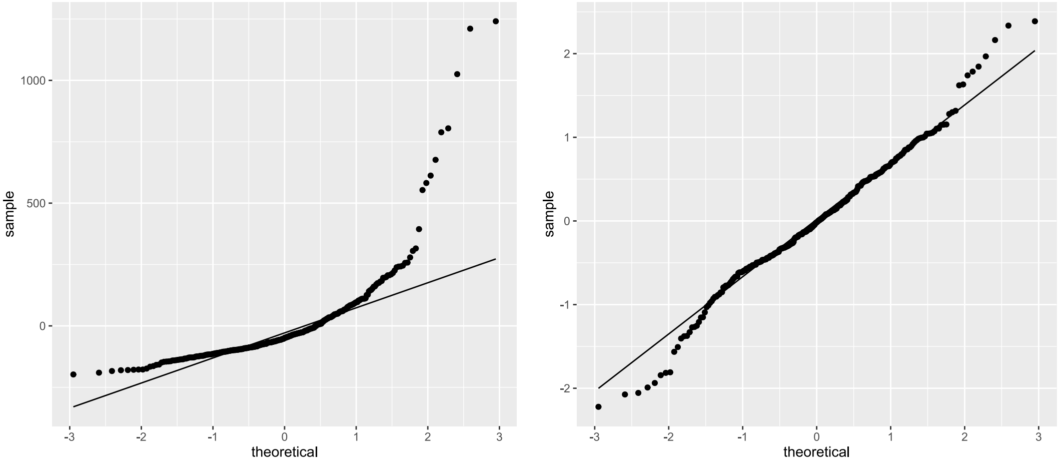


Figure 3: QQ plots of residuals for linear and log-linear model.

b) Since the log-transformed model performs better according to the residuals analysis, it will be used throughout the report. Figure 4 plots log plasma against age, in addition to the fitted line of the following log linear model:

$$\log(Y) = \beta_0 + \beta_1 X \tag{1}$$

where Y is beta-carotene blood plasma levels and X is age.

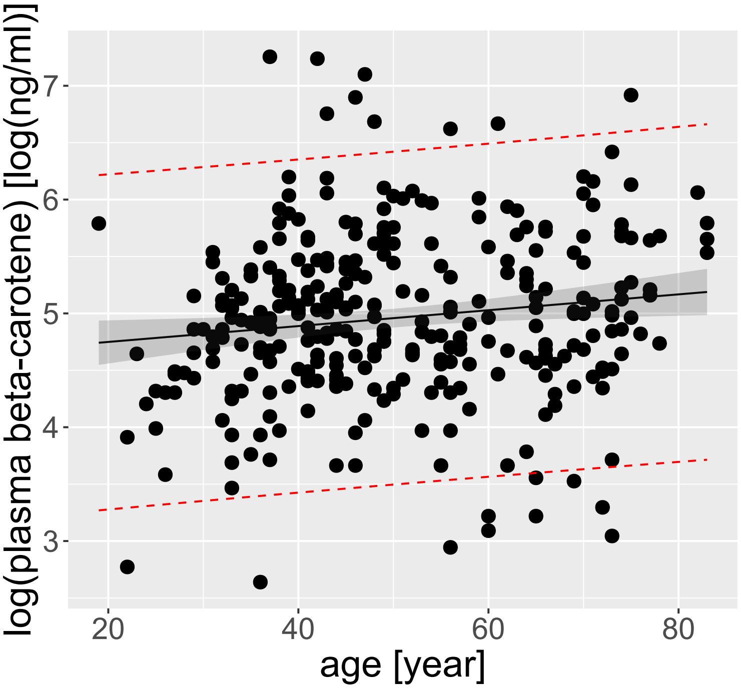


Figure 4: Log plasma against age with fitted line, confidence and prediction intervals.

Since β_1 is 0.006966, on average the yearly increase in beta-carotene plasma levels is roughly, $e^{0.006966} \approx 1.007$, with a 95% confidence interval of $(1.001, 1.013)$.
The additive difference in expected plasma beta-carotene levels can be calculated the estimated coefficients:

$$y_2 - y_1 = e^{\beta_0}(e^{\beta_1 x_2} - e^{\beta_1 x_1}) \tag{2}$$

where y_2 is the carotene level for the age x_2 and y_1 is the carotene level for the age x_1. This relation clearly shows the age dependency of the additive difference and is further exemplified in figure 5. Where, the additive difference between a 30-year old and a 31-year old is 0.865651. The same difference between a 70-year old and a 71-year old is 1.143803.

