Project 1 - Report: Linear Regression

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.

Author Contributions

Name	Roles
Mattis Ranheim	Analysis, Discussions, Programming, Visualisation, Writing
Madeleine Ekstrand	Analysis, Discussions, Programming, Visualisation, Writing
Yassin Hjuler El Mahdaoui	Analysis, Discussions, Programming, Visualisation, Writing

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:

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/height², kg/m²)
`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.

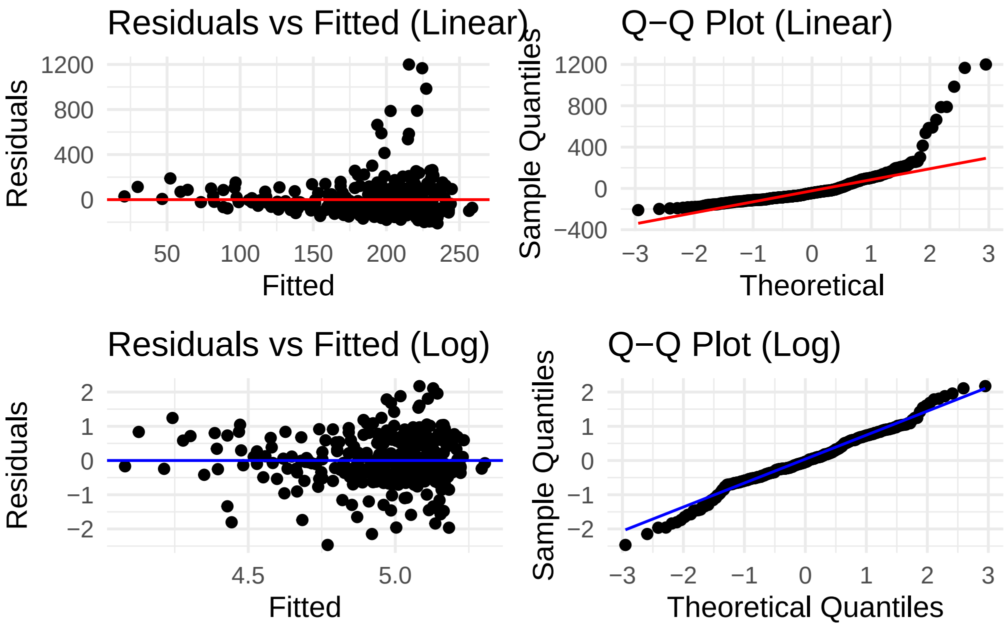


Figure 1: Residual plots and Q-Q plots for linear and log-transformed BMI model

The residual plots and Q-Q plots 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**:

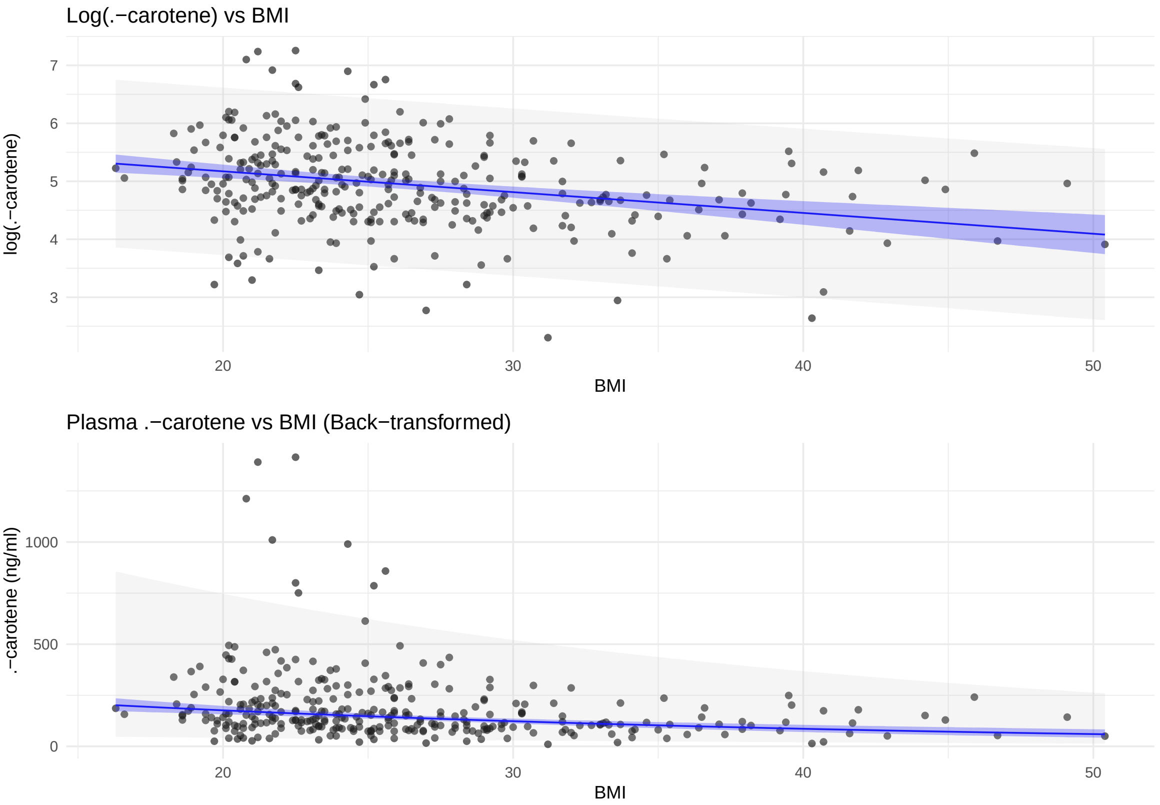
The table below shows the **β-estimates** and their associated **95% confidence intervals**. The intercept corresponds to the expected value of log(β-carotene) when BMI is zero (which is not realistic in practice, but needed for the mathematical formulation), while the slope for BMI (β₁) describes the expected change in the log(β-carotene) concentration for each one-unit increase in BMI.

