Statistical Models for Data Science

Assignment 1

U.S. Population Nutrition Survey

Submitted by: Haya Salameh

Submitted to: Prof. Philip Tzvi Reiss

Date of Submission: 08/12/2024

Table of Contents

Predictors of Body Mass Index: A Multiple Regression Analysis of the 1995-1996 U.S. Population Nutrition Survey

Introduction

This report analyzes data from a comprehensive nutrition survey conducted by the United States Department of Agriculture in 1995-1996, examining relationships among dietary habits, knowledge, attitudes, and health in U.S. households. Using a representative sample of 4,036 respondents drawn from diverse geographical locations and demographic backgrounds across the United States, the primary objective was to identify and understand the factors affecting Body Mass Index (BMI). The survey data included a wide range of variables spanning demographic characteristics (age, sex, race, region, urbanization), socioeconomic factors (income, education), dietary habits (various types of diets), lifestyle choices (exercise frequency), and self-reported health status, providing a comprehensive foundation for understanding BMI determinants in the U.S. population.

The analysis began with data preprocessing and exploratory data analysis (EDA) to understand variable distributions and relationships. This was followed by univariate linear regression models to screen potential predictors of BMI and concluded with multiple linear regression models to examine the combined effects of significant predictors, integrating demographic variables, socioeconomic factors, dietary habits, and lifestyle choices.

Preprocessing and exploratory data analysis

Dataset structure

The initial dataset comprised information from four separate sheets, containing data for males and females, with both main variables and scale variables. After merging these sheets, and only considering the variables mentioned in the instructions, the dataset included the following key variable categories:

- Identifier/index variable: Household ID (hhid).
- Demographic: age, sex, race, region, urbanization
- Socioeconomic: income, education (highest grade completed)
- Dietary: five diet-related variables (low-calorie, low-fat, low-salt, high-fiber, diabetic)
- Lifestyle: exercise frequency
- Health status: self-reported weight status
- Response variable: BMI.

Data cleaning and exploratory data analysis of predictor variables

The analysis encompassed two main categories of variables:

Categorical variables:

- Demographic: Region, urbanization, sex, race
- Health-related: Self-reported weight status
- Dietary: Five diet-specific indicators
- Behavioral: Exercise frequency

Numeric variables:

- Socioeconomic: Income, Highest grade completed
- Demographic: Age

Special considerations in variable treatment

- Education (highest grade completed):
  - Despite its ordinal nature, it is treated as a continuous variable
  - Rationale: High number of distinct levels (17) and semi-continuous numerical progression
- Exercise Frequency:
  - Maintained as an ordered categorical variable
  - Rationale: Non-uniform intervals between categories, treating it as ordered categorical avoids inappropriate assumptions about interval equality

Data Cleaning Process

- Categorical variables:
  - Encoded variable values according to provided documentation
  - Missing values counted pre- and post-preprocessing
  - Frequency distributions analyzed for completeness
- Numeric variables:
  - Range checks performed to identify outliers
  - Descriptive statistics computed

Missing Data Analysis

Post-preprocessing missing values:

1. Self-reported weight status: 20 missing values (0.50% of sample)
2. Education: 40 missing values (0.99% of sample)

3. Exercise frequency: 10 missing values (0.25% of sample)

Notable characteristics:

- Missing data concentrated in self-reported variables
- Core demographic variables showed complete coverage
- Overall low missing data rate for all variables, no need to perform imputation

Distribution Analysis

Demographic Characteristics

1. Gender Distribution:
    - Nearly equal representation of males and females

2. Racial Composition:
    - Substantial White majority (82%), Black respondents: 11.3%, Asian/Pacific Islander: 1.6%, American Indian/Alaska Native: 0.6%, Other: 4%.
    - Notable imbalance in racial representation, which should be considered when interpreting results

3. Regional Distribution:
    - South: 35%, Midwest: 26%, Northeast: 19%, West: 20%
    - Well-balanced regional representation across the US

