STA302 Final Project Part 3

Table of Contents

Table of Contents **0**

I. Introduction **1**

A. Research Question: 1

Dataset & Predictor Variable Description: 1

Chosen Linear Model: 2

B. Literature & Context: 2

Existing Literature 2

Filling In The Gaps 2

II. Methods **3**

A. Assessing Model Assumptions 3

B. Diagnostics 4

C. Model Selection 5

II. Results **7**

IV. Discussion **11**

III. Appendix **13**

Plots 13

Final Model 13

Intermediate Model 14

Response vs Predictor Plots 15

Residual vs PredictorStandardized residuals histogram & Normal QQ plot 16

Preliminary Model 23

Distribution of Predictor and Response Variables 34

Works Cited 39

I. Introduction

Background

Mental health, a critical aspect of human well-being, is influenced by a combination of various factors. Among these, biological and lifestyle markers have emerged as strong predictors of mental health outcomes.

A. Research Question:

This paper is dedicated to exploring the following research question: "To what extent can biological and lifestyle markers, such as age, gender, direct cholesterol, systolic and diastolic blood pressure, BMI, duration of sleep, physical activity and alcohol consumption predict depressive symptoms, measured by number of self-reported bad mental health days per month, among survey respondents 18 - 80 years old?".

In this study, our investigation centers solely on the categorical predictor of gender, which encompasses two distinct categories: male and female.

Dataset & Predictor Variable Description:

The data is obtained from the NHANES dataset available at https://wwwn.cdc.gov/nchs/nhanes/ . The description of variables chosen is provided here:

Table 1: variables, justification of use and role in model.

Variable Name	Justification for Use	Role in Model
Gender	Gender is a biological marker and is a predictor, which is a predictor of interest. People of one gender may be more predisposed to higher self-reported days of bad mental health.	Categorical Predictor Variable
DirectChol	Direct Cholesterol is a biological marker, which is a predictor of interest.	Predictor Variable
BPSysAve	Average Systolic Blood Pressure is a biological marker, which is a predictor of interest.	Predictor Variable
BPDiaAve	Average Diastolic Blood Pressure is a biological marker, which is a predictor of interest.	Predictor Variable
BMI	BMI is a biological marker, which is a predictor of interest.	Predictor Variable
SleepHrsNight	Self-reported hours of sleep a night is a lifestyle marker, which is a predictor of interest.	Predictor Variable

PhysActiveDays	Self-reported active days in a week is a lifestyle marker, which is a predictor of interest.	Predictor Variable
AlcoholYear	Number of estimated alcohol consumption in a year is a lifestyle marker, which is a predictor of interest.	Predictor Variable
Age	Age is a biological marker, which is a predictor of interest.	Predictor Variable
DaysMentHlthBad	The number of self-reported bad mental health days is a good measure of the well-being of the individual, which we are interested in studying.	Response Variable

Chosen Linear Model:

The linear model chosen to answer our research question is given by:
SleepHrsNight ~ Gender + SleepHrsNight + DirectChol + PhysActiveDays + Age + AlcoholYear + BMI + BPSysAve + BPDiaAve

The properties of the multiple linear regression model that were chosen to answer the research question are the coefficients for each of the predictor variables of interest. In a multiple linear regression model, each predictor variable is assigned a coefficient that quantifies its relationship with the dependent variable (in this case, the number of self-reported bad mental health days per month). The magnitude of these coefficients indicates the strength and direction of the relationship.This property is crucial for answering the research question as it directly indicates how and to what extent each biological and lifestyle marker influences the depressive symptoms.

B. Literature & Context:

Existing Literature

Previous research has examined the relationship between individual biological and lifestyle markers and depression. Notable studies include an investigation into the link between sleep duration and depression titled "Association between Sleep Duration and Depression in US Adults: A Cross-Sectional Study", a study on the levels of physical activity and depressive symptoms named "Relationship Between Physical Activity and Depression and Anxiety Symptoms: A Population Study", and research exploring the connection between race, cardiovascular risk, and depression, as seen in "History of Depression, Race, and Cardiovascular Risk in CARDIA". In much of existing research, a singular biological or lifestyle marker's relationship with depressive symptoms is explored. However, there remains a substantial gap in understanding the collective impact of these various biological/lifestyle markers on predicting depressive symptoms.

Filling In The Gaps

This paper focuses on several specific markers and seeks to answer the research question mentioned above. With existing research exploring the relationship between specific biological and lifestyle markers with depressive symptoms separately. This paper aims to provide an even more thorough and comprehensive understanding by examining the combined effect of multiple select markers rather than fixating on a singular one through the use of multiple linear regression.

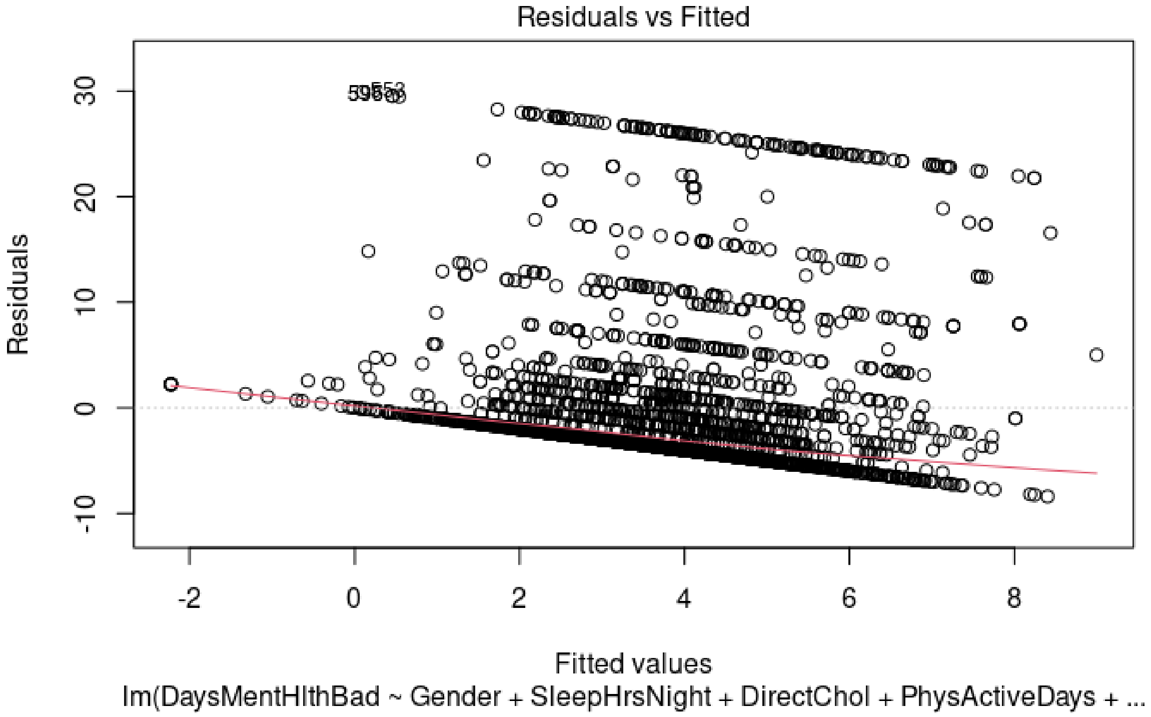
II. Methods

A. Assessing Model Assumptions

There are three model assumptions and the methods to check them are listed:

1. **Linearity Assumption**: The relationship between the response variable and all predictor variables is linear. This can be assessed by looking at plots of :
   - residuals vs fitted values and vs predictors: the model is correct if the plot has constant mean 0, and looks like a null plot.
   - Response vs fitted values: the model is correct if the plot is linear
   - Response vs predictors: the model is correct if the plot is linear

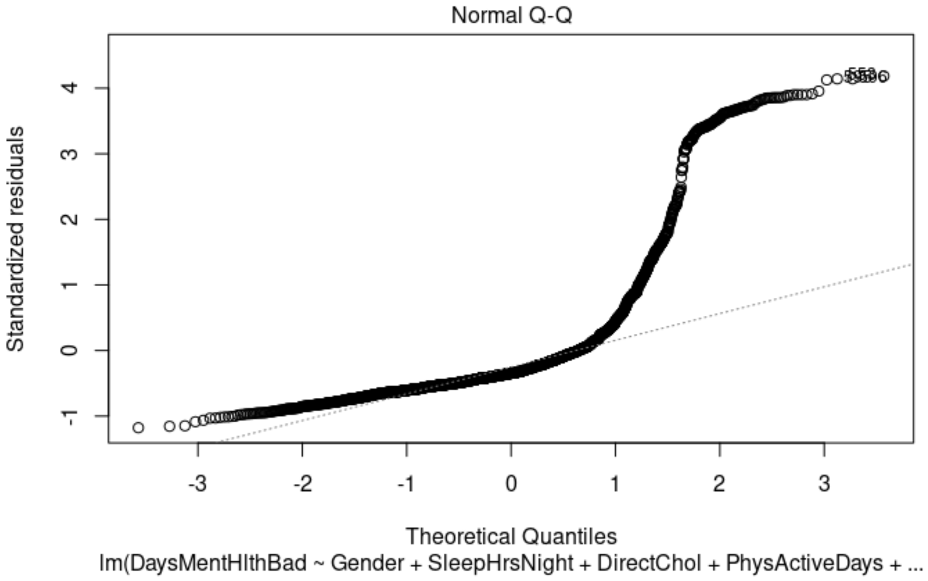
   Figure 1: Plot of Residual vs Fitted Values

   Our Model does not look like a null plot, and thus **violates this assumption.**
2. **Constant Error Variance:** This assumption can be assessed by looking at residual and standardized residual plots. The assumption holds if the plot of the residuals and standardized residuals varies without obvious patterns about 0, and have unit variance.

   We observe from figure 1 that the plot of the residuals and standardized residuals does vary with an obvious pattern, and hence **our model violates this assumption**.
3. **Independent and Normal Errors:** This can be assessed by looking at the QQ plot produced by our model, to see if the plot looks approximately like a line.

   Figure 2: QQ Plot for our model



We observe that the QQ plot doesn’t approximately look like a straight line. Thus, **our model violates this assumption.**

**all diagnostic plots for this model can be found in the appendix.*

B. Diagnostics

The issue for our model arises because the distribution of some variables in our model is extremely skewed, which can be determined from residual plots for predictors and the QQ plot for response. Once we identify problematic variables, we transform them to have stable variance or

We can fix our model to satisfy the assumptions by:

- Removing the DaysMentHlthBad variable as it’s heavily right skewed even after applying log transformation. (We use modSleepHrsNight in its place)
- applying the following transformations obtained by performing box-cox:
    - Square transformations on left skewed variable distributions where:
        - modSleepHrsNight = (SleepHrsNight)^2
    - Log transformation on right skewed variable distributions where:
        - modAlcoholYear = log(AlcoholYear + 1). We add 1 as AlcoholYear may be 0.
        - modBMI = log(BMI)
        - modBPSysAve = log(BPSysAve)
        - modDirectChol = log(DirectChol)

These transformations stabilize variance in the variables, and give us residual plots that have constant mean around 0, with no discernable pattern. This satisfies the first 2 assumptions. The QQ plot produced resembles a line and the standardized residual histogram is also approximately normal.