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

The estimate for β_1 is **−0.0359**, with a 95% confidence interval from **−0.0494 to −0.0224**, indicating a statistically significant negative association. This suggests that for every additional unit increase in BMI, the **log of plasma β-carotene decreases**. This negative association 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** 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** displays the same model but transformed back to the original β-carotene scale (ng/ml). This gives a more intuitive interpretation of the effect in absolute terms.



Log(.−carotene) vs BMI

Plasma .−carotene vs BMI (Back−transformed)

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.

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 investigate how smoking status relates to plasma β-carotene levels, we begin by converting the categorical variable smokstat into a factor variable with meaningful labels:

- **1** = Never Smoker

- **2** = Former Smoker
- **3** = Current Smoker

We then present a table showing the number of individuals in each smoking group, and the mean and standard deviation of both **plasma β-carotene (ng/ml)** and **log-transformed plasma β-carotene** for each group.

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

Given the three smoking categories — *Never smoker*, *Former smoker*, and *Current smoker* — we choose **Never smoker** as the reference category in our regression modeling. This choice is motivated by the following:

• It is the **largest group** (n = 157), ensuring stable estimates when used as baseline.

• It represents a **natural baseline** with respect to exposure — those who have not been exposed to smoking-related influences.

• Comparing other groups (former and current smokers) to this "cleanest" category allows for **straightforward interpretation** of differences in plasma β-carotene levels.

As a result, we will refer to this 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:

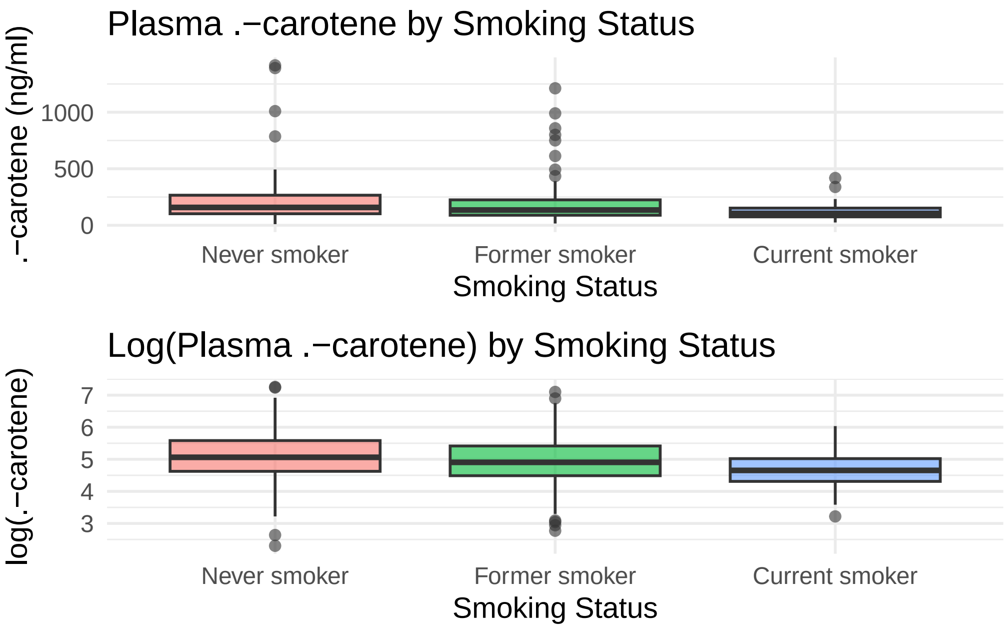


Figure 3: Box plots of β-carotene and log(β-carotene)