4. Urbanization Distribution:
    - Suburban: (43.8%), Central city: (28.7%), Nonmetropolitan: (26.4%).
    - Balanced distribution across urban categories
    - No missing values

Socioeconomic Patterns

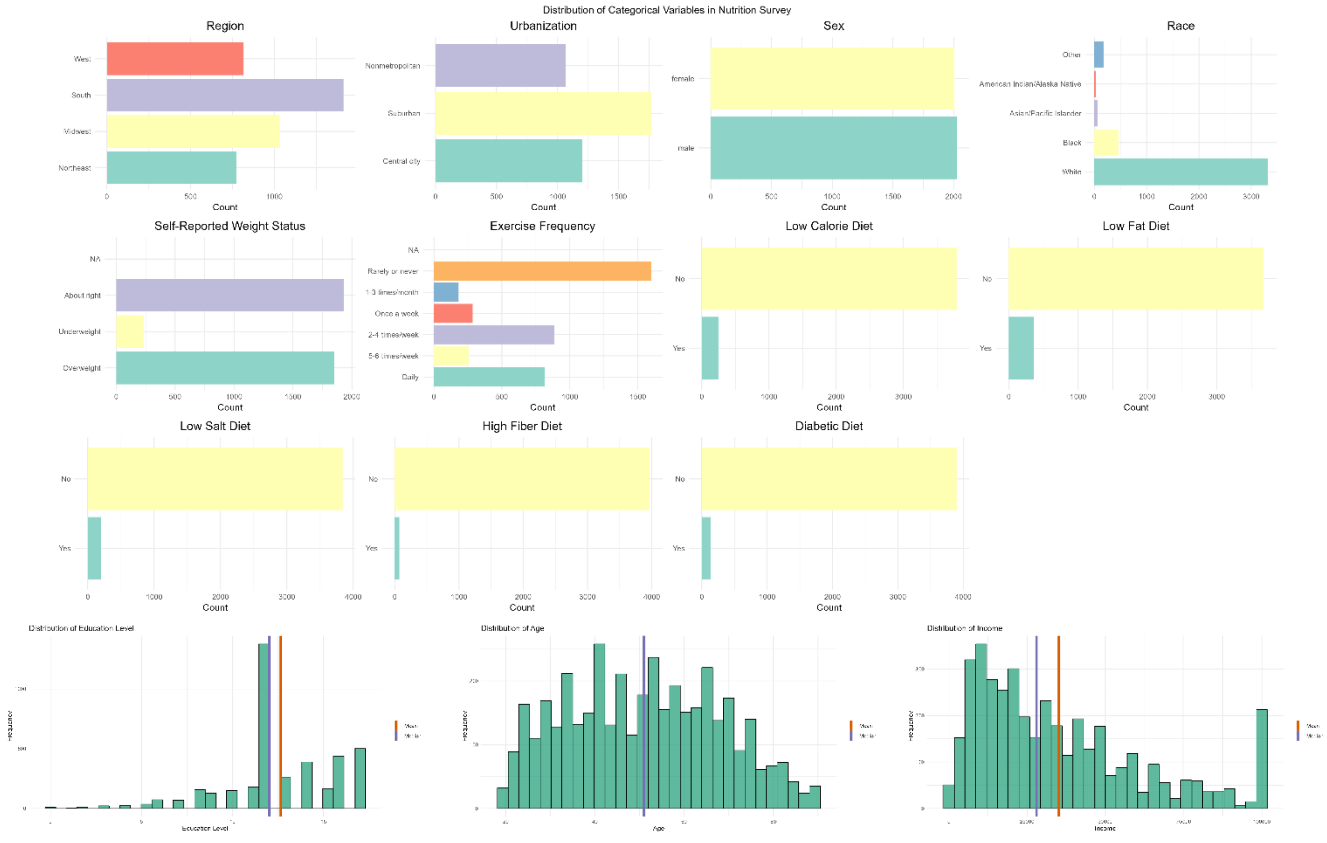
1. Income Distribution:
    - Right-skewed distribution (skewness = 0.95). Median income: 28,000. Mean 35,147.