Figure 3: QQ Plot for transformed model

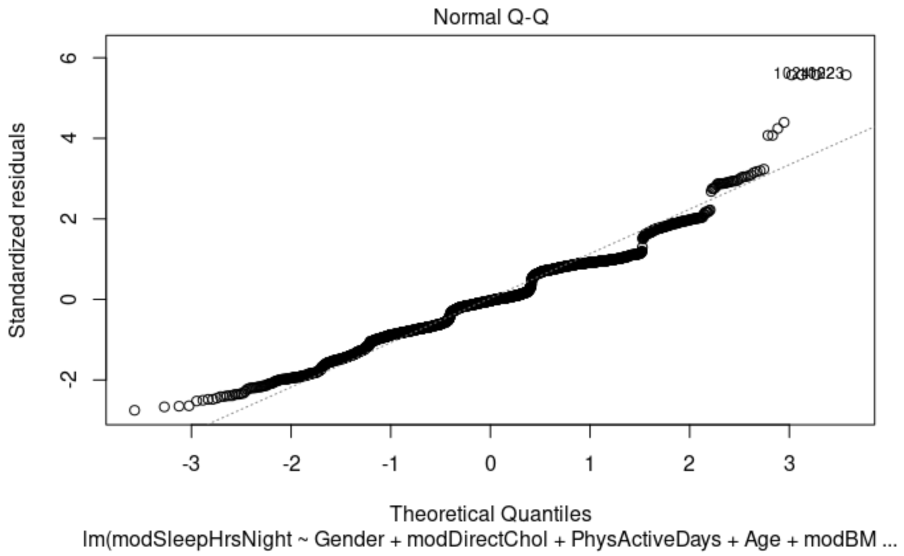
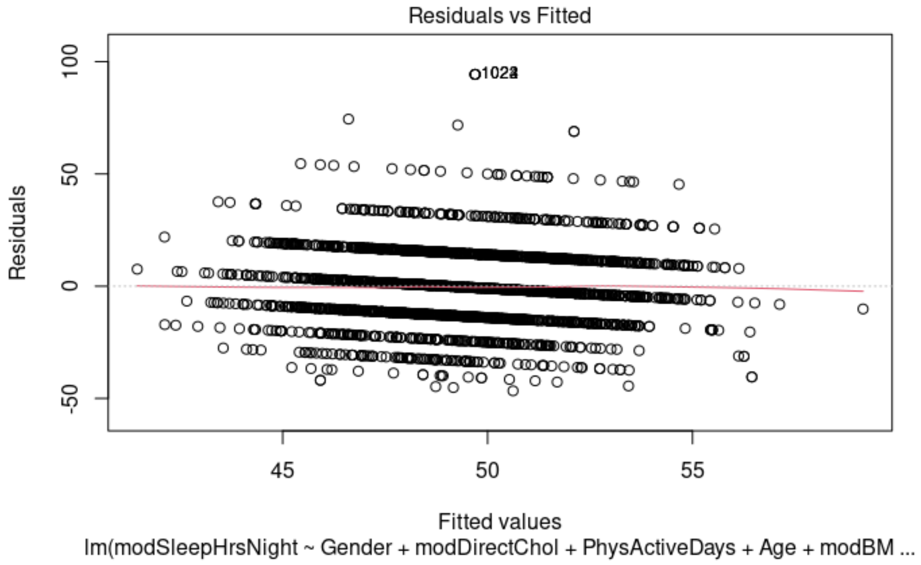


Figure 4: Residual Plot for transformed model



C. Model Selection

Now, we choose the transformed predictor variables that have a linear relationship with the transformed response variable only to be in our model using t tests, with null hypothesis being that the coefficient of a predictor variable = 0, and with alternative hypothesis being that it doesn't equal 0. It can be seen from the model summary, in the appendix (Intermediate Model), that modDirectChol, PhysActiveDays, and BPDiaAve do not have any statistically significant linear relationship with modSleepHrsNight as indicated by their p values, for the t test, which are greater than 0.05. Thus, we may remove them from our final model, and only keep Gender, Age, BMI, BPSysAve, and Alcohol year that are transformed as previously determined.

Thus, the final model is given by:
(SleepHrsNight)^2 ~ Gender + Age + log(BMI) + log(BPSysAve) + log(AlcoholYear)

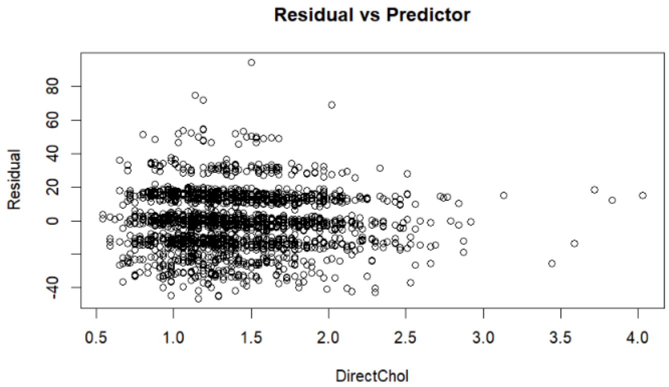
Now, we use a partial f test with a null hypothesis that coefficients of all predictor variables equal 0, and an alternative hypothesis that at least one coefficient of predictors does not equal zero. This helps us determine if the final model we've chosen is statistically significant. From the model summary, in the appendix(Final Model), we can observe the calculated test statistic and determine whether we should reject or accept the null hypothesis when the test statistic is less than 0.05 or greater than 0.05 respectively.

II. Results

Our study commenced with the objective of predicting the level of depressive symptoms among individuals aged 18 to 80, utilizing a range of biological and lifestyle markers. These markers were judiciously selected based on their recognized importance as indicators of physical and mental health, as substantiated by existing literature.

Thus, we started with this preliminary model:
DaysMentHlthBad ~ Gender + DirectChol + BPDiaAve + PhysActiveDays +Age + BMI + BPSysAve + AlcoholYear + SleepHrsNight

To ensure the robustness of our model, we began by validating the key assumptions of linear regression. This entailed a careful examination of the residuals plotted against each predictor variable. In all graphs that were plotted, large clusters of many points were observed indicating that the errors were correlated. The "Uncorrelated Error" assumption was violated.

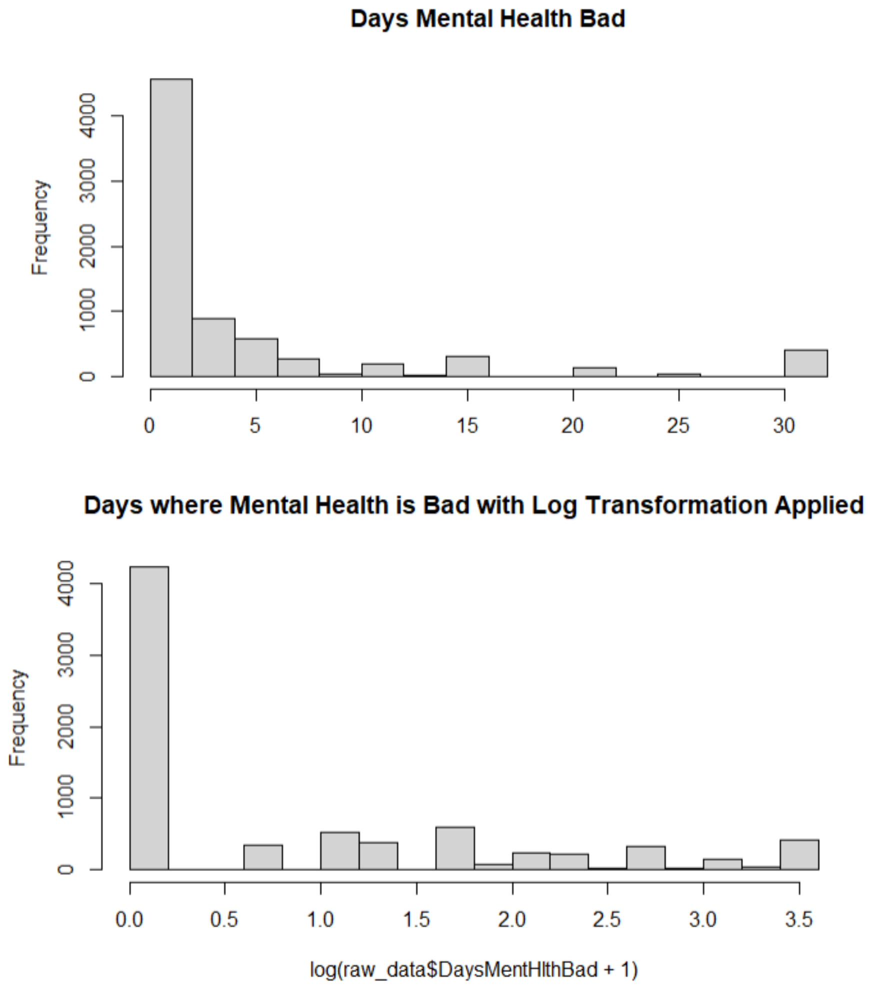


Plotting Residuals against Cholesterol levels. Refer to the Appendix, "Preliminary Model" section for the rest of the plots

Additionally, we noticed that there were stark deviations from the diagonal when a QQ plot was created to verify the Normality Assumption. This QQ plot can be found in the "Assessing Model Assumptions" section of **Methods.**

In evaluating the normality assumption critical to our regression model, a QQ plot was generated (it can be referenced in the "Assessing Model Assumptions" section of the Methods). The plot indicated stark deviations from the diagonal.

To narrow down the source of the violation, we looked at the distribution of each variable in our model to ensure that it followed a normal distribution. Although the predictor variables were normally distributed, the response variable was not, even after a Log Transformation was applied to it to correct for its right skew.



Histogram showing the right-skew of the Bad Mental Health days data. Histograms for the predictor variables can be found in Appendix

A decision was made to change the response variable to SleepHrsNight. The reason it was chosen was two-fold:

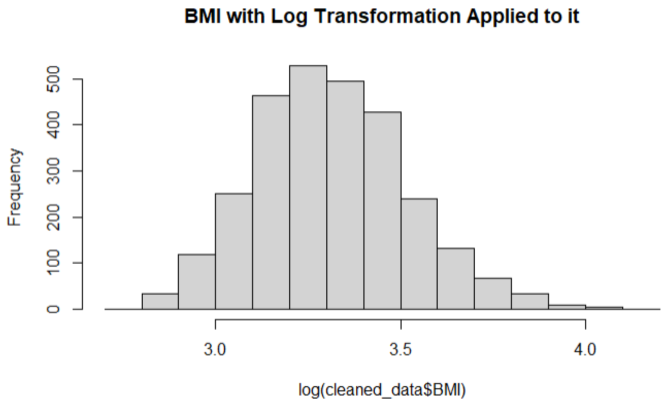
1. The link between sleep hours had already been investigated in existing literature (Dong, Lu, et al, 2021), and thus by combining multiple biological and lifestyle factors, we build upon the gaps in existing research.
2. After a Box-Cox transformation was applied to the Sleep Hours per Night distribution, it was successfully normalized.

