Association Between Dietary Patterns and Blood Lipid Profiles Among Chinese Women

Stat Project

Contents

1 PART I: PROJECT DOCUMENTATION

2 Project Overview

2.1 Project Description

Project Summary

This project analyzes the association between **dietary patterns** and **blood lipid profiles** among Chinese women, replicating and extending the methodology of a published nutritional epidemiology study. The analysis employs multivariate statistical techniques to identify dietary patterns and examine their relationships with cardiovascular biomarkers.

Core Research Question

How do dietary patterns (derived from macronutrient intake) associate with blood lipid biomarkers in Chinese women?

2.2 Project Context and Source

Repository: github.com/natej-ghodbane/projetStat

Original State: The project was provided as a foundational statistical analysis with:

- R Markdown analysis file (`test.Rmd`)
- Standalone R script (`script.R`)
- Two CSV datasets (biomarker and dietary data)
- Comprehensive README documentation

2.3 Project Objectives

2.3.1 Primary Statistical Objectives

#	Objective
1	**Pattern Identification:** Use Principal Component Analysis (PCA) to derive dietary patterns from macronutrient intake data
2	**Association Analysis:** Examine relationships between dietary patterns and four blood lipid biomarkers: • HDL-C (High-Density Lipoprotein Cholesterol) • LDL-C (Low-Density Lipoprotein Cholesterol) • TG (Triglycerides) • TC (Total Cholesterol)
3	**Quartile Comparisons:** Assess lipid level differences across dietary pattern adherence quartiles
4	**Methodological Rigor:** Apply appropriate parametric and non-parametric statistical methods with assumption checking

3 Tools and Technologies

3.1 Software Environment

Table 1: Primary Software Tools

Tool	Description
R Statistical Software	Industry-standard statistical computing environment (version 4.x) for all statistical analyses, data manipulation, and visualization
RStudio IDE	Integrated development environment providing code editing, debugging, chunk execution, and document knitting for R programming
R Markdown	Literate programming framework (`test.Rmd`) combining code, results, and narrative for reproducible research with PDF and HTML output
Git + GitHub	Version control system with remote repository (natejghodbane/projetStat) for code sharing and collaboration

3.2 R Packages and Libraries

Table 2: Key R Packages Used in Analysis

Package	Category	Purpose
`psych`	Multivariate Analysis	Principal Component Analysis with varimax rotation, factor score calculation, and dietary pattern identification
`ggplot2`	Data Visualization	Grammar-of-graphics based plotting system for scree plots, factor loadings, and publication-quality graphics
`reshape2`	Data Manipulation	Data frame reshaping and transformation using `melt()` for visualization preparation
`lmtest`	Regression Diagnostics	Statistical tests including Breusch-Pagan (homoscedasticity) and Durbin-Watson (autocorrelation)
`car`	Regression Diagnostics	Variance Inflation Factor (VIF) calculation for multicollinearity assessment (planned enhancement)
`knitr`	Results Presentation	Dynamic report generation with `kable()` for professional table formatting and document creation

3.3 Statistical Methods Toolbox

3.3.1 Implemented Techniques

Table 3: Statistical Methods and R Implementation

Method	**R Function/Package**	**Purpose**
PCA	`psych::principal()`	Dietary pattern extraction
Standardization	`scale()`	Variable normalization
Linear Regression	`lm()`	Pattern-lipid associations
ANOVA	`aov()`	Group comparisons
Post-hoc Tests	`TukeyHSD()`	Pairwise comparisons
Chi-square Test	`bartlett.test()`	Variance homogeneity
Kruskal-Wallis	`kruskal.test()`	Non-parametric ANOVA
Confidence Intervals	`confint()`	Parameter precision

4 Original Project State Assessment

4.1 Original Methodology Implemented

4.1.1 Data Processing

Implemented Steps:

1. Data loading from CSV files
2. Filtering to 2009 wave
3. Dataset merging by participant ID
4. Variable selection for PCA
5. Z-score standardization

4.1.2 Statistical Analyses Present

Multivariate Analysis:

- Principal Component Analysis with varimax rotation
- Three-component extraction based on eigenvalues > 1
- Factor score calculation
- Quartile classification

Visualizations:

- Scree plot
- Factor loadings bar chart
- Distribution histograms

Regression Analysis:

- Simple linear regression (TG $\sim$ fat intake)
- Multiple regression (4 models for each lipid biomarker)
- Diagnostic plots

Assumption Checking:

- Breusch-Pagan test (homoscedasticity)
- Durbin-Watson test (autocorrelation)
- Cook's distance (influential points)
- Residual plots

Group Comparisons:

- One-way ANOVA across quartiles
- Bartlett test for equal variances
- Kruskal-Wallis non-parametric alternative

5 PART II: STATISTICAL ANALYSIS

6 Introduction

Dietary pattern analysis, particularly using Principal Component Analysis (PCA), allows identification of dominant eating patterns within a population. Traditional nutritional analyses focus on individual nutrients; however, this approach does not reflect real-life dietary behavior where nutrients are consumed in combination. This analysis employs multivariate techniques to relate dietary patterns to blood lipid biomarkers (HDL-C, LDL-C, triglycerides, and total cholesterol).

