

# 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 author. - Output was carefully verified for correctness.

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

---

## Author Contributions

Name	Roles
Mattis Ranheim	Derivations, Analysis, Discussions, Programming, Visualisation, Writing (original draft), Writing (revision & editing), Project Management

## Introduction

Numerous observational studies have suggested that low dietary intake or low plasma concentrations of  $\beta$ -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  $\beta$ -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  $\beta$ -carotene concentrations** as the outcome of interest.

The study highlights considerable individual variation in plasma  $\beta$ -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/height <sup>2</sup> , kg/m <sup>2</sup> )
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 $\beta$ -carotene intake per day (mg)
betaplasma	<b>Plasma <math>\beta</math>-carotene concentration (ng/ml)</b> — this is the <b>response variable</b> 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  $\beta$ -carotene concentration for individual  $i$ ,  $\mathbf{x}_i$  is the vector of explanatory variables,  $\beta$  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

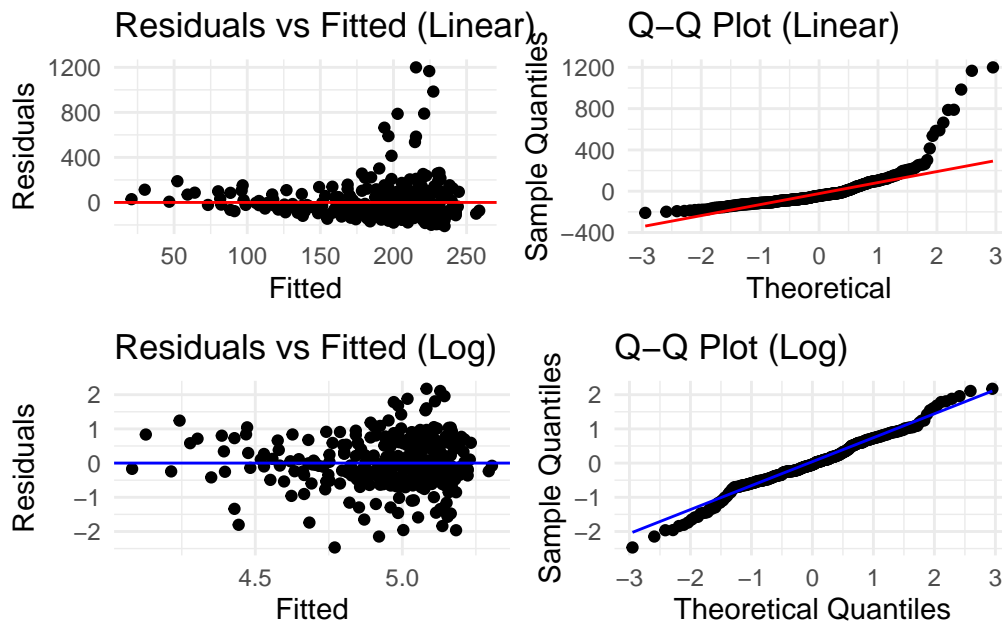
We fitted two models to examine the relationship between BMI and plasma  $\beta$ -carotene levels:

- **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.

These graphs show the residual and QQ-plots for the **Linear** and log-trans



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.

## 1.2 Model Estimates

To interpret the relationship between BMI and plasma  $\beta$ -carotene concentration, we present the coefficient estimates from the log-linear model:

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

	Estimate	2.5 %	97.5 %
Intercept	5.8896	5.5273	6.2519
$\beta_1$	-0.0359	-0.0494	-0.0224

The estimate for  $\beta_1$  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  $\beta$ -carotene decreases**, implying an **approximate 3.5% reduction** in  $\beta$ -carotene levels per BMI unit.

This negative association supports the hypothesis that higher body fat may be linked to lower concentrations of this micronutrient.

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

To investigate how plasma  $\beta$ -carotene levels relate to BMI, we fit a linear regression model where the outcome was **log-transformed  $\beta$ -carotene concentration**. This transformation helps satisfy linear regression assumptions, especially linearity and homoscedasticity.

Below, we show two plots:

- The **top plot** shows the relationship between BMI and the log-transformed  $\beta$ -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  $\beta$ -carotene scale (ng/ml). This gives a more intuitive interpretation of the effect in absolute terms.

