Linear Regression - Determinants of plasma beta-carotene levels

15 April 2025 - Group 31 - FMSN30 Linjär och Logistisk Regression : Lunds Tekniska Högskola

Name	Roles: All did discussions, programming, visualisation, writing
Mattis Ranheim	Mainly parts 1-2. Writing for part 3
Madeleine Ekstrand	Mainly part 4. Programming for part 3
Yassin Hjuler El Mahdaoui	Mainly part 5. Programming for part 3

Use of AI Tools

AI tools were used to assist in writing and coding: - **ChatGPT (OpenAI)** was used to clarify statistical concepts, draft parts of code, and suggest text structure. - All code was reviewed, understood, and adapted by the authors. Output was carefully verified for correctness.

Spelling and grammar suggestions from **RStudio Visual Editor** were used.

Introduction

Numerous observational studies have suggested that low dietary intake or low plasma concentrations of β-carotene and other carotenoids may be linked to an increased risk of developing certain types of cancer. However, relatively few studies have examined which factors actually influence plasma concentrations of these micronutrients.

In this project, we analyze data from a cross-sectional study conducted by Nierenberg et al. (1989), where the goal was to investigate the relationship between **personal characteristics**, **dietary intake**, and **plasma concentrations of β-carotene**. The study population consisted of 315 patients who underwent elective surgical procedures to biopsy or remove benign (non-cancerous) lesions in organs such as the lung, colon, breast, skin, ovary, or uterus. For this analysis, we focus exclusively on **plasma β-carotene concentrations** as the outcome of interest.

The study highlights considerable individual variation in plasma β-carotene levels and suggests that much of this variability may be explained by lifestyle and dietary factors.

Data Description

The dataset used in this project contains **315 observations** and **12 variables**, stored in the file `carotene.xlsx`. Each row corresponds to an individual patient from the study. The variables are described below:

Table 1: Description of variables in dataset

Variable	Description
`age`	Age (years)
`sex`	Sex (1 = Male, 2 = Female)
`smokstat`	Smoking status (1 = Never, 2 = Former, 3 = Current)
`bmi`	Body mass index (BMI = weight/height^2, kg/m^2)
`vituse`	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
`calories`	Daily calorie intake (MJ)
`fat`	Fat consumed per day (g)
`fiber`	Fiber consumed per day (g)
`alcohol`	Alcoholic drinks per week
`cholesterol`	Daily cholesterol intake (mg)
`betadiet`	Dietary β-carotene intake per day (mg)
`betaplasma`	**Plasma β-carotene concentration (ng/ml) —** this is the **response variable** we aim to model

Our objective is to model how `betaplasma` varies as a function of the other variables using a **linear regression model** of the form:

$$Y_i = \mathbf{x}_i\beta + \varepsilon_i$$

where Y_i is the plasma β-carotene concentration for individual i, $\mathbf{x}_i$ is the vector of explanatory variables, β is the vector of unknown regression coefficients, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are the error terms. To satisfy the linear model assumptions (e.g., normality and homoscedasticity of residuals), we may need to apply **suitable transformations** to the response and/or predictor variables throughout the analysis.

1. Testing Model Assumptions: Linear vs Log-transformed

We fitted two models to examine the relationship between BMI and plasma β-carotene levels:

- **Linear BMI model**: `betaplasma ~ bmi`

- **Log-transformed BMI model**: `log(betaplasma) ~ bmi`

The aim is to assess whether a log-transformation of the outcome variable improves model fit and better satisfies the assumptions of linear regression — particularly **normality of residuals** and **constant variance (homoscedasticity)**.

Below, we compare the two models using **residual plots** and **Q-Q plots** for both. A good model should show no patterns in the residuals vs fitted plot, and the residuals should lie close to the theoretical line in the Q-Q plot.

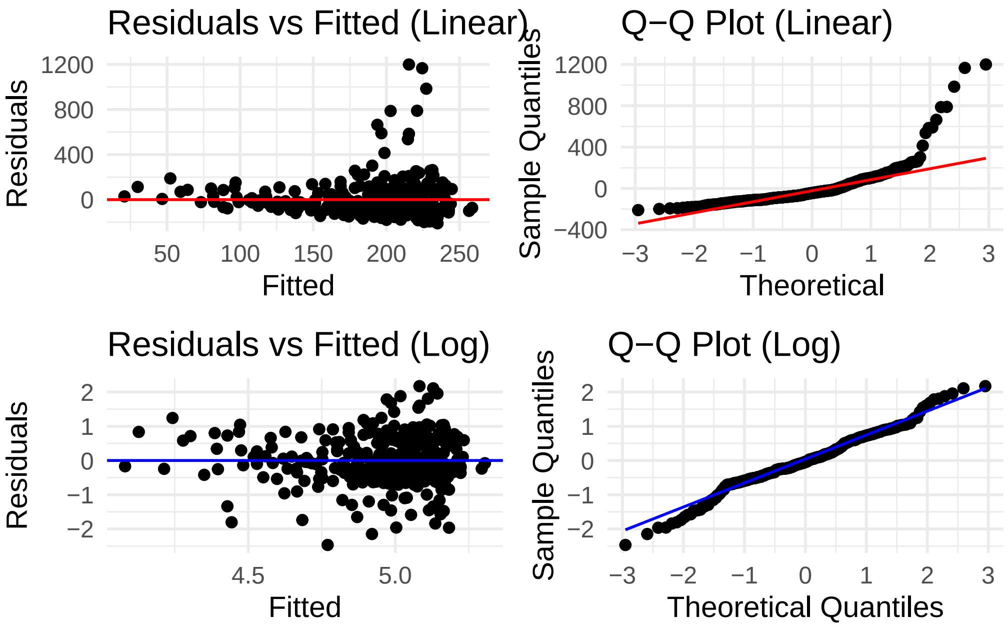

Figure 1: Residual plots and Q-Q plots for linear and log-transformed BMI model

The residual plots and Q-Q plots in figure 1 show that the log-transformed model produces more homoscedastic residuals and better alignment with the normal distribution in the Q-Q plot. In contrast, the residuals of the linear model display signs of heteroscedasticity and heavier tails.

This suggests that the log transformation stabilizes the variance and brings the residuals closer to normality. Therefore, the log-transformed model is more suitable for satisfying the assumptions of linear regression. This model will hence force be referred to as **Model 1.**

1.2 Model Estimates

To interpret the relationship between BMI and plasma β-carotene concentration, we present the coefficient estimates from the **Model 1**:

The table below shows the **β-estimates** and their associated **95% confidence intervals**. The intercept corresponds to the expected value of log(β-carotene) when BMI is hypthetically zero, while the slope for BMI (β1) describes the expected change in the log(β-carotene) concentration for each one-unit increase in BMI.

Table 2: Parameters in Model 1

	Estimate	2.5 %	97.5 %
Intercept	5.8896	5.5273	6.2519
β_1	-0.0359	-0.0494	-0.0224

Table 2 shows a significant negative β_1, meaning a negative association. Which supports the hypothesis that higher body fat may be linked to lower plasma concentrations of this micronutrient.

1.2.2 Linear Model confidence and prediction intervals with Log Transformation and Back-Transformation

- The **top plot** in figure 2 shows the relationship between BMI and the log-transformed β-carotene levels, with fitted line, 95% confidence interval, and 95% prediction interval.
- The **bottom plot** in figure 2 displays the same model but transformed back to the original β-carotene scale (ng/ml).

