Table of contents

1 Project 1 - Report: Linear Regression

1.1 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.

1.2 Author Contributions

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

2 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.

3 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)

Variable	Description
`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.

4 1. Testing Model Assumptions: Linear vs Log-transformed

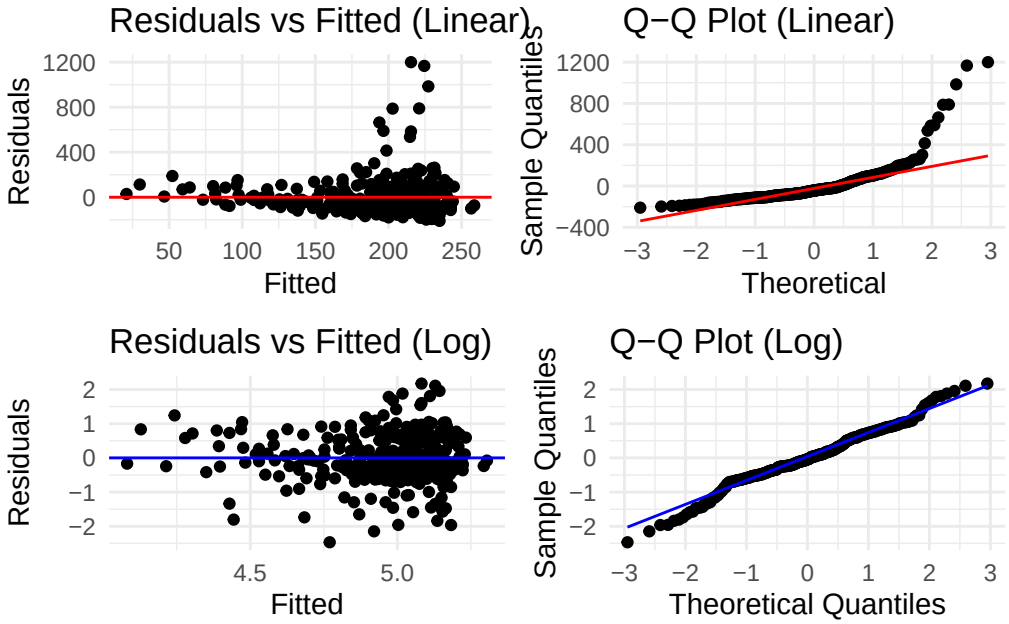
We fitted two models to examine the relationship between BMI and plasma β-carotene levels:

- **Linear model**: `betaplasma ~ bmi`
- **Log-transformed 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.