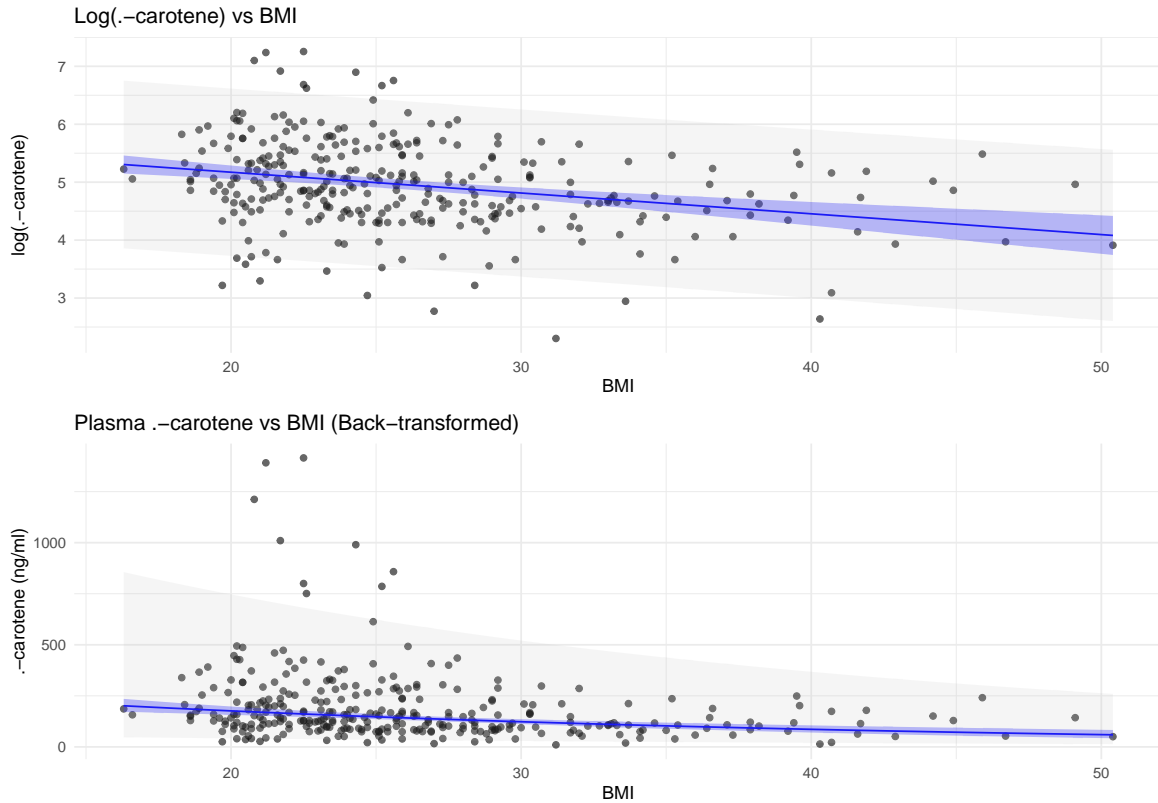


Figure 1: Log-scale and back-transformed plots with 95% CI and prediction intervals.

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

These results suggest that BMI is negatively associated with  $\beta$ -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 $\beta$ -carotene for Changes in BMI

We interpret the log-linear regression model by expressing expected **percentage changes** in plasma  $\beta$ -carotene (ng / ml) for three different BMI changes:

1. **BMI increased by 1 unit**
2. **BMI decreased by 1 unit**
3. **BMI decreased by 10 units**

From the calculations, we can conclude the following:

- A **1-unit increase in BMI** is associated with a **3.5% decrease** in plasma  $\beta$ -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  $\beta$ -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  $\beta$ -carotene becomes **multiplicative**, not additive.

#### 1.4 Hypothesis Test for Linear Relationship Between BMI and log( $\beta$ -carotene)

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

- **Null hypothesis:**  $H_0 : \beta_1 = 0$  (no relationship between BMI and log( $\beta$ -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.

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

Since the **P-value is far less than 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  $\beta$ -carotene levels.

The **negative value of the t-statistic** ( $t = -5.23$ ) along with the **negative slope coefficient** (from earlier results) suggests an **inverse relationship**: as BMI increases, the expected log( $\beta$ -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)** strengthens the evidence against the null hypothesis.

Thus, we conclude that there is a **statistically and practically significant linear relationship** between BMI and log( $\beta$ -carotene).

## 2. Plasma $\beta$ -carotene and Smoking Habits

To investigate how smoking status relates to plasma  $\beta$ -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  $\beta$ -carotene (ng/ml)** and **log-transformed plasma  $\beta$ -carotene** for each group.