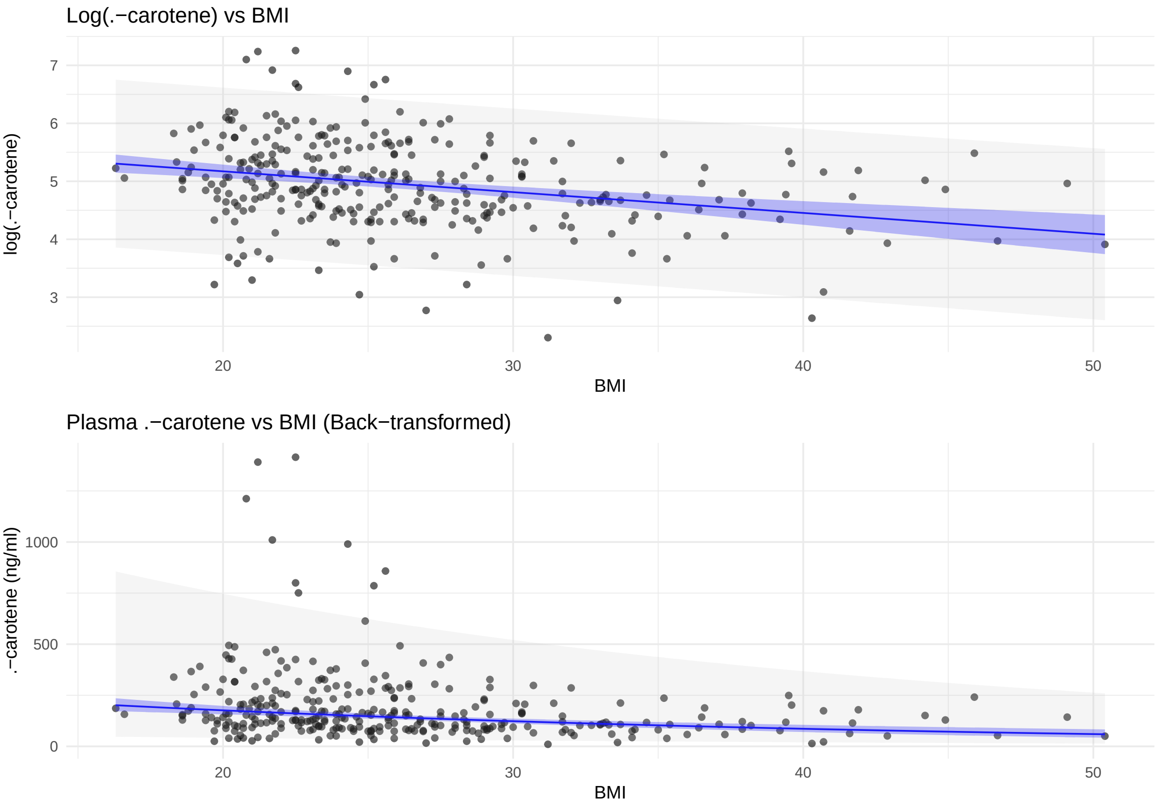

Figure 2: Log-scale and back-transformed plots with 95% CI and prediction intervals.

The log-scale plot shows a clearer linear trend and narrower intervals, while the back-transformed version reveals the actual magnitude of plasma β-carotene and illustrates the asymmetry introduced by the exponential function.

These results suggest that BMI is negatively associated with β-carotene levels, and that using a log transformation is appropriate to model this relationship under the assumptions of linear regression.

1.3 Estimated Changes in Plasma -carotene for Changes in BMI

We interpret the log-linear regression model by expressing expected **percentage changes** in plasma β-carotene (ng/ml) for three different BMI changes. From the calculations, we can conclude the following:

- A **1-unit increase in BMI** is associated with a **3.5% decrease** in plasma β-carotene concentration, with a 95% CI ranging from **−4.8% to −2.2%**.

- A **1-unit decrease in BMI** leads to an **estimated 3.7% increase** in plasma β-carotene concentration, with a 95% CI ranging from **5.1% to 2.3%**.
- A **10-unit decrease in BMI** is associated with a **43% increase**, with a 95% CI ranging from **25% to 64%**.

This nonlinear interpretation stems from the exponential structure of the log-linear model: the relationship between BMI and β-carotene becomes **multiplicative**, not additive.

1.4 Hypothesis Test for Linear Relationship Between BMI and log(-carotene)

We test whether there is a statistically significant linear relationship between BMI and plasma β-carotene concentration on the log scale, based on the following hypotheses

- **Null hypothesis:** $H_0 : \beta_1 = 0$ (no relationship between BMI and log((β-carotene))
- **Alternative hypothesis**: $H_1 : \beta_1 \neq 0$

We use a **t-test** on the slope coefficient in the linear model. The test statistic follows a **t-distribution** with (n - 2) degrees of freedom.

Table 3: Hypothesis test summary for model 1

Test statistic (t)	-5.23
Degrees of freedom	313
Two-sided P-value	3.111e-07

Since the P-value is far below 0.05, we reject the null hypothesis at the 5% significance level. This indicates that BMI is a statistically significant predictor of log-transformed plasma β-carotene levels. The negative t-statistic (t = –5.23) and the negative slope coefficient support an inverse relationship: as BMI increases, the expected log(β-carotene) concentration tends to decrease.

With 313 degrees of freedom, the model has a strong basis for inference, and the very low p-value (3.111e-07) provides robust evidence against the null hypothesis. We therefore conclude that there is both a statistically and practically significant linear relationship between BMI and log(β-carotene).

2. Plasma -carotene and Smoking Habits

To examine how smoking status relates to plasma β-carotene levels, we first recode smokstat as a factor with the levels: 1 = Never smoker, 2 = Former smoker, and 3 = Current smoker. We summarize the data by smoking group in the table 4

Table 4: Summary of Smokstat categories, including frequency in data, mean and standard deviation beta-carotene and log(beta-carotene)

Smokstat	**Count**	**Mean β**	**S.D. β**	**Mean log(β)**	**S.D. log(β)**
Never Smoker	157	206.1146	193.14184	5.050849	0.7453628
Former Smoker	115	193.4696	191.63952	4.941126	0.7975007
Current Smoker	43	121.3256	78.81163	4.613638	0.6243772

We select **Never smoker** as the reference category for regression modeling. This group is the largest, offers stable baseline estimates, and represents the cleanest comparison point in terms of smoking exposure. We will refer to this model as **Model 2** in subsequent chapters

2.1.3 Boxplots for Smoking Categories

To further assess differences between groups and evaluate the need for transformation, we present boxplots of plasma β-carotene and log-transformed β-carotene across smoking status:

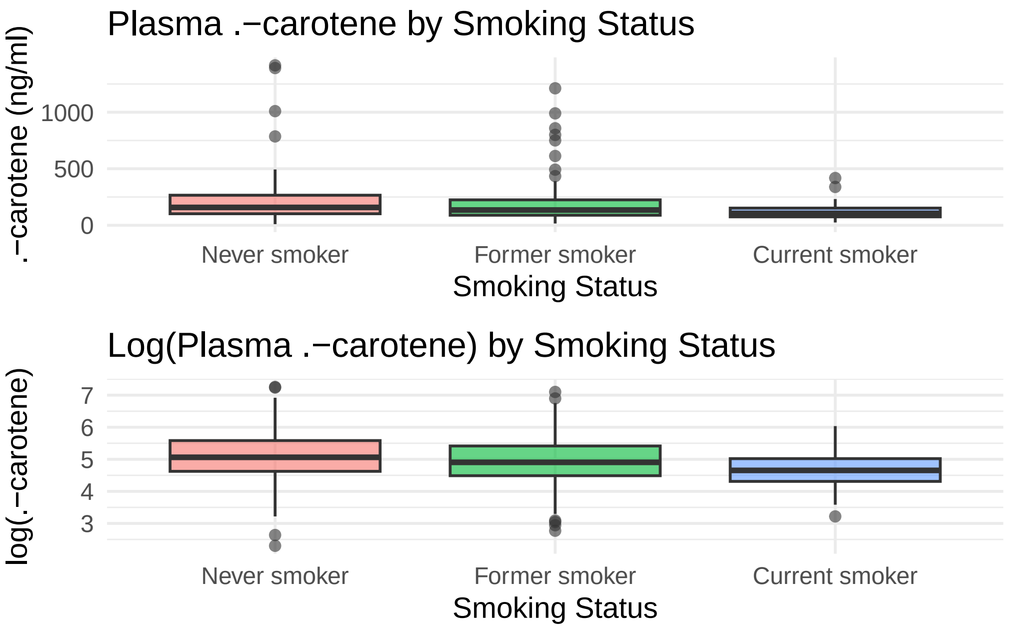

Figure 3: Box plots of β-carotene and log(β-carotene)

The boxplots in figure 3 reveal substantial skewness and variability in the raw β-carotene values, especially among non-smokers and former smokers. The **log-transformed plot** shows a more symmetric and homoscedastic distribution across groups, supporting the continued use of **log(β-carotene)** as the dependent variable in regression modeling.

2.2 Modeling -carotene and Smoking Status

Comparing Reference Categories in Categorical Regression

We investigate how plasma β-carotene levels (on the log scale) differ across smoking status categories by fitting two versions of a linear regression model:

- In **Model 2**, the reference category is **"Never smoker"**
- In **Model 2B**, the reference category is **"Current smoker"**

Table 5: Parameter estimate and standard error in Model 2

Never Smoker = Reference	Estimate	Standard Error
Intercept	5.0508487	0.05986447
$\beta_{FormerSmoker}$	-0.1097225	0.09206715

Never Smoker = Reference	Estimate	Standard Error
$\beta_{CurrentSmoker}$	-0.4372105	0.12910704

Table 6: Parameter estimate and standard error in Model 2 with current smoker as reference

Current Smoker = Reference	Estimate	Standard Error
Intercept	4.6136382	0.1143891
$\beta_{FormerSmoker}$	0.3274879	0.1340801
$\beta_{NeverSmoker}$	0.4372105	0.1291070

- In both models, the intercept represents the mean log(β-carotene) for the reference category, while the other coefficients show differences relative to that group.
- In **Model 2,** the intercept corresponds to "Never smoker", and the coefficients reflect how much lower former and current smokers are on average. In **Model 2B**, the intercept corresponds to "Current smoker", and the coefficients show how much higher the other groups are.
- The standard error of the intercept is larger in **Model 2B**, likely due to fewer observations in the "Current smoker" group, resulting in a less precise estimate.

These results reinforce our decision to choose Never Smoker as the reference in **Model 2.**

2.3 Predicted -carotene Levels by Smoking Group

To further understand the relationship between smoking habits and plasma β-carotene concentration, we now compute predicted values for each smoking group using both **Model 2 and 2B**. We calculate the expectedand 95% confidence interval for **log(β-carotene)** and **β-carotene (ng/ml)**.

Table 7: Estimate and confidence interval for Log(beta-carotene) and beta-carotene

Log(beta-carotene)	2.5 %	Estimate	97.5 %
Never Smoker	4.933	5.051	5.169
Former Smoker	4.803	4.941	5.079
Current Smoker	4.389	4.614	4.839
beta-carotene	**2.5 %**	**Estimate**	**97.5 %**
Never Smoker	138.8	156.2	175.7
Former Smoker	121.9	139.9	160.6
Current Smoker	80.5	100.9	126.3

The predicted values and their 95% confidence intervals are **identical** for both models and we therefore only include one *table 7*. This confirms that **the predictions and their confidence intervals are invariant to the choice of reference level**. Changing the reference category affects the interpretation of the regression coefficients, but not the actual fitted values or predictions.

2.4 Testing for Differences Between Smoking Groups

To evaluate whether smoking status has a statistically significant effect on plasma β-carotene levels (on the log scale), we applied an **ANOVA Global F-test** to compare the group means across the three smoking categories (Never smoker, Former smoker, and Current smoker).

The hypotheses for the test are:

- **Null hypothesis (H_0):** $\mu_1 = \mu_2 = \mu_3$ - All groups have the same mean log plasma β-carotene level.

- **Alternative hypothesis (H_1):** At least one group has a different mean.

The test result gave the following:

- **F-statistic:** 5.75

- **Degrees of freedom:** 312

- **P-value:** 0.00353

Since the p-value is below the significance level of 0.05, we **reject the null hypothesis**. This indicates that there is a statistically significant difference in mean log plasma β-carotene levels among the different smoking status categories.

3.1 Multiple Linear Regression

In this section, we recode the variables *sex* and *vituse* as categorical (factor) variables with meaningful labels. This is essential for regression modeling, where we interpret coefficients relative to a reference category.

The variable `sex` is originally coded numerically (1 = male, 2 = female). We convert it into a factor with labels `"male"` and `"female"`. Similarly, the variable `vituse` (vitamin use) is coded as 1 = Yes, fairly often, 2 = Yes, not often, 3 = No. We recode this to `"often"`, `"seldom"` and `"no"` respectively.

	Male	**Female**	**Often**	**Seldom**	**No**
Frequency	42	273	122	82	111

The frequency table for sex shows that the majority of individuals in the dataset are female. Therefore, setting "female" as the reference category ensures that comparisons are made against the most common group, improving interpretability and often resulting in lower standard errors.

For vituse, the most common category is "often" (vitamin use fairly often). However, from a domain knowledge perspective, "no" (no vitamin use) makes the most sense as a reference group, because it represents the absence of intervention. This allows interpretation of coefficients as the effect of vitamin use relative to no supplementation, which aligns well with research goals.

3.2 Pairwise Correlation Analysis and Outlier Detection

To examine potential multicollinearity and other problems among the continuous predictors, we calculate all pairwise Pearson correlations between the following variables:

- bmi, age, calories, fat, cholesterol, fiber, alcohol, and betadiet

We focus particularly on correlations stronger than ±0.6, which there was shown to be 3 combinations of. These might indicate collinearity issues if both variables are included in the same regression model.

- **Fat** and **Calories**: r ≈ 0.87
- **Cholesterol** and **Calories**: r ≈ 0.66
- **Cholesterol** and **Fat**: r ≈ 0.71

The correlation between these variables is shown in scatterplots in figure 4.