4.1 1.2 Model Estimates

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

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 describes the expected **multiplicative change** in β-carotene concentration for each one-unit increase in BMI.

```
            Estimate   2.5 %  97.5 %
(Intercept)   5.8896  5.5273  6.2519
bmi          -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**, implying an **approximate 3.5% reduction** in β-carotene levels per BMI unit.

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

4.1.1 1.2.2 Linear Model confidence and prediction intervals with Log Transformation and Back-Transformation

To investigate how plasma β-carotene levels relate to BMI, we fit a linear regression model where the outcome was **log-transformed β-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 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.

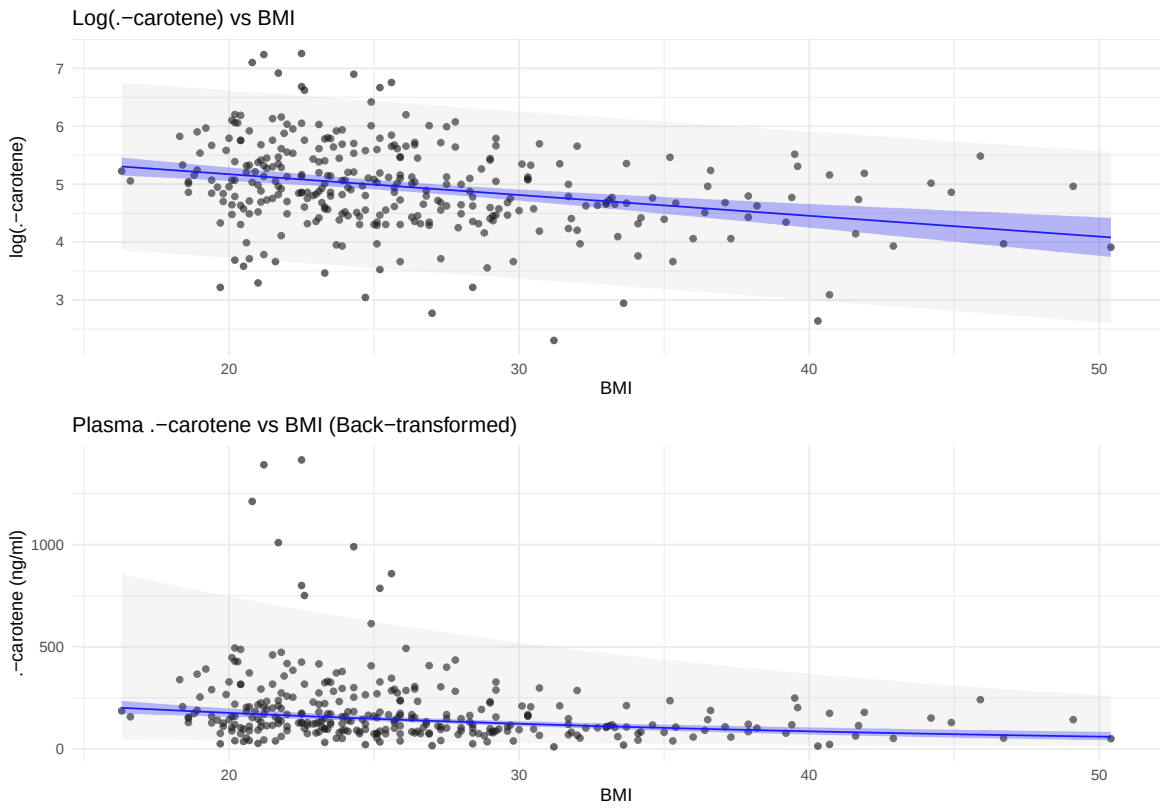


Figure 1: Log-scale and back-transformed plots with 95% CI and prediction intervals.

The log-scale plot shows a clearer linear trend and narrower intervals, while the back-transformed version reveals the actual magnitude of plasma 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.

4.2 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:

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

```
Estimated % change in plasma  -carotene (95% CI)
```

	BMI.Change	Estimate....	X95..CI.Lower....	X95..CI.Upper....
1	+1	-3.52	-4.82	-2.21
2	-1	3.65	2.26	5.06
3	-10	43.17	25.09	63.86

From the table, 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.

4.3 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 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 β-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(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)** strengthens the evidence against the null hypothesis.

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

5 2.1 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 frequency table showing the number of individuals in each smoking group, and calculate the mean and standard deviation of both **plasma β-carotene (ng/ml)** and **log-transformed plasma β-carotene** for each group.

These summaries help us understand both the distribution of the smoking status variable and the differences in β-carotene levels across categories.

```
[1] "Frequency table for smokstat:"
```

```
  Never smoker  Former smoker Current smoker
           157            115             43
```