The boxplots 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 2A**, the reference category is **“Never smoker”**
- In **Model 2B**, the reference category is **“Current smoker”**

Never Smoker = Reference	Estimate	Standard Error
Intercept	5.0508487	0.05986447
$\beta_{FormerSmoker}$	-0.1097225	0.09206715
$\beta_{CurrentSmoker}$	-0.4372105	0.12910704

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) level for the **reference category**. The other coefficients represent the difference in mean log(β-carotene) compared to the reference.
- In Model 2A, the intercept corresponds to "Never smoker", and the coefficients show how much lower (on average) former or current smokers are. In Model 2B, the intercept corresponds to "Current smoker", and the coefficients show how much higher never and former smokers are.
- One reason the standard error for the intercept is higher in the model where Never Smoker is the reference is because that answer has fewer observations, leading to 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 versions. We calculate:

- The **expected log(β-carotene)** level with 95% confidence intervals.
- The **expected β-carotene (ng/ml)** level on the original scale by back-transforming the predictions (i.e., applying the exponential function).

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

Interpretation

The predicted values and their 95% confidence intervals are **identical** for both models and we therefore only include 2 tables above.

- For **Never smokers**, the estimated plasma β-carotene level is approximately **156.2 ng/ml** with a 95% CI of **138.8 to 175.7**.
- For **Former smokers**, the level is approximately **139.9 ng/ml** with a CI of **121.9 to 160.6**.
- For **Current smokers**, it is around **100.9 ng/ml**, with a CI of **80.5 to 126.3**.

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.

These values also align well with the group-wise means from section 2(a), reinforcing that the model provides an appropriate summary of the data.

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 might indicate collinearity issues if both variables are included in the same regression model.

The correlation table revealed three particularly strong linear relationships which are plotted with scatterplots below.

- **Fat** and **Calories**: $r \approx 0.87$
- **Cholesterol** and **Calories**: $r \approx 0.66$
- **Cholesterol** and **Fat**: $r \approx 0.71$

These variables are highly interrelated, and care should be taken when including them in the same multiple regression model, due to potential multicollinearity.

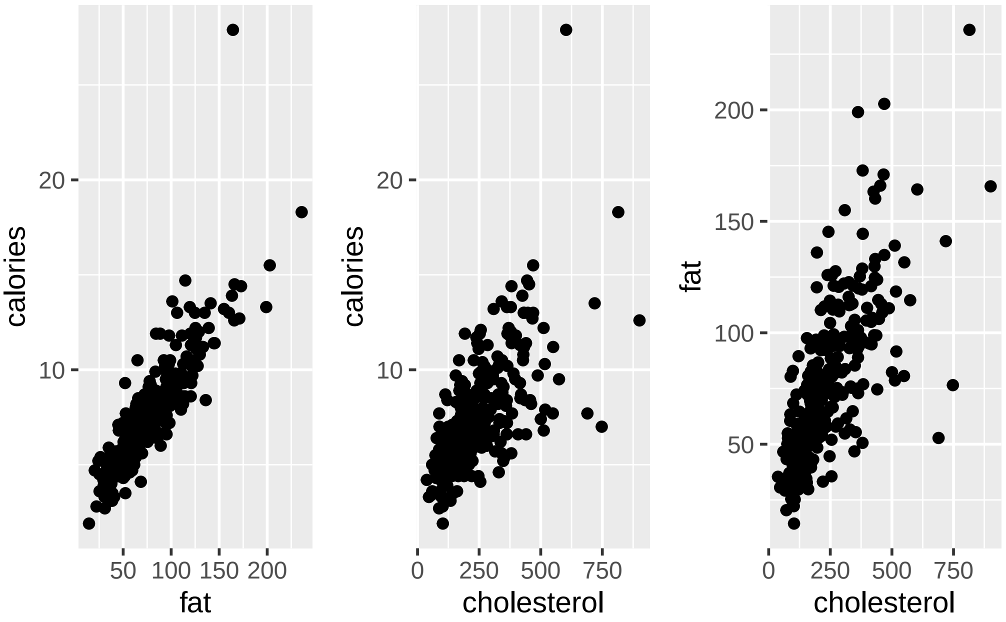


Figure 4: Correlation scatterplots of variables with abs(cor) > 0.6

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:

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

A GVIF-adjusted 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.

Calories and **fat** both exceed the GVIF threshold of 2.24. The strongest multicollinearity is found between these variables, 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.

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.

