Exploratory Data Analysis

Objective:

The aim of this hypothesis test is to assess the existence of a significant link between Body Mass Index (BMI) and Systolic Blood Pressure in the NHANES dataset.

The aim of this hypothesis test is to assess the existence of a significant link between Body Mass Index (BMI) and Systolic Blood Pressure in the NHANES dataset.

Data Loading and Cleaning:

The code loads the NHANES dataset and removes duplicate entries based on the "ID" variable.

It uses the nhanesA package to load specific tables related to demographics, dietary information, examinations, laboratory results, and questionnaires for the year 2010.

Variable Selection:

The code creates a new data frame (selected_data) with selected variables related to age, gender, height, weight, average systolic and diastolic blood pressure, diabetes status, physical activity, and BMI.

Handling Missing Values:

The code defines a function (replace_na_with_mode) to replace missing values in categorical columns with the mode.

It then applies this function to replace missing values in the selected data.

Handling Numeric Missing Values:

The code replaces missing numeric values (e.g., blood pressure, age, height, weight, BMI) with their respective column means.

Adjusting for Age:

The code builds an adjusted linear regression model by including age as an additional predictor.

It prints the summary of the adjusted model, residuals, and other relevant information.

Dataset Description:

The NHANES dataset is sourced, and potential duplicates are removed. Selected variables include Age, Gender, Height, Weight, Body Mass Index (BMI), Systolic Blood Pressure (BPSysAve), Diabetes status, and Physical Activity level.

Data Preprocessing:

- Missing values in categorical columns are replaced with the mode.
- Missing values in numeric columns (BPSysAve, BPDiaAve, Age, Height, Weight, BMI) are imputed with the mean.

Descriptive Statistics:

Descriptive statistics for all selected variables are computed. This includes measures such as mean, standard deviation, minimum, and maximum values.

The code prints the structure and summary statistics of the selected_data dataframe.

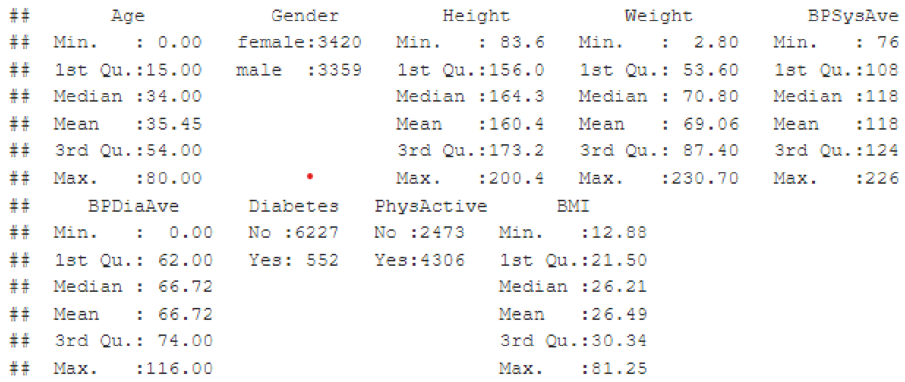
It creates a scatter plot of BMI against systolic blood pressure and calculates the correlation between the two variables.

Null Hypothesis (H0):

There is no link between BMI and Systolic Blood Pressure. In the regression model, the coefficients for BMI are equal to zero.

Alternative Hypothesis (H1):