```
[1] "Summary statistics by smoking category:"
```

```
# A tibble: 3 x 6
  smokstat       Count Mean_beta SD_beta Mean_log SD_log
  <fct>          <int>     <dbl>   <dbl>    <dbl>  <dbl>
1 Never smoker     157      206.   193.      5.05  0.745
2 Former smoker    115      193.   192.      4.94  0.798
3 Current smoker    43      121.    78.8     4.61  0.624
```

These summaries will later guide our modeling choices, including the selection of a reference category for regression and whether a log transformation of the response variable remains appropriate when including smokstat as a predictor.

Reference Category Choice and Distribution Plots

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:

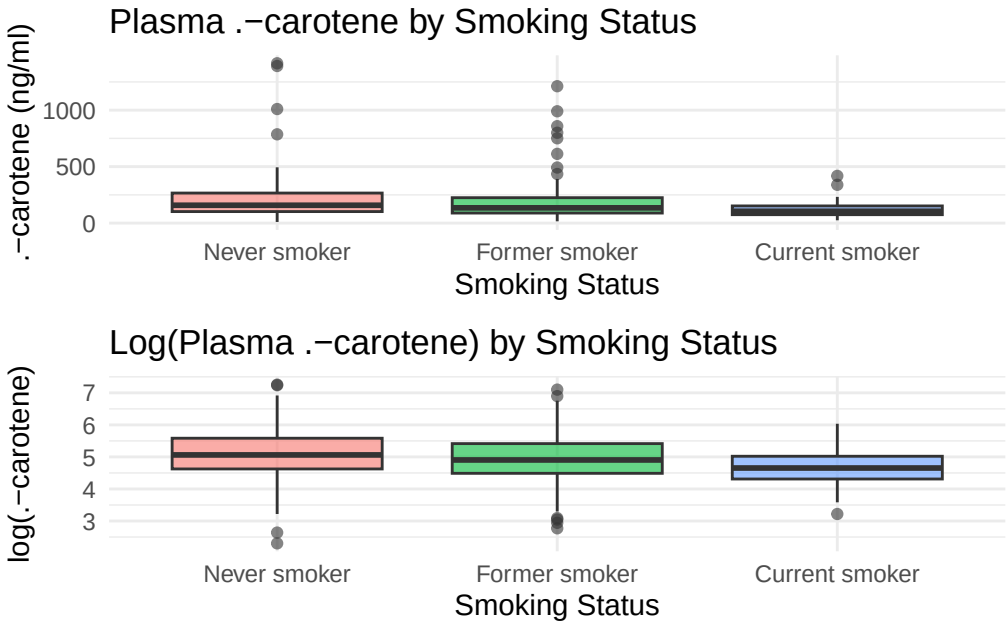
• It is the **largest group** (n = 157), ensuring stable estimates when used as baseline.

• 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.

5.0.1 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:



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.

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

```
Model with 'Never smoker' as reference:

                                Estimate Std. Error   t value      Pr(>|t|)
(Intercept)                    5.0508487 0.05986447 84.371387 7.460874e-217
smokstat_neverFormer smoker   -0.1097225 0.09206715 -1.191766  2.342586e-01
smokstat_neverCurrent smoker -0.4372105 0.12910704 -3.386418  7.987562e-04
```

```
Model with 'Current smoker' as reference:

                               Estimate Std. Error   t value      Pr(>|t|)
(Intercept)                   4.6136382  0.1143891 40.332837 8.462413e-126
smokstat_currentNever smoker  0.4372105  0.1291070  3.386418  7.987562e-04
smokstat_currentFormer smoker 0.3274879  0.1340801  2.442479  1.514140e-02
```

5.1.1 Interpretation of Dummy Coding

• 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 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.

5.1.2 Which Reference Category Is More Reasonable?

We argue that **"Never smoker" is the more reasonable reference category**, for several reasons:

1. **Interpretability**: Using a baseline group that is not affected by the exposure (smoking) allows for clearer interpretation of how smoking status impacts β-carotene levels. It's more intuitive to interpret how smoking decreases levels relative to a clean baseline (non-smoker).
2. **Scientific Relevance**: "Never smokers" likely represent the physiological baseline, whereas "Current smokers" are more likely to be a group with altered biology.
3. **Model Stability**: Model A exhibits **smaller standard errors** for the intercept and slightly more balanced standard errors overall. This suggests more stable estimation around the baseline.

As a result, we will refer to **Model A (with "Never smoker" as the reference)** as **Model 2(b)** in subsequent analyses.

5.2 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:

- **Model A**: Uses "Never smoker" as the reference category.
- **Model B**: Uses "Current smoker" as the reference category.

For each model, 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).

We compute these predictions for: - Never smokers - Former smokers - Current smokers

```
           Group Log_ModelA_Estimate Log_ModelA_Lower Log_ModelA_Upper
1   Never smoker               5.051            4.933            5.169
2  Former smoker               4.941            4.803            5.079
3 Current smoker               4.614            4.389            4.839
  Log_ModelB_Estimate Log_ModelB_Lower Log_ModelB_Upper Beta_ModelA_Estimate
1               5.051            4.933            5.169                156.2
2               4.941            4.803            5.079                139.9
3               4.614            4.389            4.839                100.9
  Beta_ModelA_Lower Beta_ModelA_Upper Beta_ModelB_Estimate Beta_ModelB_Lower
1             138.8             175.7                156.2             138.8
2             121.9             160.6                139.9             121.9
3              80.5             126.3                100.9              80.5
  Beta_ModelB_Upper
1             175.7
2             160.6
3             126.3
```

5.2.1 Interpretation

We see that the predicted values and their 95% confidence intervals are **identical** for both models:

- 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**, 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.

5.3 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 fit the model:

```
Analysis of Variance Table

Response: log(betaplasma)
                  Df Sum Sq Mean Sq F value  Pr(>F)
smokstat_current   2   6.47  3.2352    5.75 0.00353 **
Residuals        312 175.55  0.5626
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


F-statistic: 5.75

Degrees of freedom: 2 and 312

P-value: 0.00353
```

We then 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:** 2 and 312

• **P-value:** 0.00353

5.3.1 Conclusion

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.

6 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
    42    273
```

```
 often seldom     no
   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.

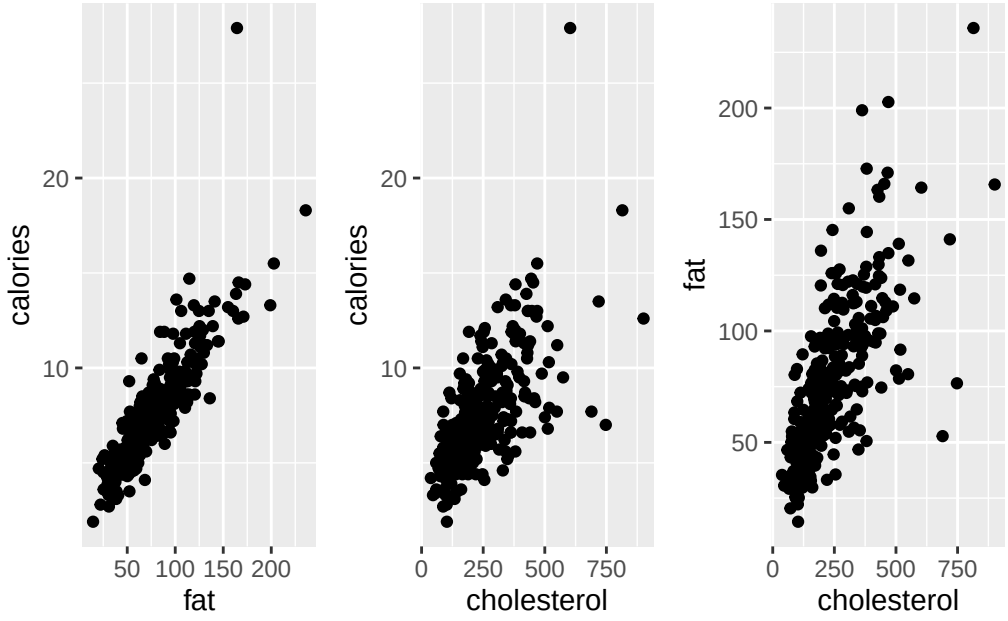
6.1 3.2 Pairwise Correlation Analysis and Outlier Detection

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 results are visualized in scatterplots for the highly correlated pairs.

```
# A tibble: 3 x 8
  var1        var2          cor statistic        p conf.low conf.high method
  <chr>       <chr>       <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
1 calories    fat          0.87      31.5 6.27e-99    0.842     0.896 Pearson
2 calories    cholesterol  0.66      15.5 1.13e-40    0.592     0.718 Pearson
3 cholesterol fat          0.71      17.8 1.45e-49    0.650     0.761 Pearson
```




```
# A tibble: 1 x 12
    age sex   smokstat       bmi vituse calories   fat fiber alcohol cholesterol
  <dbl> <fct> <fct>        <dbl> <fct>     <dbl> <dbl> <dbl>   <dbl>       <dbl>
1    65 male  Current smo~  23.4 no         27.9  164.  11.3     203         603
# i 2 more variables: betadiet <dbl>, betaplasma <dbl>
```

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

- **Fat and Calories**: r 0.87
- **Cholesterol and Calories**: r 0.66
- **Cholesterol and Fat**: r 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.

6.1.1 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. According to the assignment, 12 bottles of vodka per week equates to about 12 MJ/day. 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.

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.

6.2 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:

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

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.