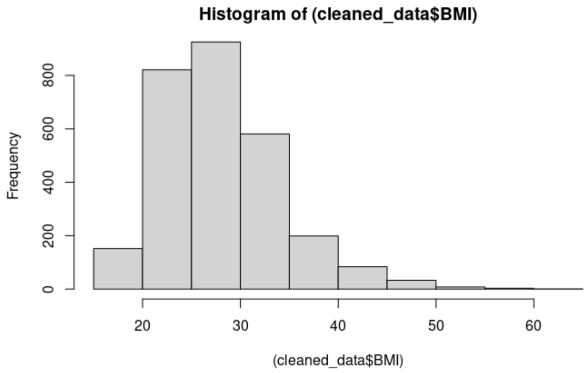
Hence, our research question changed to *"To what extent can biological and lifestyle markers, such as age, gender, direct cholesterol, systolic and diastolic blood pressure, BMI, physical activity and alcohol consumption predict hours of sleep at night".*

Therefore, our intermediate model is given by:

SleepHrsNight ~ Gender + DirectChol + BPDiaAve + PhysActiveDays +Age + BMI + BPSysAve
+ AlcoholYear

Again, the linear regression assumptions were verified through “Residual vs Predictor plots”, “Residual vs Fitted Value plots” and QQ plots.
Log transformations were applied to the AlcoholYear, BMI, BPSysAve and DirectChol variables to correct for their right-skewness, as seen in the Diagnostics sections (under Methods).



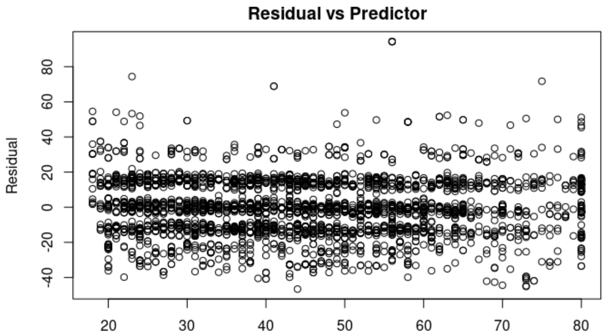


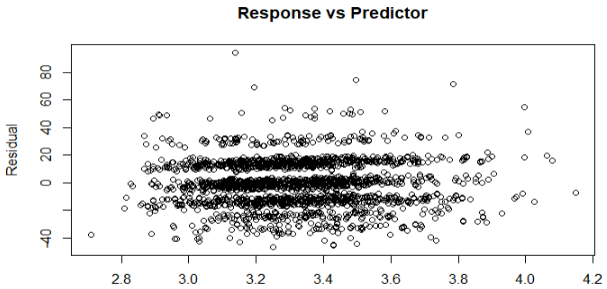
Above is an example of the linear regression assumption checking plots, together with the distribution of the BMI predictor variable before & after a Log transformation was applied to it. To look at the rest of the assumption-checking plots for the updated research question, look at the Intermediate Model section in the ***Appendix.***

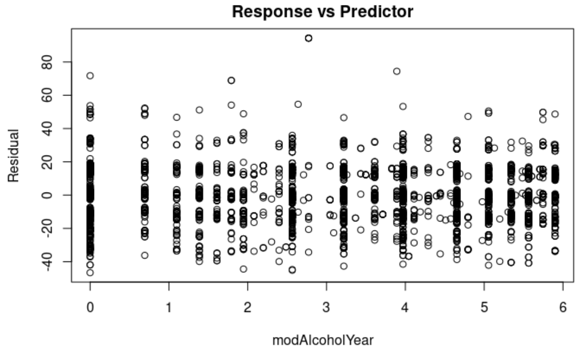
From the snippet of residual vs predictor graphs below, we can see that all 4 assumptions have been satisfied. This was the case for all the other residual vs predictor graphs too

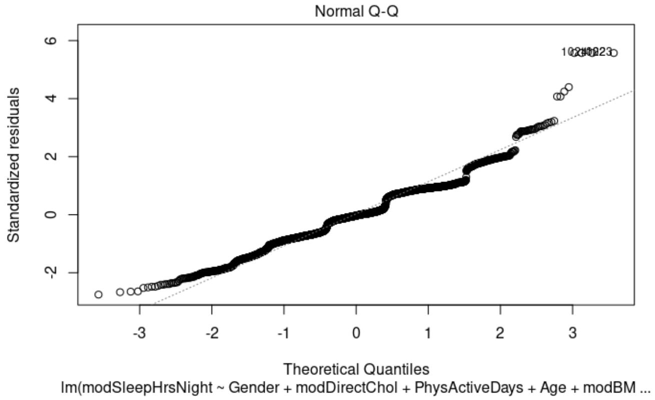
1. No systematic patterns such as curves or increasing/ decreasing spreads can be observed, hence both the linearity and constant variance assumptions are satisfied
2. There are no large patterns of many points across a particular sequence so we have no correlated errors

3. There are no Stark deviations from the diagonal line on the QQ plot, as there were with our preliminary model so our normality assumption is satisfied too.



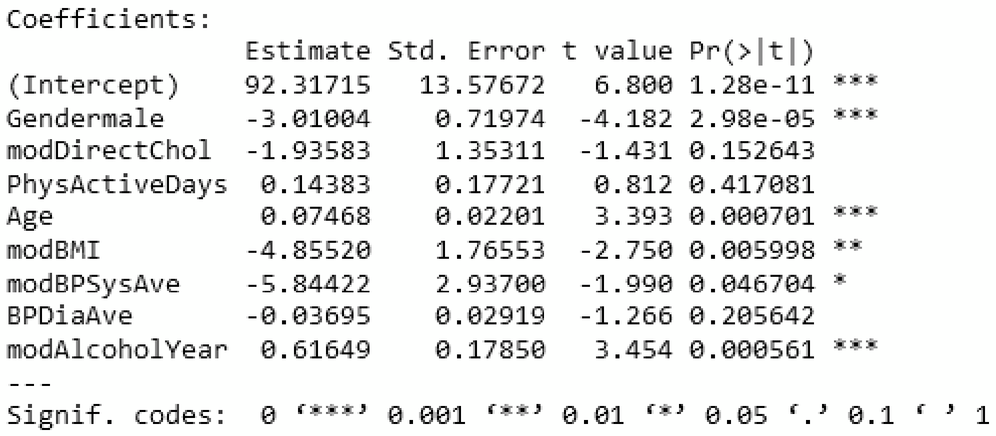






Thus, the model so far is given by:
(SleepHrsNight)^2 ~ Gender + log(DirectChol) + BPDiaAve + PhysActiveDays +Age + log(BMI) +
+log(BPSysAve) + log(AlcoholYear)

We then proceeded to perform T-tests on the individual coefficients of each of the predictor variables to see if they were statistically significant in predicting the response variable, getting the following output:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    92.31715   13.57672   6.800 1.28e-11 ***
Gendermale     -3.01004    0.71974  -4.182 2.98e-05 ***
modDirectChol  -1.93583    1.35311  -1.431 0.152643    
PhysActiveDays  0.14383    0.17721   0.812 0.417081    
Age             0.07468    0.02201   3.393 0.000701 ***
modBMI         -4.85520    1.76553  -2.750 0.005998 ** 
modBPSysAve    -5.84422    2.93700  -1.990 0.046704 *  
BPDiaAve       -0.03695    0.02919  -1.266 0.205642    
modAlcoholYear  0.61649    0.17850   3.454 0.000561 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the model summary, the following predictor variables were found to not have a statistically significant linear relationship with modSleepHrsNight at the 5% significance level:

1. modDirectChol (t = -1.43, p = 0.153)
2. PhysActiveDays (t = 0.812, p = 0.417)
3. BPDiaAve (t = -1.27, p = 0.206)

Therefore we removed them from our model and ended up with this final regression fit:

(SleepHrsNight)^2 ~ Gender + Age + log(BMI) + log(BPSysAve) + log(AlcoholYear)

IV. Discussion

Why is it important?

Understanding the impact of these markers on mental health is crucial in addressing mental health challenges globally. The model contributes to understanding the possible causes of bad mental health and its determinants. It highlights various marks that can influence an individual's mental health. Based on the implications and analysis, individuals facing this can take a practical approach to improving their mental health.

Limitations of the analysis:

- The model is only able to predict the average duration of sleep that a person estimates that they get, and isn't able to directly make claims about the amount of self-reported days where they had bad mental health. However, in Dong, Lu, et al 2021, a direct correlation is established between too much or too little sleep and mental health issues. This helps us make educated guesses about the lifestyle factors that affect mental health and sleep.
- The transformations applied to the predictor variables makes this model harder to interpret intuitively.

Why were some limitations not addressed?

1. We weren't able to address this limitation because the MentlHlthBad variable was severely skewed even after log transformation, and had to pivot to somewhat answer our research question.
2. Transformations need to be applied to the raw predictor variables to allow us to be able to perform multiple linear regression that offers accurate insight, even if we have to compromise on interpretability.

Summary and Observations made on the Linear Model:

Despite the limitation in directly analyzing the relationship, the model was able to provide us a valuable insight into the biological and lifestyle markers. Upon further analysis of the model, the model implies that:

- Based on the analysis, on average, men had lesser quantity of sleep than women, people with high BMI and Systolic Blood pressure had lesser quantity of sleep, and there was a positive correlation between age and alcohol consumption with quantity of sleep.

- Sleep, which is correlated to mental health, is highly affected by all these lifestyle and biological factors. The observation that lifestyle and biological factors affect sleep aligns with the findings of [Dong, Lu, et al]. This was another reason why we changed the response variable from MentHlthBadDays to the Sleep duration. The extent to which sleep duration plays a major role as a predictor upon predicting MentHlthBadDays was rather unexpected.

In conclusion, the linear model emphasizes the relationship of sleep with gender, age, BMI, Systolic BP, and alcohol consumption. This helps us make educated guesses about people's mental health, aligning with the objective of the research question.

Contributions

Table 2: Contributions

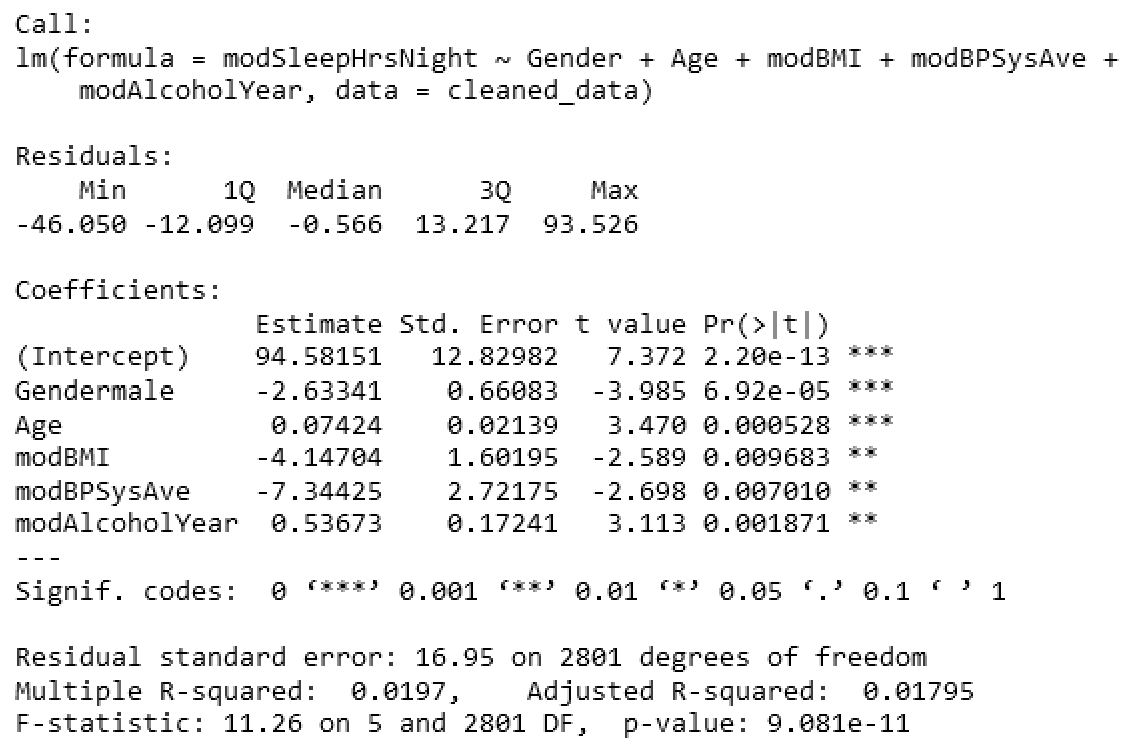
Contributor	Contribution
Happy Nasit	Responsible for the Interpretation and Discussion Section.
Kevin Z Shen	Responsible for the Introduction Section and Works Cited page.
Khushil Nagda	Responsible for the Results Section.
Madhav Kanna Thenappan	Responsible for performing R analysis and methods section, document layout and formatting.