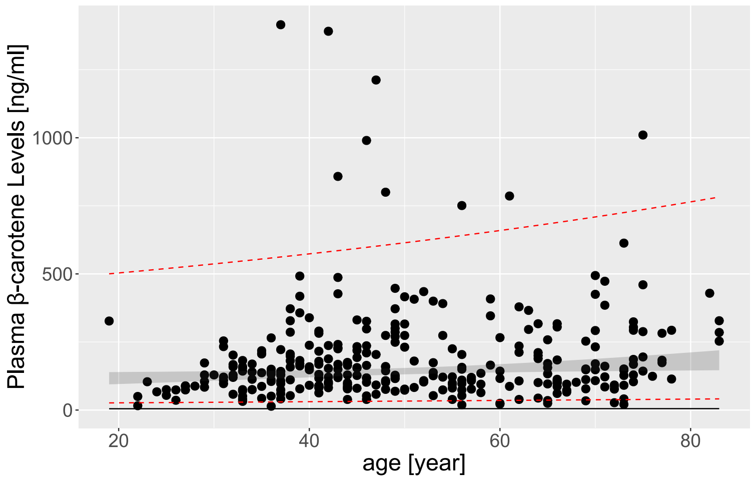


Figure 5: The beta-carotene plasma levels against age with the log transformed models' fitted line, confidence and prediction intervals.

c) The width of the prediction interval for beta-carotene level for a 30-year-old is 2.933221. The same width for a 70-year-old is 2.93291, which shows that they have no substantial difference.

Part 2. Plasma Beta-carotene and the background variables

Other than age, there are also the variables smoking, sex, and BMI whose impact on beta-carotene levels can be measured. Smoking is a categorical variable with options *never*, *former*, and *current*. For sex there is only *female* or *male*, while BMI can be used as a continuous variable. It can, however, also be used as a categorical variable with the options *underweight*, *normal*, *overweight*, and *obese*.

a) After the categorical variables are turned into factors, their frequency tables can be output from R:

Table 1: Summary frequency table for all categorical variables.

Gender	Male				Female			
BMI	Underweight	Normal	Overweight	Obese	Underweight	Normal	Overweight	Obese
Never Smoked	0	5	7	1	2	75	28	38
Former Smoker	0	11	7	4	0	48	30	15
Current Smoker	0	3	2	2	2	18	15	1

For each categorical variable, it is most appropriate to choose the category with the most patients as a reference. For BMI this is *normal*, for sex this is *female*, and for the

Table 2: Frequency table for type of smokers among patients.

Smoker	Never	Former	Current
Amount	156	115	43

Table 3: Frequency table for sex among patients.

Sex	Male	Female
Amount	42	272

Table 4: Frequency table for patients' BMI categories.

BMI	Underweight	Normal	Overweight	Obese
Amount	3	160	89	61

smoking category this is *never*. These can be designated to be the reference category by using the `re-level` function in R.

b) To illustrate that re-leveling to these more populated categories impacts the model's accuracy, a model is fitted where beta-carotene plasma levels depend only on the categorical variable BMI. In the model, as mentioned earlier, the default *underweight* BMI is used as the reference category before it is changed to *normal*. As can be seen from tables 5 and 6, re-leveling leads to lower errors throughout all categories apart from underweight, which is to be expected if it was first used as a reference category.

Table 5: The impact factor of the BMI variable where underweight is taken as a reference.

	Estimate	Std. Error
(Intercept)	5.3602	0.3615
Normal	-0.2324	0.3660
Overweight	-0.4870	0.3695
Obese	-0.7421	0.3732

Table 6: The impact factor of the BMI variable where normal is taken as a reference.

	Estimate	Std. Error
(Intercept)	5.128	0.05716
Underweight	0.2324	0.3660
Overweight	-0.2545	0.09560
Obese	-0.5097	0.1088

c) Given the findings in (b), all categorical variables were re-leveled to their most populace category and subsequently fit with a model that incorporates age, smoking status, BMI, and sex:

$$log(Y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 \tag{3}$$

Where Y is the beta-carotene blood plasma levels, β_0 is the intercept. Thus the mean of a normal BMI individual, the subsequent subscripts are for age, underweight individuals, overweight individuals, obese individuals, former smokers, current smokers, and male sex, respectively.

The resulting β and e^{β} coefficient estimates, alongside their confidence intervals at a 5% significance level can be found in tables 7 and 8, respectively. This new model has a p-value of 3.786×10^{-9}, from which it can be concluded that this model is significantly better than a model with only an intercept. Further testing with the model reveals that the Underweight β coefficient has a p-value of 0.388193, and thus, it is concluded that it is insignificant compared to all other variables. Additionally, an analysis of variance reveals that BMI and smoking status are significant to determining the beta-carotene plasma levels, whereas sex and age are not. Moreover, when comparing this new model to the one used in Part 1, using an analysis of variance test, it is found that the new model performs significantly better, as the p-value of the test is 1.763×10^{-8}.

Table 7: The β coefficient estimates alongside their confidence intervals at a 5% significance level.

Parameter	β estimate	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	4.896	4.583	5.208
Age	0.007462	0.001832	0.01309
Underweight	0.30667	-0.3917	1.005
Overweight	-0.2173	-0.3996	-0.03512
Obese	-0.5477	-0.7549	-0.3406
Former Smoker	-0.1082	-0.2796	0.06320
Current Smoker	-0.4493	-0.6946	-0.2040
Male	-0.3391	-0.5792	-0.09895

Table 8: The e^β coefficient estimates alongside their confidence intervals at a 5% significance level.