2. Education:
    - Mode value: 12 years (high school completion)
    - Mean: 12.65 years

- 0 is the minimum and it occurs 10 times, but it was not mentioned in the instructions that this is missing data.

Health and Lifestyle Characteristics

1. Exercise Frequency:
    - "Rarely or never" category: 40%, 2-4 times a week: 22%, Daily exercise: 20%, Once a week: 7%, 5-6 times/week: 6%, 1-3 times/month: 4%.
    - A large number of respondents never or rarely exercise.
2. Dietary Patterns:
    - Low adherence to specific dietary restrictions – high imbalance, in general, 90% answer no.



Distributions of predictor variables

Predictor variables interactions

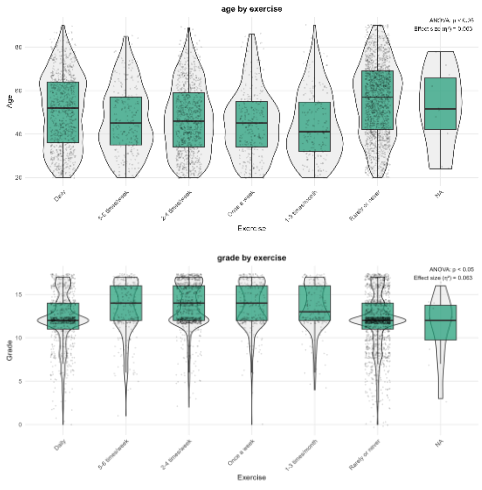
To assess interaction between predictor variables, I used appropriate statistical measures for different variable types (correlation coefficients, η^2, and Cramer's V). This helped distinguish meaningful associations from those that were merely statistically significant due to the large sample size.

This methodological approach provided a statistical framework for evaluating relationships between variables. While these relationships between predictors may not directly inform BMI, they reveal underlying patterns in the dataset that could be relevant for the subsequent regression analysis of BMI.

<u>Numeric-numeric variable interactions</u>

- Method: Pearson correlation coefficients
- Implementation: Used the ggpairs() function from GGally to visualize correlations
- Interpretation threshold: |r| > 0.3 considered noteworthy
- Results showed a moderate positive correlation only between income and education (r = 0.44, p<0.001)