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 back-transformed domain (i.e., original β-carotene scale).

The table below presents both the log-scale β-estimates and their exponentiated versions (which represent multiplicative effects on the geometric mean of β-carotene), along with 95% confidence intervals:

	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 the table and discussion below.

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

This is a **t-test** on the BMI coefficient in **Model 3.3**, testing:

- **Null hypothesis**: $H_0 : \beta_{BMI} = 0$
- **Alternative hypothesis**: $H_1 : \beta_{BMI} \neq 0$

We extract the result from the model summary:

- **Test statistic**: t = -4.918
- **Degrees of freedom**: 302
- **P-value**: 1.44e-06

Conclusion: Since the P-value is far below 0.05, we **reject the null hypothesis**. BMI is a statistically significant predictor of log(β-carotene) even after adjusting for other variables.

(ii) Is this model significantly better than the model from chapter 1 which only used bmi?

This is an **F-test** comparing **Model 1** to the full **Model 3**

- **Null hypothesis:** $H_0 : \beta_{\text{age}} = \beta_{\text{fat}} = \cdots = \beta_{\text{vituse}} = 0$ - All variables except BMI have no added value
- **Alternative hypothesis:** H_1 : At least one of the additional predictors improves the model fit
- **F-statistic**: 6.2597
- **Degrees of freedom**: 302 and 313
- **P-value**: 2.633e-09

Conclusion: We **reject the null hypothesis**. The full model is significantly better than using BMI alone.

(iii) Is Model 3.3 significantly better than the Categorical Smoking Model?

This is an **F-test** comparing **Model 2** to the full **Model 3**

- **Null hypothesis:** $H_0 : \beta_{\text{bmi}} = \beta_{\text{age}} = \cdots = \beta_{\text{vituse}} = 0$ - All variables except SmokStat have no added value
- **Alternative hypothesis:** H_1 : At least one of the additional predictors improves the model fit
- **F-statistic**: 8.6928
- **Degrees of freedom**: 302 and 312
- **P-value**: 1.63e-12

Conclusion: Again, we **reject the null hypothesis**. **Model 3** adds substantial explanatory power beyond smokstat alone.

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.

These plots help identify potential problems with the model assumptions.

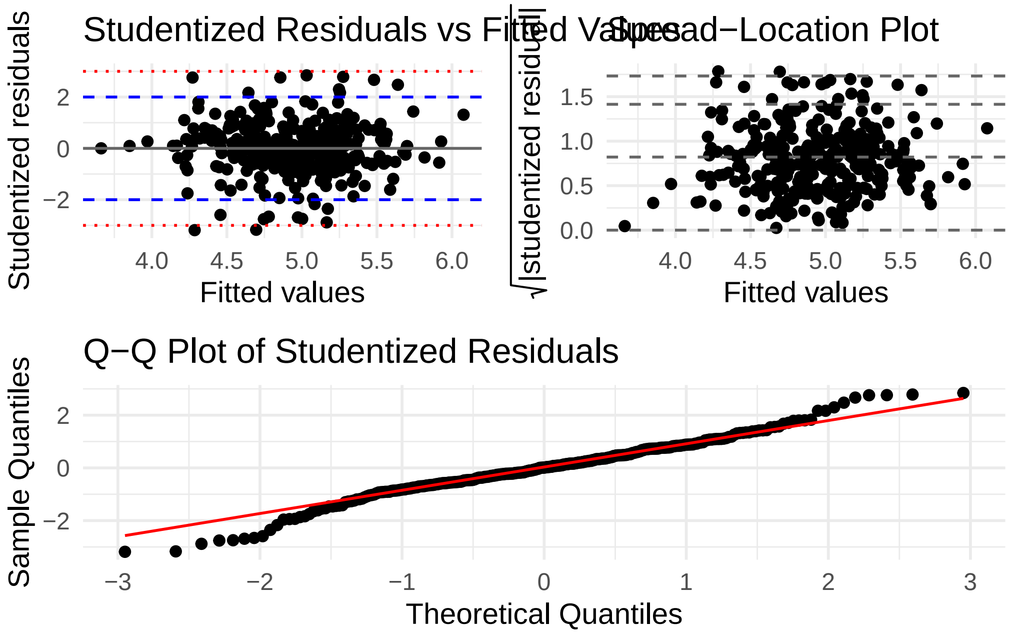


Figure 5: Diagnostic plots for Model 3: studentized residuals vs fitted values, spread-location plot, and Q-Q plot.

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.