Parameter	e^β estimate	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	133.7143955	97.8021406	182.8133767
Age	1.0074904	1.0018332	1.0131795
Underweight	1.3589125	0.6759210	2.7320397
Overweight	0.8046530	0.6706075	0.9654925
Obese	0.5782537	0.4700572	0.7113545
Former Smoker	0.8974655	0.7561136	1.0652425
Current Smoker	0.6380574	0.4992583	0.8154443
Male	0.7124070	0.5603093	0.9057922

d) Plotting this model, shown in figure 6, reveals that there are no underweight males included in the dataset, nor are there any underweight former smokers of any sex. Hence, the predictive power of the model within this sub-group is unreliable.

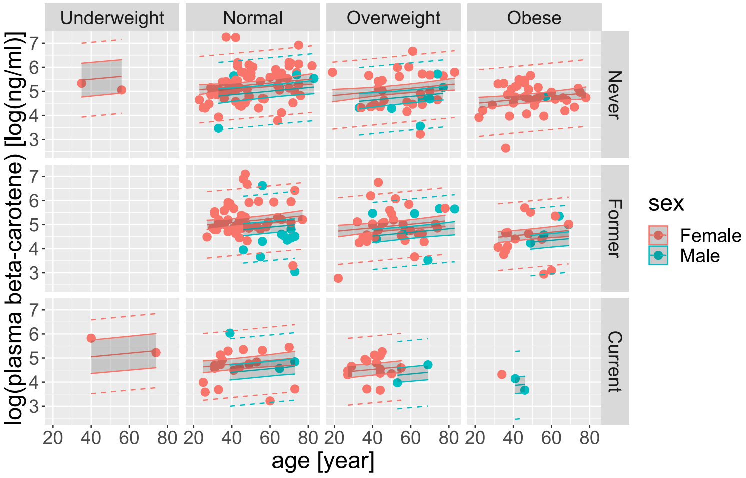


Figure 6: Log plasma against age seperated according to BMI and smoking status with fitted lines, confidence and prediction intervals for each sex.

From the data, it is concluded that for an underweight male who formerly smoked and is 62 years of age, their beta-carotene plasma levels would be 5.210 according to the model. This predicted value has an interval width of 1.475 across the confidence interval, this is within the average width for male confidence intervals, and thus it is a reasonable prediction.

Furthermore, as BMI increases, the plasma beta-carotene levels decrease, which is emphasized by the fact that the coefficient is positive for underweight individuals. In contrast, overweight and obese individuals have negative coefficients, with the obese coefficient being the most negative value.

e) The model fitted with age, sex, smokstat, and quetelet (instead of the categorical BMI) have similar β coefficients to the model with BMI in section (c), with the exception that the β coefficients associated with categorical BMI, $beta_{2,3,4}$, are replaced by a single coefficient associated with the quetelet; this coefficient is always ”on” as the data is numerical rather than categorical. In table 9 and 10 are the β and e^{β} of the model along with their confidence intervals. These tables can be compared with table 7 and table 8 to show that the other coefficients other than BMI and quetelet are not too different for the two models.

Table 9: The β coefficient estimates alongside their confidence intervals at a 5% significance level when using quetelet instead of BMI.

Parameter	β estimate	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	5.705	5.241	6.169
Age	0.0074	0.0019	0.0130
Quetelet	-0.0371	-0.0499	-0.0242
Former Smoker	-0.1149	-0.2840	0.0542
Current Smoker	-0.4514	-0.6912	-0.2116
Male	-0.3436	-0.5816	-0.1056

Table 10: The e^{β} coefficient estimates alongside their confidence intervals at a 5% significance level when using quetelet instead of BMI.

Parameter	e^{β} estimate	Lower Confidence Interval	Upper Confidence Interval
(Intercept)	300.3	188.9	477.5
Age	1.007	1.002	1.013
Quetelet	0.9636	0.9513	0.9761
Former Smoker	0.8915	0.7528	1.056
Current Smoker	0.6367	0.5010	0.809
Male	0.7092	0.5590	0.8998

Both models with discrete BMI and continuous BMI (Quetelet) are used to predict the average of log plasma beta-carotene of male and female both 30-year-old former smoker with BMI of 20, or normal for the discrete BMI. The result is presented in table 11 along with the confidence intervals.

Table 11: The average beta-carotene and their confidence interval for a normal, formerly smoking male and female that are 30 years of age using discrete (BMI category) and continuous models (quetelet).

	Male	Female
BMI	106.9 (79.55, 143.8)	150.1 (124.9 , 180.5)
Quetelet	113.1 (84.22, 151.9)	159.5 (132.4, 192.1)

The confidence intervals for the man are wider than the corresponding intervals for the woman, as there are fewer samples for men in this study.

For an obese individual, the plasma beta-carotene levels decrease the predicted values along with the confidence intervals are reported in table 12.

Table 12: The average beta-carotene and their confidence interval for an obese, formerly smoking male and female that are 30 years of age using discrete (BMI category) and continuous models (quetelet).