7 Data and Methods

7.1 Data Loading

```
bio  <- read.csv("biomarker.csv")
diet <- read.csv("c12diet.csv")

str(bio)
```

```
## 'data.frame':    9549 obs. of  49 variables:
##  $ IDind       : num  2.11e+11 2.11e+11 2.11e+11 2.11e+11 2.11e+11 ...
##  $ UREA        : num  7.24 6.06 7.31 5.92 5.53 NA 5.38 6.46 4.71 6.15 ...
##  $ UA          : int  452 200 339 366 279 NA 341 236 197 398 ...
##  $ APO_A       : num  1.95 1.79 1.35 1.63 2.39 NA 2.18 1.41 1.35 1.79 ...
##  $ LP_A        : int  344 400 10 215 77 NA 187 118 79 34 ...
##  $ HS_CRP      : int  1 1 1 3 2 NA 1 1 3 1 ...
##  $ CRE         : num  83 71 100 75 70 NA 67 72 76 89 ...
##  $ HDL_C       : num  1.15 1.4 0.98 0.84 1.68 NA 1.42 1.19 1.03 1.02 ...
##  $ LDL_C       : num  3.71 3.54 2.66 2.59 2.58 NA 3.98 1.83 1.33 2.42 ...
##  $ APO_B       : num  1.47 1.22 0.94 1.19 0.89 NA 1.41 0.68 0.58 1.32 ...
##  $ MG          : num  0.97 0.84 0.91 0.98 0.93 NA 0.93 0.88 0.85 1.04 ...
##  $ FET         : num  154.6 45.9 168.8 106 32.1 ...
##  $ INS         : num  12.08 11.21 12.34 18.32 5.57 ...
##  $ HGB         : int  142 152 166 134 124 NA 134 120 122 157 ...
##  $ WBC         : num  4.2 4 7.5 5.4 4.4 NA 5.8 4.2 5.6 4.5 ...
##  $ RBC         : num  4.3 4.9 5.6 4.5 4.1 NA 4.8 4.1 4.1 5.3 ...
##  $ PLT         : int  173 221 166 393 253 NA 250 190 22 213 ...
##  $ Glu_field   : num  5.02 5.79 4.98 5.99 NA NA 4.83 4.81 5.06 6.01 ...
##  $ Y48_2       : int  51 13 21 17 NA NA 16 16 7 29 ...
##  $ Y48_3       : num  81.5 76.2 85.5 82 NA NA 72 74.6 75.6 70.5 ...
##  $ Y48_4       : num  51 45.4 46 45.6 NA NA 46.5 46.8 46.7 43.4 ...
##  $ Y48_5       : num  6.63 6.01 4.35 5.52 NA NA 6.52 3.85 2.69 6.06 ...
##  $ Y48_6       : num  2.55 1.32 0.8 3.17 NA NA 1.27 0.99 0.39 4.01 ...
##  $ HbA1c       : num  5.2 5.4 5 5.5 NA NA 5.1 5.7 5.4 5.9 ...
##  $ TP          : num  79.2 75.6 84.1 79.8 79.3 NA 71 74.9 78.5 67.7 ...
##  $ ALB         : num  51.8 45.4 46.5 48.6 47.1 NA 45 45.1 48.5 45.1 ...
##  $ GLUCOSE     : num  5.18 5.3 4.49 5.99 4.56 ...
##  $ TG          : num  3.13 1.24 0.94 3.16 1.26 NA 1.26 0.97 0.45 4.92 ...
##  $ TC          : num  6.4 5.65 4.03 5.27 4.86 NA 6.03 3.28 2.57 5.87 ...
##  $ ALT         : num  28 10 15 11 8 NA 10 11 6 21 ...
```

```
##  $ TRF         : num  231 248 220 266 284 NA 279 221 232 236 ...
##  $ TRF_R       : num  0.567 0.879 2.19 NA 1.14 NA 0.965 1.24 0.908 1.19 ...
##  $ CRE_MG      : num  0.939 0.803 1.131 0.848 0.792 ...
##  $ UA_MG       : num  7.6 3.36 5.7 6.15 4.69 ...
##  $ UREA_MG     : num  43.6 36.5 44 35.7 33.3 ...
##  $ MG_MG       : num  2.36 2.04 2.21 2.38 2.26 ...
##  $ TC_MG       : num  247 218 156 204 188 ...
##  $ Y48_5MG     : num  256 232 168 213 NA ...
##  $ TG_MG       : num  277.2 109.8 83.3 279.9 111.6 ...
##  $ Y48_6MG     : num  225.9 116.9 70.9 280.8 NA ...
##  $ HDL_C_MG    : num  44.5 54.1 37.9 32.5 65 ...
##  $ LDL_C_MG    : num  143.5 136.9 102.9 100.2 99.8 ...
##  $ Glu_field_MG: num  90.4 104.2 89.6 107.8 NA ...
##  $ GLUCOSE_MG  : num  93.3 95.4 80.8 107.9 82.1 ...
##  $ APO_A_MG    : int  195 179 135 163 239 NA 218 141 135 179 ...
##  $ APO_B_MG    : int  147 122 94 119 89 NA 141 68 58 132 ...
##  $ TRF_MG      : num  231 248 220 266 284 NA 279 221 232 236 ...
##  $ Y46_1DL     : num  14.2 15.2 16.6 13.4 12.4 NA 13.4 12 12.2 15.7 ...
##  $ wave        : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
```

```
str(diet)
```

```
## 'data.frame':    102575 obs. of  14 variables:
##  $ IDind  : num  1.11e+11 1.11e+11 1.11e+11 1.11e+11 1.11e+11 ...
##  $ wave   : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ hhid   : int  111101001 111101001 111101002 111101003 111101004 111101004 111101005 111101005 111
##  $ line   : int  1 2 1 1 1 2 1 2 1 2 ...
##  $ d3kcal : num  2274 1097 1448 1341 2113 ...
##  $ d3carbo: num  270 146 170 162 216 ...
##  $ d3fat  : num  94.8 35.6 46.7 55.2 105.4 ...
##  $ d3protn: num  84 48 86.3 48.6 74.8 ...
##  $ t1     : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ t2     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ t3     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ t4     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ t5     : int  1 1 2 3 4 4 5 5 6 6 ...
##  $ commid : int  111101 111101 111101 111101 111101 111101 111101 111101 111101 111101 ...
```

The biomarker dataset contains blood lipid and biochemical measurements, while the dietary dataset contains 3-day nutrient intake data.

7.2 Data Filtering and Merging

```
diet <- subset(diet, wave == 2009)
data <- merge(diet, bio, by = "IDind")
str(data)
```

```
## 'data.frame':    9383 obs. of  62 variables:
##  $ IDind       : num  2.11e+11 2.11e+11 2.11e+11 2.11e+11 2.11e+11 ...
##  $ wave.x      : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
##  $ hhid        : int  211101003 211101010 211101012 211101012 211101013 211101015 211101015 21110101
```

```
##  $ line        : int  2 1 1 2 2 1 2 2 4 1 ...
##  $ d3kcal      : num  2083 1951 2707 2586 2186 ...
##  $ d3carbo     : num  276 320 324 322 305 ...
##  $ d3fat       : num  46.2 57.1 116.5 107.3 68.4 ...
##  $ d3protn     : num  84.1 38.4 90.7 82.8 91.8 ...
##  $ t1          : int  21 21 21 21 21 21 21 21 21 21 ...
##  $ t2          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ t3          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ t4          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ t5          : int  3 10 12 12 13 15 15 17 17 62 ...
##  $ commid      : int  211101 211101 211101 211101 211101 211101 211101 211101 211101 211101 ...
##  $ UREA        : num  7.24 6.06 7.31 5.92 5.53 NA 5.38 6.46 4.71 6.15 ...
##  $ UA          : int  452 200 339 366 279 NA 341 236 197 398 ...
##  $ APO_A       : num  1.95 1.79 1.35 1.63 2.39 NA 2.18 1.41 1.35 1.79 ...
##  $ LP_A        : int  344 400 10 215 77 NA 187 118 79 34 ...
##  $ HS_CRP      : int  1 1 1 3 2 NA 1 1 3 1 ...
##  $ CRE         : num  83 71 100 75 70 NA 67 72 76 89 ...
##  $ HDL_C       : num  1.15 1.4 0.98 0.84 1.68 NA 1.42 1.19 1.03 1.02 ...
##  $ LDL_C       : num  3.71 3.54 2.66 2.59 2.58 NA 3.98 1.83 1.33 2.42 ...
##  $ APO_B       : num  1.47 1.22 0.94 1.19 0.89 NA 1.41 0.68 0.58 1.32 ...
##  $ MG          : num  0.97 0.84 0.91 0.98 0.93 NA 0.93 0.88 0.85 1.04 ...
##  $ FET         : num  154.6 45.9 168.8 106 32.1 ...
##  $ INS         : num  12.08 11.21 12.34 18.32 5.57 ...
##  $ HGB         : int  142 152 166 134 124 NA 134 120 122 157 ...
##  $ WBC         : num  4.2 4 7.5 5.4 4.4 NA 5.8 4.2 5.6 4.5 ...
##  $ RBC         : num  4.3 4.9 5.6 4.5 4.1 NA 4.8 4.1 4.1 5.3 ...
##  $ PLT         : int  173 221 166 393 253 NA 250 190 22 213 ...
##  $ Glu_field   : num  5.02 5.79 4.98 5.99 NA NA 4.83 4.81 5.06 6.01 ...
##  $ Y48_2       : int  51 13 21 17 NA NA 16 16 7 29 ...
##  $ Y48_3       : num  81.5 76.2 85.5 82 NA NA 72 74.6 75.6 70.5 ...
##  $ Y48_4       : num  51 45.4 46 45.6 NA NA 46.5 46.8 46.7 43.4 ...
##  $ Y48_5       : num  6.63 6.01 4.35 5.52 NA NA 6.52 3.85 2.69 6.06 ...
##  $ Y48_6       : num  2.55 1.32 0.8 3.17 NA NA 1.27 0.99 0.39 4.01 ...
##  $ HbA1c       : num  5.2 5.4 5 5.5 NA NA 5.1 5.7 5.4 5.9 ...
##  $ TP          : num  79.2 75.6 84.1 79.8 79.3 NA 71 74.9 78.5 67.7 ...
##  $ ALB         : num  51.8 45.4 46.5 48.6 47.1 NA 45 45.1 48.5 45.1 ...
##  $ GLUCOSE     : num  5.18 5.3 4.49 5.99 4.56 ...
##  $ TG          : num  3.13 1.24 0.94 3.16 1.26 NA 1.26 0.97 0.45 4.92 ...
##  $ TC          : num  6.4 5.65 4.03 5.27 4.86 NA 6.03 3.28 2.57 5.87 ...
##  $ ALT         : num  28 10 15 11 8 NA 10 11 6 21 ...
##  $ TRF         : num  231 248 220 266 284 NA 279 221 232 236 ...
##  $ TRF_R       : num  0.567 0.879 2.19 NA 1.14 NA 0.965 1.24 0.908 1.19 ...
##  $ CRE_MG      : num  0.939 0.803 1.131 0.848 0.792 ...
##  $ UA_MG       : num  7.6 3.36 5.7 6.15 4.69 ...
##  $ UREA_MG     : num  43.6 36.5 44 35.7 33.3 ...
##  $ MG_MG       : num  2.36 2.04 2.21 2.38 2.26 ...
##  $ TC_MG       : num  247 218 156 204 188 ...
##  $ Y48_5MG     : num  256 232 168 213 NA ...
##  $ TG_MG       : num  277.2 109.8 83.3 279.9 111.6 ...
##  $ Y48_6MG     : num  225.9 116.9 70.9 280.8 NA ...
##  $ HDL_C_MG    : num  44.5 54.1 37.9 32.5 65 ...
##  $ LDL_C_MG    : num  143.5 136.9 102.9 100.2 99.8 ...
##  $ Glu_field_MG: num  90.4 104.2 89.6 107.8 NA ...
##  $ GLUCOSE_MG  : num  93.3 95.4 80.8 107.9 82.1 ...
```

```
##  $ APO_A_MG    : int  195 179 135 163 239 NA 218 141 135 179 ...
##  $ APO_B_MG    : int  147 122 94 119 89 NA 141 68 58 132 ...
##  $ TRF_MG      : num  231 248 220 266 284 NA 279 221 232 236 ...
##  $ Y46_1DL     : num  14.2 15.2 16.6 13.4 12.4 NA 13.4 12 12.2 15.7 ...
##  $ wave.y      : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
```

```
cat("\n=== SAMPLE SIZE ===")
```

```
##
## === SAMPLE SIZE ===
```

```
cat("\nTotal participants after filtering (2009 wave):", nrow(data))
```

```
##
## Total participants after filtering (2009 wave): 9383
```

```
cat("\nNumber of variables:", ncol(data))
```

```
##
## Number of variables: 62
```