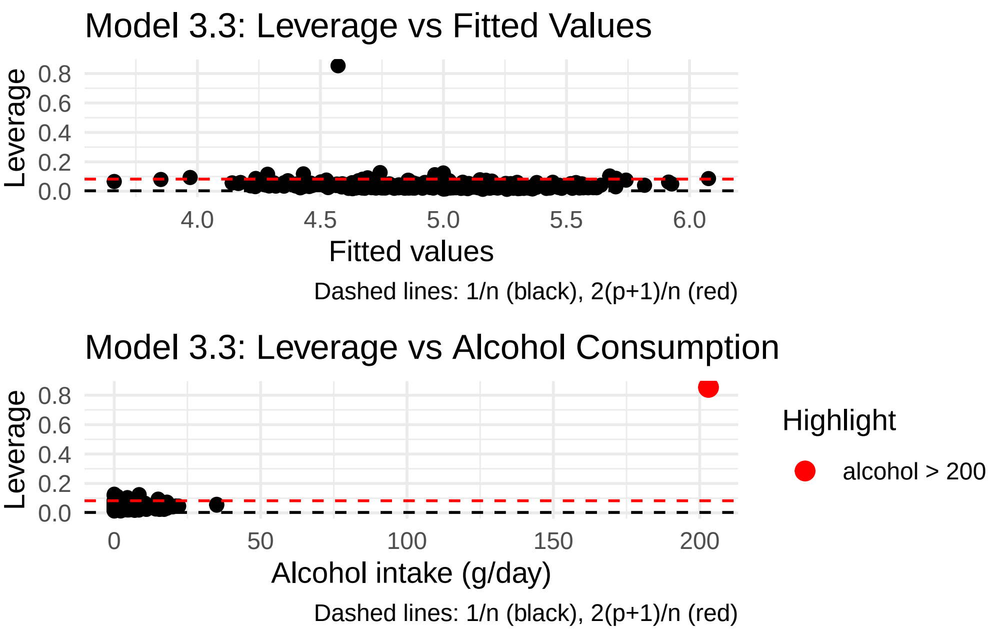
In the **Q-Q plot**, the points generally follow the red reference line, though some deviation is observed in the tails. This indicates slight departure from normality, but not to a degree that would seriously undermine the model.

Conclusion: Overall, the diagnostic plots 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.

We include horizontal reference lines at (1/n) (black) and (2(p+1)/n) (red), where (n) is the number of observations and (p+1) is the number of model parameters (including the intercept). Observations with leverage values above the red line should be carefully examined.


Model 3.3: Leverage vs Fitted Values

Dashed lines: 1/n (black), 2(p+1)/n (red)

Model 3.3: Leverage vs Alcohol Consumption

Dashed lines: 1/n (black), 2(p+1)/n (red)

Figure 6: Leverage of observations

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)

This high-leverage point is likely caused by the fact that this individual's alcohol consumption is extremely distant from all other observations, making them an outlier in the predictor space. As leverage is determined by distance in multivariate predictor space, this makes sense.

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.