<u>Numeric-categorical variable interactions</u>

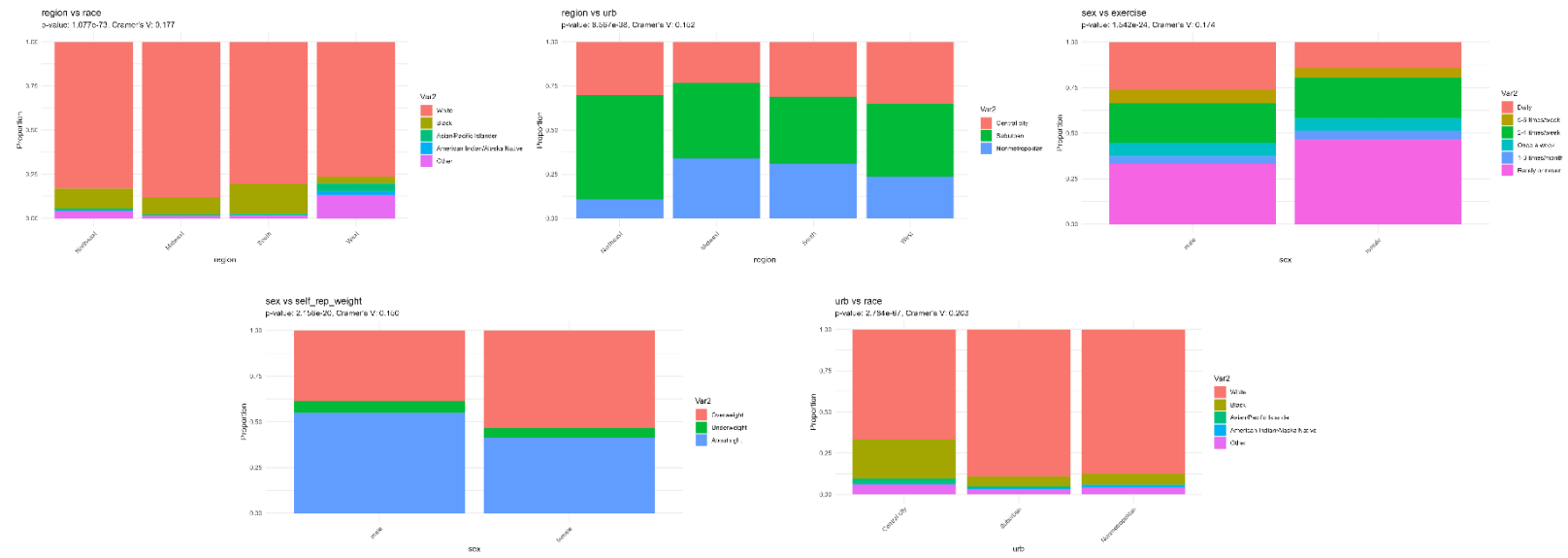


- Method: One-way ANOVA with eta-squared (η^2) effect size
- Implementation: Calculated ANOVA test for all numeric-categorical pairs with aov() function.
- Only reported relationships meeting both statistical significance ($p < 0.05$) and effect size criteria of $\eta^2 \geq 0.06$ (for medium effect size).
- Key findings included:
    - Age-education level: $\eta^2 = 0.063$, Indicates that education level is associated with different exercise patterns.
    - Age-exercise: $\eta^2 = 0.063$, Shows that exercise frequency varies with age.

<u>Categorical-categorical variable interactions</u>

- Method: Chi-square tests with Cramer's V effect size

- Implementation:
- Applied to non-diet categorical variables only due to high imbalance in diet variables
- Only reported relationships meeting both statistical significance ($p < 0.05$) and effect size criteria of Cramer's $V > 0.15$ (for medium effect size). No further comparisons were conducted, see figure for a comparison between levels proportions.
- Significant associations found (see figure ahead):
    1. Urbanization-Race: V = 0.203. Indicates strong geographic patterns in racial distribution across urban/suburban/rural areas
    2. Region-Race: V = 0.177. Shows significant geographic clustering of racial groups across US regions
    3. Sex-Exercise: V = 0.174. Reveals notable differences in exercise patterns between males and females
    4. Region-Urbanization: V = 0.152. Demonstrates the relationship between regional location and urban/rural setting
    5. Sex-Self-reported Weight: V = 0.150. Indicates gender differences in self-perception of weight status.



BMI Distribution Analysis and Transformation

The analysis of Body Mass Index (BMI) proceeded in several stages, examining data quality, distribution characteristics, and the need for transformation.

Initial Data Quality:

- From 4,036 total observations:
  - 85 extreme values (BMI ≥ 99) identified and treated as missing
  - Final missing count of 85 observations (2.1% of data)
- Cases with missing BMI were removed for subsequent analyses

Original BMI Distribution:

The detailed descriptive analysis after removing extreme values revealed:

- Central tendency and spread:
  - Mean: 26.5
  - Standard deviation: 5.37
  - Range: 15.19 - 64.37
  - Interquartile range (IQR): 6.3
  - Percentage of outliers outside IQR: 5.1%
- Distribution shape:
  - Strong positive skewness (1.26)
  - High kurtosis (6.32)
  - Heavy right tail evident in density plot
  - Significant deviation from normality in Q-Q plots
  - Shapiro-Wilk test rejected normality ($p < 0.001$)