III. Appendix

Plots

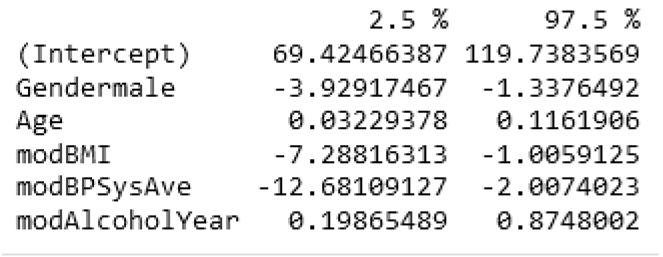
Final Model

Fig 10: Summary of Final Model



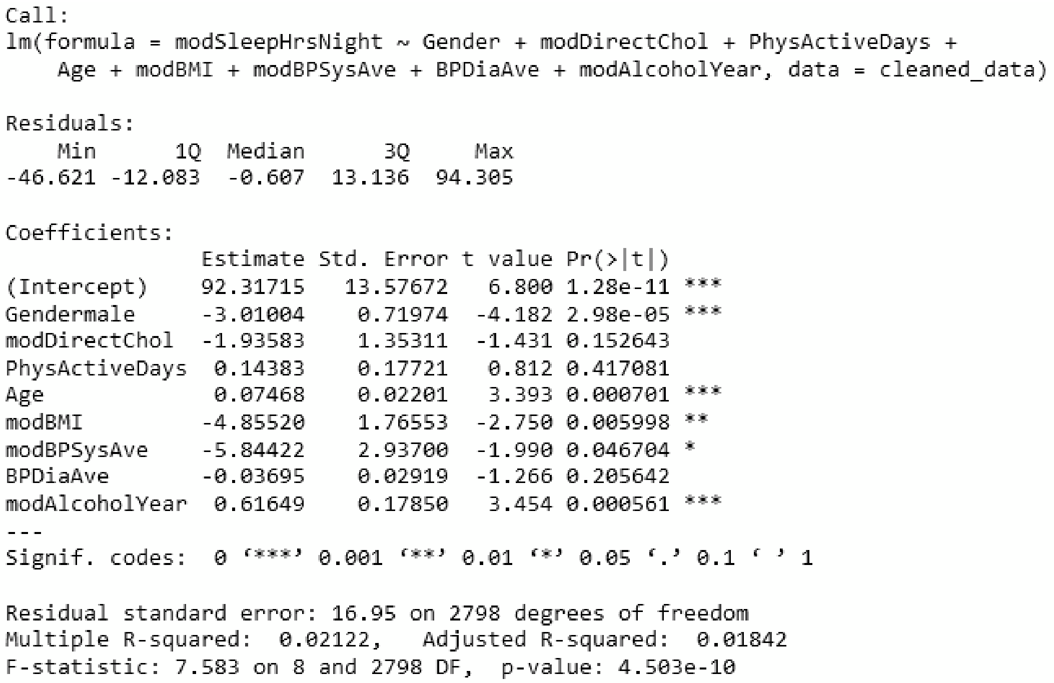
```
Call:
lm(formula = modSleepHrsNight ~ Gender + Age + modBMI + modBPSysAve +
    modAlcoholYear, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-46.050 -12.099  -0.566  13.217  93.526

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    94.58151   12.82982   7.372 2.20e-13 ***
Gendermale     -2.63341    0.66083  -3.985 6.92e-05 ***
Age             0.07424    0.02139   3.470 0.000528 ***
modBMI         -4.14704    1.60195  -2.589 0.009683 **
modBPSysAve    -7.34425    2.72175  -2.698 0.007010 **
modAlcoholYear  0.53673    0.17241   3.113 0.001871 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.95 on 2801 degrees of freedom
Multiple R-squared:  0.0197,    Adjusted R-squared:  0.01795
F-statistic: 11.26 on 5 and 2801 DF,  p-value: 9.081e-11
```

Fig11: Confidence Intervals

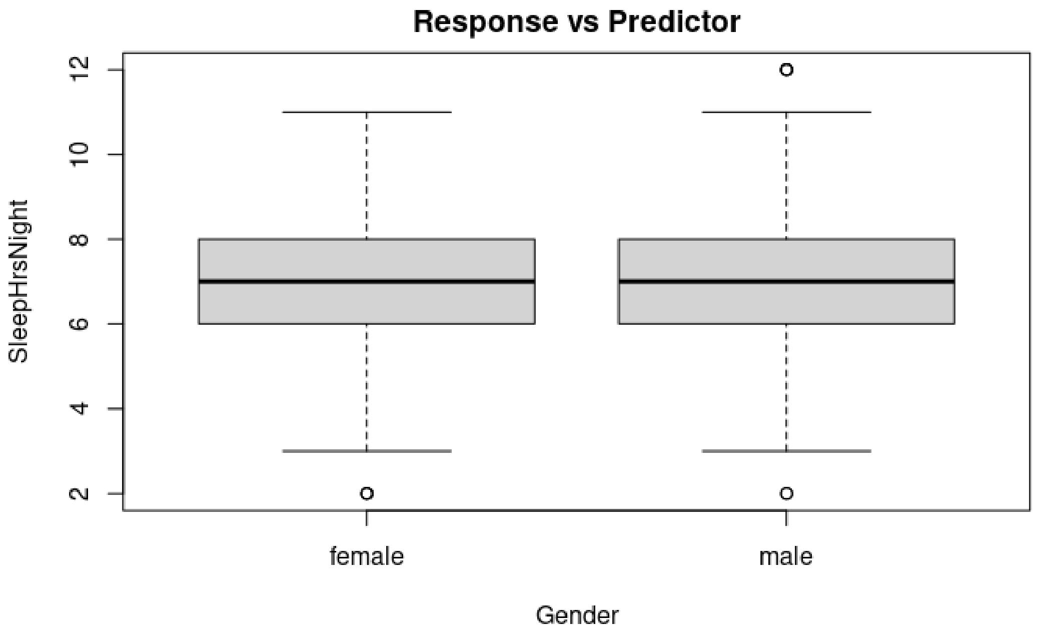


```
                      2.5 %      97.5 %
(Intercept)     69.42466387 119.7383569
Gendermale      -3.92917467  -1.3376492
Age              0.03229378   0.1161906
modBMI          -7.28816313  -1.0059125
modBPSysAve    -12.68109127  -2.0074023
modAlcoholYear   0.19865489   0.8748002
```