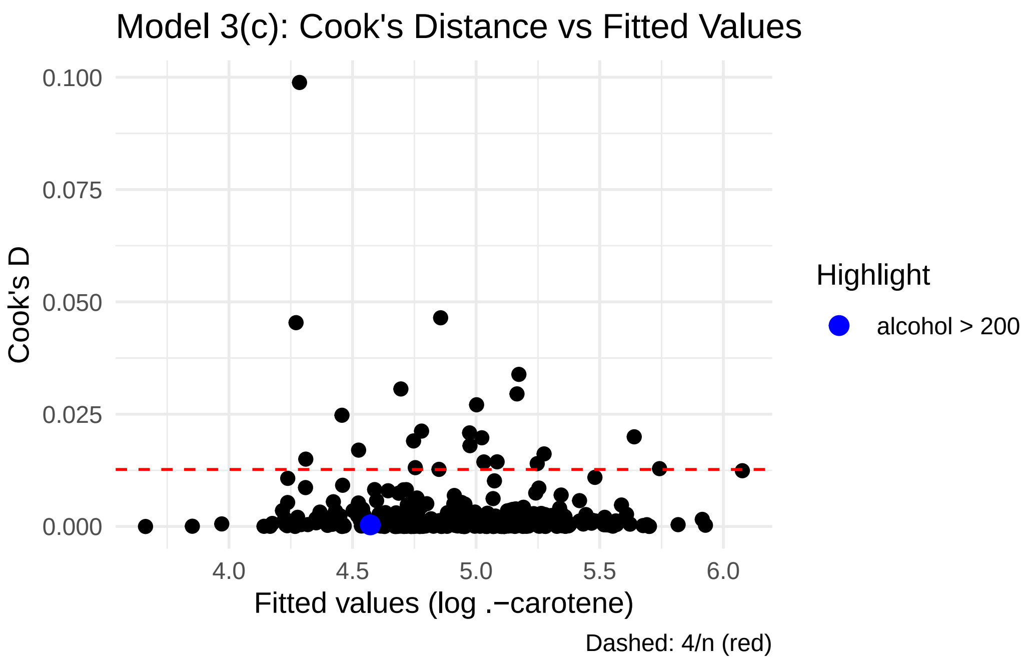


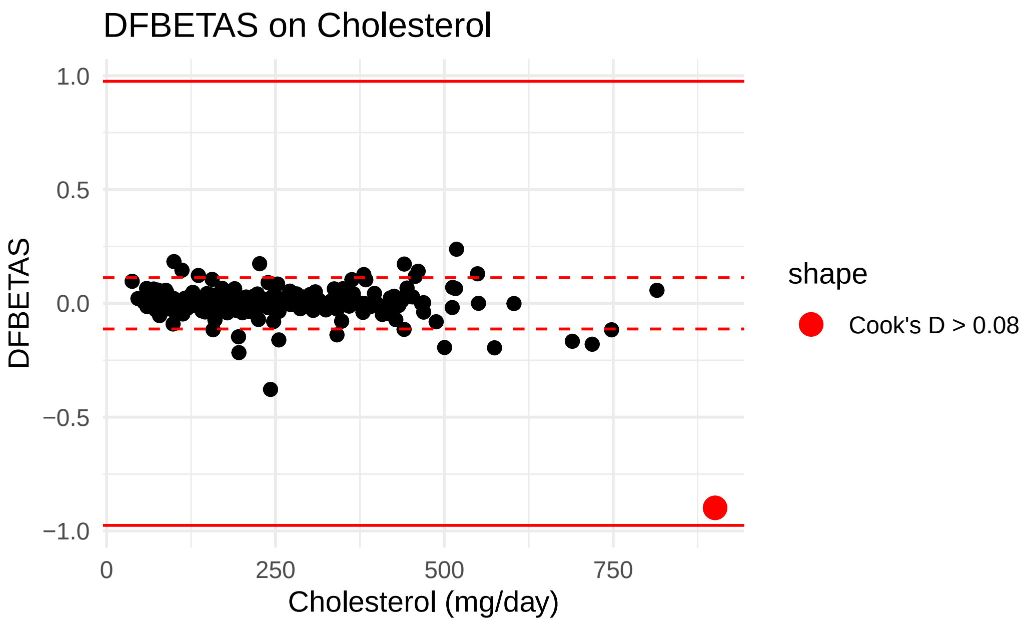
Figure 7: Cooks Distance of observations

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**.

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:



Solid: sqrt(F−threshold), Dashed: ±2/.n

Figure 8: DFBETAS for the cholesterol coefficient plotted against cholesterol values. The red lines indicate common influence thresholds used to assess whether observations unduly influence the regression estimates.

The plot 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 further investigate the reason behind the large influence of the observation with the highest Cook's distance, we examine the relationship between cholesterol intake and log-transformed plasma β-carotene.

The following plot highlights this influential observation and shows how it diverges from the general data pattern.