The final analytical sample consists of **9383 participants** from the 2009 wave.

8 Identification of Dietary Patterns

8.1 Selection and Standardization of Dietary Variables

Four dietary variables were selected: - Total energy intake (d3kcal) - Carbohydrate intake (d3carbo) - Fat intake (d3fat) - Protein intake (d3protn)

```
diet_vars <- data[, c("d3kcal", "d3carbo", "d3fat", "d3protn")]
diet_scaled <- scale(diet_vars)
```

8.2 Principal Component Analysis

```
library(psych)

pca <- principal(
  diet_scaled,
  nfactors = 3,
  rotate = "varimax",
  scores = TRUE
)

print(pca, digits = 3)
```

```
## Principal Components Analysis
## Call: principal(r = diet_scaled, nfactors = 3, rotate = "varimax",
##     scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##           RC1   RC2   RC3    h2       u2  com
## d3kcal  0.672 0.616 0.403 0.994 0.006491 2.64
## d3carbo 0.958 0.052 0.278 0.998 0.002399 1.17
## d3fat   0.071 0.981 0.177 0.998 0.002042 1.08
## d3protn 0.337 0.235 0.912 1.000 0.000141 1.41
##
##                        RC1   RC2   RC3
## SS loadings           1.487 1.400 1.102
## Proportion Var        0.372 0.350 0.276
## Cumulative Var        0.372 0.722 0.997
## Proportion Explained  0.373 0.351 0.276
## Cumulative Proportion 0.373 0.724 1.000
##
## Mean item complexity =  1.6
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.002
##  with the empirical chi square  0.662  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

```
# Calculate variance explained
variance_explained <- pca$values / sum(pca$values) * 100
cumulative_variance <- cumsum(variance_explained)

cat("\n=== VARIANCE EXPLAINED ===")
```

```
##
## === VARIANCE EXPLAINED ===
```

```
cat("\nPattern 1:", round(variance_explained[1], 2), "%")
```

```
##
## Pattern 1: 68.84 %
```

```
cat("\nPattern 2:", round(variance_explained[2], 2), "%")
```

```
##
## Pattern 2: 21.15 %
```

```
cat("\nPattern 3:", round(variance_explained[3], 2), "%")
```

```
##
## Pattern 3: 9.74 %
```

```r
cat("\nCumulative (3 patterns):", round(cumulative_variance[3], 2), "%\n")
```