Gender	Male	Female
BMI	61.84 (44.71, 85.54)	86.80 (68.84, 109.4)
Quetelet	72.49 (54.25, 96.86)	102.2 (85.02, 122.9)

The relative difference between a healthy individual and an obese one is shown in table 13 and for the BMI model was calculated as,

$$\frac{Y_O - Y_N}{Y_N} = \frac{e^{\alpha+\beta_4 x_O} - e^{\alpha}}{e^{\alpha}} \tag{4}$$

Where β_4 is the coefficient associated with obese individuals and α are all other relevant parameters. Consequently, equation 4 can be expressed as,

$$\frac{Y_O - Y_N}{Y_N} = e^{\beta_4} - 1 \tag{5}$$

As x_O is equal to 1 as it is a categorical variable in the discrete BMI model. Similarly, the relative difference for the quetelet model is:

$$\frac{Y_O - Y_N}{Y_N} = \frac{e^{\alpha+\beta_2 x_O} - e^{\alpha+\beta_2 x_N}}{e^{\alpha+\beta_2 x_N}} = e^{\beta_2(x_O - x_N)} - 1 \tag{6}$$

Table 13: The relative difference and their confidence intervals between an obese and normal individual who are, formerly smoking and 30 years of age using discrete (BMI category) and continuous models (quetelet).

	Male	Female
BMI	0.01617 (0.02237, 0.01169)	0.01152 (0.01453, 0.009137)
Quetelet	0.01379 (0.01843, 0.01032)	0.009784 (0.01176, 0.008138)

f) When both quetelet and BMI are used in the model, the coefficients of BMI and quetelet are not significant. The lack of significance likely arises since both variables, describe the same statistic and are thus, collinear, which allows "important" variables to be replaced, as their effects are distributed across to β values.

Part 3. Model validation and selection

In this part, the best model is determined. Furthermore, problematic observations are investigated.

a) The models in 2 c) and e) only differ in how they use the BMI-variable: categorical or continuous. As previously stated, this difference went almost unnoticed by the other coefficients of the two models. It is, however, still possible to compare the two versions. Because they are not nested, partial F-tests cannot be used. Instead, the fraction of the variability of the observations explained by the regression model can be calculated. The variability explained in the model is called the coefficient of determination, R^2, and it increases with more parameters. Because our two models do not have an equal number of parameters, it is better to calculate R^2_{adj}. According to this, the best model would be the one with the higher R^2_{adj}. For the categorical model, $R^2_{adj,cat} = 0.1385$, and for the continuous model, $R^2_{adj,con} = 0.1506$, favouring the second of the two.

A similar method compares the Akaike Information Criterion (AIC) or the Schwarz Bayesian Criterion (BIC) for the two models. This is justified here because $n \gg p$ and yields $AIC_{cat} = 671.6$, and $AIC_{con} = 665.2$, $BIC_{cat} = 705.3$, and $BIC_{con} = 691.4$. For both these criteria, the model with a lesser value should be preferred, which is the second one in both cases. Together with the R^2 result, we conclude to choose the continuous model as our preferred *Background* model. The model from Part 1 with only age dependency will be referred to as the *Age* model.

b) Before the model can be extended to include all dietary variables, their correlations with each other need to be analyzed. Including two highly correlated variables would not add new information but instead increase the noise. After all, it can be expected that, for example, eating more fat is correlated with higher calorie intake (unless patients would eat little else but fat). This hypothesis is what calculations in R confirm: fat and calorie intake have a correlation of 0.8718, and fat and cholesterol have a correlation of 0.7098; All other correlations are below 0.7.