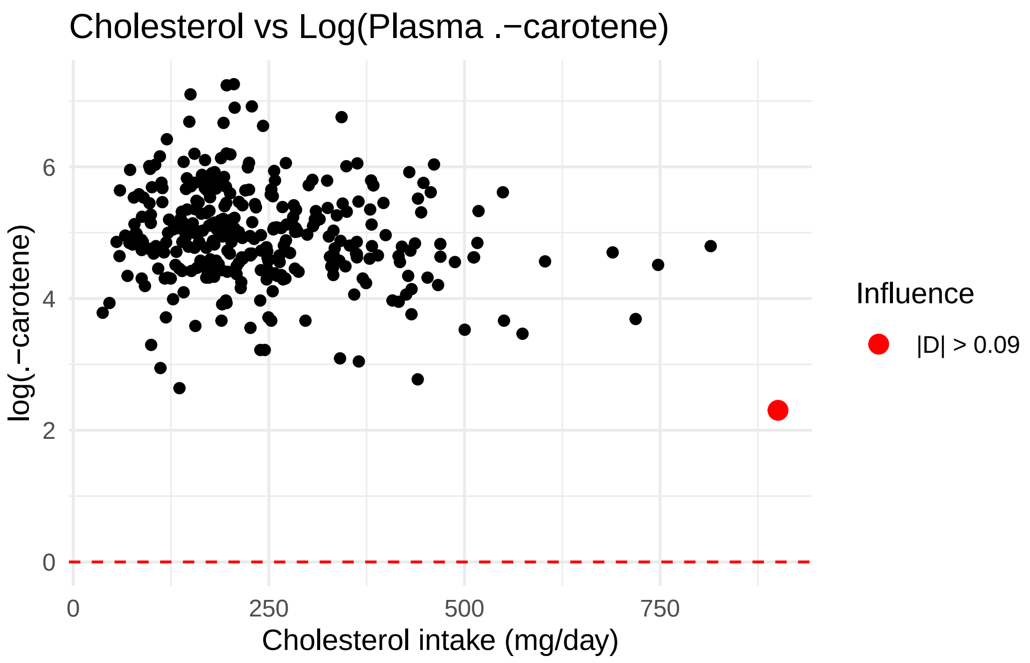


Figure 9: Scatterplot of observations comparing cholesterol intake to Log(Plasma β-carotene)

This plot illustrates that the most influential observation has an exceptionally high cholesterol value and a relatively low plasma β-carotene level. It clearly lies far from the bulk of the data and thus heavily influences the fitted relationship between cholesterol and β-carotene. Its position in this predictor–response space gives it both high leverage and high impact on the slope estimate for cholesterol, as shown previously in the DFBETAS plot. This confirms that its influence stems not only from extremeness in the predictor but also from its strong deviation from the overall regression trend.

4.1 Reestimation of Model 3.3 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 3.3** 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, **"Model 3.3.1",** can be seen in the table below.

Table 13: *Reestimated coefficients for model 3.3 on the reduced dataset*

Parameter	**β**	exp(**β**)	**2.5** % **exp(β)**	97.5 % **exp(β)**
intercept	5.2135731366	183.7494485	107.6307964	313.700734
bmi	-0.0318708725	0.9686317	0.9564787	0.9809390
age	0.0060732191	1.0060917	1.0005044	1.0117102
fat	-0.0015391882	0.9984620	0.9952115	1.0017231
cholesterol	-0.0003615077	0.9996386	0.9987949	1.0004829
fiber	0.0231324057	1.0234020	1.0064598	1.0406294
alcohol	0.0015117870	1.0015129	0.9951960	1.0078700
betadiet	0.0476578519	1.0488117	0.9897400	1.1114091
smokstat former	-0.0841591231	0.9192850	0.7799284	1.0835416
smokstat current	-0.2936459897	0.7455404	0.5867189	0.9473540
sex male	-0.2348952060	0.7906537	0.6183010	1.011050
vituse often	0.2817784870	1.3254851	1.1101790	1.582547
vituse seldom	0.2607167010	1.2978599	1.0682971	1.576753

Furthermore, we plotted the Cook's distance plot for the model on the reduced dataset, which can be seen below. 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 resulted in no datapoints having a particularly large Cook's distance value. We can draw the conclusion that there don't seem to be any more outlier datapoints to be taken care of.

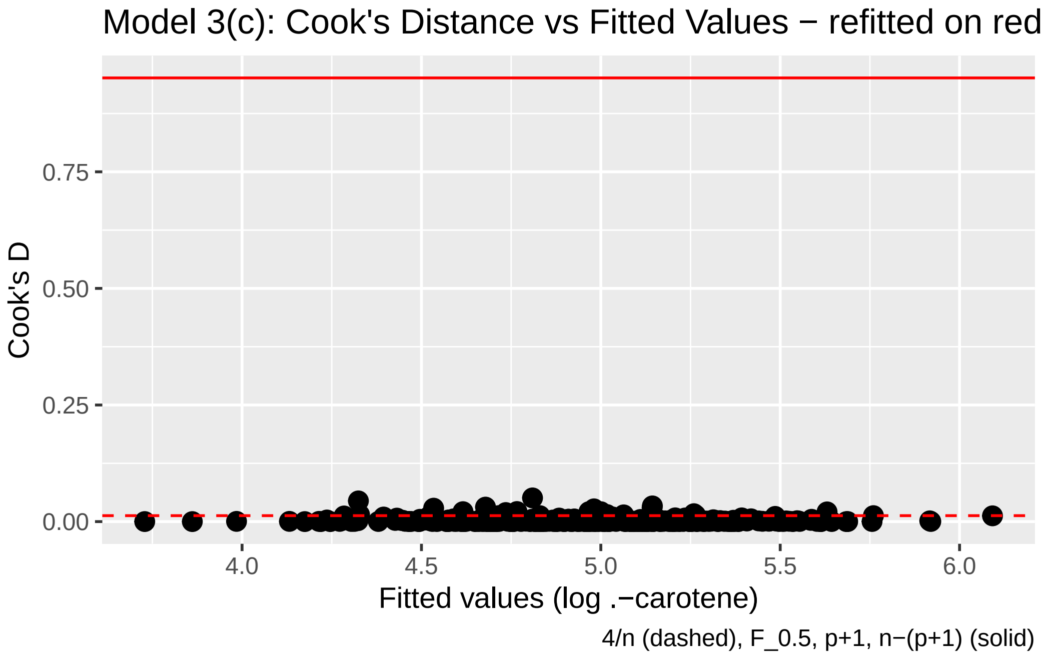


Figure 10: Cook's distance for reduced dataset

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: $\rho = 0.71$

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 an effect on correlations.