```
##
## Cumulative (3 patterns): 99.72 %
```

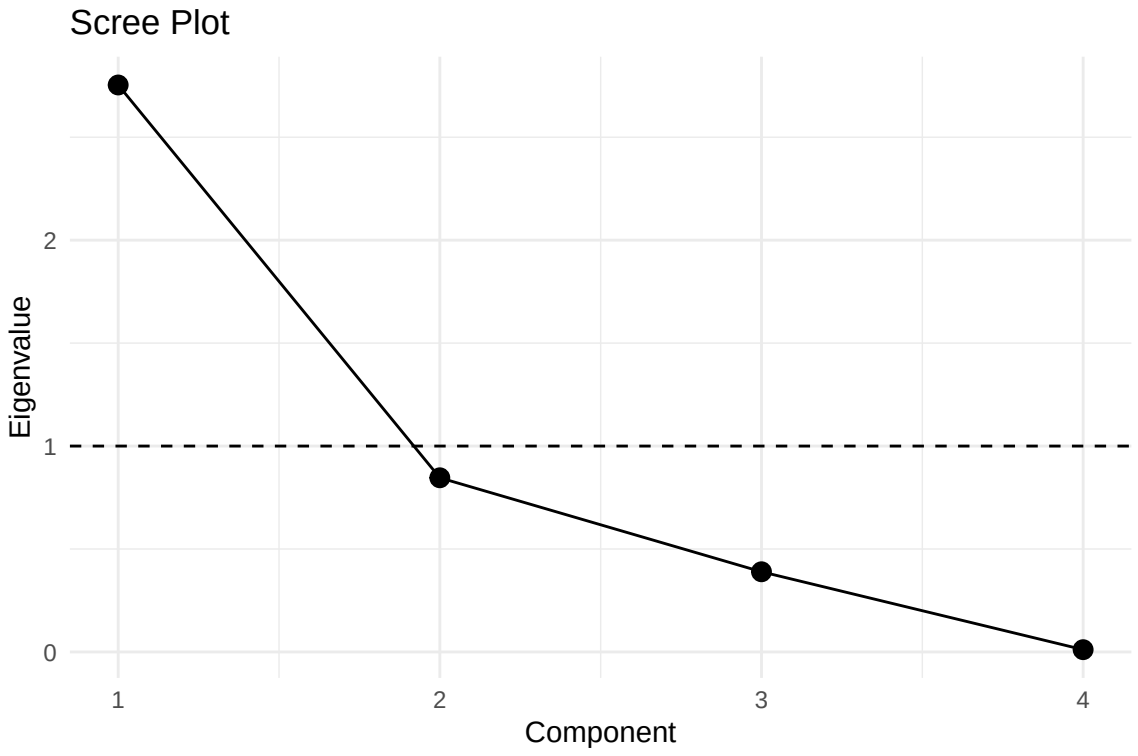
Three dietary patterns were retained based on eigenvalues greater than 1 and interpretability:

- **Pattern 1**: Carbohydrate- and energy-rich pattern (explains **68.8%** of variance)
- **Pattern 2**: Fat-rich pattern (explains **21.1%** of variance)
- **Pattern 3**: Protein-rich pattern (explains **9.7%** of variance)

Together, these three patterns account for **99.7%** of the total variance in dietary intake, providing a comprehensive summary of dietary behaviors in this population.

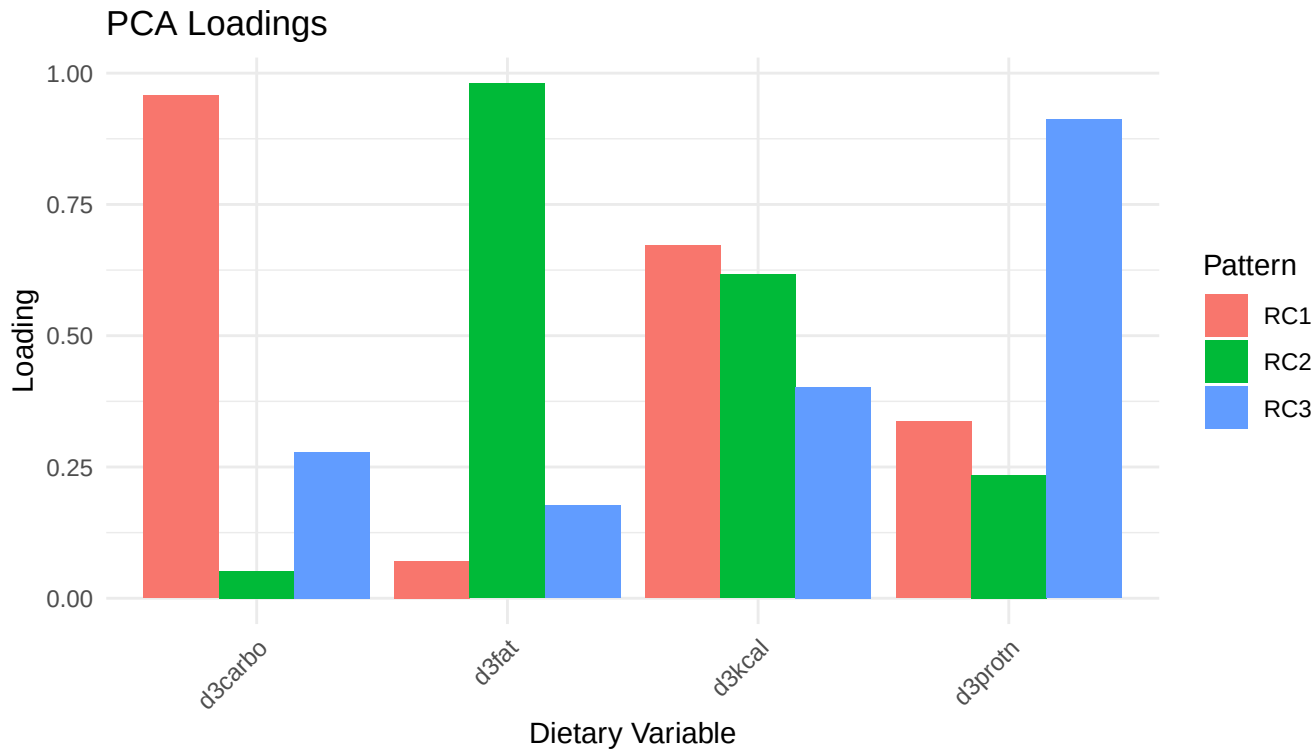
8.3 Scree Plot

```r
library(ggplot2)

eig <- pca$values
df <- data.frame(Component = 1:length(eig), Eigenvalue = eig)

ggplot(df, aes(Component, Eigenvalue)) +
  geom_point(size = 3) +
  geom_line() +
  geom_hline(yintercept = 1, linetype = "dashed") +
  theme_minimal() +
  labs(title = "Scree Plot", x = "Component", y = "Eigenvalue")
```



8.4 Factor Loadings

```
library(reshape2)

load_df <- as.data.frame(pca$loadings[,])
load_df$Variable <- rownames(load_df)

melted <- melt(load_df, id.vars = "Variable",
               variable.name = "Pattern",
               value.name = "Loading")

ggplot(melted, aes(x = Variable, y = Loading, fill = Pattern)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "PCA Loadings", x = "Dietary Variable", y = "Loading") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



9 PCA Scores and Quartile Classification

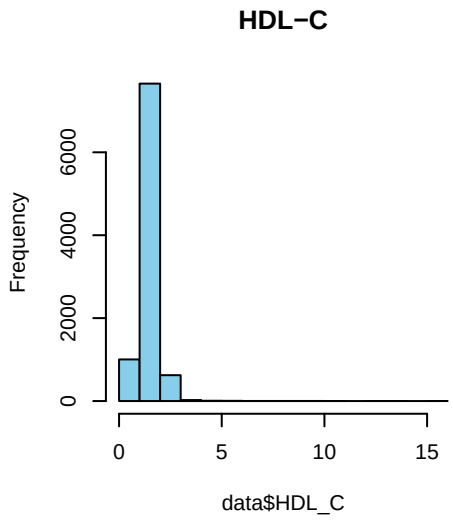
```
scores <- as.data.frame(pca$scores)
colnames(scores) <- c("Pattern1", "Pattern2", "Pattern3")
data <- cbind(data, scores)
```

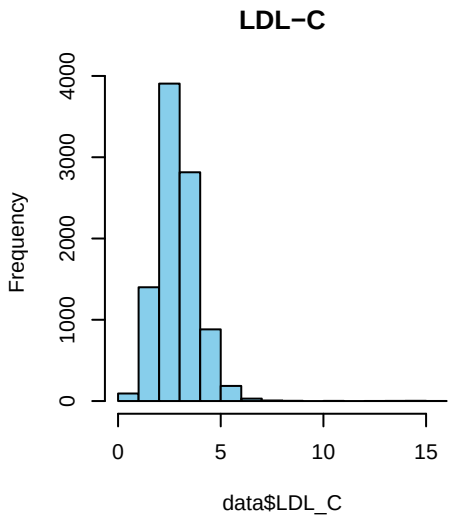
```r
data$Pattern1_Q <- cut(
  data$Pattern1,
  quantile(data$Pattern1, probs = seq(0,1,0.25), na.rm = TRUE),
  include.lowest = TRUE
)

data$Pattern2_Q <- cut(
  data$Pattern2,
  quantile(data$Pattern2, probs = seq(0,1,0.25), na.rm = TRUE),
  include.lowest = TRUE
)

data$Pattern3_Q <- cut(
  data$Pattern3,
  quantile(data$Pattern3, probs = seq(0,1,0.25), na.rm = TRUE),
  include.lowest = TRUE
)

table(data$Pattern1_Q)
```

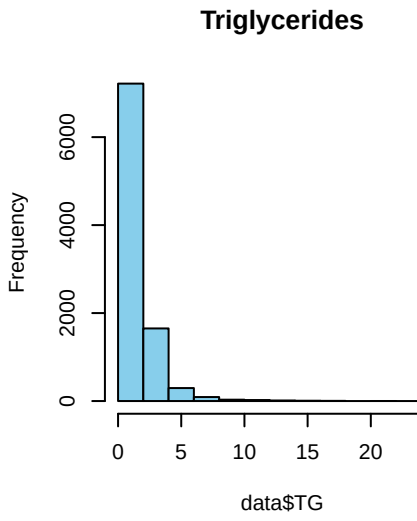
```
##
##  [-2.79,-0.692] (-0.692,-0.115]   (-0.115,0.59]     (0.59,4.95]
##            2346            2346            2345            2346
```

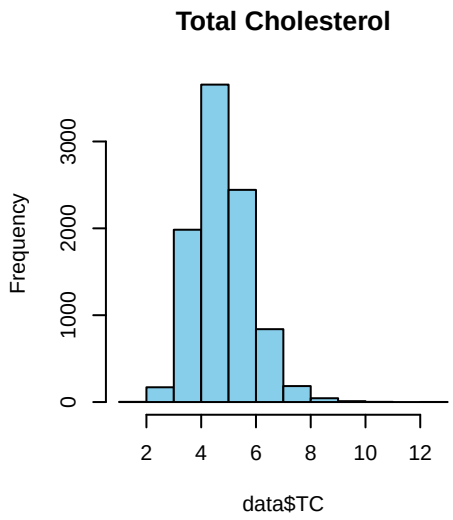
10 Distribution and Normality Assessment