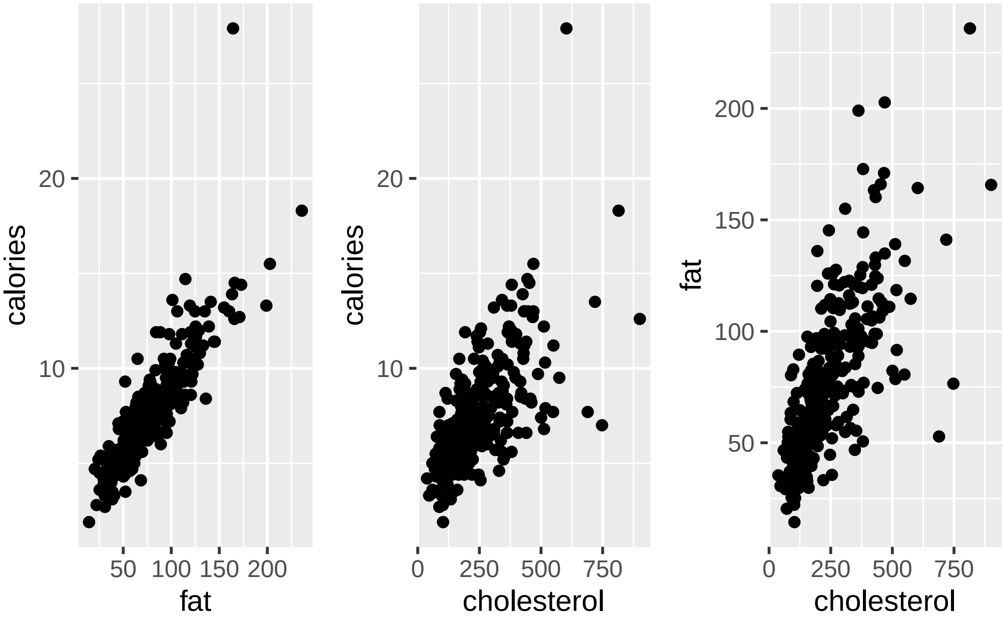

Figure 4: Correlation scatterplots of variables with abs(cor) > 0.6

3.2.2 Outlier Analysis

We also identify the individual who reportedly consumes **over 200 alcoholic drinks per week**, which is flagged as a potential outlier. We examine whether this person is also extreme in other nutritional dimensions:

This individual not only consumes **203 alcoholic drinks per week**, but also has:

- **Cholesterol intake:** 603 mg/day
- **Fat intake**: 164.3 g/day
- **Calorie intake:** 27.9 MJ/day

All of these values are among the highest in the dataset, suggesting this person is an outlier in multiple nutritional variables. This may impact model fitting or residual diagnostics if not accounted for properly.

3.3 Assessing Multicollinearity with VIF

We now examine whether multicollinearity is an issue in a model where log plasma β-carotene is regressed on all available predictors: bmi, age, calories, fat, cholesterol, fiber, alcohol, betadiet, smokstat, sex, and vituse.

To do this, we compute the Generalized Variance Inflation Factor (GVIF) for each variable. We focus on the adjusted GVIF metric:

$$\mathrm{GVIF}^{1/(2\cdot\mathrm{Df})}$$

An adjusted GVIF value above **2.24** indicates that more than 80% of the variance in that variable can be explained by the remaining variables, which suggests problematic multicollinearity.

Table 9: GVIF and adjusted GVIF for all variables

Variable	GVIF	Adjusted GVIF
bmi	1.069660	1.034244
age	1.307586	1.143497
calories	13.210244	3.634590
fat	8.175794	2.859334
cholesterol	2.195956	1.481876
fiber	2.504249	1.582482
alcohol	2.564752	1.601484
betadiet	1.338719	1.157030
smokstat	1.178201	1.041849
sex	1.287887	1.134851
vituse	1.149879	1.035531

As seen in table 9, **calories** and **fat** both exceed the adjusted GVIF threshold of 2.24. This is likely due to their strong correlation with each other and with cholesterol.

To address this, we remove **calories**—the most problematic variable—and refit the model to see whether multicollinearity improves.

Table 10: GVIF and adjusted GVIF for variables without calories

Table 11: As indicated in table 10, after removing calories, all GVIF-adjusted values drop below the threshold, indicating no serious multicollinearity remains. The values for fat and cholesterol remain the highest, but are now within an acceptable range.

Variable	GVIF	Adjusted GVIF
bmi	1.067334	1.033119
age	1.219329	1.104232
fat	2.244437	1.498144
cholesterol	2.129296	1.459211
fiber	1.465013	1.210377
alcohol	1.124159	1.060264
betadiet	1.338004	1.156721
smokstat	1.177041	1.041593
sex	1.287797	1.134811
vituse	1.115890	1.027792

We conclude that removing calories substantially improves the multicollinearity profile of the model. We call this **Model 3.**

3.4 Hypothesis Testing and Model Comparison

We now use **Model 3** to test specific hypotheses about the relationships between log-transformed plasma β-carotene and various explanatory variables. First, we interpret the estimated regression coefficients and their confidence intervals, both on the log scale and in the original β-carotene scale which represent multiplicative effects on the geometric mean of β-carotene)

*Table 11: Parameter estimates and confidence interval in log and original scale for **Model 3***

	Log-scale β	**exp(β)**	**2.5 % exp(β)**	**97.5 % exp(β)**
Intercept	5.2702253504	192.3287664	111.8394983	330.7449956
BMI	-0.0320296627	0.9684779	0.9561460	0.9809688
Age	0.0060022886	1.0060203	1.0003499	1.0117229
Fat	-0.0012301144	0.9987706	0.9954761	1.0020761
Cholesterol	-0.0007325298	0.9992677	0.9984439	1.0000923
Fiber	0.0227289930	1.0229893	1.0058037	1.0404685
Alcohol	0.0018263703	1.0018280	0.9954176	1.0082798
Betadiet	0.0558579827	1.0574475	0.9972458	1.1212835
Smokstat Former	-0.0718425845	0.9306774	0.7877884	1.0994836
Smokstat Current	-0.2728505836	0.7612065	0.5971060	0.9704062
Sex male	-0.2010518069	0.8178701	0.6378043	1.0487723

Vituse often	0.2657965768	1.3044697	1.0899781	1.5611700
Vituse Selfom	0.2670512397	1.3061074	1.0719915	1.5913527

We then examine whether the overall model is significant and conduct **three hypothesis tests**, summarized in table 12

(i) Is there a significant relationship between log plasma β-carotene and BMI in Model 3, adjusting for other variables?

(ii) Is Model 3 significantly better than the Model 1?

(iii) Is Model 3 significantly better than the Model 2?

Table 12: Hypotheses tests testing significance of Model 3 compared to Model 1 and 2

Test	Type	Null Hypthesis	Alternative Hypothesis	Test statistic	P-value	Distribution
i	t-test	$H_0 : \beta_{BMI} = 0$	$H_1 : \beta_{BMI} \neq 0$	-4.918	1.44e-06	t(302)
ii	Partial F-test	$H_0 : \beta_{\text{age}} == \cdots = \beta_{\text{vituse}} = 0$	H_1 : At least one of predictors improves fit	6.2597	2.633e-09	F(11, 302)
iii	Partial F-test	$H_0 : \beta_{\text{age}} == \cdots = \beta_{\text{vituse}} = 0$	H_1 : At least one of predictors improves fit	8.6928	1.63e-12	F(10, 302)

Based on the results in table 12, we can conclude that all null hypotheses should be rejected and that atleast some of the variables introduced in model 3 add value compared to using only BMI or Smokstat.