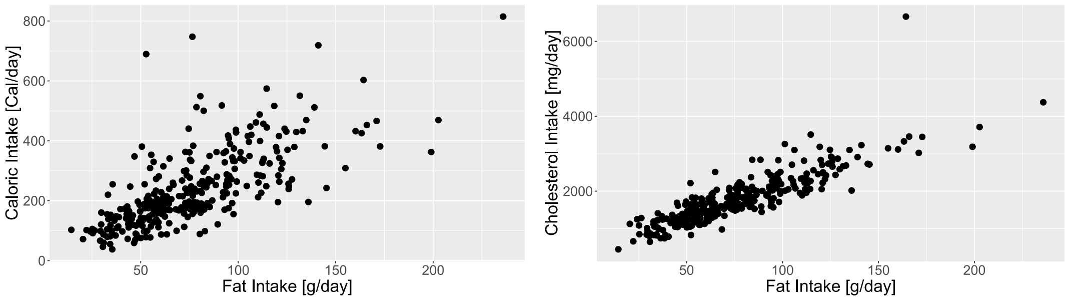


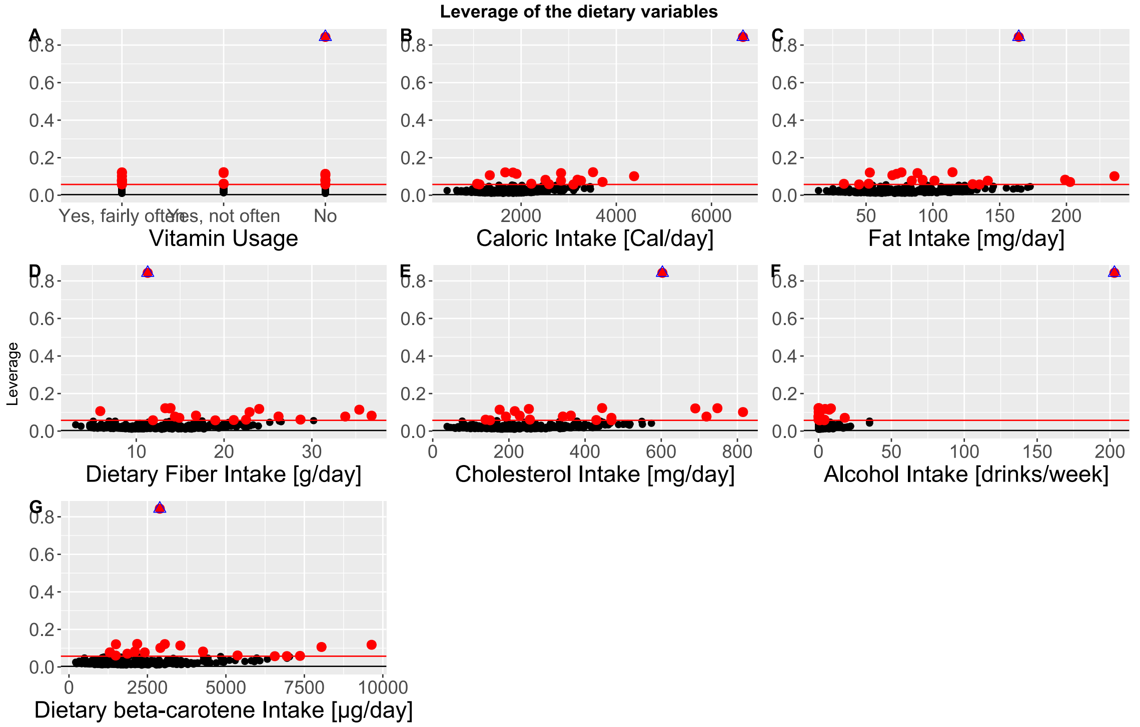
Figure 7: Plots of calorie intake and cholesterol levels as functions of fat intake.

One more categorical variable is vitamin use. From the frequency table 14, it can be seen that *Yes, fairly often* is a suitable reference category.

Table 14: Frequency table for vitamin use among patients.

Vitamin use	Female	Male
Yes, fairly often	108	13
Yes, not often	77	5
No	87	24

c) Once a model with the covariates vitamin-, calories-, fat-, fiber-, alcohol-, cholesterol-, and dietary beta-carotene- intake is fitted, the leverages for all these can be calculated. The result is seen in figure 8, which indicates that one individual is an extreme outlier. This patient consumes 203 alcoholic drinks per week and can thus safely be excluded, as this is an illogical level of alcohol intake.



y = 1/n (black) and 2(p+1)/n (red)

Figure 8: The leverages for all dietary covariates where one alcoholic individual's leverage is highlighted with a blue triangle in each plot. In red are all the leverages $> 2(p+1)/n$, the latter indicated by a red horizontal.

It is tempting to use the logarithm of alcohol consumption, but this is a bad idea as we would lose 110 non-drinkers' data. That said, we generally take the log (or any transform) to normalize the distribution of the variable and stabilize its variance. We can see that the alcohol intake is indeed non-normally distributed, but its p-value is 0.1180 and, thus, rather irrelevant. Consequently, eliminating this individual from the dataset is likely a better recourse. However, it was decided to keep this outlier in the model and investigate the potential problem it causes.

The extreme alcoholic is extreme also in caloric, fat, and cholesterol intake per day, as can be seen from figure 9 (but these variables are, of course, correlated), which indicates that the patient's extremeness is not just a result of an experimental error.