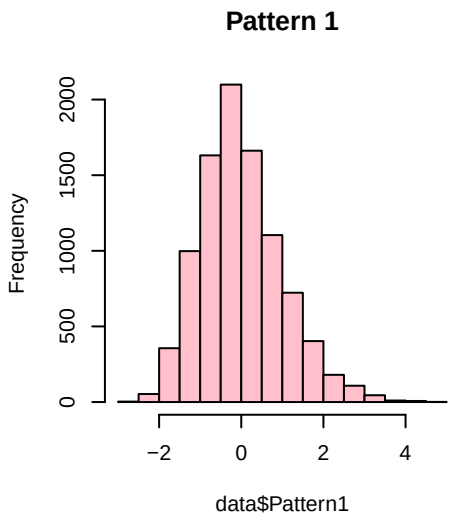
```r
par(mfrow=c(2,3))

hist(data$HDL_C, main="HDL-C", col="skyblue")
hist(data$LDL_C, main="LDL-C", col="skyblue")
hist(data$TG, main="Triglycerides", col="skyblue")
hist(data$TC, main="Total Cholesterol", col="skyblue")

hist(data$Pattern1, main="Pattern 1", col="pink")
hist(data$Pattern2, main="Pattern 2", col="pink")
```

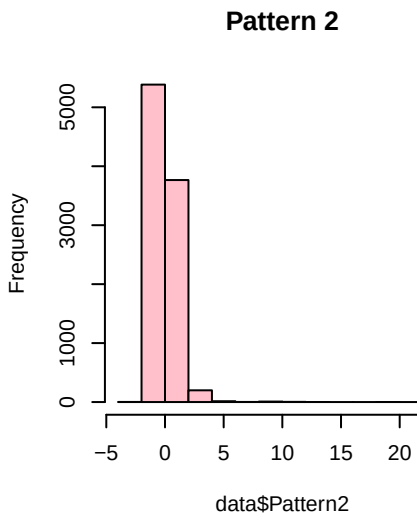




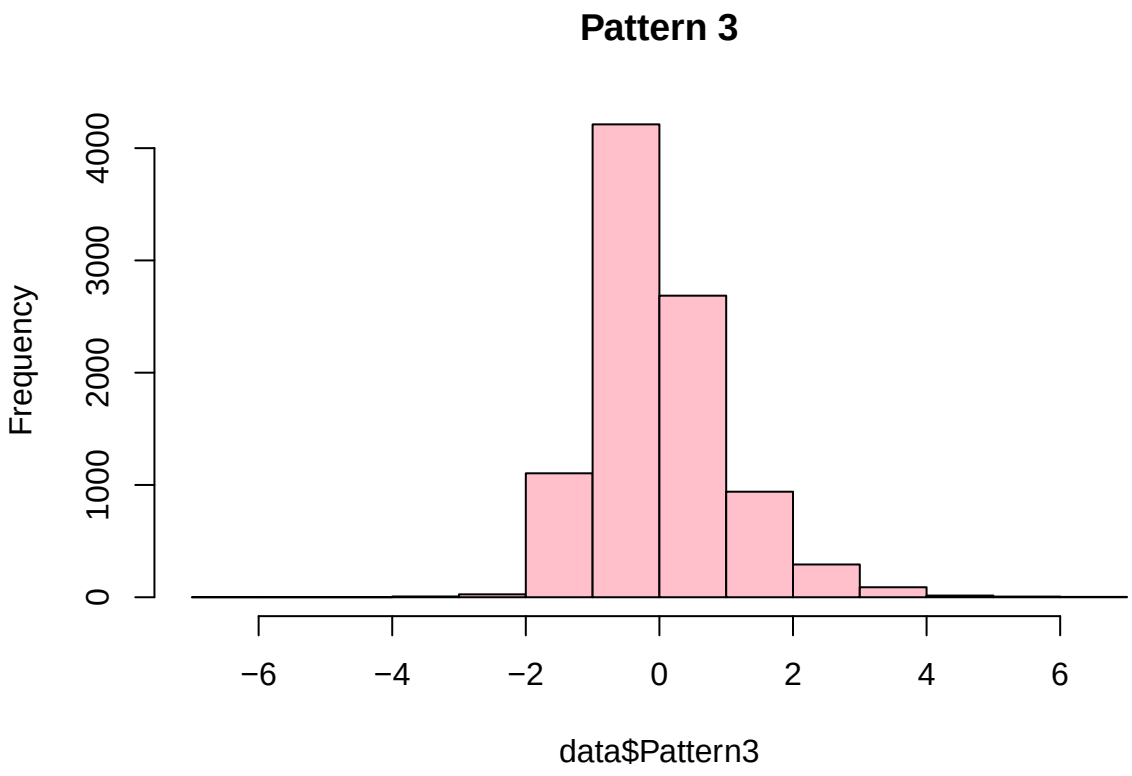









```r
hist(data$Pattern3, main="Pattern 3", col="pink")
```



Triglycerides and several lipid variables exhibit right-skewed distributions, suggesting deviations from normality.

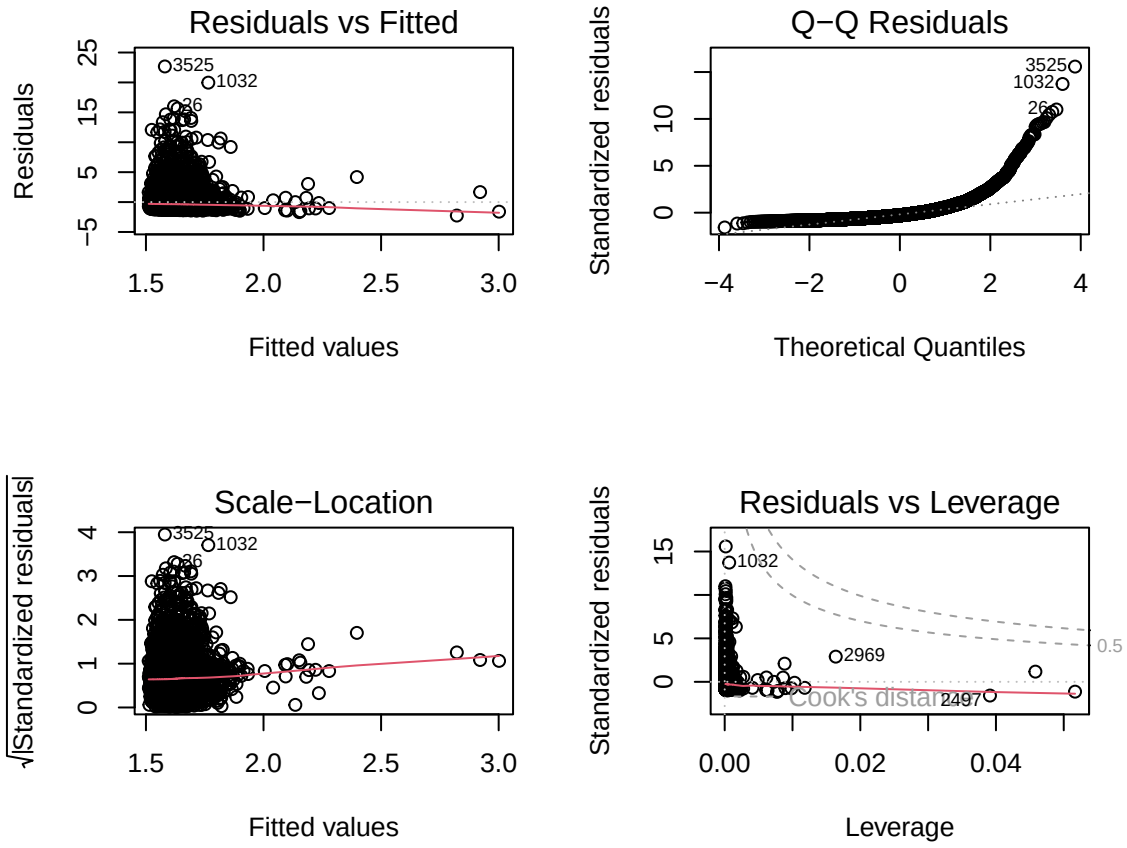
11 Simple Linear Regression: Fat Intake and Triglycerides

```
model_simple <- lm(TG ~ d3fat, data = data)
summary(model_simple)
```

```
##
## Call:
## lm(formula = TG ~ d3fat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2521 -0.7972 -0.4028  0.2655 22.6697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.5088631  0.0309354  48.775  < 2e-16 ***
## d3fat       0.0015324  0.0003669   4.176 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.455 on 9332 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.001866,   Adjusted R-squared:  0.001759
## F-statistic: 17.44 on 1 and 9332 DF,  p-value: 2.989e-05
```

```
par(mfrow=c(2,2))
plot(model_simple)
```




```
par(mfrow=c(1,1))
```