3.5 Residual Diagnostics for Model 3

To assess the adequacy of **Model 3**, we visually inspect the **studentized residuals** using three standard diagnostic plots:

- A **residuals vs fitted** plot to detect non-linearity and outliers.
- A **spread-location (scale-location)** plot to assess the assumption of **constant variance** (homoscedasticity).
- A **Q-Q plot** to evaluate the **normality** of residuals.

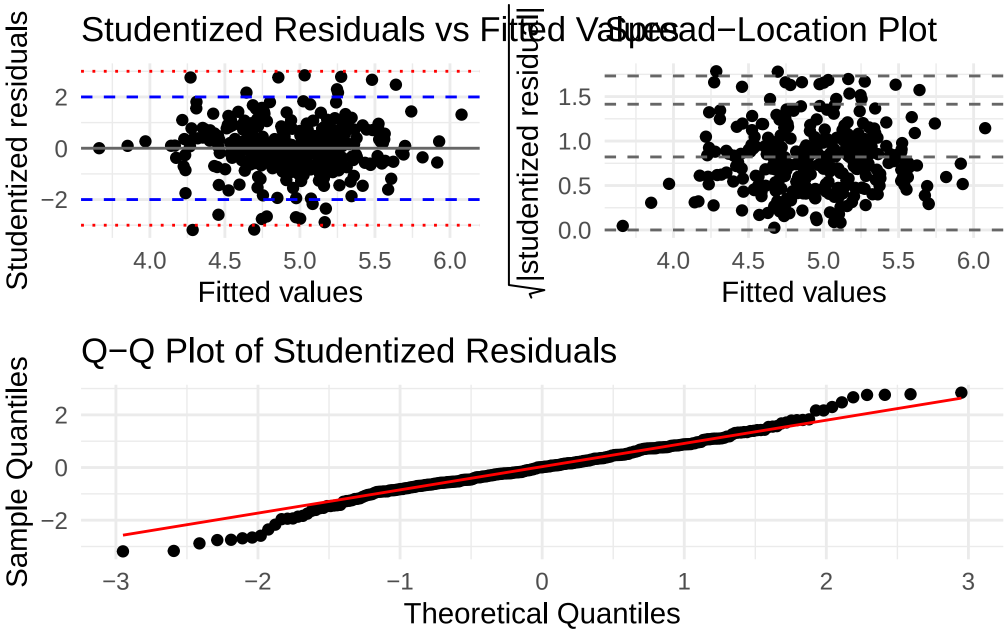

Figure 5: Diagnostic plots for Model 3: studentized residuals vs fitted values, spread-location plot, and Q-Q plot.

The **residuals vs fitted** plot shows no clear pattern, suggesting that the linearity assumption is reasonable. A few residuals fall beyond ±3, but these are not numerous enough to suggest a systemic issue.

The **spread-location plot** indicates that the variance of residuals remains fairly constant across the range of fitted values, supporting the assumption of homoscedasticity.

In the **Q-Q plot**, the points generally follow the red reference line, though some deviation is observed in the tails. This indicates slight departure from normality, but not to a degree that would seriously undermine the model.

Conclusion: Overall, the diagnostic plots in figure 5 support the adequacy of **Model 3**. The residuals show no major violations of linearity, constant variance, or normality assumptions.

3.6. Leverage Analysis for Model 3

We now calculate the leverage values for each observation in **Model 3** and inspect them visually in two diagnostic plots. High-leverage points can exert substantial influence on the model estimates and are often located far from the "center" of the predictor space.

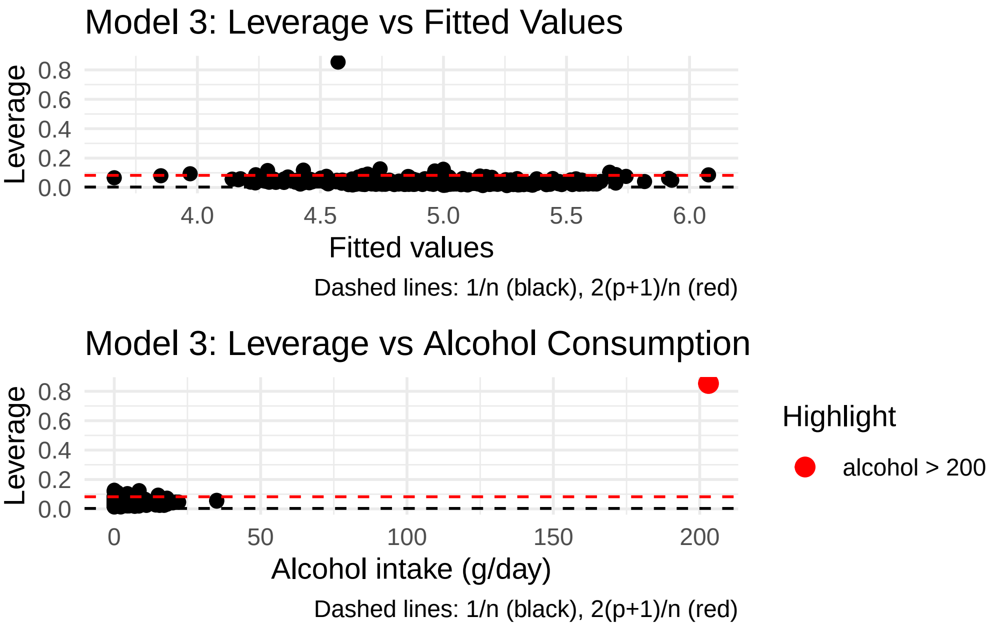

Figure 6: Leverage of observations

In figure 6, the top plot reveals one observation with leverage well above the red line, marking it as potentially influential. When plotted against alcohol consumption, this same observation clearly stands out — it corresponds to the individual consuming over 200 grams of alcohol daily (as previously identified in chapter 3.2.2).

3.7 Influence Diagnostics: Cook's Distance and DFBETAS for Model 3

To evaluate the influence of individual observations on the parameter estimates of **Model 3**, we calculate **Cook's distance** for each observation and visualize it against the fitted values.