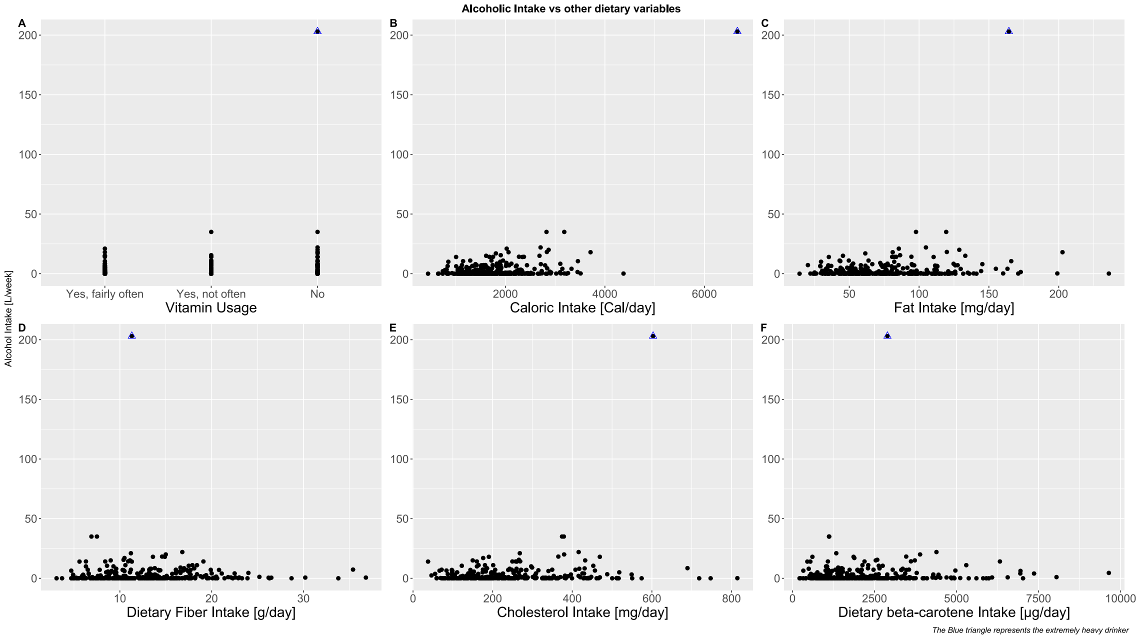


Figure 9: The alcohol consumption against each of the other dietary variables.

d) Figure 10 lends itself to a visual inspection of the studentized residuals. The outlier does not behave extremely in the covariates (i.e., no heavy alcoholic, no one very obese, etc.), but it is saved for future reference.

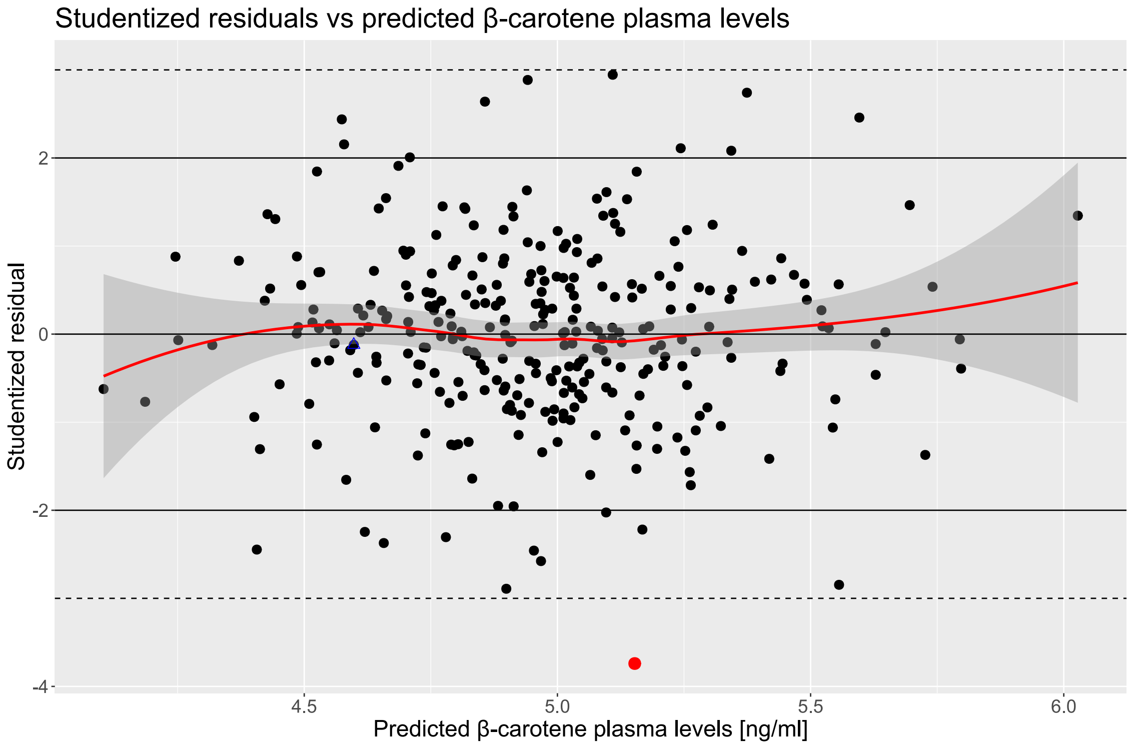


Figure 10: The studentized residuals against predicted values with one residuals < -3 marked in red and the alcoholic with a blue triangle.

e) Naturally, some of the patients affected the models' prediction more than others. Hence, their Cook's distance is plotted in figure 11. The largest residual did not contribute more than all the others. Even the alcoholic did not have the strongest contribution.