Moreover, we computed and compared the Generalized Variance Inflation Factor (GVIF) between the model estimated on the two different datasets. Once again it was **calories** and **fat** that had values exceeding threshold of **2.24** for the adjusted GVIF. 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 is also expected, as one data point is not likely to effect these values.

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 the table 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:

- *bmi* → *fiber* → *cholesterol* → *vituse_no* and *vituse_not_often*

And for the model without outlier in the data:

- *bmi* → *fiber* → *fat* → *vituse_no* and vituse_not_often

The parameter estimates and confidence intervals for in the case of the full data set as well as the reduced dataset are seen below (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 either model only includes one 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.

Parameter	**β**	exp(**β**)	**2.5 % exp(β)**	**97.5 % exp(β)**
intercept	5.720511083	305.0607947	201.5795579	461.6643146
bmi	-0.029134120	0.9712862	0.9589201	0.9838118
vituse not often	-0.049477374	0.9517267	0.7846682	1.1543525
vituse no	-0.325908184	0.7218715	0.6041441	0.8625399
fiber	0.032224213	1.0327490	1.0177981	1.0479196
cholesterol	-0.001203526	0.9987972	0.9982105	0.9993843

Parameter	**β**	exp(**β**)	**2.5 % exp(β)**	**97.5 % exp(β)**
intercept	5.727915461	307.3279630	202.8998257	465.5029965
bmi	-0.029604312	0.9708296	0.9586857	0.9831273
vituse not often	-0.071990181	0.9305400	0.7689271	1.1261207
vituse no	-0.343359645	0.7093830	0.5947563	0.8461016
fiber	0.034661934	1.0352697	1.0200571	1.0507091
fat	-0.003876143	0.9961314	0.9938012	0.9984670

Parameter estimates with confidence intervals for models 4b by stepwise selection. With outlier (above) and without outlier (below).

5.1 Fine-tuning the Model

To determine whether it is necessary to have three categories in the categorical x-variables “vituse” and “smokstat” we look at the confidence interval for the beta-variables in **model 4**. The confidence interval for the category “seldom” in “vituse” includes zero. From this result, the conclusion that “vituse” should be only two categories (“yes” and “no”) is drawn.

Now we present the new beta estimates confidence intervals for this new **Model 5**:

Table 16: As we can see in the confidence intervals for the new beta estimates for the updated variables, no interval contains zero which means that all the variables included in the model has a significant effect on the outcome.

	2.5%	**97.5%**
Intercept	4.629663852	5.57369646
age	0.003870043	0.01451842
sexmale	-0.577195070	-0.11476421
bmi	-0.043363058	-0.01848219
vituseno	-0.462672341	-0.14454230
fiber	0.013935129	0.04213240

5.2 compare the different models

After refitting the models for the new data without the outlier. We summarize the models in the table below:

	Number of beta parameters	**R^2**	**R^2 adjusted**	**Residual Standard Deviation**	**BIC**	**AIC**
Model 1	1	0.07836045	0.07540648	0.7189481	699.1165	687.8683
Model 2	2	0.04020327	0.03403094	0.7348585	717.604	702.6064
Model 3	12	0.2532462	0.2182811	0.66107	707.7871	647.7968
Model 4	6	0.21026	0.1974396	0.6698244	673.6166	647.3708
Model 5	5	0.21026	0.1974396	0.6698244	673.6166	647.3708

Among the five models, Model 3 shows the highest explanatory power with an R^2 of 0.253 and an adjusted R^2 of 0.219, indicating it captures more variability in the response variable than the others. However, this model also includes 12 parameters, raising concerns about potential overfitting. While its AIC is the lowest (647.80), its BIC is slightly worse than that of Models 4 and 5.

Models 4 and 5, on the other hand, provide a more parsimonious fit with only 5–6 parameters. They achieve similar adjusted R^2 and nearly identical AIC values compared to Model 3. However, they outperform Model 3 in terms of BIC, which penalizes complexity more strongly. Their lower BIC values (673.62 vs. 707.79) suggest they strike a better balance between goodness of fit and simplicity.

Conclusion

While Model 3 provides the best fit in terms of raw explanatory power, Model 5 (or Model 4) 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 BIC, Model 5 is preferred for its interpretability and efficient use of predictors.