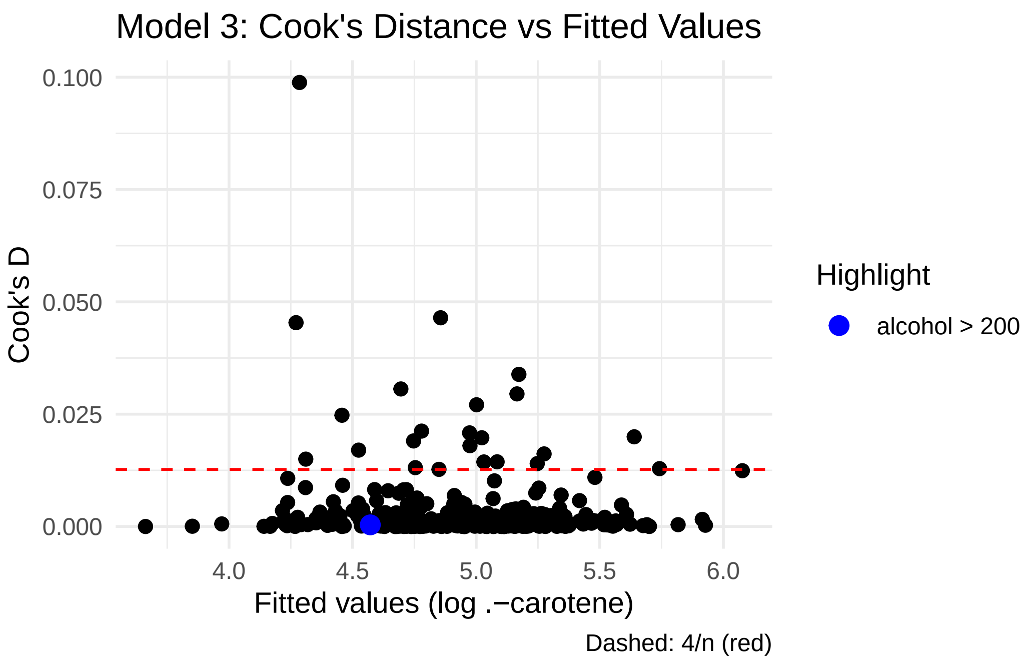

Figure 7: Cooks Distance of observations

The individual with extreme alcohol consumption—over 200g / day—has the **largest leverage**, but has an average and very acceptable **Cook's distance** of 0.000365. This suggests that while the observation is structurally far from the center of the covariate space (i.e., high leverage), it does **not unduly influence the fitted model**, and therefore there is no immediate reason to exclude it.

Instead, we shift our focus to the observation with the **largest Cook's Distance**, which has a value of approximately 0.099. To better understand **which coefficients** are most affected by this point, we inspect the corresponding **DFBETAS**.

Table 13: DFBETAS for observation with largest Cook's distance

Term	**DFBETA**
cholesterol	-0.89848002
betadiet	0.27943057
sexfemale	0.27184195
(Intercept)	0.19952345
fat	0.18684595
VItuseseldom	0.06405373
vituseoften	-0.17770415
smokstatCurrent smoker	0.17106852

smokstatFormer smoker	0.14759318
alcohol	0.09788456

The only DFBETA that is particularly high is cholesterol. We therefore plot the DFBETAS for Cholesterol on all observations:

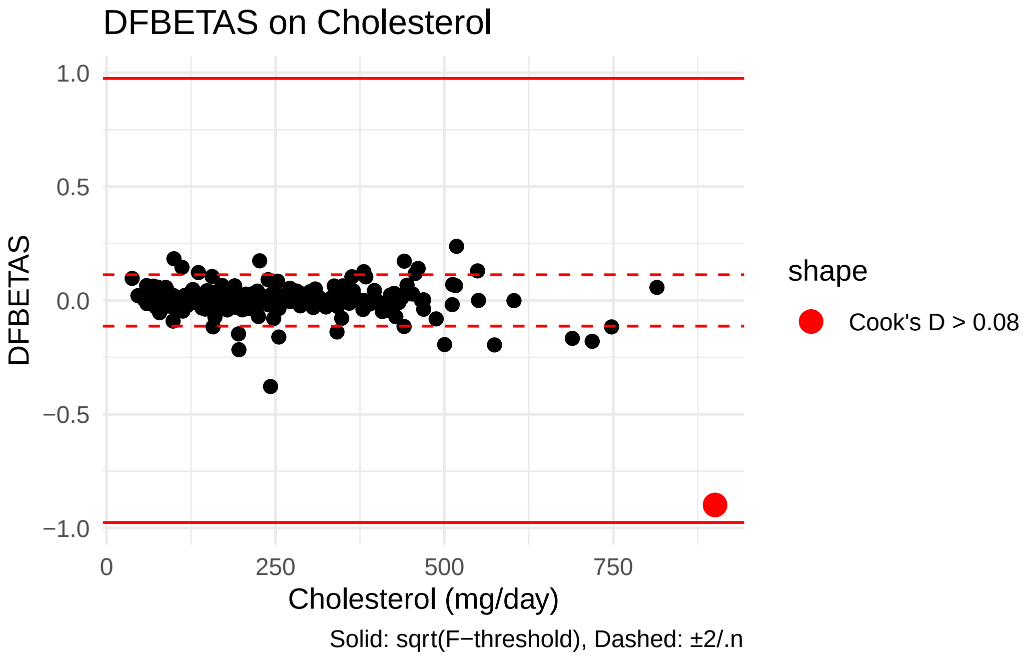

Figure 8: DFBETAS for the cholesterol coefficient plotted against cholesterol values

Figure 8 shows that one observation exerts a substantially larger influence on the cholesterol coefficient than all others, as its DFBETA approaches the common cutoff threshold of ±1. This observation corresponds to the highest Cook's distance and has an extremely high cholesterol intake. Such influential observations may unduly affect the model estimates.

3.7.2 Influence of Outlier in Cholesterol on -Carotene Estimates

To understand the impact of the observation with the highest Cook's distance, we examine the relationship between cholesterol and log-transformed plasma β-carotene.

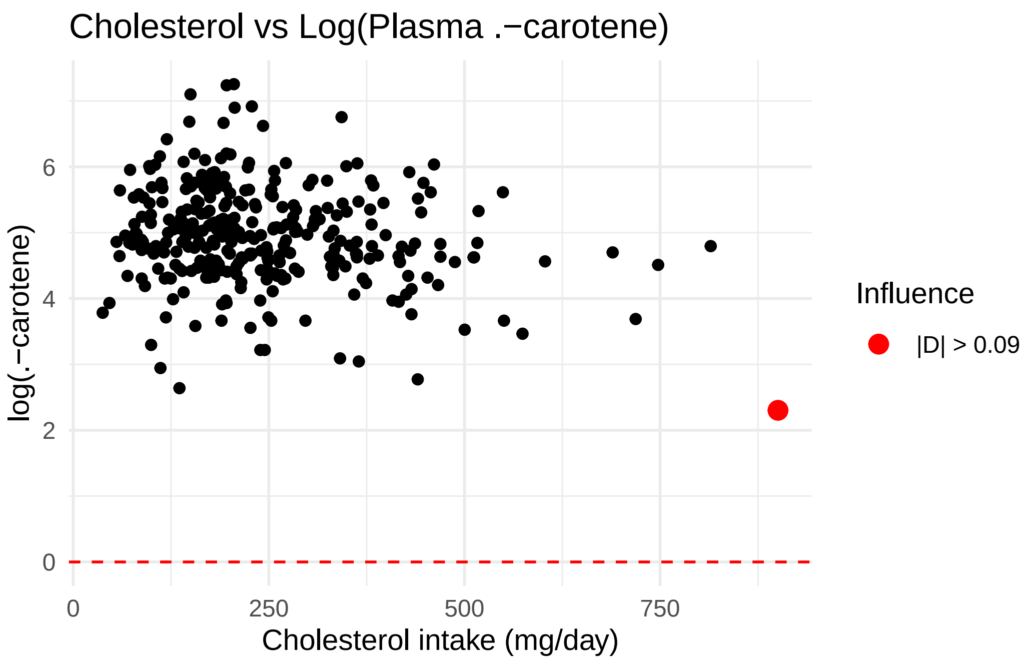

Figure 9: Scatterplot of observations comparing cholesterol intake to Log(Plasma β-carotene)

Figure 9 shows that this individual has an extremely high cholesterol intake combined with a low β-carotene level, placing them far from the main data cluster. This outlier exerts strong influence on the fitted regression line for cholesterol due to both high leverage and a large deviation from the trend, as confirmed in the DFBETAS plot. Its position explains its strong effect on the slope estimate.