```
                GVIF Df GVIF^(1/(2*Df))
bmi         1.069660  1        1.034244
age         1.307586  1        1.143497
calories   13.210244  1        3.634590
fat         8.175794  1        2.859334
cholesterol 2.195956  1        1.481876
fiber       2.504249  1        1.582482
alcohol     2.564752  1        1.601484
betadiet    1.338719  1        1.157030
smokstat    1.178201  2        1.041849
sex         1.287887  1        1.134851
vituse      1.149879  2        1.035531
```

As seen above, **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 (as observed in task 3(b)).

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

```
               GVIF Df GVIF^(1/(2*Df))
bmi         1.067334  1        1.033119
age         1.219329  1        1.104232
fat         2.244437  1        1.498144
cholesterol 2.129296  1        1.459211
fiber       1.465013  1        1.210377
alcohol     1.124159  1        1.060264
betadiet    1.338004  1        1.156721
```

```
smokstat    1.177041  2        1.041593
sex         1.287797  1        1.134811
vituse      1.115890  2        1.027792
```

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(c)**

6.3 3.4 Hypothesis Testing and Model Comparison

We now use **Model 3(c)** 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:

```
          (Intercept)                   bmi                   age
         5.3239510026          -0.0320296627          0.0060022886
                  fat           cholesterol                 fiber
        -0.0012301144          -0.0007325298          0.0227289930
              alcohol              betadiet  smokstatFormer smoker
         0.0018263703           0.0558579827         -0.0718425845
smokstatCurrent smoker              sexfemale           vituseseldom
        -0.2728505836           0.2010518069          0.0012546629
             vituseno
        -0.2657965768

          (Intercept)                   bmi                   age
            205.1930006             0.9684779             1.0060203
                  fat           cholesterol                 fiber
              0.9987706             0.9992677             1.0229893
              alcohol              betadiet  smokstatFormer smoker
              1.0018280             1.0574475             0.9306774
smokstatCurrent smoker              sexfemale           vituseseldom
              0.7612065             1.2226881             1.0012555
               vituseno
              0.7665951
```

```
                             2.5 %      97.5 %
(Intercept)            108.6693300 387.4521681
bmi                      0.9561460   0.9809688
age                      1.0003499   1.0117229
fat                      0.9954761   1.0020761
cholesterol              0.9984439   1.0000923
fiber                    1.0058037   1.0404685
alcohol                  0.9954176   1.0082798
betadiet                 0.9972458   1.1212835
smokstatFormer smoker    0.7877884   1.0994836
smokstatCurrent smoker   0.5971060   0.9704062
sexfemale                0.9534958   1.5678791
vituseseldom             0.8258204   1.2139594
vituseno                 0.6405452   0.9174496
```

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(c), testing:

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

We extract the result from the model summary:

```
Call:
lm(formula = log(betaplasma) ~ bmi + age + fat + cholesterol +
    fiber + alcohol + betadiet + smokstat + sex + vituse, data = study_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05637 -0.37058  0.00815  0.41219  1.86637

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.3239510  0.3230131  16.482  < 2e-16 ***
bmi                    -0.0320297  0.0065122  -4.918 1.44e-06 ***
age                     0.0060023  0.0028724   2.090  0.03749 *
fat                    -0.0012301  0.0016790  -0.733  0.46435
```

```
cholesterol            -0.0007325  0.0004191  -1.748  0.08154 .
fiber                   0.0227290  0.0086094   2.640  0.00872 **
alcohol                 0.0018264  0.0032621   0.560  0.57598
betadiet                0.0558580  0.0297868   1.875  0.06172 .
smokstatFormer smoker  -0.0718426  0.0847032  -0.848  0.39702
smokstatCurrent smoker -0.2728506  0.1233885  -2.211  0.02776 *
sexfemale               0.2010518  0.1263674   1.591  0.11265
vituseseldom            0.0012547  0.0978899   0.013  0.98978
vituseno               -0.2657966  0.0912869  -2.912  0.00386 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6718 on 302 degrees of freedom
Multiple R-squared:  0.2511,    Adjusted R-squared:  0.2214
F-statistic: 8.439 on 12 and 302 DF,  p-value: 8.684e-14
```

- `**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(b)** (log(β-carotene) ~ bmi) to the full **Model 3(c)**.

- **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

```
Analysis of Variance Table

Model 1: log(betaplasma) ~ bmi + age + fat + cholesterol + fiber + alcohol +
    betadiet + smokstat + sex + vituse
Model 2: log(betaplasma) ~ bmi
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    302 136.31
2    313 167.39 -11   -31.079 6.2597 2.633e-09 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **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(c) significantly better than the model from chapter 2, which only used smokstat?