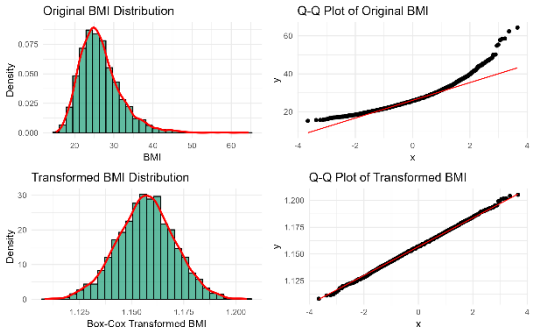
Transformation Analysis:

To address the non-normality, a Box-Cox transformation was applied:

- A Box-Cox transformation from the MASS library was applied using the model $bmi\_sp \sim self\_rep\_weight + diet\_low\_cal + race$
- Optimal lambda identified as -0.78, rounded to -0.8 (several other models were tried and the optimal lambda stayed between -0.75 - -0.8)
- Transformation formula: $box\_cox\_bmi = \frac{(bmi^{-0.8} - 1)}{-0.8}$

Comparison between original values and transformed values

Metric	Original	Transformed
Skewness	1.26	-0.02000
Kurtosis	6.32	3.04000
Shapiro-Wilk p-value	0.00	0.47786



Transformation Results:

The transformation substantially improved the distribution's properties:

- Skewness reduced from 1.26 to 0.02, indicating improvement in symmetry.
- Kurtosis improved from 6.32 to 3.06, very close to the normal distribution's value of 3.
- Shapiro-Wilk test showed marked improvement (p = 0.4), no longer rejecting normality.
- Q-Q plot showed much better alignment with the theoretical normal line, with points following the diagonal more closely throughout the distribution

The transformed BMI values were used as the response variable in subsequent analyses, as they better satisfy the normality assumptions of linear regression.

Examining Relationships with Body Mass Index

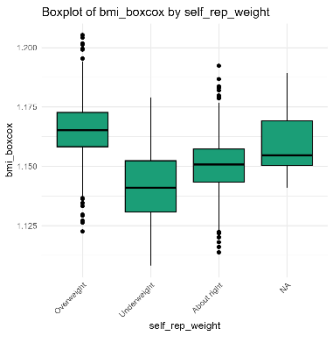
Visual Analysis of BMI Relationships:

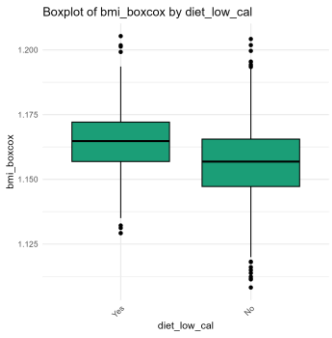
Exploratory visualization through box plots and scatter plots revealed several notable patterns in how BMI relates to various predictors. See code for all plots, after assessing significance with linear models the most important plots will be included in the summary.

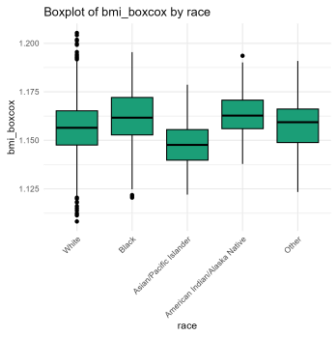
Statistical Validation Through Regression

Individual regression analyses quantified these relationships, identifying significant predictors in order of statistical significance:

- Self-reported weight status – negative beta with the highest r squared 0.34 (medium effect size).
- Race (“Black”) – negative beta with the second highest r squared 0.02 (small effect size).
- Low-calorie diet adherence (“No” level - lack of adherence) - negative beta







- The following variables showed significant coefficients but the r squared was small: education (grade), sex, income, other dietary habits (low-fat, low-salt, diabetic diet), exercise frequency, and age.