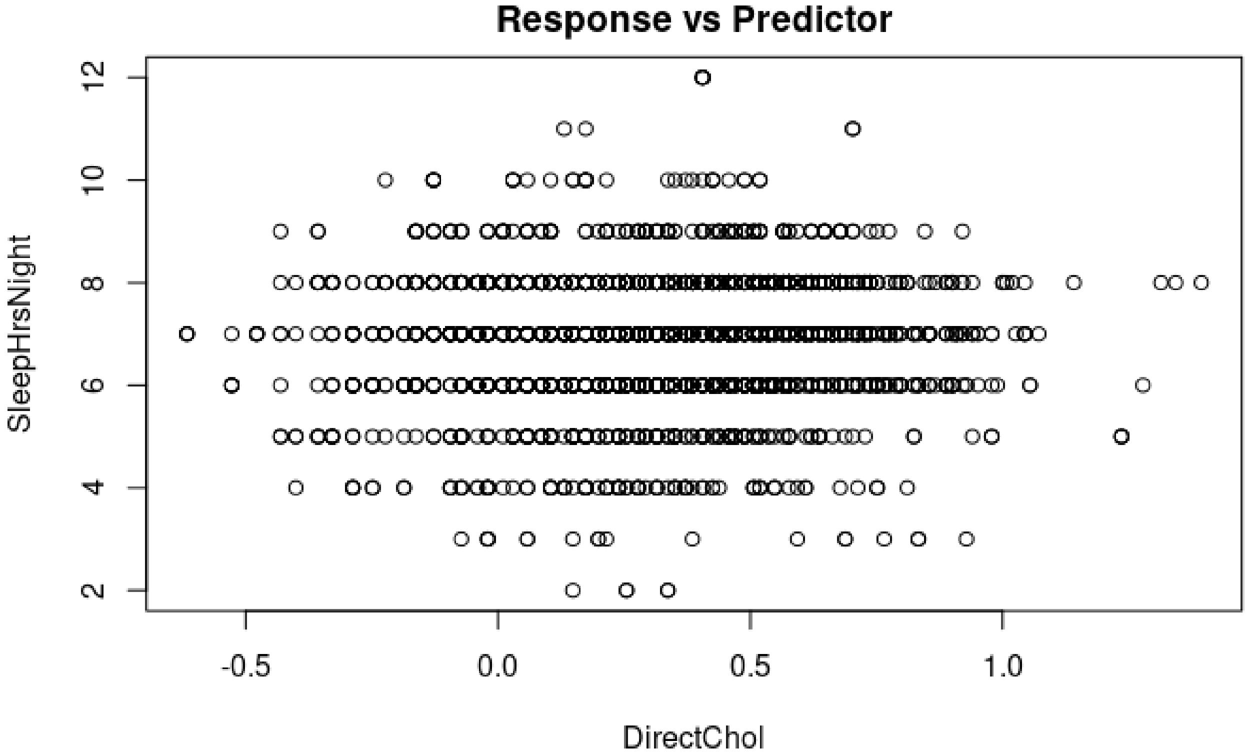
Intermediate Model

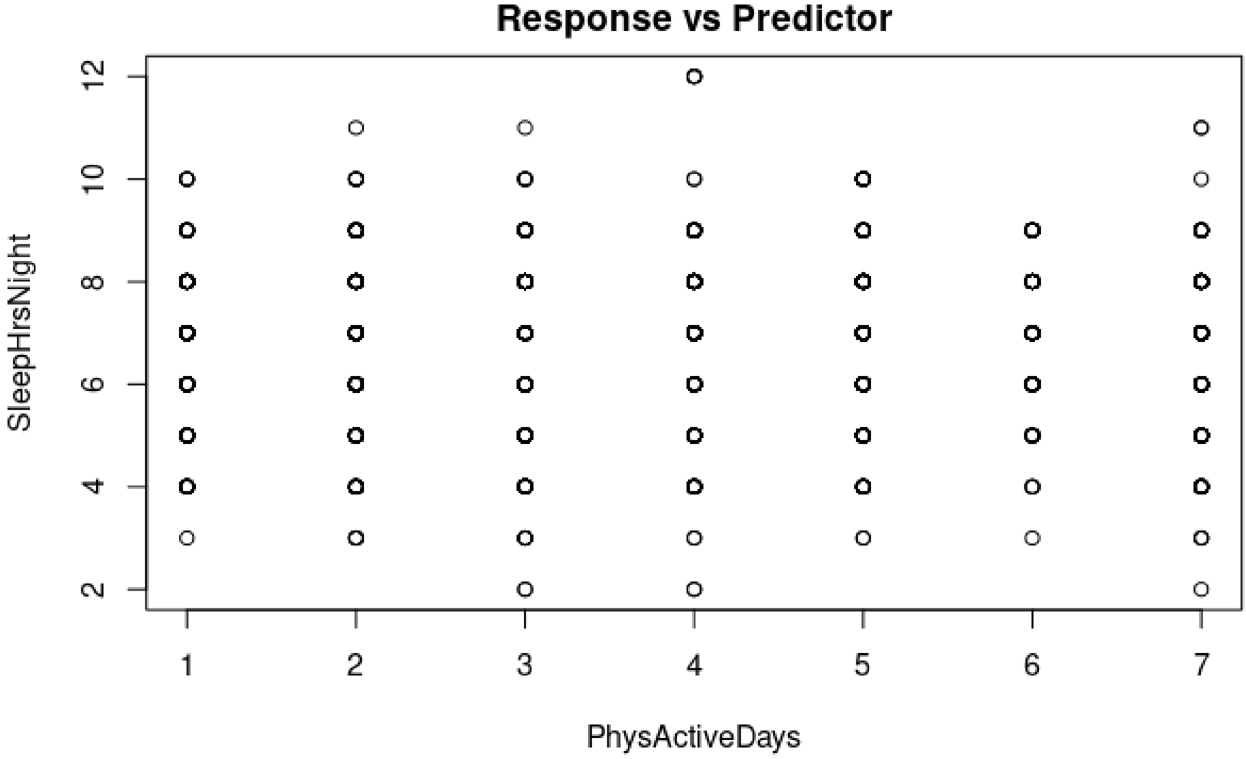


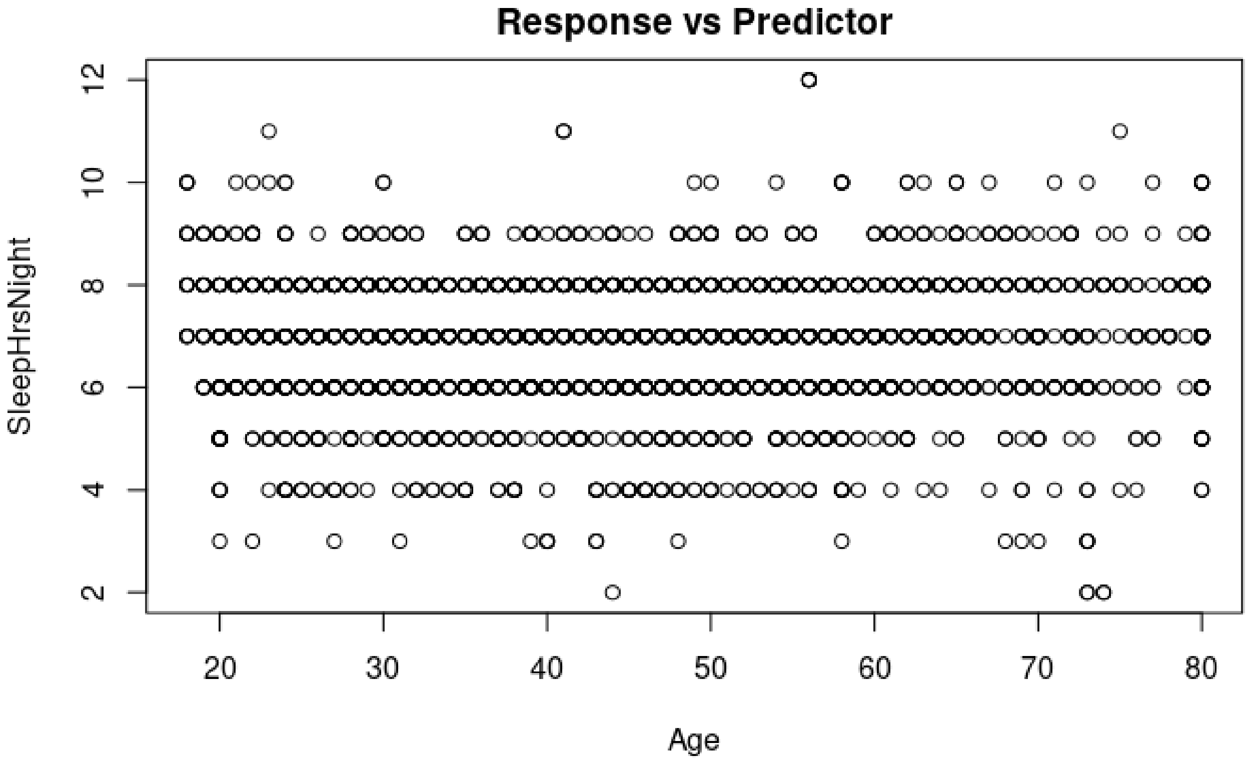
```
Call:
lm(formula = modSleepHrsNight ~ Gender + modDirectChol + PhysActiveDays +
    Age + modBMI + modBPSysAve + BPDiaAve + modAlcoholYear, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-46.621 -12.083  -0.607  13.136  94.305

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    92.31715   13.57672   6.800 1.28e-11 ***
Gendermale     -3.01004    0.71974  -4.182 2.98e-05 ***
modDirectChol  -1.93583    1.35311  -1.431 0.152643
PhysActiveDays  0.14383    0.17721   0.812 0.417081
Age             0.07468    0.02201   3.393 0.000701 ***
modBMI         -4.85520    1.76553  -2.750 0.005998 **
modBPSysAve    -5.84422    2.93700  -1.990 0.046704 *
BPDiaAve       -0.03695    0.02919  -1.266 0.205642
modAlcoholYear  0.61649    0.17850   3.454 0.000561 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.95 on 2798 degrees of freedom
Multiple R-squared:  0.02122,	Adjusted R-squared:  0.01842
F-statistic: 7.583 on 8 and 2798 DF,  p-value: 4.503e-10
```