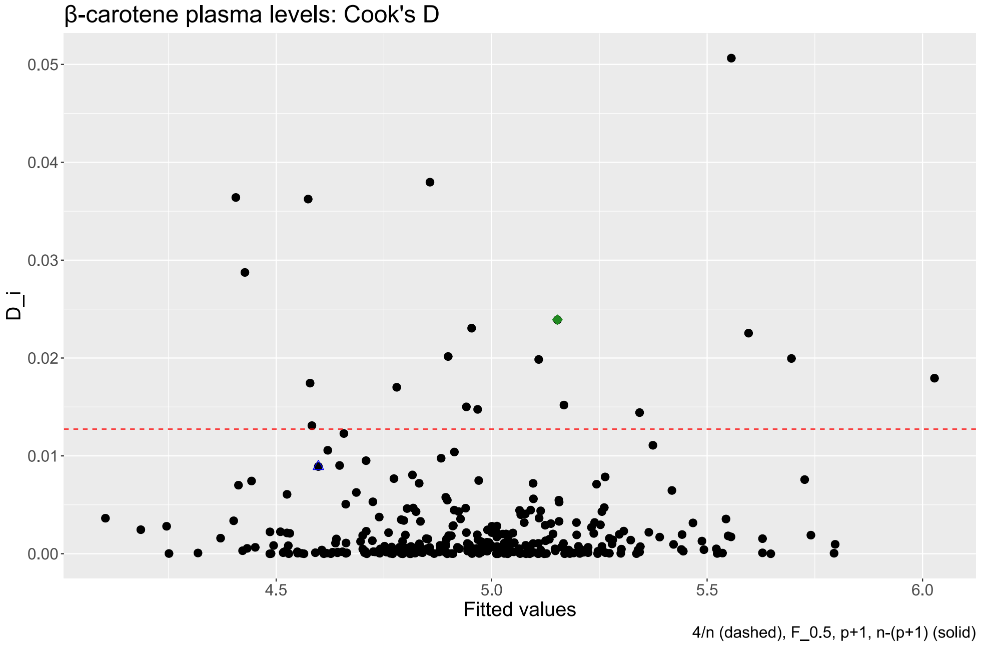


Figure 11: Cook's distance vs. fitted values. Only the dashed horizontal line can be seen. The line marking the upper quartile is way above the highest contributor and not shown here. Marked in blue is the candidate with the highest leverage, an alcoholic. The green diamond highlights the patient with the biggest studentized residual.

From figure 12 it can be seen clearly that the alcoholic only disproportionally affects the parameter taking into account the contribution from alcohol. However, in this case, its influence is not really worrying. The residual outlier greatly influences only the parameters accounting for the impact of often taking vitamins.

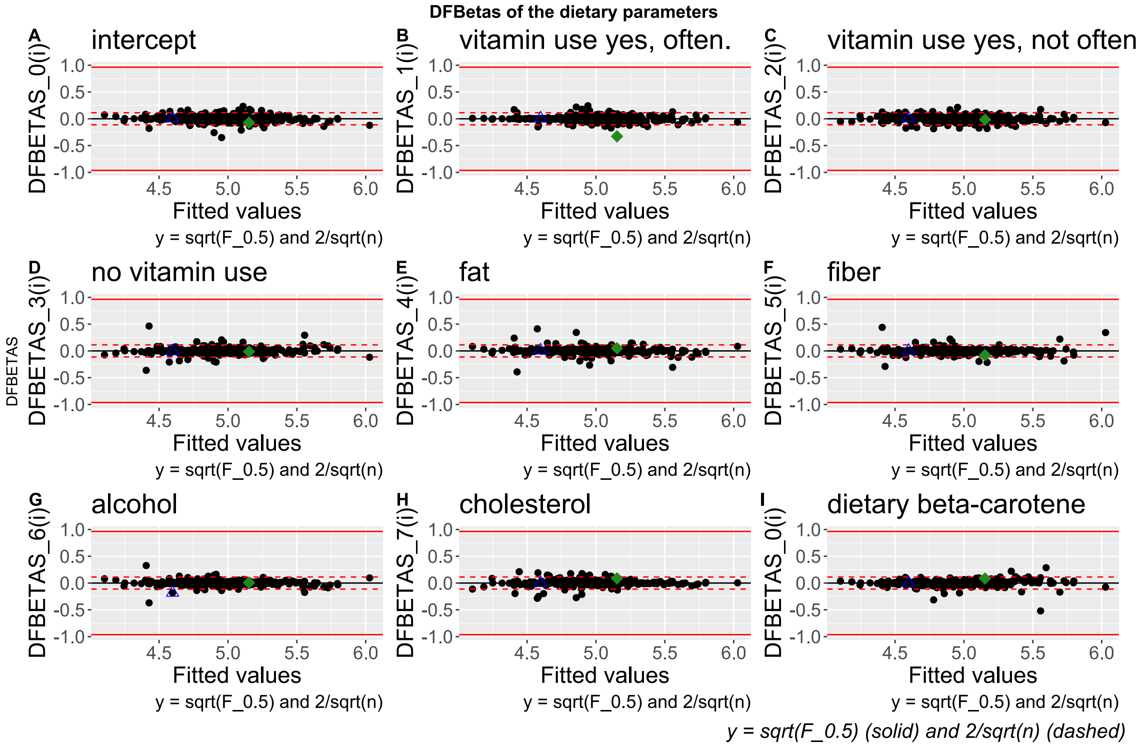


Figure 12: DFBETAS for the different β-parameters. The alcoholic (the blue triangle) only greatly influences the parameter that accounts for the contribution from alcohol. The residual outlier is marked in green; it influences more the parameter from frequent vitamin users.

f) Starting from the model with all the dietary variables, we use backward elimination to improve the model. In each step, the variable in the model whose deletion would cause the largest decrease in AIC is removed. The procedures stop once no such variable is found anymore.