- **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

```
Analysis of Variance Table

Model 1: log(betaplasma) ~ bmi + age + fat + cholesterol + fiber + alcohol +
    betadiet + smokstat + sex + vituse
Model 2: log(betaplasma) ~ smokstat_current
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    302 136.31
2    312 175.55 -10   -39.236 8.6928 1.632e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **F-statistic**: 6.6928
- **Degrees of freedom**: 10 and 302
- **P-value**: 1.63e-12

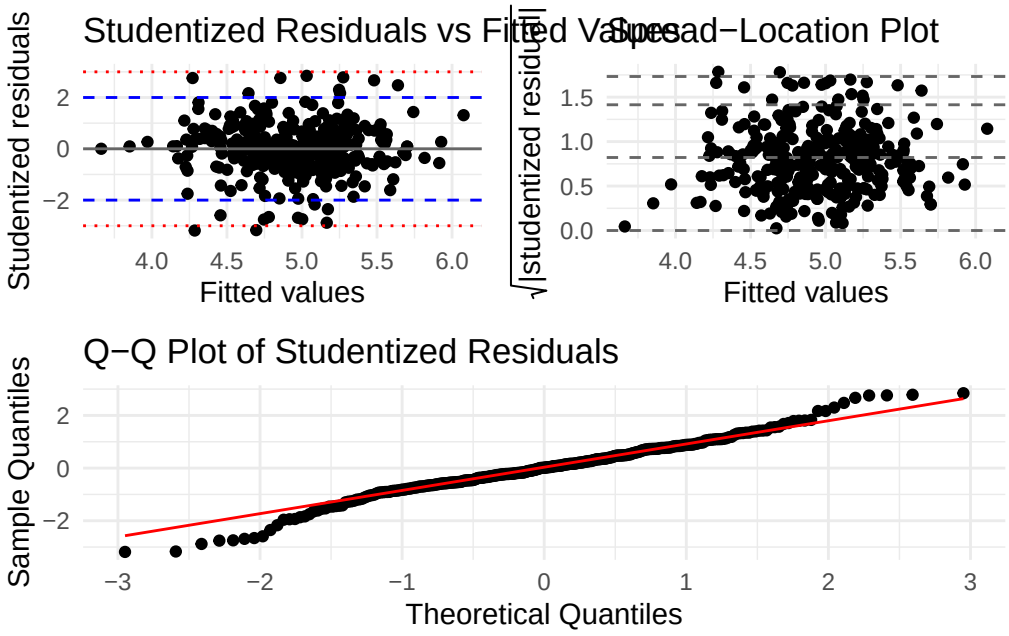
Conclusion: Again, we **reject the null hypothesis**. Model 3(c) adds substantial explanatory power beyond smokstat alone.

6.4 3.5 Residual Diagnostics for Model 3(c)

To assess the adequacy of Model 3(c), 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 ± 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.