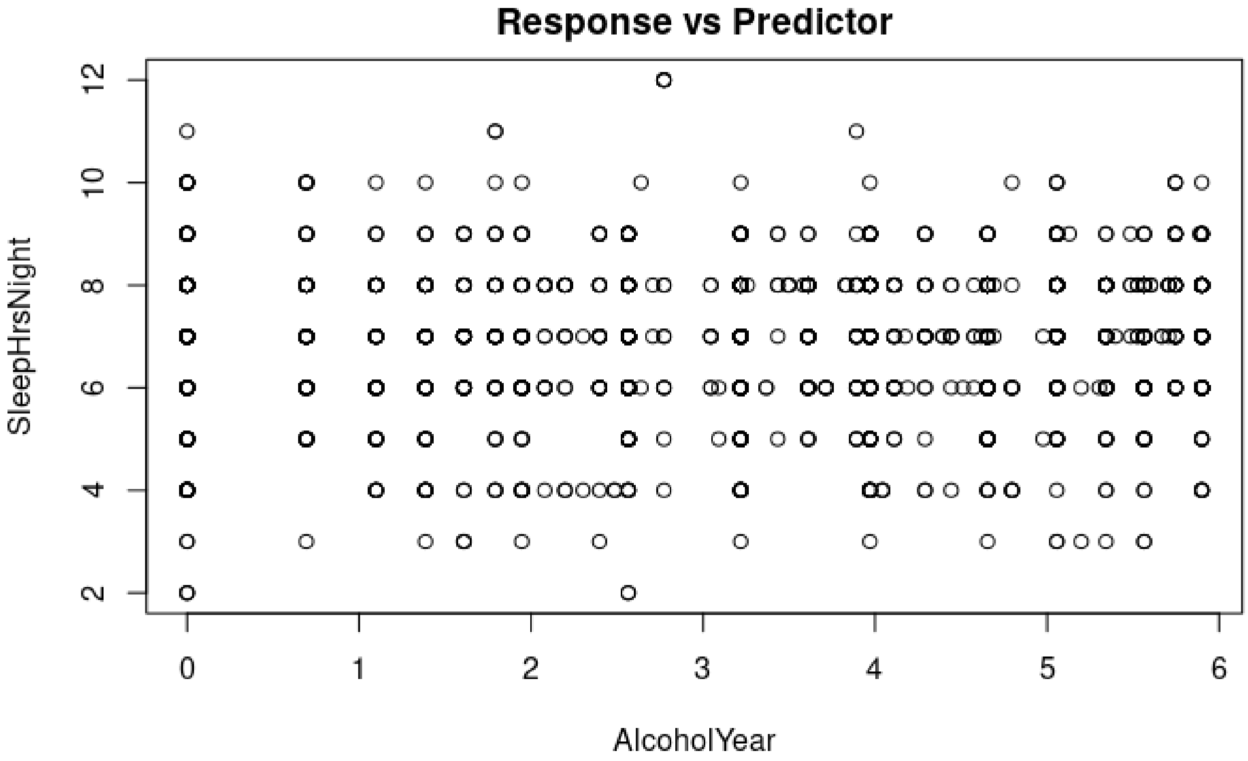


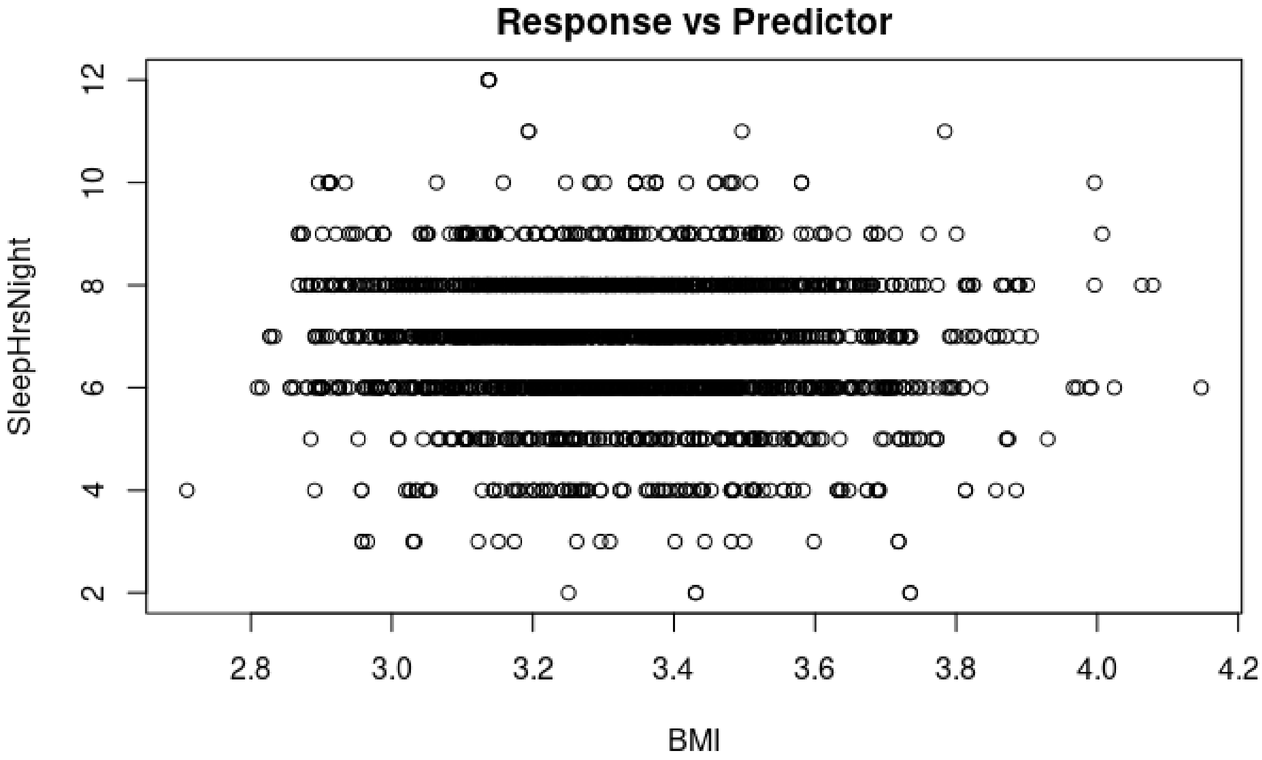
Response vs Predictor Plots

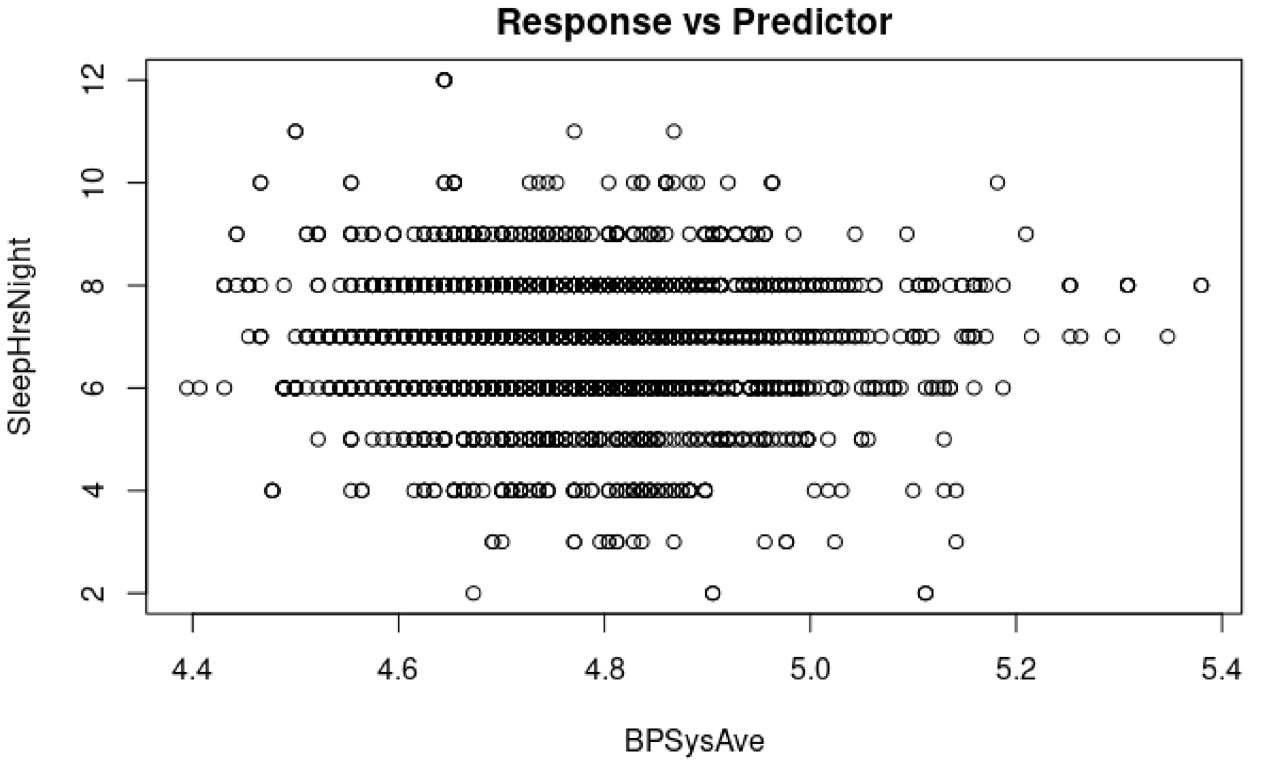


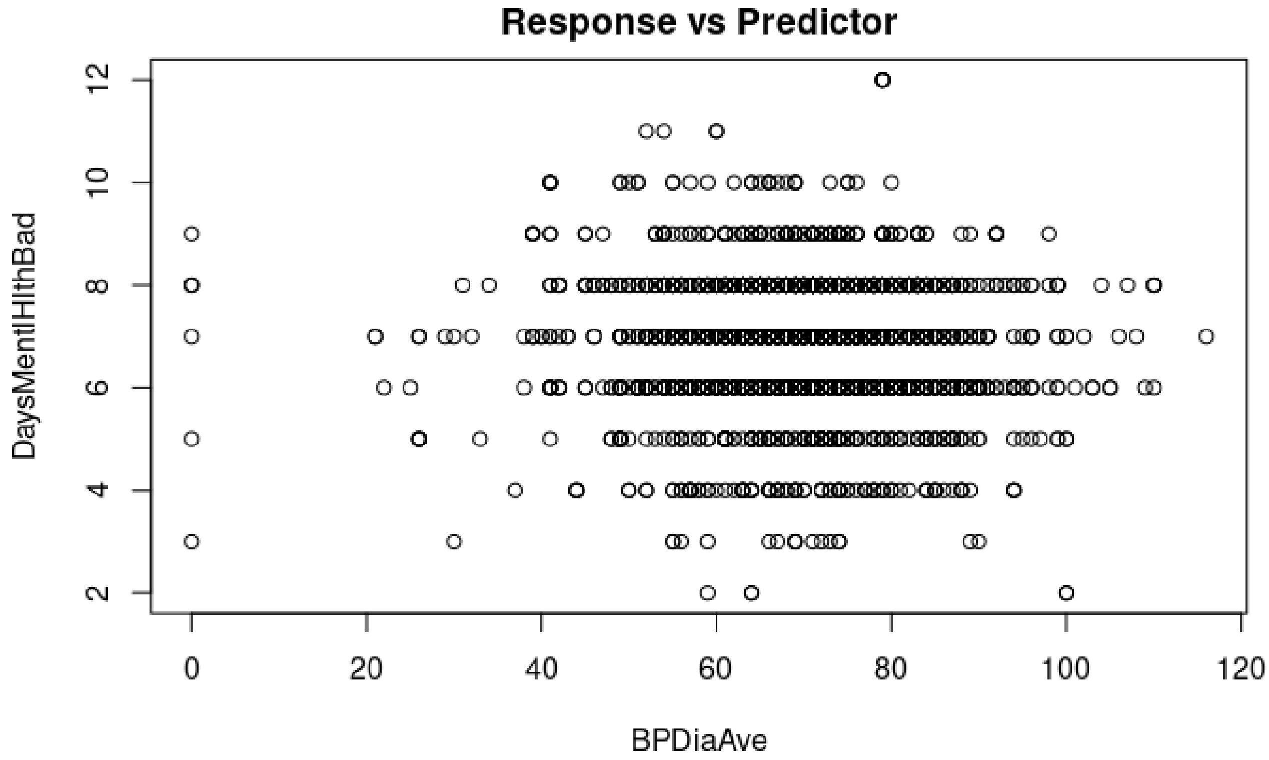



Response vs Predictor
SleepHrsNight
2
4
6
8
10
12
20
30
40
50
60
70
80
Age

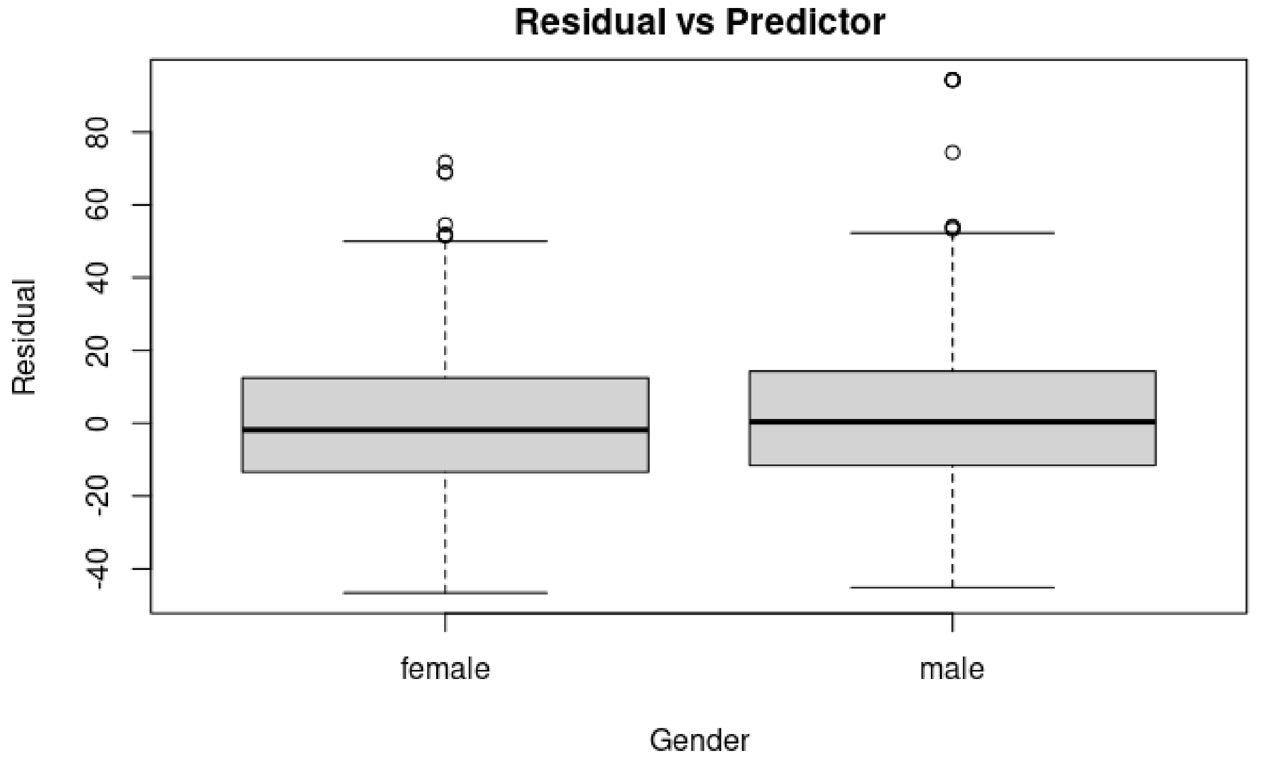

Response vs Predictor