4.1 Reestimation of Model 3 with outlier removed

In this part of the assignment, the outlier data point mentioned in **chapter 3.7** is removed from the data set and **Model 3** is reestimated to see whether we can see any difference in the diagnostic plots. Other than the removed data point, the data remains the same with the same reference categories for *sex*, *smokstat* and *vituse*, as the one data point removed didn't change effect which of the categorical values were more frequent. The new coefficient estimates of model fitted on the reduced dataset, can be seen in table Y.

Table 14: Parameters in model 3 reestimated on the reduced dataset and scaled to original domain along with confidence intervals in original domain.

Parameter	β	2.5 % β	97.5% β	exp(β)	**2.5** % **exp(β)**	97.5 % **exp(β)**
intercept	5.2135731364	4.6787068184	5.7484394548	183.7494485	107.6307964	313.700734
bmi	-0.0318708725	-0.0444967240	-0.0192450208	0.9686317	0.9564787	0.9809390
age	0.0060732191	0.0005042334	0.0116422049	1.0060917	1.0005044	1.0117102
fat	-0.0015391882	-0.0047999751	0.0017215986	0.9984620	0.9952115	1.0017231
cholesterol	-0.0003615077	-0.0012058459	0.0004828305	0.9996386	0.9987949	1.0004829
fiber	0.0231324057	0.0064390421	0.0398257693	1.0234020	1.0064598	1.0406294
alcohol	0.0015117870	-0.0048155988	0.0078391728	1.0015129	0.9951960	1.0078700
betadiet	0.0476578519	-0.0103130060	0.1056287098	1.0488117	0.9897400	1.1114091
smokstat former	-0.0841591231	-0.2485531536	0.0802349075	0.9192850	0.7799284	1.0835416
smokstat current	-0.2936459897	-0.5332095117	-0.0540824677	0.7455404	0.5867189	0.9473540
sex male	-0.2348952060	-0.4807798872	0.0109894751	0.7906537	0.6183010	1.011050
vituse often	0.2817784870	0.1045212894	0.4590356845	1.3254851	1.1101790	1.582547
vituse seldom	0.2607167010	0.0660658913	0.4553675107	1.2978599	1.0682971	1.576753

Furthermore, we plotted the Cook's distance plot for the model 3 on the reduced dataset, which can be seen in figure 10. We see a clear difference between this plot and the one in chapter 3.7 with the full data set. The removal of the outlier datapoint has of course resulted in the data point with largest Cook's Distance being removed, although there are still datapoints exceeding the threshold limit, which means that these could also be considered outliers.

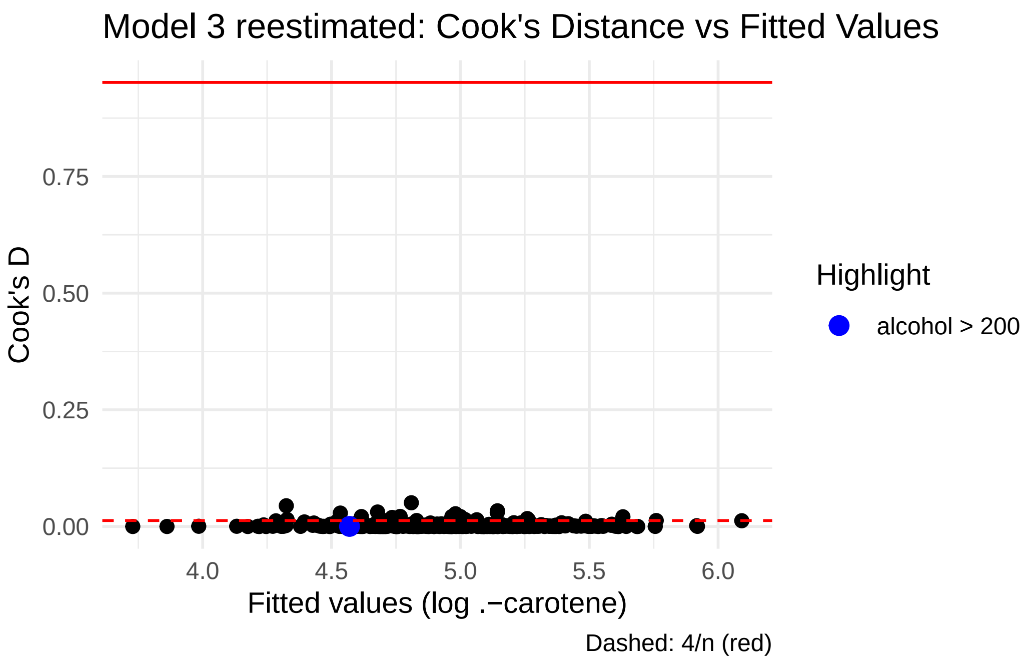

Figure 10: Cook's distance for reduced dataset

Next we checked the pairwise correlations between the continuous variables on the dataset without outlier. Once again, we focus on the variables having correlation $|\rho| > 0.6$. Like the case with the outlier, it is once again the correlations between calories-fat, calories-cholesterol and cholesterol-fat that proved problematic. The correlations between these were:

- calories / fat : $\rho = 0.87$
- calories / cholesterol: $\rho = 0.66$
- cholesterol / fat: $\rho = 0.7$

Compared to the results on the full data set, the correlations seem to be unchanged. This is also quite expected, as the omission of one single data point is unlikely to have a large effect on correlations.

Moreover, we computed and compared the adjusted Generalized Variance Inflation Factor (GVIF) on the model estimated on the the reduced dataset. Once again it was **calories** and **fat** that had values exceeding threshold of **2.24** for the adjusted GVIF, which can be seen in table 15. After removing calories, which had the highest values once again, none of the rest of the variables proved to have a value larger than the threshold. As with the pairwise correlations, the similarities between GVIF for the full data set and the reduced dataset is also expected, as one data point is not likely to effect these values.

Table 15: Adjusted GVIF on model 3 reestimated on the reduced dataset. Before and after removing calories.

variable	adjusted GVIF (all variables)	adjusted GVIF (calories removed)
bmi	1.033159	1.031988
age	1.142792	1.103411
calories	3.633360	(removed)
fat	2.831670	1.484160
cholesterol	1.482826	1.454680
fiber	1.586636	1.210335
alcohol	1.603541	1.060644
betadiet	1.160541	1.159959
smokstat	1.042013	1.041745
sex	1.138658	1.138658
vituse	1.036193	1.028028

4.2 Stepwise variable selection

We performed 2 stepwise variable selections for **model 3** with BIC as criterion, for which the only difference is the dataset used to estimate these models. For both of these, we started with the null model as the smallest (which is just the intercept), and then the largest allowed model being all variables seen in table 14 in *Chapter 4.1*. For the two models, the variables added were **different**. In neither of the two case were any variables removed after being added (no backward steps). The selection order for the model with outlier in the data was:

- *null model (intercept)* → *bmi* → *fiber* → *cholesterol* → *vituseoften* and *vituseseldom*

And for the model without outlier in the data:

- *null model (intercept)* → *bmi* → *fiber* → *fat* → *vituseoften* and *vituseseldom*

The parameter estimates and confidence intervals for in the case of the full data set as well as the reduced dataset are seen in Table 16 and Table 17 (we call these **model 4**). The interesting thing is that the variables that differ between the models are **cholesterol** and **fat**. As seen in *3.2* and *4.1*, these variables had high correlation, so it is not a surprise that a stepwise selection would only include on of these. Also interesting is that in the case without outlier, the stepwise selection chose fat instead of cholesterol. This also makes sense, as we found that the the outlier was an outlier exactly from the extreme value in cholesterol.