The identification of these significant relationships was the first step in screening predictors for the final multiple regression model.

Building a Multiple Regression Model for BMI Analysis

Model Development Process

Step 1: Initial variable screening based on statistical significance in individual regression

I began by using individual linear regression analyses to identify significant predictors of BMI (using the Box-Cox transformed BMI as response variable). 10 Variables were screened based on a significance threshold of $p < 0.05$, which identified several important predictors:

Formula: $bmi\_boxcox \sim self\_rep\_weight + diet\_low\_cal + race + sex + diet\_diabetic + income + diet\_low\_fat + diet\_low\_salt + exercise + age$

The initial multiple regression model incorporating all these significant predictors explained approximately 41% of the variance in BMI with Adjusted R-squared: 0.4129224, AIC: -24317.58, BIC: -24186.01.

Step 2: Backward step selection using AIC

Backward Selection Using AIC:

Started with all significant predictors from Step 1.

Applied backward elimination using AIC criterion (k=2) with stepAIC() function.

Resulted in a model retaining 7 predictors:

Self-reported weight status, diet (low-calorie, diabetic), race, sex, income, age

Model updated formula: $bmi\_boxcox \sim self\_rep\_weight + diet\_low\_cal + race + sex + diet\_diabetic + income + exercise + age$

Adjusted R-squared: 0.4101.

Model comparison with the model from step 1:

- Original Linear Model AIC: -24533.89
- Backward Model AIC: -24536.66
- Original Linear Model BIC: -24408.39
- Backward Model AIC: -24423.71

The model after Backward step selection using AIC shows better AIC and BIC. So I will continue with this model and refine it to include interaction terms (if they improve the model AIC and BIC).

Step 3: Backward step selection using BIC with Interactions:

Built a full model including all two-way interactions between the predictors

Used BIC criterion for more stringent variable selection (using AIC results in the inclusion of many interaction terms).

Applied backward elimination using BIC criterion (k=log(n)) with stepAIC() function.

Final Model Formula: $bmi\_boxcox \sim self\_rep\_weight + diet\_low\_cal + race + sex + diet\_diabetic + income + self\_rep\_weight{:}\,sex + self\_rep\_weight{:}\,diet\_diabetic + sex{:}\,diet\_diabetic + sex{:}\,income$

Adjusted R-squared: 0.4269

Comparison between model in step 2 (without interaction) and model in step 3:

Base Model AIC: -24579.89

Final Model with Interactions AIC: -24695.7

Base Model BIC: -24498.29

Final Model with Interactions BIC: -24582.72

Given that the model from step 3 yielded better (lower) AIC, BIC and better (higher) Adjusted R-squared, this will be the final model.

Final Model Assessment

The final model was evaluated using multiple diagnostic approaches:

Baseline (Reference Categories)

The intercept (1.169) represents the predicted transformed BMI for the reference case:

- Overweight individual (reference for self-reported weight)
- Following a low-calorie diet
- White race
- Male
- Following a diabetic diet
- Zero income

Main Effects

1. **Self-Reported Weight Status** (ref: Overweight)
    - Being "About right" significantly decreased BMI ($\beta = -0.01379$, $p < 0.001$)
    - Being "Underweight" showed no significant independent effect ($\beta = -0.00617$, $p = 0.105$)
2. **Diet**
    - Not following a low-calorie diet decreased BMI ($\beta = -0.00388$, $p < 0.001$)
    - Not following a diabetic diet showed no significant independent effect ($\beta = 0.00084$, $p = 0.622$)
3. **Race** (ref: White)
    - Black individuals showed higher BMI ($\beta = 0.00561$, $p < 0.001$)
    - American Indian/Alaska Native showed higher BMI ($\beta = 0.00618$, $p = 0.004$)
    - Asian/Pacific Islander ($\beta = -0.00089$, $p = 0.513$) and Other races ($\beta = 0.00148$, $p = 0.080$) showed no significant differences
4. **Sex and Income**

