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 (noncancerous) 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)
`smokstat`	Smoking status (1 = Never, 2 = Former, 3 = Current)
`bmi`	Body mass index (BMI = weight/height², kg/m^2)
`vituse`	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
`calories`	Daily calorie intake (MJ)
`fat`	Fat consumed per day (g)
`fiber`	Fiber consumed per day (g)
`alcohol`	Alcoholic drinks per week
`cholesterol`	Daily cholesterol intake (mg)
`betadiet`	Dietary β-carotene intake per day (mg)

Variable	Description
`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.

3.1 Testing Model Assumptions: Linear vs Log-transformed

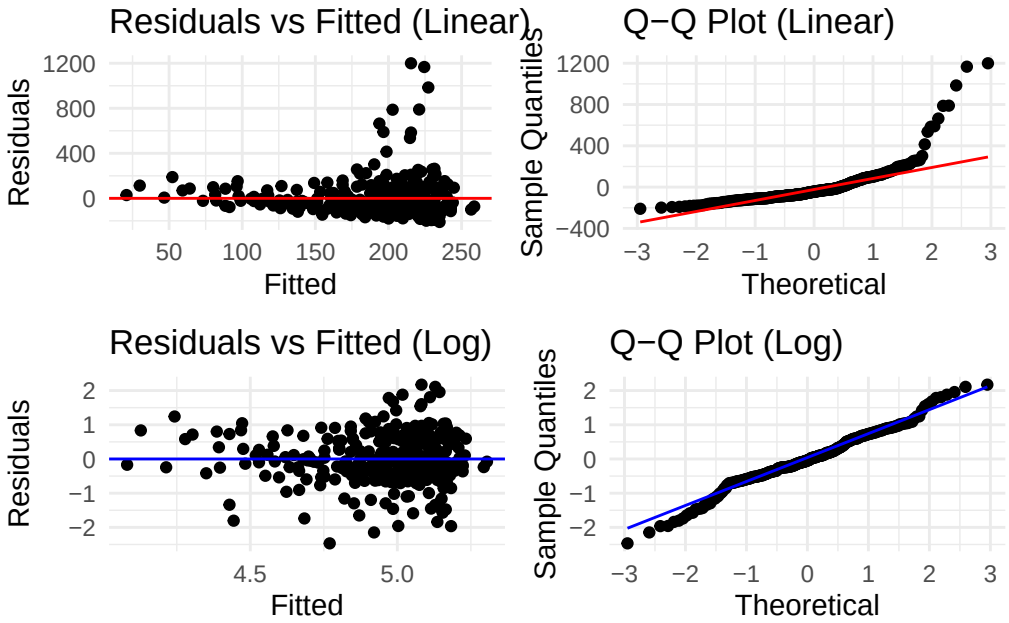
We fitted two models to examine the relationship between BMI and plasma β-carotene levels:

- **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.

3.1.1 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.

Table 3: β estimates and 95% confidence intervals for log-linear model

	Estimate	2.5 %	97.5 %
(Intercept)	5.8896	5.5273	6.2519
bmi	-0.0359	-0.0494	-0.0224

The estimate for **BMI** 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.

3.2 Linear Model 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.

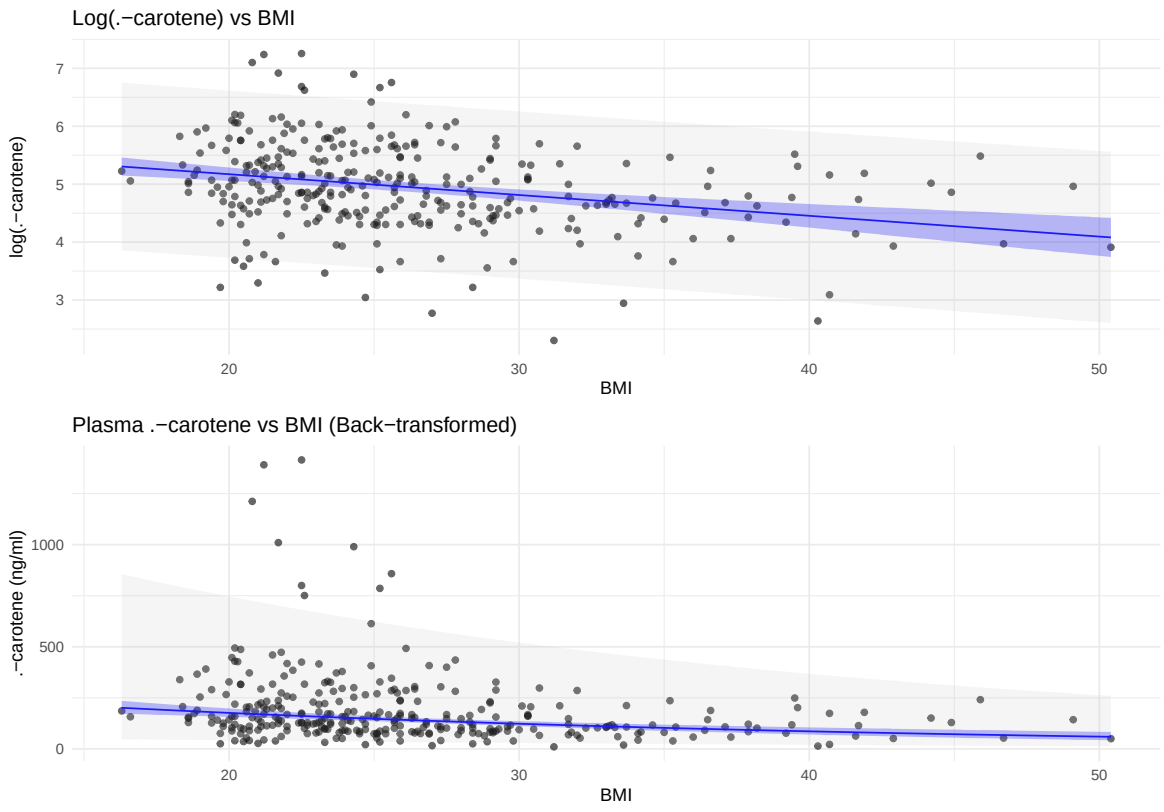


Figure 1: Log-scale and back-transformed plots with 95% CI and prediction intervals.

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.

3.2.1 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**

Table 4: Estimated percentage change in plasma β-carotene for selected changes in BMI (with 95% CI)

BMI.Change	Estimate….	X95..CI.Lower….	X95..CI.Upper….
+1	-3.52	-4.82	-2.21
−1	3.65	2.26	5.06
−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, 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 β-carotene concentration.
- 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.

3.2.2 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).

```
[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
```