Table 16. Parameter estimates with confidence intervals for the model in 4.2 by stepwise selection, ***with*** *outlier data included.*

Parameter	β	2.5 % β	97.5 % β	exp(β)	**2.5 % exp(β)**	**97.5 % exp(β)**
intercept	5.394602900	4.975754756	5.8134510435	220.2146824	144.8581165	334.7724487
bmi	-0.029134120	-0.041947529	-0.0163207109	0.9712862	0.9589201	0.9838118
fiber	0.032224213	0.017641547	0.0468068789	1.0327490	1.0177981	1.0479196
cholesterol	-0.001203526	-0.001791134	-0.0006159179	0.9987972	0.9982105	0.9993843
vituseoften	0.325908184	0.147873864	0.5039425031	1.3852882	1.1593666	1.6552342
vituseseldom	0.276430810	0.080235963	0.4726256565	1.3184157	1.0835427	1.6042007

Table 17. Parameter estimates with confidence intervals for the model in 4.2 by stepwise selection, ***without*** *outlier data.*

Parameter	β	2.5 % β	97.5 % β	exp(β)	**2.5 % exp(β)**	**97.5 % exp(β)**
intercept	5.384555816	4.962661493	5.806450139	218.0132446	142.9738138	332.4369237
bmi	-0.029604312	-0.042191947	-0.017016678	0.9708296	0.9586857	0.9831273
fiber	0.034661934	0.019858580	0.049465288	1.0352697	1.0200571	1.0507091
fat	-0.003876143	-0.006218125	-0.001534161	0.9961314	0.9938012	0.9984670
vituseoften	0.343359645	0.167115797	0.519603492	1.4096757	1.1818911	1.6813608
vituseseldom	0.271369464	0.077805293	0.464933634	1.3117596	1.0809122	1.5919085

5.1 Fine-tuning the Model

To determine whether it is necessary to have three categories in the categorical x-variable *vituse* in model 4, we perform a partial F-test between model 4 and the same model but where *vituse often* and *vituse seldom* have been grouped together to *vituse yes*, this variable also being the reference. The P-value for this test is P = 0.4583 > 0.05, which means having three variables in model 4 is not significantly better. The estimates and confidence intervals of the new **model 5**, having only *vituse yes*, can be seen in Table 18.

Table 18. Estimates and confidence intervals for Model 5.

Variable	**β**	**2.5% β**	**97.5% β**
intercept	5.392206843	4.97111136	5.813302322
bmi	-0.029937735	-0.04248504	-0.017390426

fiber	0.034965232	0.02019464	0.049735830
fat	-0.003907911	-0.00624665	-0.001569172
vituse yes	0.313699756	0.15606821	0.471331300

5.2 Compare the different models

After refitting the models for the new data without the outlier, we summarize the model parameters in table 19 and the model statistics in table Q. The estimations of **Model 5** is left out, as it can be seen in table 18 above. For the sake of a readable table and the fact that they don't add anything to the discussion, we omit the estimate and confidence in the original domain, i.e. expontential transformation of the model estimates.

Table 19. Model parameters and confidence intervals for Model 1-4, estimated on the reduced dataset.

model / parameters	**β**	**2.5% β**	**97.5% β**
Model 1			
Intercept	5.86865	5.51229560	6.22501439
bmi	-0.03478	-0.04806868	-0.02149404
Model 2			
intercept	5.0685	4.9526992	5.18423224
smokstat former	-0.1273	-0.3050524	0.05037317
smokstat current	-0.4548	-0.7038710	-0.20578403
Model 3			
intercept	5.2135731	4.6787068184	5.7484394548
bmi	-0.0318709	-0.0444967241	-0.0192450208
age	0.0060732	0.0005042334	0.0116422049
fat	-0.0015392	-0.0047999751	0.0017215986
cholesterol	-0.0003615	-0.0012058459	0.0004828305
fiber	0.0231324	0.0064390421	0.0398257693
alcohol	0.0015118	-0.0048155988	0.0078391728
betadiet	0.0476579	-0.0103130060	0.1056287098
smokstat former	-0.0841591	-0.2485531536	0.0802349075
smokstat current	-0.2936460	-0.5332095117	-0.0540824677
sex male	-0.2348952	-0.4807798872	0.0109894751
vituse often	0.2817785	0.1045212894	0.4590356845
vituse seldom	0.2607167	0.0660658913	0.4553675107
Model 4			
intercept	5.384556	4.962661493	5.806450139
bmi	-0.029604	-0.042191947	-0.017016678
fiber	0.034662	0.019858580	0.049465288

model / parameters	**β**	**2.5% β**	**97.5% β**
fat	-0.003876	-0.006218125	-0.001534161
vituse often	0.343360	0.167115797	0.519603492
vituse seldom	0.271369	0.077805293	0.464933634

Table 20. Summary of model parameters, confidence intervals, R^2, R^2_{adj}, residual standard deviation, BIC and AIC values.

Table 21: So which model is the best? First and foremost, a criteria to be looked at is whether or not there are models with insignificant variables. Looking at table 19, both Model 2 and especially model 3 have variable(s) that are insignificant, seen by the confidence interval including 0. This means that these models can be further simplified.

	Number of parameters (incl. intercept)	R^2	R^2_{adj}	**Residual Std.dev.**	**BIC**	**AIC**
Model 1	2	0.07836045	0.07540648	0.7189481	699.1165	687.8683
Model 2	3	0.04020327	0.03403094	0.7348585	717.604	702.6064
Model 3	13	0.2463894	0.2163451	0.661888	699.1584	646.6669
Model 4	6	0.1988381	0.1858322	0.6746509	678.1254	651.8797
Model 5	5	0.1974038	0.1870142	0.6741609	672.9377	650.4413

Secondly, looking now at the models' statistics, the R^2 is difficult to interpretet, as a higher model order can always achieve higher scores, explaining more data variability but through overfitting. Therefore, we instead look at the adjusted R^2 , which normalizes the regular R^2 . Here, we see model 3 gets the best score, followed models 5 and 4, in that order

Moreover, looking at the standard deviation of the model residuals, the lowest (i.e. best) score is also for model 3 followed again by models 5 and 4. However, for both AIC and BIC Model 5 gets the best scores out of all models, followed by model 4. So if we are looking to get as close as possible to the underlying model generating the data, or using the models for prediction of future data, then Model 5 is best.

Since, model 3 has a couple of insignificant parameters, and the fact that we might be interested in finding a "correct model" and actually using our model for predictions, then Model 5 is the best choice in our opinion.

Conclusion

While Model 3 provides the best fit in terms of raw explanatory power, Model 5 offers the most reasonable trade-off between accuracy and complexity. Given the similar adjusted R^2 and residual standard deviation to Model 3, but with far fewer parameters and better AIC and BIC, Model 5 is preferred for its interpretability and efficient use of predictors.