Smokstat	Count	Mean $\beta$	S.D. $\beta$	Mean $\log(\beta)$	S.D. $\log(\beta)$
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

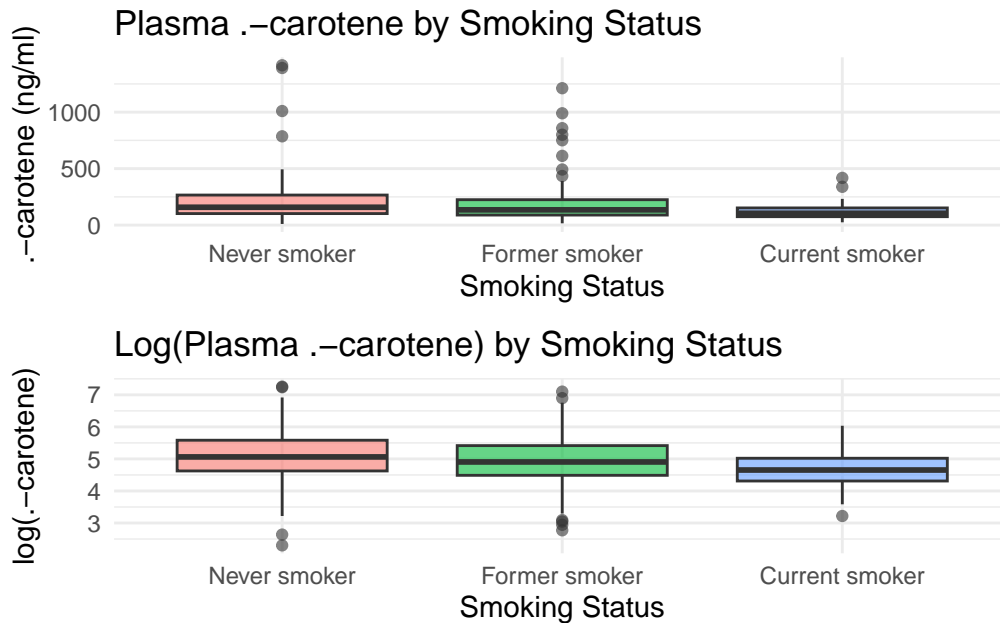
Given the three smoking categories — *Never smoker*, *Former smoker*, and *Current smoker* — we choose **Never smoker** as the reference category in our regression modeling. This choice is motivated by the following:

- 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  $\beta$ -carotene levels.

As a result, we will refer to this as **Categorical Smoking Model** 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  $\beta$ -carotene and log-transformed  $\beta$ -carotene across smoking status:



The boxplots reveal substantial skewness and variability in the raw  $\beta$ -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  $\log(\beta\text{-carotene})$  as the dependent variable in regression modeling.

## 2.2 Modeling $\beta$ -carotene and Smoking Status

### Comparing Reference Categories in Categorical Regression

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

- In **Model A**, the reference category is “**Never smoker**”
- In **Model B**, 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  $\log(\beta\text{-carotene})$  level for the **reference category**. The other coefficients represent the difference in mean  $\log(\beta\text{-carotene})$  compared to the reference.
- In Model A, the intercept corresponds to “Never smoker”, and the coefficients show how much lower (on average) former or current smokers are. In Model B, 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.

## 2.3 Predicted $\beta\text{-carotene}$ Levels by Smoking Group

To further understand the relationship between smoking habits and plasma  $\beta\text{-carotene}$  concentration, we now compute predicted values for each smoking group using both model versions. We calculate:

- The **expected  $\log(\beta\text{-carotene})$**  level with 95% confidence intervals.
- The **expected  $\beta\text{-carotene}$  (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 %
<b>Never Smoker</b>	4.933	5.051	5.169
<b>Former Smoker</b>	4.803	4.941	5.079
<b>Current Smoker</b>	4.389	4.614	4.839

beta-carotene	2.5 %	Estimate	97.5 %
<b>Never Smoker</b>	138.8	156.2	175.7
<b>Former Smoker</b>	121.9	139.9	160.6
<b>Current Smoker</b>	80.5	100.9	126.3

## Interpretation

The predicted values and their 95% confidence intervals are **identical** for both models:

- For **Never smokers**, the estimated plasma  $\beta$ -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**, as expected. 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  $\beta$ -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  $\beta$ -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:** 2 and 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  $\beta$ -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.