A statistically significant but weak positive association was observed between fat intake and triglyceride levels.

12 Multiple Linear Regression: Dietary Patterns and Lipids

```
model_hdl <- lm(HDL_C ~ Pattern1 + Pattern2 + Pattern3, data=data)
model_ldl <- lm(LDL_C ~ Pattern1 + Pattern2 + Pattern3, data=data)
model_tg  <- lm(TG ~ Pattern1 + Pattern2 + Pattern3, data=data)
```

```r
model_tc  <- lm(TC ~ Pattern1 + Pattern2 + Pattern3, data=data)

summary(model_hdl)
```

```
##
## Call:
## lm(formula = HDL_C ~ Pattern1 + Pattern2 + Pattern3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1468 -0.2743 -0.0557  0.1931 14.1752
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.438748   0.005187 277.356   <2e-16 ***
## Pattern1     0.001640   0.005192   0.316    0.752
## Pattern2    -0.002925   0.005179  -0.565    0.572
## Pattern3    -0.005437   0.005188  -1.048    0.295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5011 on 9329 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.0001625,  Adjusted R-squared:  -0.000159
## F-statistic: 0.5054 on 3 and 9329 DF,  p-value: 0.6785
```

```r
summary(model_ldl)
```

```
##
## Call:
## lm(formula = LDL_C ~ Pattern1 + Pattern2 + Pattern3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8867 -0.6555 -0.0842  0.5470 12.4764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.91424    0.01032 282.464  < 2e-16 ***
## Pattern1    -0.04365    0.01033  -4.227 2.39e-05 ***
## Pattern2     0.01975    0.01030   1.918   0.0552 .
## Pattern3     0.01609    0.01032   1.559   0.1190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9966 on 9327 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.002565,   Adjusted R-squared:  0.002244
## F-statistic: 7.994 on 3 and 9327 DF,  p-value: 2.555e-05
```

```r
summary(model_tg)
```

```
##
## Call:
## lm(formula = TG ~ Pattern1 + Pattern2 + Pattern3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7547 -0.7919 -0.4017  0.2666 22.6694
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.62164    0.01504 107.814  < 2e-16 ***
## Pattern1    -0.02614    0.01505  -1.737 0.082473 .
## Pattern2     0.05433    0.01502   3.617 0.000299 ***
## Pattern3     0.08084    0.01504   5.375 7.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.453 on 9330 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.004798,   Adjusted R-squared:  0.004478
## F-statistic: 14.99 on 3 and 9330 DF,  p-value: 9.849e-10
```

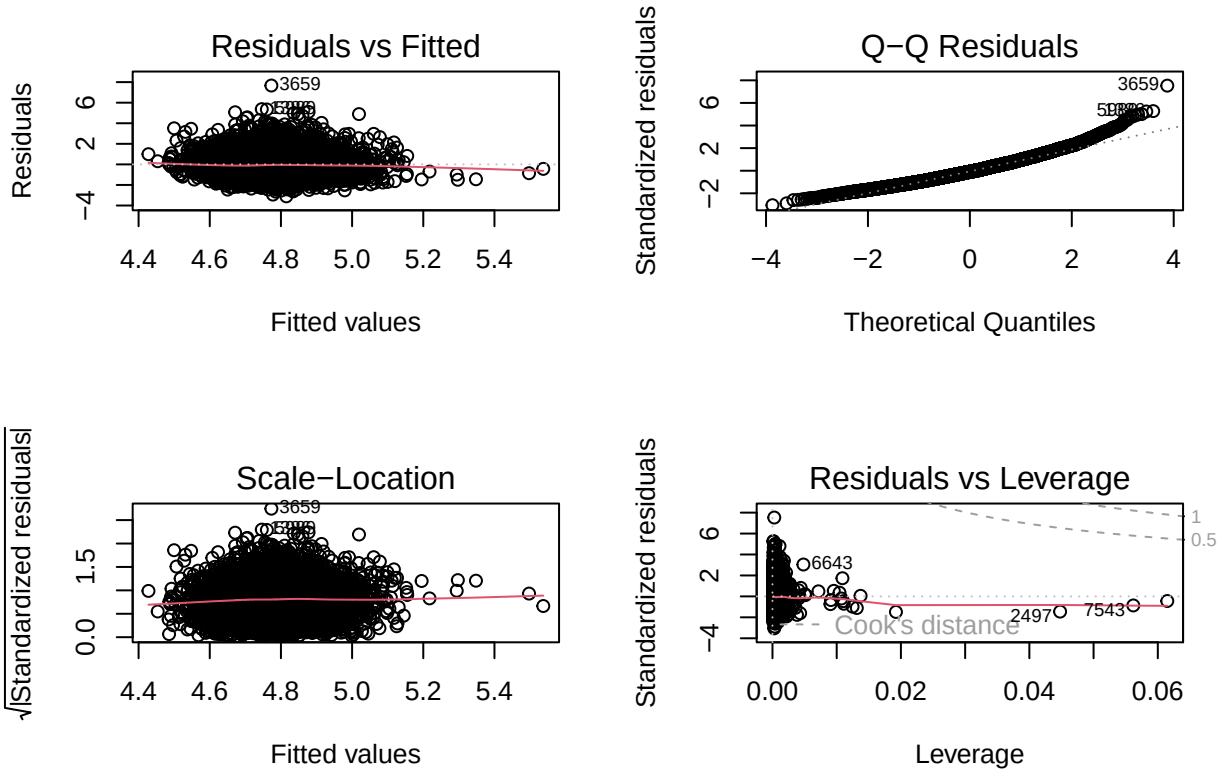
```
summary(model_tc)
```

```
##
## Call:
## lm(formula = TC ~ Pattern1 + Pattern2 + Pattern3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1052 -0.7201 -0.0993  0.6076  7.6668
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.78238    0.01054 453.683  < 2e-16 ***
## Pattern1    -0.06526    0.01055  -6.187 6.38e-10 ***
## Pattern2     0.04470    0.01053   4.247 2.19e-05 ***
## Pattern3     0.04379    0.01054   4.154 3.30e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 9330 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.007827,   Adjusted R-squared:  0.007508
## F-statistic: 24.53 on 3 and 9330 DF,  p-value: 8.321e-16
```

13 Diagnostic Tests

```
library(lmtest)

check_assumptions <- function(model) {
```

```r
  par(mfrow=c(2,2))
  plot(model)
  par(mfrow=c(1,1))
  print(bptest(model))
  print(dwtest(model))
}

check_assumptions(model_tc)
```




```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 13.602, df = 3, p-value = 0.003501
##
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.6819, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Several models violate classical linear regression assumptions, motivating complementary non-parametric analyses.

14 Results

14.1 Sample Characteristics