SleepHrsNight
2
4
6
8
10
12
0
1
2
3
4
5
6
AlcoholYear

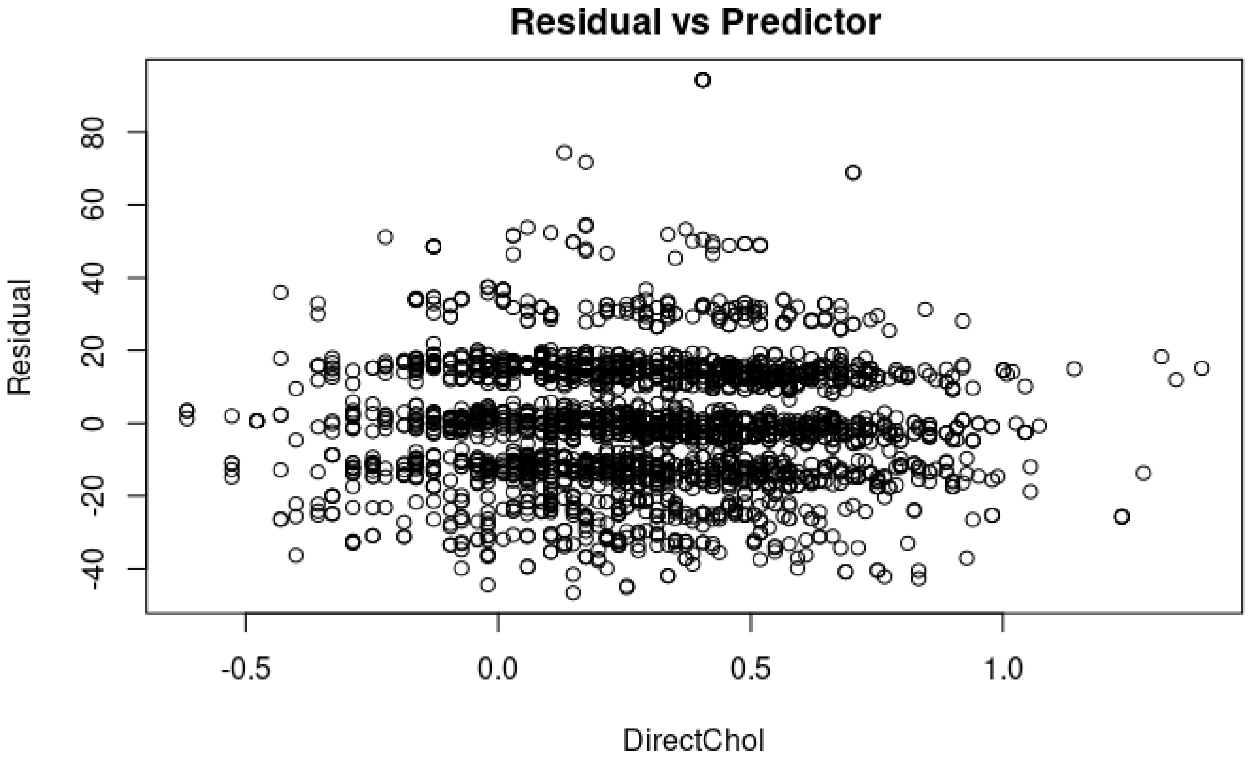

Response vs Predictor
SleepHrsNight
2
4
6
8
10
12
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
BMI

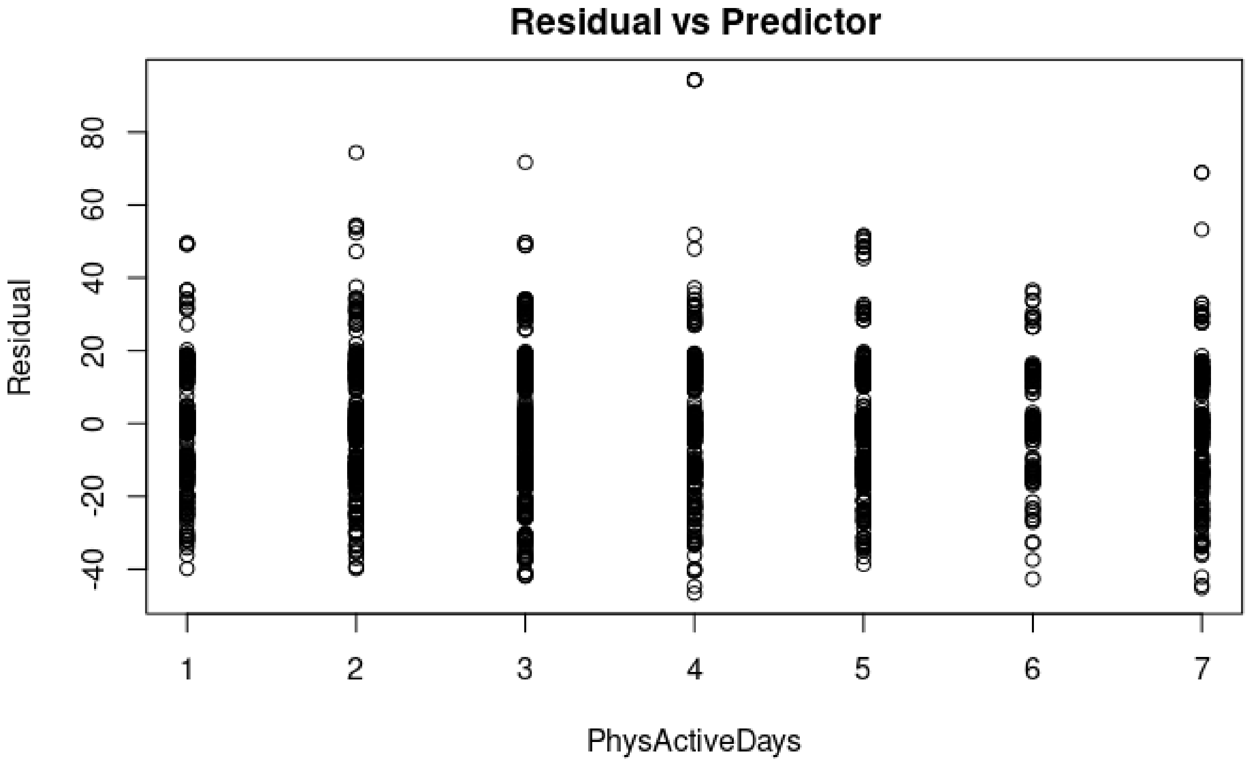

Response vs Predictor
SleepHrsNight
2
4
6
8
10
12
4.4
4.6
4.8
5.0
5.2
5.4
BPSysAve