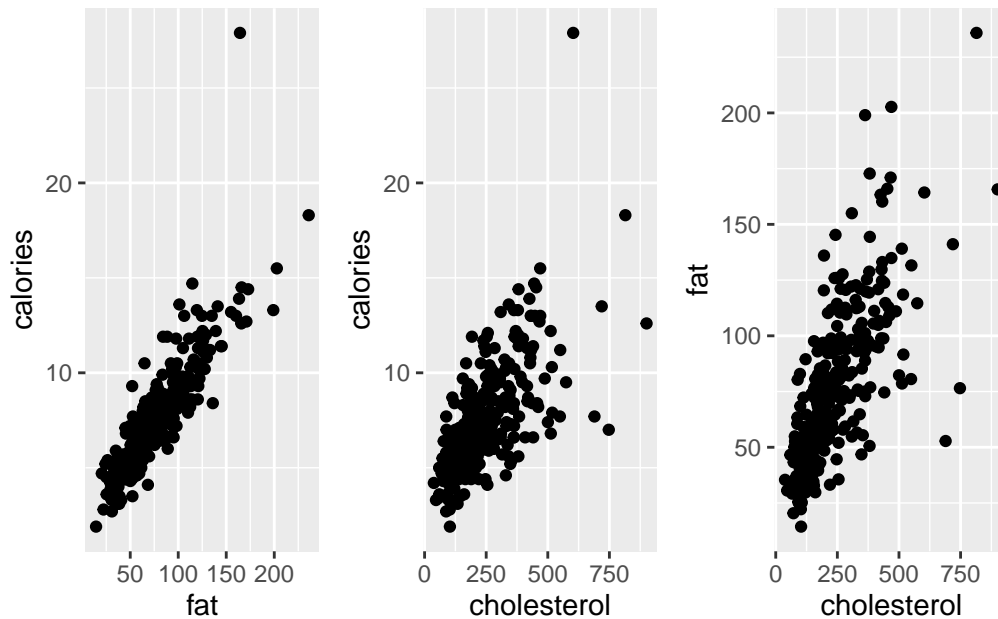
For `vituse`, the most common category is "often" (vitamin use fairly often). However, from a domain knowledge perspective, "no" (no vitamin use) makes the most sense as a reference group, because it represents the absence of intervention. This allows interpretation of coefficients as the effect of vitamin 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  $\pm 0.6$ , which might indicate collinearity issues if both variables are included in the same regression model. The results are visualized in scatterplots for the highly correlated pairs.



The correlation table and scatterplots reveal three particularly strong linear relationships:

- **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.

### 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  $\beta$ -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:

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

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**

### 3.4 Hypothesis Testing and Model Comparison

We now use **Model 3.3** to test specific hypotheses about the relationships between log-transformed plasma  $\beta$ -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  $\beta$ -carotene scale).

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

	Log-scale $\beta$	$\exp(\beta)$	2.5 % $\exp(\beta)$	97.5 % $\exp(\beta)$
Intercept	5.2702253504	194.4597791	85.3053474	443.2852903
BMI	-0.0316141867	0.9688803	0.9565275	0.9813927
Age	0.0054407584	1.0054556	0.9998229	1.0111200
Fat	0.0013103650	0.9986905	0.9953907	1.0020012
Cholesterol	0.0007114657	0.9992888	0.9984636	1.0001147
Fiber	0.0230049533	1.0232716	1.0060481	1.0407900
Alcohol	0.0018106489	1.0018123	0.9953881	1.0082780
Betadiet	0.0550358051	1.0565784	0.9963092	1.1204936
Smokstat	0.0690739529	0.9332577	0.7897174	1.1028880
Former Smokstat	0.2715769150	0.7621767	0.5975556	0.9721492
Current Sex	0.2180853684	1.2436932	0.9703005	1.5941174
Vituse	0.1307604385	0.8774279	0.8019314	0.9600319

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  $\beta$ -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  $\log(\beta\text{-carotene})$  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 **Log-transformed BMI model** ( $\log(\beta\text{-carotene}) \sim \text{bmi}$ ) to the full **Model 3.3**

- **Null hypothesis:**  $H_0 : \beta_{age} = \beta_{fat} = \dots = \beta_{vituse} = 0$  - 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:** 6.6928
- **Degrees of freedom:** 10 and 302
- **P-value:** 1.63e-12

**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?**

- **Null hypothesis:**  $H_0 : \beta_{\text{bmi}} = \beta_{\text{age}} = \dots = \beta_{\text{vituse}} = 0$  - 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:** 6.6928
- **Degrees of freedom:** 10 and 302
- **P-value:** 1.63e-12