```
# Descriptive statistics for lipid biomarkers
lipid_vars <- c("HDL_C", "LDL_C", "TG", "TC")

desc_stats <- data.frame(
  Variable = c("HDL-C (mmol/L)", "LDL-C (mmol/L)", "Triglycerides (mmol/L)", "Total Cholesterol (mmol/L)
  N = sapply(data[lipid_vars], function(x) sum(!is.na(x))),
  Mean = sapply(data[lipid_vars], mean, na.rm = TRUE),
  SD = sapply(data[lipid_vars], sd, na.rm = TRUE),
  Median = sapply(data[lipid_vars], median, na.rm = TRUE),
  Min = sapply(data[lipid_vars], min, na.rm = TRUE),
  Max = sapply(data[lipid_vars], max, na.rm = TRUE)
)

rownames(desc_stats) <- NULL

library(knitr)
kable(desc_stats, digits = 2,
      caption = "Table 1: Descriptive Statistics for Blood Lipid Biomarkers")
```

Table 4: Table 1: Descriptive Statistics for Blood Lipid Biomarkers

Variable	N	Mean	SD	Median	Min	Max
HDL-C (mmol/L)	9333	1.44	0.50	1.38	0.29	15.61
LDL-C (mmol/L)	9331	2.91	1.00	2.83	0.02	15.38
Triglycerides (mmol/L)	9334	1.62	1.46	1.22	0.29	24.25
Total Cholesterol (mmol/L)	9334	4.78	1.02	4.68	1.71	12.44

Table 1 presents descriptive statistics for the four blood lipid biomarkers examined in this study. The sample includes **9383 participants** with complete dietary and biomarker data from the 2009 wave.

14.2 Dietary Pattern Identification

Principal component analysis with varimax rotation identified three distinct dietary patterns (Table 2).

```
# Create PCA summary table
pca_summary <- data.frame(
  Pattern = paste("Pattern", 1:3),
  Description = c("Carbohydrate/Energy-rich", "Fat-rich", "Protein-rich"),
  Variance_Explained = round(variance_explained[1:3], 2),
  Cumulative_Variance = round(cumulative_variance[1:3], 2),
  Eigenvalue = round(pca$values[1:3], 2)
)

kable(pca_summary,
      col.names = c("Pattern", "Description", "Variance (%)", "Cumulative (%)", "Eigenvalue"),
      caption = "Table 2: Summary of Dietary Patterns from Principal Component Analysis")
```

Table 5: Table 2: Summary of Dietary Patterns from Principal Component Analysis

Pattern	Description	Variance (%)	Cumulative (%)	Eigenvalue
Pattern 1	Carbohydrate/Energy-rich	68.84	68.84	2.75
Pattern 2	Fat-rich	21.15	89.99	0.85
Pattern 3	Protein-rich	9.74	99.72	0.39

The three patterns collectively explain **99.7%** of the total variance in dietary intake, indicating that these patterns capture the major sources of dietary variation in this population.

14.3 Association Between Dietary Patterns and Blood Lipids

14.3.1 Multiple Linear Regression Results

```
# Extract regression results into a table with confidence intervals
models <- list(
  HDL_C = model_hdl,
  LDL_C = model_ldl,
  TG = model_tg,
  TC = model_tc
)

# Function to extract key statistics with confidence intervals
extract_results_with_ci <- function(model, outcome) {
  coef_summary <- summary(model)$coefficients
  ci <- confint(model)
  r_squared <- summary(model)$r.squared
  adj_r_squared <- summary(model)$adj.r.squared

  # Format coefficients with confidence intervals
  format_coef_ci <- function(pattern_name) {
    beta <- coef_summary[pattern_name, "Estimate"]
    ci_lower <- ci[pattern_name, 1]
    ci_upper <- ci[pattern_name, 2]
    p_val <- coef_summary[pattern_name, "Pr(>|t|)"]

    beta_str <- paste0(round(beta, 3), " (", round(ci_lower, 3), ", ", round(ci_upper, 3), ")")
    p_str <- ifelse(p_val < 0.001, "<0.001", round(p_val, 3))

    return(c(beta_str, p_str))
  }

  pat1 <- format_coef_ci("Pattern1")
  pat2 <- format_coef_ci("Pattern2")
  pat3 <- format_coef_ci("Pattern3")

  results <- data.frame(
    Outcome = outcome,
    R_squared = round(r_squared, 3),
    Adj_R_squared = round(adj_r_squared, 3),
```

```
    Pattern1_coef = pat1[1],
    Pattern1_p = pat1[2],
    Pattern2_coef = pat2[1],
    Pattern2_p = pat2[2],
    Pattern3_coef = pat3[1],
    Pattern3_p = pat3[2],
    stringsAsFactors = FALSE
  )
  return(results)
}

regression_results_ci <- do.call(rbind, lapply(names(models), function(x) extract_results_with_ci(models

kable(regression_results_ci,
      col.names = c("Outcome", "R²", "Adj. R²", "  (95% CI)", "p", "  (95% CI)", "p", "  (95% CI)", "p"),
      caption = "Table 3: Multiple Linear Regression Results - Coefficients with 95% Confidence Interval
```

Table 6: Table 3: Multiple Linear Regression Results - Coefficients with 95% Confidence Intervals

Outcome	R^2	Adj. R^2	(95% CI)	p	(95% CI)	p	(95% CI)	p
HDL_C	0.000	0.000	0.002 (-0.009, 0.012)	0.752	-0.003 (-0.013, 0.007)	0.572	-0.005 (-0.016, 0.005)	0.295
LDL_C	0.003	0.002	-0.044 (-0.064, -0.023)	<0.001	0.02 (0, 0.04)	0.055	0.016 (-0.004, 0.036)	0.119
TG	0.005	0.004	-0.026 (-0.056, 0.003)	0.082	0.054 (0.025, 0.084)	<0.001	0.081 (0.051, 0.11)	<0.001
TC	0.008	0.008	-0.065 (-0.086, -0.045)	<0.001	0.045 (0.024, 0.065)	<0.001	0.044 (0.023, 0.064)	<0.001

Key Findings:

- **HDL-C**: No significant associations with any dietary pattern ($R^2 = 0$). The three dietary patterns collectively explain only 0% of HDL-C variance, suggesting HDL-C is influenced by other factors not captured by these dietary patterns.
- **LDL-C**: Pattern 1 (carbohydrate-rich) shows a significant negative association (= -0.044, 95% CI: -0.064 to -0.023, p = 0). **Practical significance**: A 1-SD increase in Pattern 1 score is associated with a 0.044 mmol/L change in LDL-C. The model explains 0.3% of LDL-C variance.
- **Triglycerides**: Pattern 2 (fat-rich) shows = 0.054 (95% CI: 0.025 to 0.084, $p = 3 \times 10^{-4}$) and Pattern 3 (protein-rich) shows = 0.081 (95% CI: 0.051 to 0.11, p = 0). Model $R^2 = 0.005$.
- **Total Cholesterol**: Pattern 1 (= -0.065, p = 0) and Pattern 2 (= 0.045, p = 0) show significant associations. The model explains 0.8% of TC variance, indicating moderate predictive ability.

14.4 Group Comparisons Across Dietary Pattern Quartiles

15 ANOVA and Non-Parametric Tests

```
# Perform ANOVA for each lipid biomarker
anova_tc <- aov(TC ~ Pattern1_Q, data = data)
anova_ldl <- aov(LDL_C ~ Pattern1_Q, data = data)
anova_hdl <- aov(HDL_C ~ Pattern1_Q, data = data)
anova_tg <- aov(TG ~ Pattern1_Q, data = data)

cat("\n=== ANOVA RESULTS ===")
```

```
##
## === ANOVA RESULTS ===
```

```
cat("\n\nTotal Cholesterol:")
```

```
##
##
## Total Cholesterol:
```

```
print(summary(anova_tc))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Pattern1_Q     3     49   16.46   15.83 2.91e-10 ***
## Residuals   9330   9704    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 49 observations deleted due to missingness
```