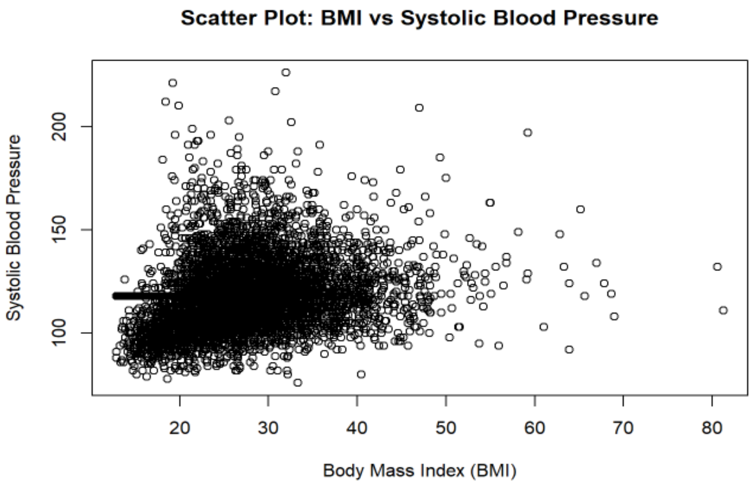
There is a link between BMI and Systolic Blood Pressure. At least one of the BMI coefficients is not equal to zero.



```
##       Age           Gender         Height          Weight          BPSysAve
##  Min.   : 0.00   female:3420   Min.   : 83.6   Min.   :  2.80   Min.   : 76
##  1st Qu.:15.00   male  :3359   1st Qu.:156.0   1st Qu.: 53.60   1st Qu.:108
##  Median :34.00                 Median :164.3   Median : 70.80   Median :118
##  Mean   :35.45                 Mean   :160.4   Mean   : 69.06   Mean   :118
##  3rd Qu.:54.00                 3rd Qu.:173.2   3rd Qu.: 87.40   3rd Qu.:124
##  Max.   :80.00                 Max.   :200.4   Max.   :230.70   Max.   :226
##     BPDiaAve       Diabetes    PhysActive       BMI
##  Min.   :  0.00   No :6227   No :2473   Min.   :12.88
##  1st Qu.: 62.00   Yes: 552   Yes:4306   1st Qu.:21.50
##  Median : 66.72                         Median :26.21
##  Mean   : 66.72                         Mean   :26.49
##  3rd Qu.: 74.00                         3rd Qu.:30.34
##  Max.   :116.00                         Max.   :81.25
```

Scatter Plot:

A scatter plot is created to visualize the relationship between BMI and Systolic Blood Pressure.



Correlation value = 0.228619577716269"

Regression Analysis:

The code builds a simple linear regression model with BMI as the predictor and systolic blood pressure as the response variable.

It plots the regression line on the scatter plot and prints the model summary.

A simple linear regression model is built with BMI as the predictor for Systolic Blood Pressure.

Hypothesis Testing:

The code provides an explanation of the hypothesis test for the simple regression model.

It formulates null and alternative hypotheses and then adjusts the model for age to see if it affects the association between BMI and systolic blood pressure

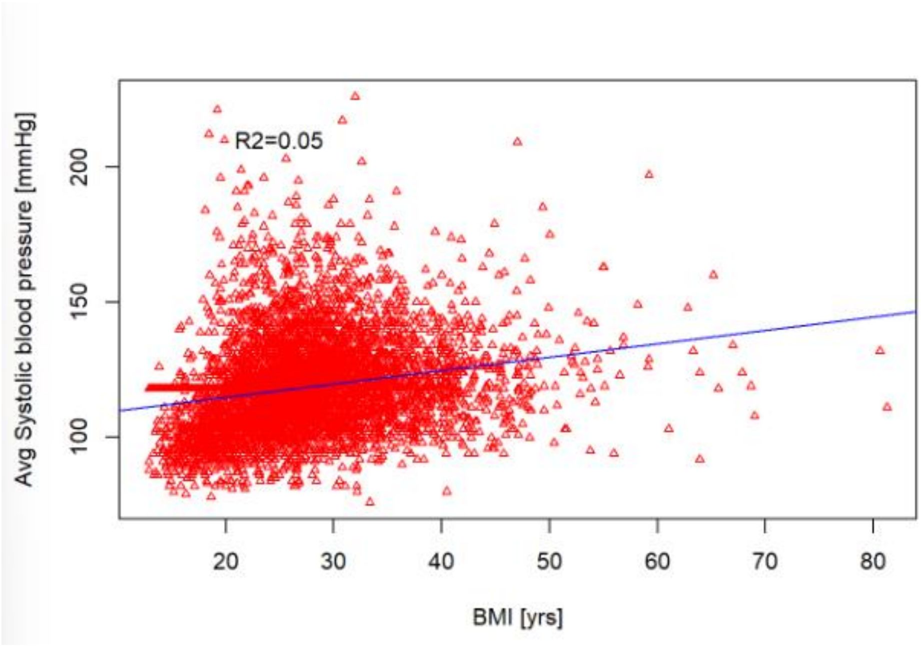
A hypothesis test is performed to evaluate the link between BMI and Systolic Blood Pressure. The null hypothesis assumes no link, while the alternative hypothesis suggests a link.