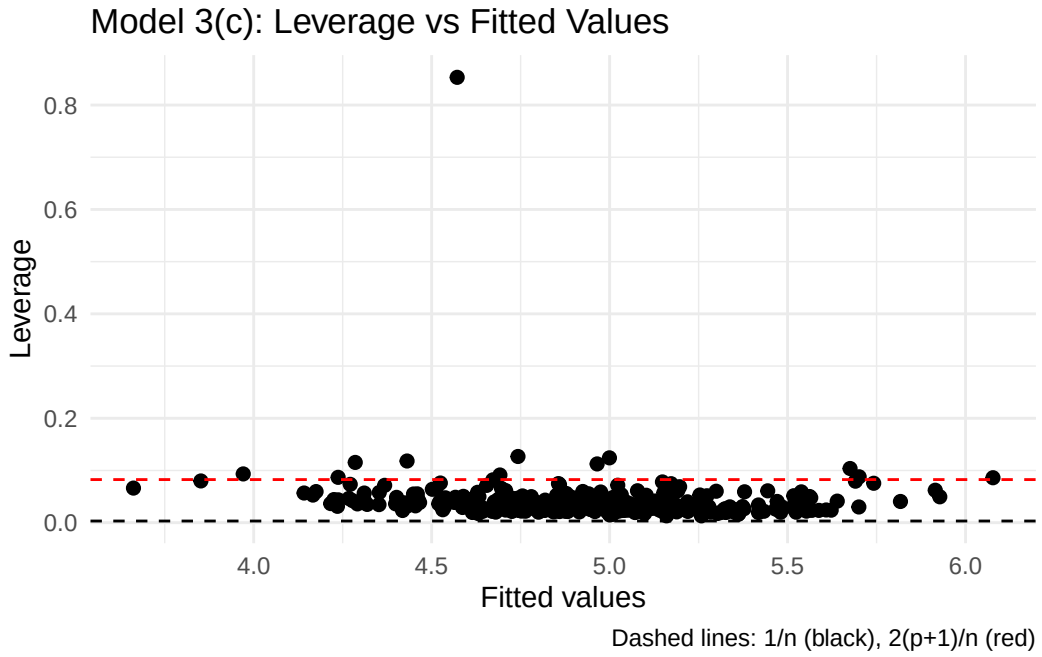
In the **Q-Q plot**, the points generally follow the red reference line, though some deviation is observed in the tails. This indicates slight departure from normality, but not to a degree that would seriously undermine the model.

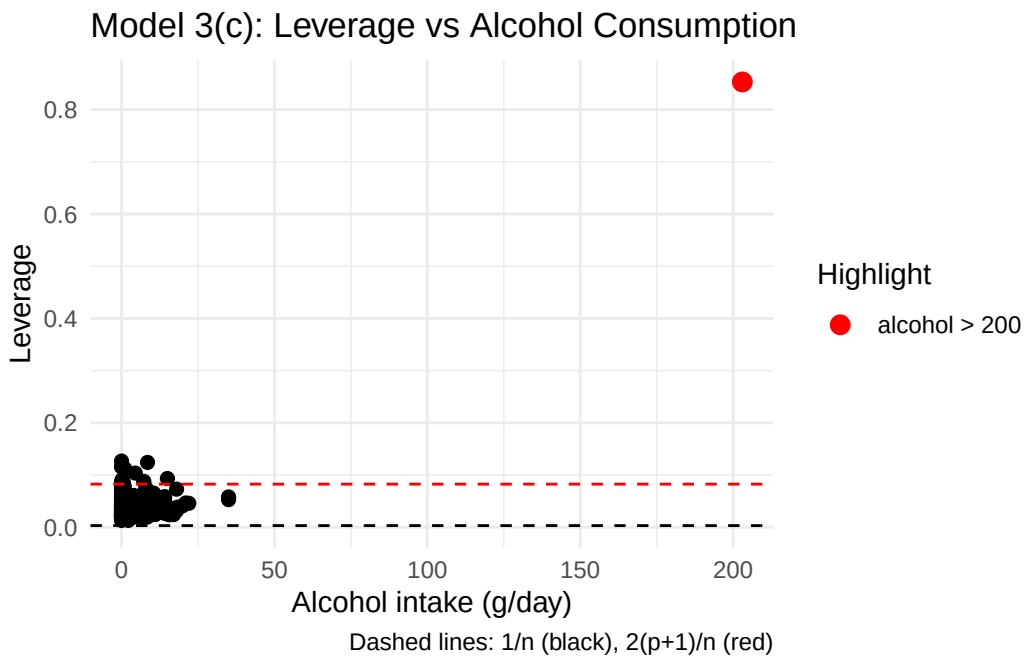
Conclusion: Overall, the diagnostic plots support the adequacy of Model 3(c). The residuals show no major violations of linearity, constant variance, or normality assumptions.

6.5 3(f). Leverage Analysis for Model 3(c)

We now calculate the leverage values for each observation in Model 3(c) 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-left 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.