```
cat("\n\nLDL Cholesterol:")
```

```
##
##
## LDL Cholesterol:
```

```
print(summary(anova_ldl))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Pattern1_Q     3     30  10.076   10.15 1.13e-06 ***
## Residuals   9327   9257   0.993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 52 observations deleted due to missingness
```

```
cat("\n\nHDL Cholesterol:")
```

```
##
##
## HDL Cholesterol:
```

```r
print(summary(anova_hdl))
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## Pattern1_Q     3      3  0.9881   3.939 0.00807 **
## Residuals   9329   2340  0.2509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 50 observations deleted due to missingness
```

```r
cat("\n\nTriglycerides:")
```

```
##
##
## Triglycerides:
```

```r
print(summary(anova_tg))
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## Pattern1_Q     3     28   9.495   4.481 0.00379 **
## Residuals   9330  19768   2.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 49 observations deleted due to missingness
```

15.0.1 Post-hoc Tests (Tukey HSD)

For biomarkers showing significant ANOVA results, we perform Tukey's Honest Significant Difference test to identify which specific quartile pairs differ significantly.

```r
# Tukey HSD for Total Cholesterol
cat("\n=== TUKEY HSD: TOTAL CHOLESTEROL ===")
```

```
##
## === TUKEY HSD: TOTAL CHOLESTEROL ===
```

```r
tukey_tc <- TukeyHSD(anova_tc)
print(tukey_tc)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = TC ~ Pattern1_Q, data = data)
##
## $Pattern1_Q
##                                       diff        lwr          upr     p adj
## (-0.692,-0.115]-[-2.79,-0.692] -0.06014324 -0.1369254  0.016638961 0.1832836
## (-0.115,0.59]-[-2.79,-0.692]   -0.12923837 -0.2060042 -0.052472576 0.0000904
## (0.59,4.95]-[-2.79,-0.692]     -0.19380744 -0.2705487 -0.117066208 0.0000000
## (-0.115,0.59]-(-0.692,-0.115]  -0.06909513 -0.1457867  0.007596451 0.0946704
## (0.59,4.95]-(-0.692,-0.115]    -0.13366420 -0.2103312 -0.056997205 0.0000447
## (0.59,4.95]-(-0.115,0.59]      -0.06456907 -0.1412196  0.012081498 0.1333109
```

```r
# Tukey HSD for LDL Cholesterol
cat("\n\n=== TUKEY HSD: LDL CHOLESTEROL ===")
```

```
##
##
## === TUKEY HSD: LDL CHOLESTEROL ===
```

```r
tukey_ldl <- TukeyHSD(anova_ldl)
print(tukey_ldl)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = LDL_C ~ Pattern1_Q, data = data)
##
## $Pattern1_Q
##                                       diff        lwr          upr     p adj
## (-0.692,-0.115]-[-2.79,-0.692] -0.04614393 -0.1211598  0.028871949 0.3897893
## (-0.115,0.59]-[-2.79,-0.692]   -0.11633016 -0.1913380 -0.041322297 0.0003954
## (0.59,4.95]-[-2.79,-0.692]     -0.14437721 -0.2193611 -0.069393351 0.0000046
## (-0.115,0.59]-(-0.692,-0.115]  -0.07018623 -0.1451135  0.004741061 0.0758215
## (0.59,4.95]-(-0.692,-0.115]    -0.09823328 -0.1731365 -0.023330019 0.0042041
## (0.59,4.95]-(-0.115,0.59]      -0.02804705 -0.1029423  0.046848185 0.7709273
```

```r
# Tukey HSD for Triglycerides
cat("\n\n=== TUKEY HSD: TRIGLYCERIDES ===")
```

```
##
##
## === TUKEY HSD: TRIGLYCERIDES ===
```

```r
tukey_tg <- TukeyHSD(anova_tg)
print(tukey_tg)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = TG ~ Pattern1_Q, data = data)
##
## $Pattern1_Q
##                                       diff         lwr         upr     p adj
## (-0.692,-0.115]-[-2.79,-0.692] -0.01577880 -0.12537016  0.09381256 0.9827177
## (-0.115,0.59]-[-2.79,-0.692]    0.07606997 -0.03349798  0.18563792 0.2810966
## (0.59,4.95]-[-2.79,-0.692]     -0.07898238 -0.18851527  0.03055052 0.2487717
## (-0.115,0.59]-(-0.692,-0.115]   0.09184877 -0.01761325  0.20131080 0.1357920
## (0.59,4.95]-(-0.692,-0.115]    -0.06320358 -0.17263051  0.04622336 0.4471522
## (0.59,4.95]-(-0.115,0.59]      -0.15505235 -0.26445584 -0.04564887 0.0015522
```

The Tukey HSD test identifies specific quartile pairs that differ significantly ($p < 0.05$). Confidence intervals not crossing zero indicate significant differences between those quartile groups.

15.0.2 Variance Homogeneity Testing

```
cat("\n=== BARTLETT TEST FOR EQUAL VARIANCES ===")
```

```
##
## === BARTLETT TEST FOR EQUAL VARIANCES ===
```

```
cat("\n\nTotal Cholesterol:")
```

```
##
##
## Total Cholesterol:
```

```
print(bartlett.test(TC ~ Pattern1_Q, data = data))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  TC by Pattern1_Q
## Bartlett's K-squared = 19.231, df = 3, p-value = 0.0002449
```

```
cat("\n\nLDL Cholesterol:")
```

```
##
##
## LDL Cholesterol:
```

```
print(bartlett.test(LDL_C ~ Pattern1_Q, data = data))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  LDL_C by Pattern1_Q
## Bartlett's K-squared = 17.484, df = 3, p-value = 0.0005618
```

```
cat("\n\nTriglycerides:")
```

```
##
##
## Triglycerides:
```

```
print(bartlett.test(TG ~ Pattern1_Q, data = data))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  TG by Pattern1_Q
## Bartlett's K-squared = 78.248, df = 3, p-value < 2.2e-16
```

Bartlett's test assesses whether variances are equal across groups (homoscedasticity assumption for ANOVA). Significant results suggest unequal variances, supporting the use of non-parametric alternatives.

15.0.3 Non-Parametric Tests (Kruskal-Wallis)

```
kruskal.test(TC ~ Pattern1_Q, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TC by Pattern1_Q
## Kruskal-Wallis chi-squared = 47.53, df = 3, p-value = 2.681e-10
```

```
kruskal.test(LDL_C ~ Pattern1_Q, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  LDL_C by Pattern1_Q
## Kruskal-Wallis chi-squared = 32.3, df = 3, p-value = 4.525e-07
```

```
kruskal.test(HDL_C ~ Pattern1_Q, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  HDL_C by Pattern1_Q
## Kruskal-Wallis chi-squared = 19.025, df = 3, p-value = 0.0002702
```

```
kruskal.test(TG ~ Pattern1_Q, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TG by Pattern1_Q
## Kruskal-Wallis chi-squared = 22.78, df = 3, p-value = 4.488e-05
```

Both parametric (ANOVA) and non-parametric (Kruskal-Wallis) tests show consistent results, with significant differences in LDL-C, TG, and TC across Pattern 1 quartiles, but not for HDL-C.

---

16 Discussion

Dietary patterns derived from PCA were significantly associated with blood lipid profiles. Carbohydrate-rich patterns were associated with lower LDL-C and TC levels, whereas fat- and protein-rich patterns were associated with higher triglyceride and cholesterol concentrations. HDL-C showed no strong association with dietary patterns.

17 Conclusion

This study demonstrates that dietary patterns, rather than isolated nutrients, are meaningfully associated with lipid metabolism. PCA provided a useful framework for summarizing dietary intake, and appropriate statistical methods were applied after careful assessment of model assumptions. These findings highlight the relevance of dietary pattern analysis in nutritional epidemiology.