P- value obtained from Hypothesis test is 2.2e- 16 < alpha

R -Squared = 0.05227

Adjusted R-squared: 0.05213

There is evidence to support a link between BMI and systolic blood pressure

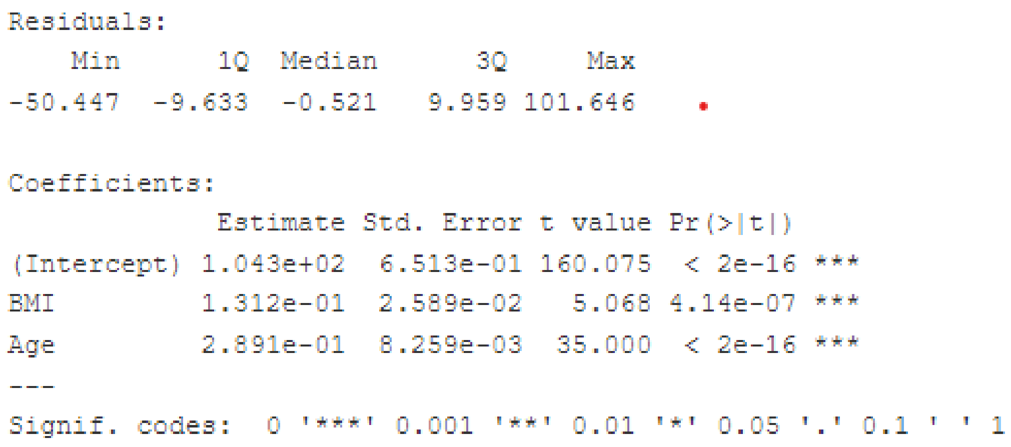


Hypothesis Test: Simple Regression Model

Hypothesis: There exists a significant association between BMI and Systolic BP

Null Hypothesis (H0): There is no significant association between BMI and Systolic BP. Mathematically, the coefficients for BMI and age in the regression model are equal to zero.

