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:

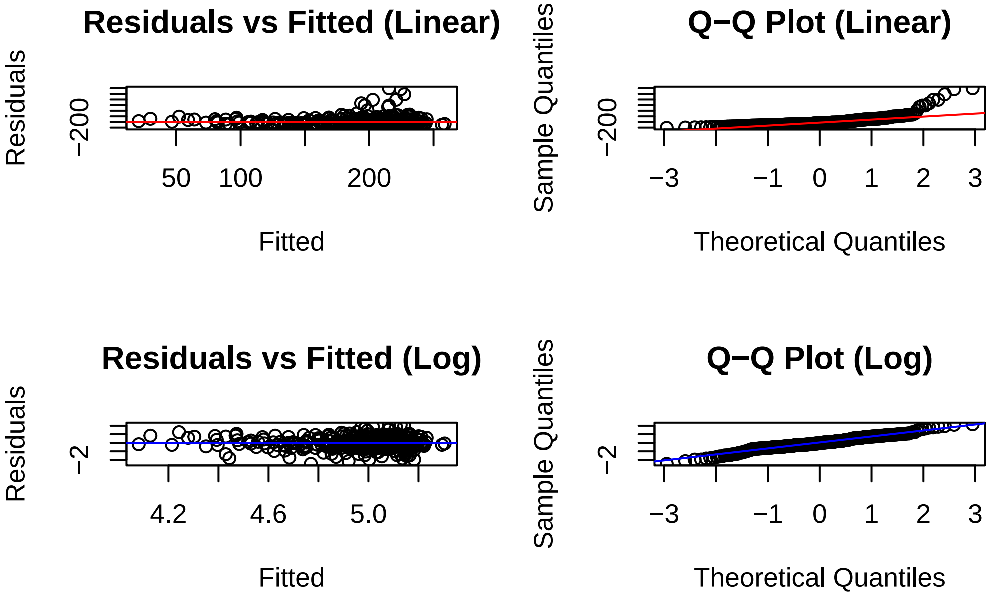
Variable	Description
`age`	Age (years)
`sex`	Sex (1 = Male, 2 = Female)
`smokstat`	Smoking status (1 = Never, 2 = Former, 3 = Current)
`bmi`	Body mass index (BMI = weight/height², kg/m²)
`vituse`	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
`calories`	Daily calorie intake (MJ)
`fat`	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.



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 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.