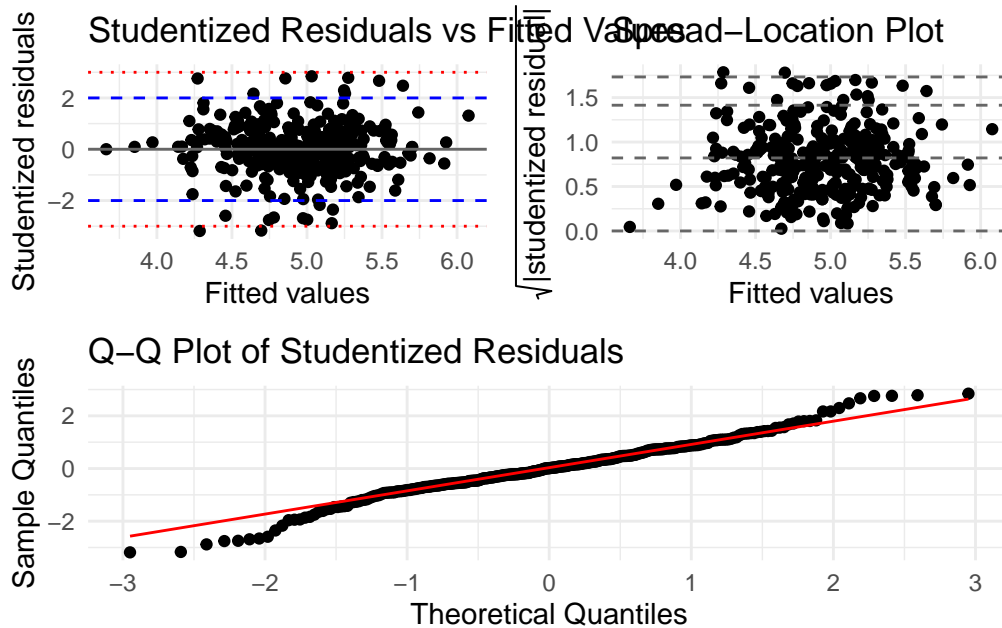
**Conclusion:** Again, we **reject the null hypothesis**. Model 3(c) adds substantial explanatory power beyond smokstat alone.

### 3.5 Residual Diagnostics for Model 3.3

To assess the adequacy of Model 3.3, 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.



The **residuals vs fitted** plot shows no clear pattern, suggesting that the linearity assumption is reasonable. A few residuals fall beyond  $\pm 3$ , 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.

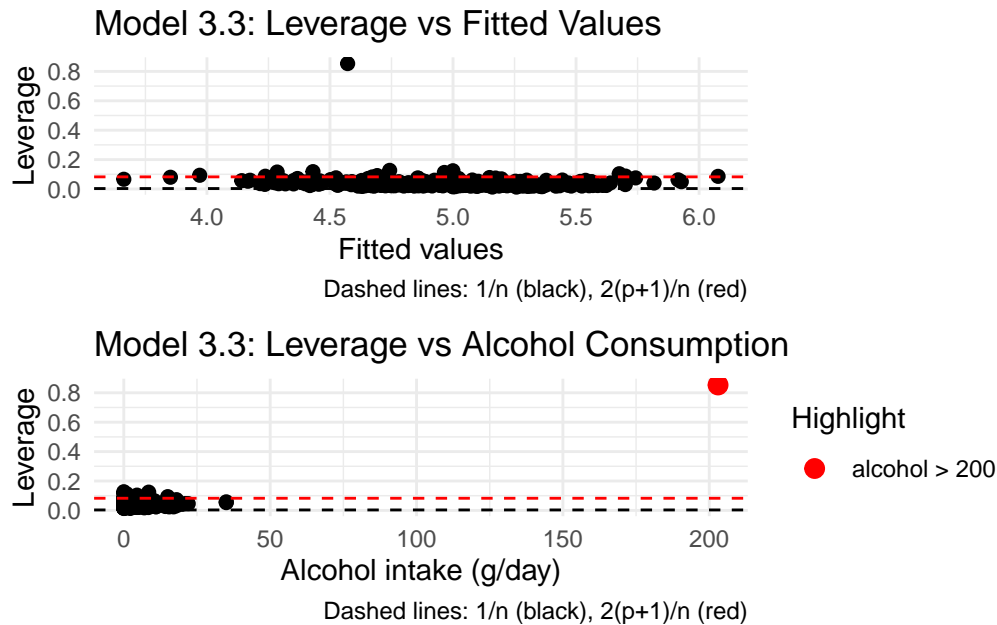
**Conclusion:** Overall, the diagnostic plots support the adequacy of Model 3.3. The residuals show no major violations of linearity, constant variance, or normality assumptions.