Alternative Hypothesis (H1): There is a significant association between BMI and Systolic BP. At least one of the coefficients for BMI or age is not equal to zero.



```
Residuals:
    Min      1Q  Median      3Q     Max
-50.447  -9.633  -0.521   9.959 101.646

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.043e+02  6.513e-01 160.075  < 2e-16 ***
BMI         1.312e-01  2.589e-02   5.068 4.14e-07 ***
Age         2.891e-01  8.259e-03  35.000  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 14.38 on 6776 degrees of freedom

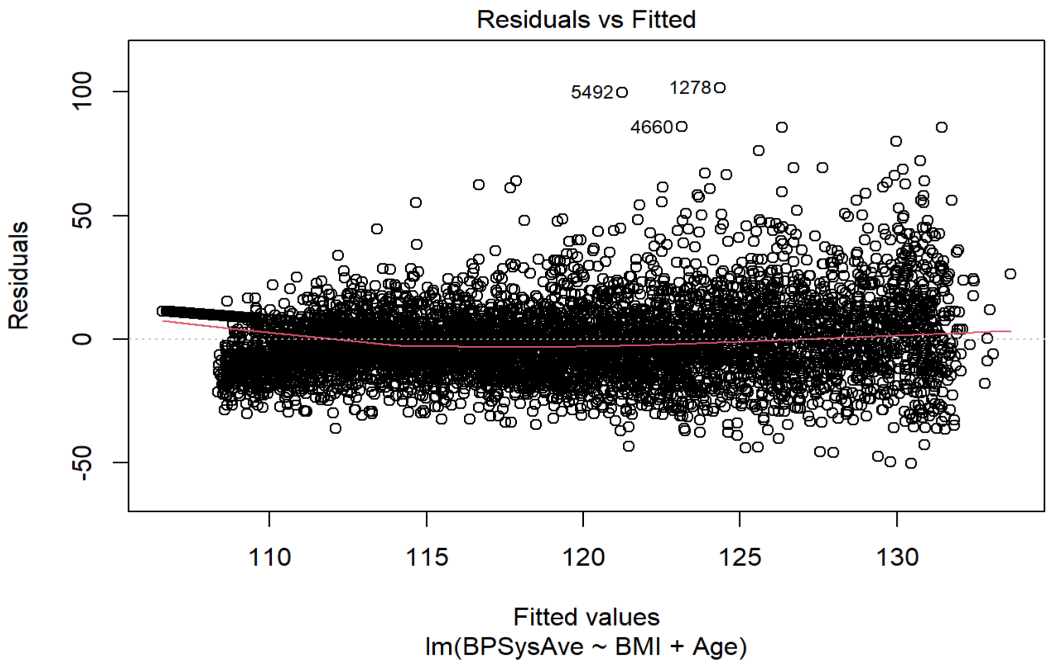
Multiple R-squared: 0.1974

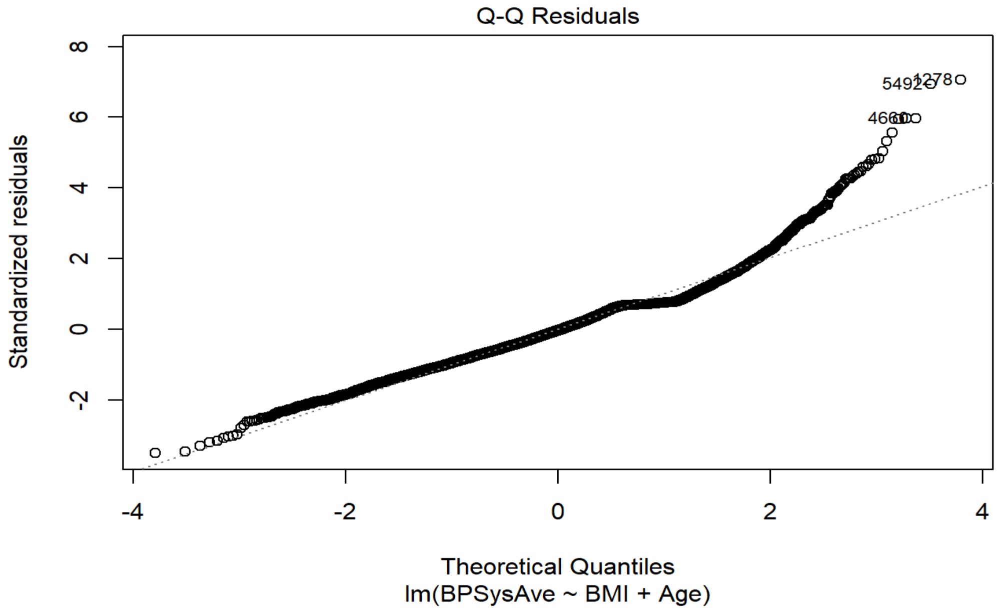
Adjusted R-squared: 0.1971

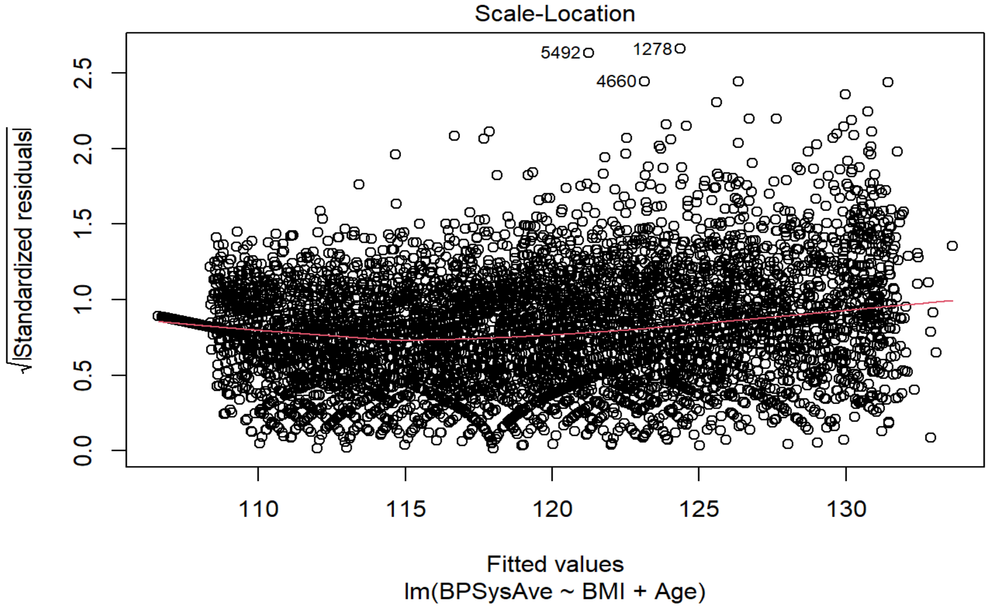
F-statistic: 833.1 on 2 and 6776 DF