- Being female increased BMI (β = 0.00663, p < 0.001)
- Income alone showed no significant effect (β = -1.151e-09, p = 0.896)

Interaction Effects

1. **Sex × Self-Reported Weight**
    - Female × Underweight: Additional decrease in BMI (β = -0.00597, p < 0.001)
    - Female × About right: Additional decrease in BMI (β = -0.00523, p < 0.001) These interactions suggest that weight status affects BMI differently for males and females.
2. **Weight Status × Diabetic Diet**
    - Underweight × No diabetic diet: Substantial decrease in BMI (β = -0.01471, p < 0.001)
    - About right × No diabetic diet: No significant effect (β = 0.00095, p = 0.639)
3. **Sex × Diet and Income**
    - Female × No diabetic diet: Decreased BMI (β = -0.00639, p < 0.001)
    - Female × Income: Negative relationship (β = -8.361e-08, p < 0.001), indicating that higher income is associated with lower BMI among females

Model Fit

- Residual standard error: 0.01044.
- Small residuals range (-0.044 to 0.042) indicates good model fit
- Lower AIC (-24695.7) and BIC (-24582.72) compared to the base model, supporting the inclusion of interaction terms.

Key Insights

1. **Complex Gender Dynamics**
    - Women's BMI is influenced by multiple interacting factors
    - Income has a particularly strong moderating effect on female BMI
    - Weight status perceptions have different impacts across genders

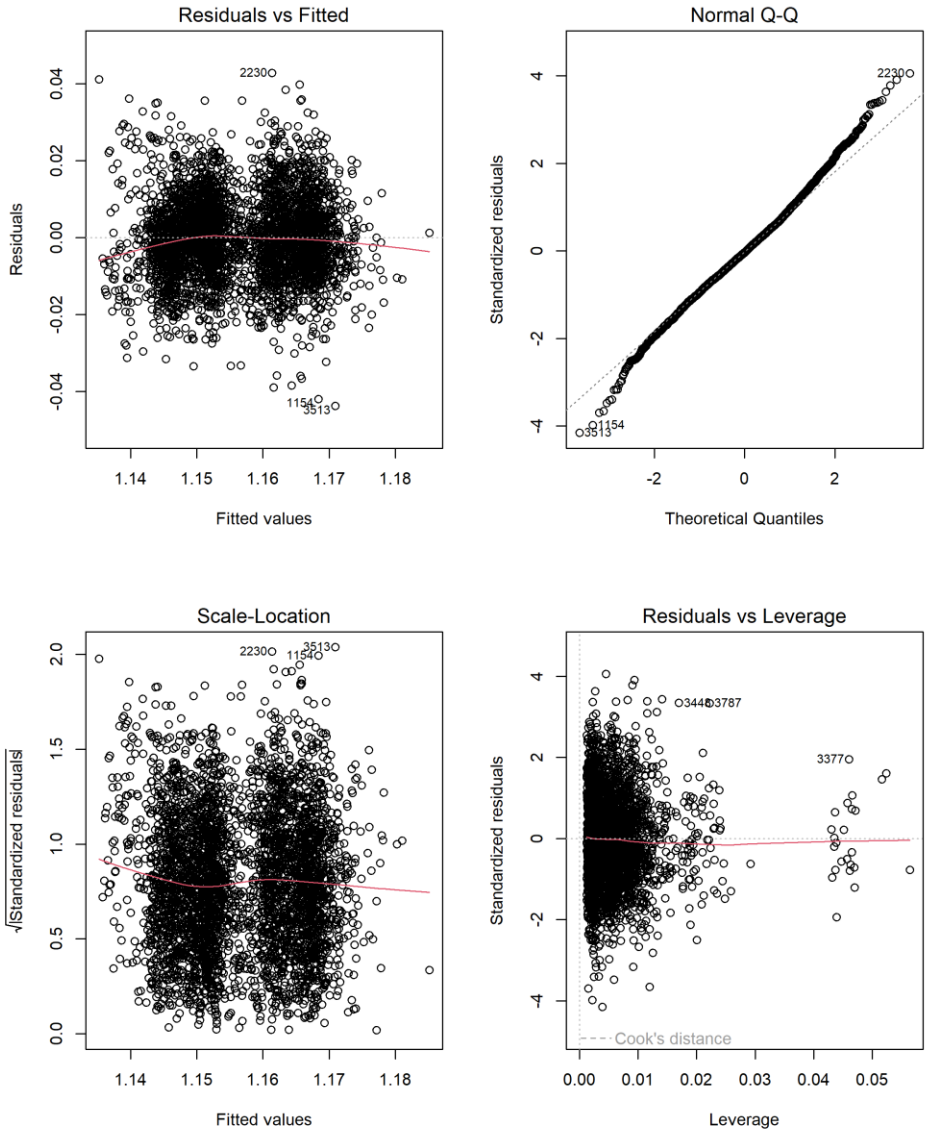
2. **Racial Differences**
    - Significant racial disparities persist even when controlling for other factors
    - Black and American Indian/Alaska Native individuals show consistently higher BMI
3. **Dietary Impacts**
    - Low-calorie diet shows direct effects
    - Diabetic diet's impact is moderated by both gender and self-reported weight status
4. **Socioeconomic Influence**
    - Income's effect is primarily evident through its interaction with gender
    - Suggests different socioeconomic influences across gender groups