### 3.6. Leverage Analysis for Model 3.3

We now calculate the leverage values for each observation in Model 3.3 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.



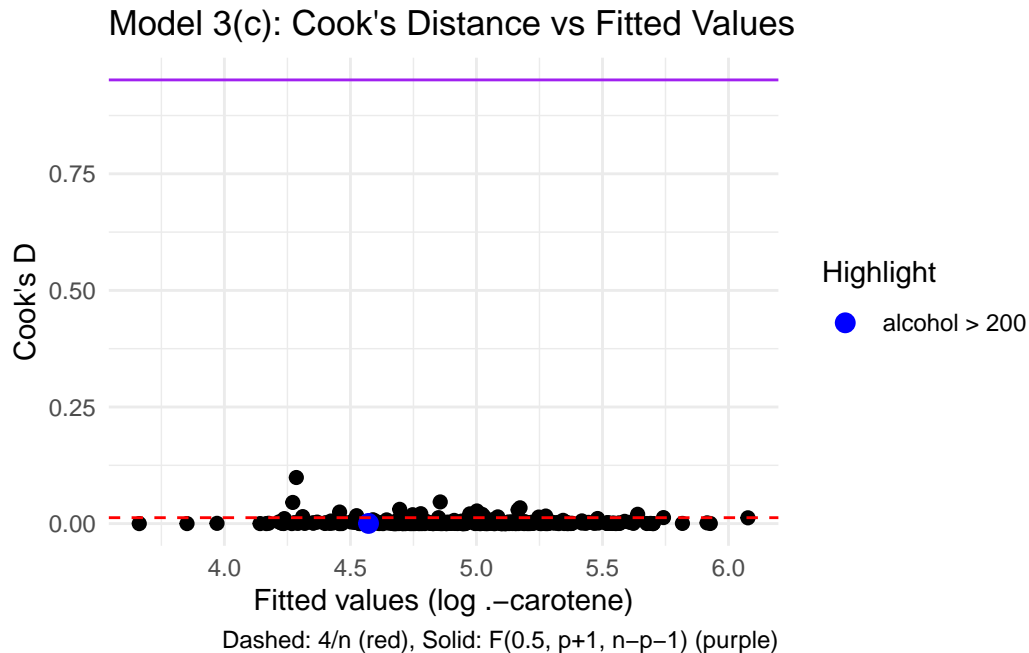


The top plot reveals one observation with leverage well above the red line, marking it as potentially influential. When plotted against alcohol consumption, this same observation clearly stands out — it corresponds to the individual consuming over 200 grams of alcohol daily (as previously identified 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 3.3

To evaluate the influence of individual observations on the parameter estimates of **Model 3.3**, we calculate **Cook's distance** for each observation and visualize it against the fitted values.



```
# A tibble: 1 x 15
  age sex    smokstat      bmi vituse calories  fat fiber alcohol cholesterol
<dbl> <fct> <fct>      <dbl> <fct>      <dbl> <dbl> <dbl> <dbl>      <dbl>
1   40 female Never smok~  31.2 often      12.6  166.  14.4      0      901.
# i 5 more variables: betadiet <dbl>, betaplasma <dbl>, yhat_model3c <dbl>,
#   r_model3c <dbl>, cooksD <dbl>
```

The plots show that **no observations exceed common Cook's distance thresholds** (either  $4/n$  or the F-distribution-based cutoff). The individual with extreme alcohol consumption—over 200g/day—has the **largest leverage**, but has an acceptable and average Cook's distance. This indicates that although the observation has **high leverage**, it **does not exert undue influence** on the parameter estimates of the model.

We identify the observation with the largest Cook's distance of 0.0988. For this observation we find the DFBETAS.

```
# A tibble: 13 x 2
  term          dfbeta
<chr>         <dbl>
1 cholesterol -0.898
2 betadiet     0.279
3 sexfemale   -0.272
```

4	vituseseldom	0.231
5	(Intercept)	0.200
6	fat	0.187
7	vituseno	0.178
8	smokstatCurrent smoker	0.171
9	smokstatFormer smoker	0.148
10	alcohol	0.0979
11	fiber	-0.0476
12	age	-0.0251
13	bmi	-0.0247

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