This procedure first eliminates fat and then cholesterol. Since these two variables are highly correlated, it is not surprising that both are eliminated. This model is given the name *Dietary* and has the form:

$$Y = \beta_0 + \beta_1 x_{vitamin,notoften} + \beta_2 x_{novitamin} + \beta_3 x_{calories} + \beta_4 x_{fiber} + \beta_5 x_{alcohol} + \beta_6 x_{betadiet}. \tag{7}$$

Table 15 presents the resulting β estimates together with the e^{β} estimates and their 96 % confidence intervals.

g) The recently found *Dietary* model can now be combined with the *Background* model, which is done in a stepwise procedure starting from the *Dietary* model. At each new step, the importance of all previously included and excluded variables is checked. Each step can be either backward or forward, depending on which action causes the largest AIC or BIC decrease. The procedure is carried out with both

Table 15: The estimated parameters for the step-wise reduced dietary model.

	β	e^{β}	confidence interval
β_0	5.002371	148.7654250	(113.2728984, 195.3790536)
β_1	-0.09391361	0.9103614	(0.7487516, 1.1068530)
β_2	-0.3867932	0.6792316	(0.5664008, 0.8145389)
β_3	-0.0003092907	0.9996908	(0.9995398, 0.9998417)
β_4	0.04318153	1.0441274	(1.0246356, 1.0639901)
β_5	0.007076871	1.0071020	(0.9996837, 1.0145753)
β_6	0.0004459385	1.0000446	(0.9999847, 1.0001045)

Table 16: Combination of *Dietary* and *Background* step-wise gives two different results depending on whether AIC or BIC is chosen as the criteria.

	AIC		BIC	
	e^{β}	confidence interval	e^{β}	confidence interval
(Intercept)	247.571659	(145.4203385, 421.4797416)	316.3236842	(208.3397089, 480.2765336)
vitamin: Yes, not often	0.9783293	(0.809265, 1.1827124)	0.9331349	(0.7712844, 1.1289489)
vitamin: No	0.7537073	(0.6318503, 0.8990653)	0.7051766	(0.5915484, 0.8406312)
calories	0.9998720	(0.9997384, 1.0000055)	0.9997823	(0.9996571, 0.9999075)
fiber	1.0269571	(1.0088069, 1.0454339)	1.0415512	(1.0249823, 1.0583879)
betadiet	1.0000476	(0.9999900, 1.0001051)		
quetelet	0.9676049	(0.9556065, 0.9797539)	0.9703891	(0.9582834, 0.9826477)
former smoker	0.9183594	(0.7796123, 1.0817991)		
current smoker	0.7513332	(0.5918417, 0.9538051)		
male	0.7890731	(0.6203950, 1.0036128)		
age	1.0060314	(1.0004039, 1.0116904)		

criteria, and their final results are compared to each other in table 16. It is expected that the AIC criteria would lead to more β parameters because the BIC is stricter on the significance of the variables. For this reason, the AIC criterium leads to a model with higher predictive power, but it might be less "true" as compared to the BIC result.

h) To determine the optimal model, the coefficient of determination, alongside its adjusted value, is computed for each of the models and is shown in table 17. The best-in-class model explains as much of the variability as is practical, which is to say that it maximizes the coefficient of determination, R^2. However, for models with differing numbers of parameters, the R^2 value always increases when adding covariates. Hence, the adjusted coefficient of determination, $R^2_{adj.}$, should be used in the comparison.

Table 17: The coefficient of determination (R^2) and the adjusted coefficient of determination ($R^2_{adj.}$ of the five models.

Model	R^2	$R^2_{adj.}$
Age	0.01847	0.01532
Background	0.1642	0.1506
Dietary	0.1660	0.1441
Step AIC	0.2440	0.2190
Step BIC	0.2016	0.1886

From table 17, the variability of the logarithmically transformed beta-carotene plasma levels of the background models' variables is approximately 16.42%, whereas that of the dietary models' variables is roughly 16.60%. However, both models are eclipsed by the Step AIC model, whose adjusted coefficient of determination is 0.2190. Thus, it is selected as the *Final* model.

Conclusion

In conclusion, by residual analysis, t-testing, partial F-tests, global F-tests the *Background* model was selected starting from the *Age* model. From that point, the models were altered by stepwise regression using both AIC and BIC as their criteria. Finally, the R^2 and $R^2_{adj.}$ values were computed, and the *Final* model was determined to be the Step AIC model, which incorporates the vitamin usage, caloric intake, fiber intake, dietary beta-carotene consumed, quetelet, smoking status, sex, and age of an individual.