Model coefficients:

```
Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                1.169e+00  1.821e-03 641.724  < 2e-16 ***
self_rep_weightUnderweight                -6.168e-03  3.804e-03  -1.621 0.105020
self_rep_weightAbout right                -1.379e-02  2.019e-03  -6.831 9.76e-12 ***
diet_low_calNo                            -3.878e-03  7.272e-04  -5.333 1.02e-07 ***
raceBlack                                  5.613e-03  5.385e-04  10.423  < 2e-16 ***
raceAsian/Pacific Islander                -8.891e-04  1.359e-03  -0.654 0.513033
raceAmerican Indian/Alaska Native          6.178e-03  2.147e-03   2.877 0.004031 **
raceOther                                  1.484e-03  8.483e-04   1.750 0.080272 .
sexfemale                                  6.632e-03  1.929e-03   3.437 0.000594 ***
diet_diabeticNo                            8.408e-04  1.704e-03   0.493 0.621831
income                                    -1.151e-09  8.810e-09  -0.131 0.896043
self_rep_weightUnderweight:sexfemale      -5.972e-03  1.481e-03  -4.031 5.66e-05 ***
self_rep_weightAbout right:sexfemale      -5.233e-03  6.978e-04  -7.499 7.88e-14 ***
self_rep_weightUnderweight:diet_diabeticNo -1.471e-02  3.766e-03  -3.906 9.53e-05 ***
self_rep_weightAbout right:diet_diabeticNo  9.490e-04  2.026e-03   0.468 0.639457
sexfemale:diet_diabeticNo                 -6.394e-03  1.916e-03  -3.337 0.000854 ***
sexfemale:income                          -8.361e-08  1.267e-08  -6.597 4.76e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01044 on 3915 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.4293,    Adjusted R-squared:  0.4269
F-statistic:   184 on 16 and 3915 DF,  p-value: < 2.2e-16
```

Model diagnostic plots:



Some points have high leverage and standardized residuals.

A better model will include removing outliers based on the cook's distance.

Code:

- All of the analysis was done in the markdown notebook. I will attach an .rmd version as well as HTML version for convenient viewing.
- All plots included in the analysis can be found in the code.
- The code roughly follows the structure of the tasks.