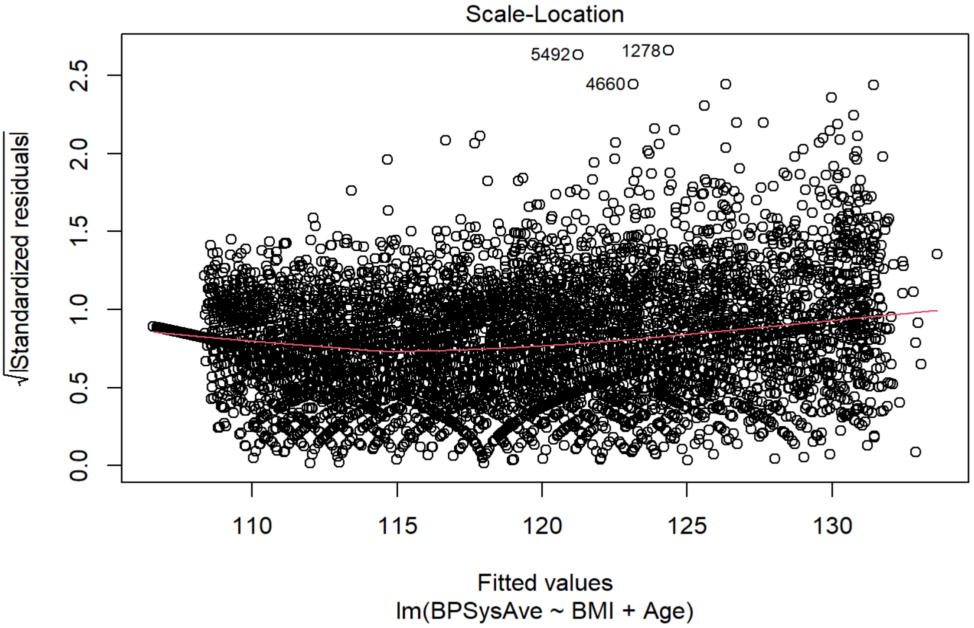
p-value: < 2.2e-16

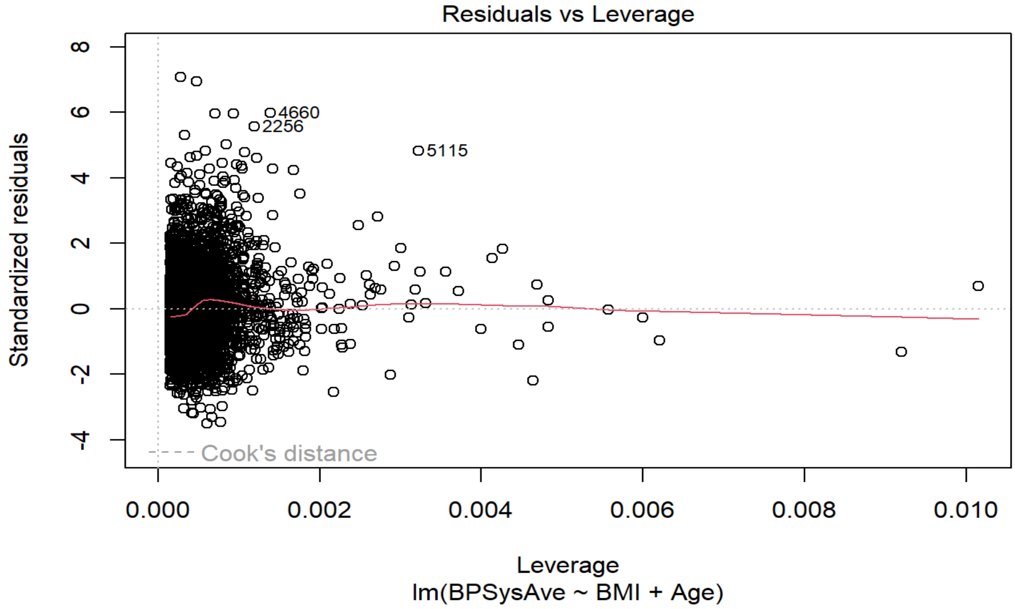
Adjusted Model




Q-Q Residuals
Standardized residuals
-2
0
2
4
6
8
5492
1278
4660
-4
-2
0
2
4
Theoretical Quantiles
lm(BPSysAve ~ BMI + Age)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
5492
1278
4660
110
115
120
125
130
Fitted values
lm(BPSysAve ~ BMI + Age)

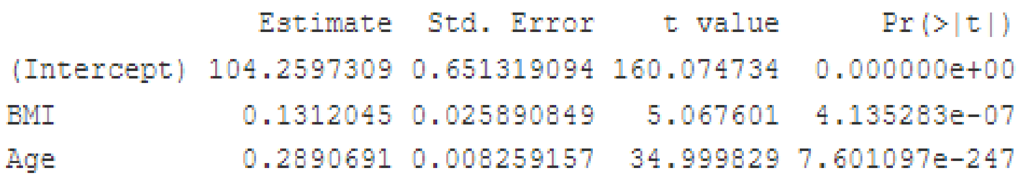

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
110
115
120
125
130
5492
1278
4660
Fitted values
lm(BPSysAve ~ BMI + Age)


Residuals vs Leverage
Standardized residuals
-4
-2
0
2
4
6
8
4660
2256
5115
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
lm(BPSysAve ~ BMI + Age)

P values = 4.135283e-07, 7.601097e-247 < alpha

R-Squared = 0.1971323

Which is an improvement in model fitting



```
               Estimate  Std. Error     t value       Pr(>|t|)
(Intercept) 104.2597309 0.651319094 160.074734  0.000000e+00
BMI           0.1312045 0.025890849   5.067601  4.135283e-07
Age           0.2890691 0.008259157  34.999829 7.601097e-247
```

Conclusion

In conclusion, systolic blood pressure and BMI are significantly correlated. Even after accounting for age, the relationship is still statistically significant. This implies that age does not lessen the observed association, even though BMI is a significant predictor of systolic blood pressure. Deeper insights into these relationships might be obtained through additional analyses or stratifications.