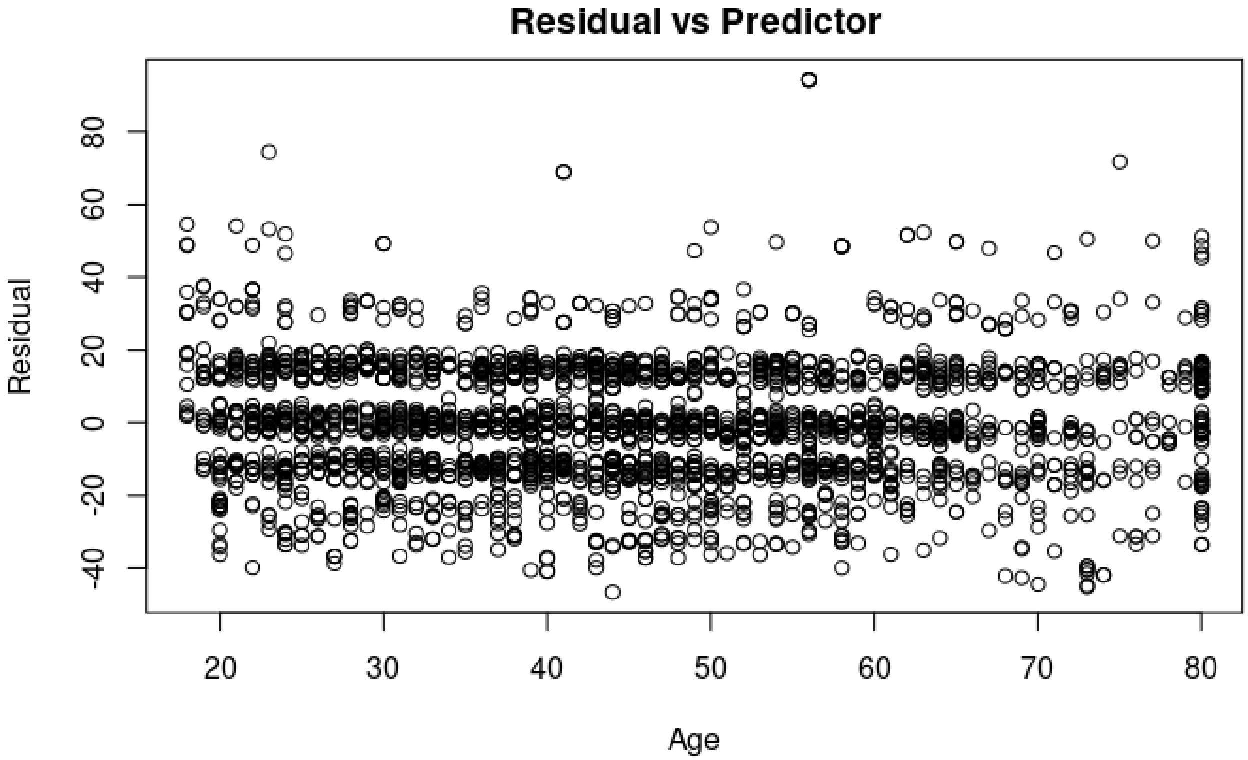


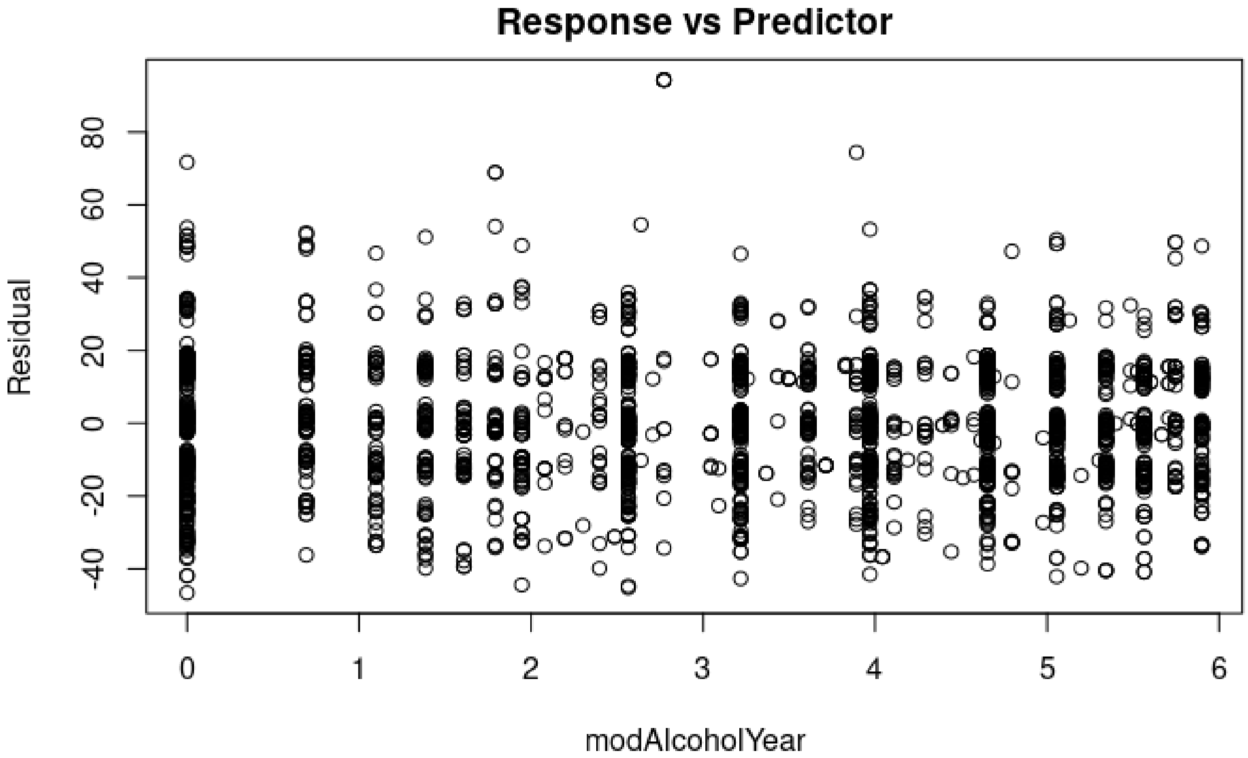
Residual vs Predictor

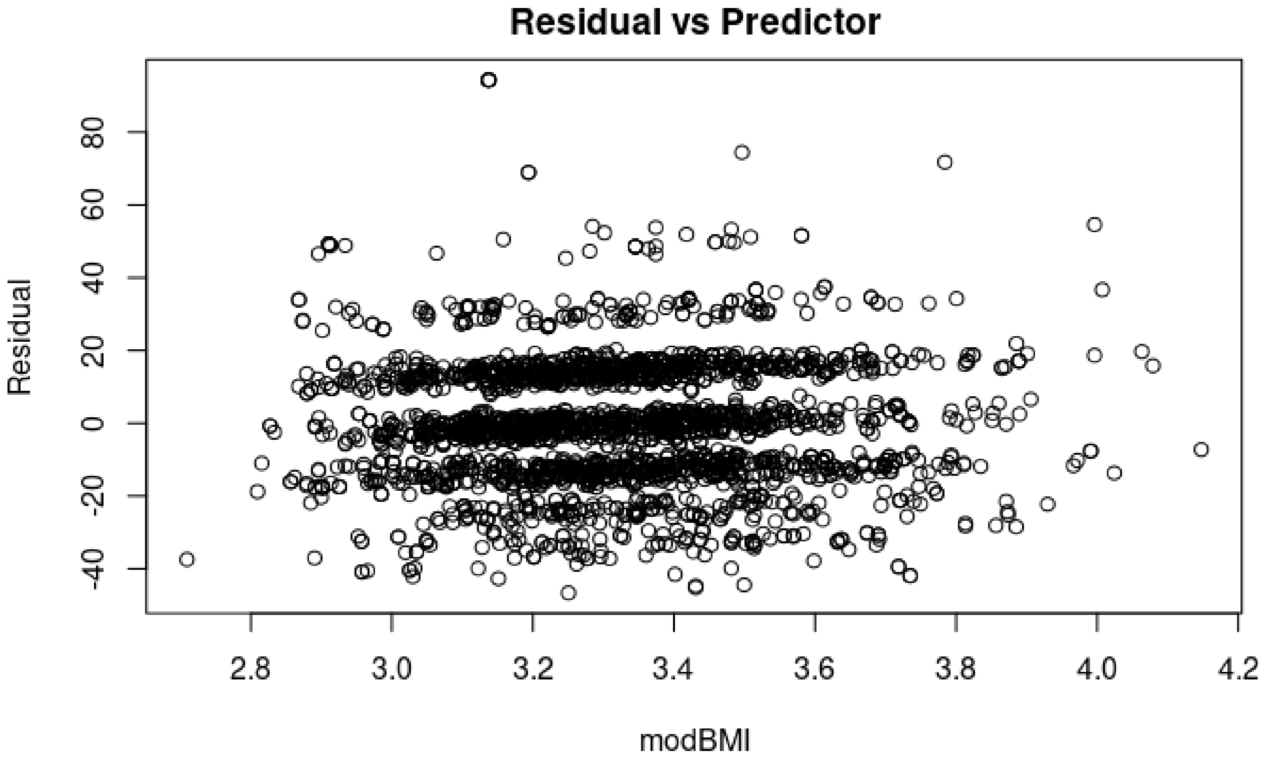


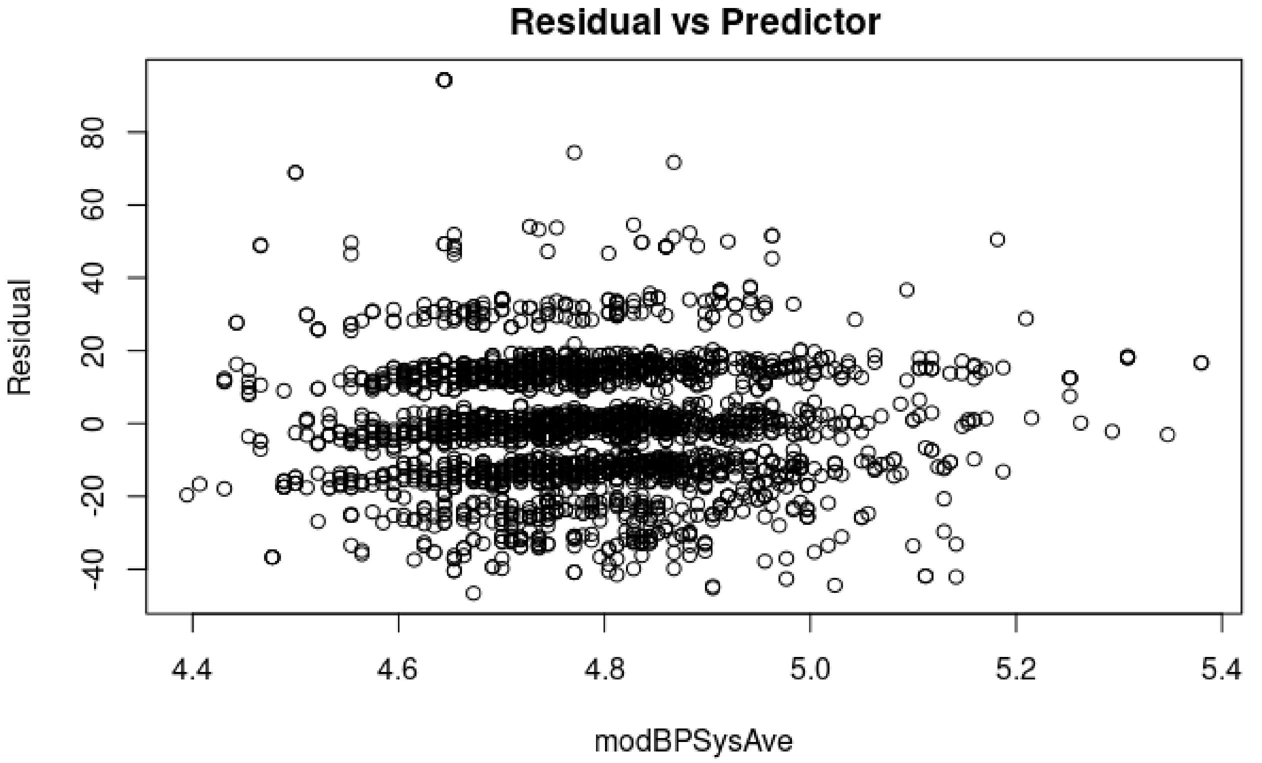

Residual vs Predictor
Residual
80
60
40
20
0
-20
-40
-0.5
0.0
0.5
1.0
DirectChol

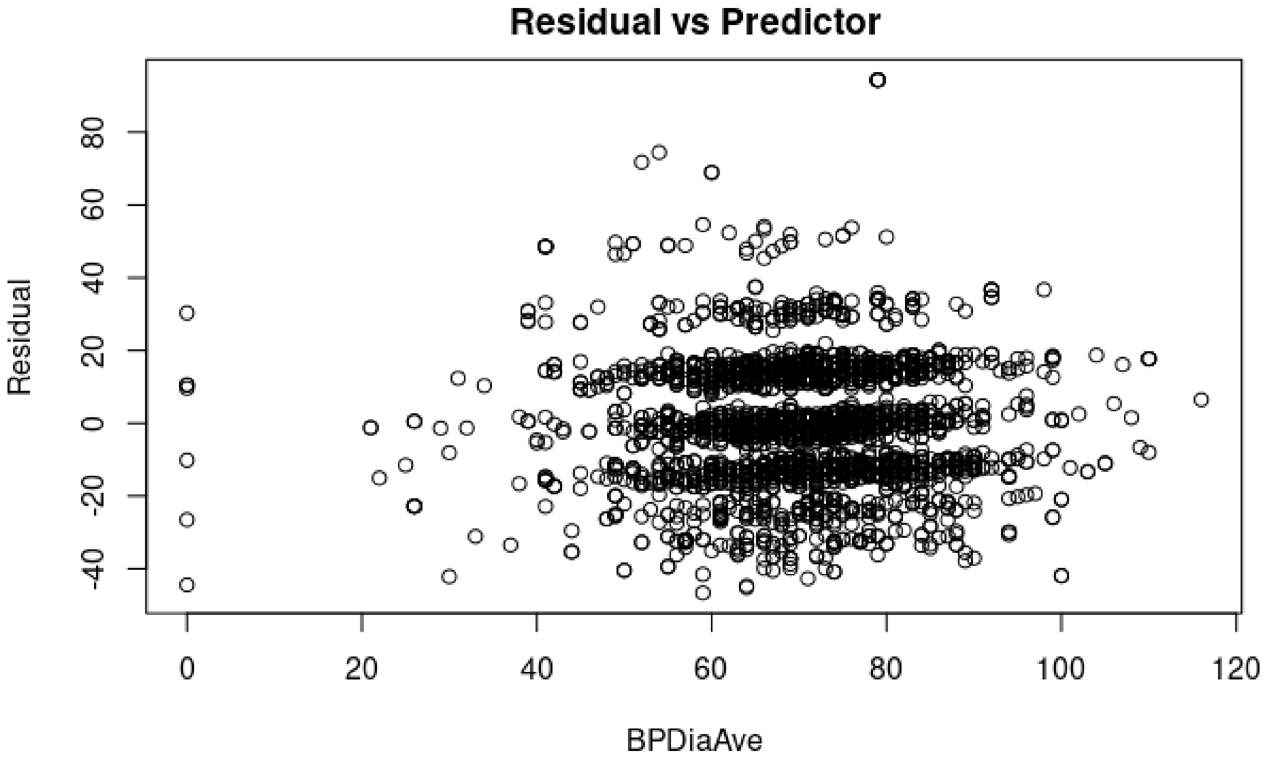

Residual vs Predictor
Residual
80
60
40
20
0
-20
-40
1
2
3
4
5
6
7
PhysActiveDays


Residual vs Predictor
Residual
80
60
40
20
0
-20
-40
20
30
40
50
60
70
80
Age

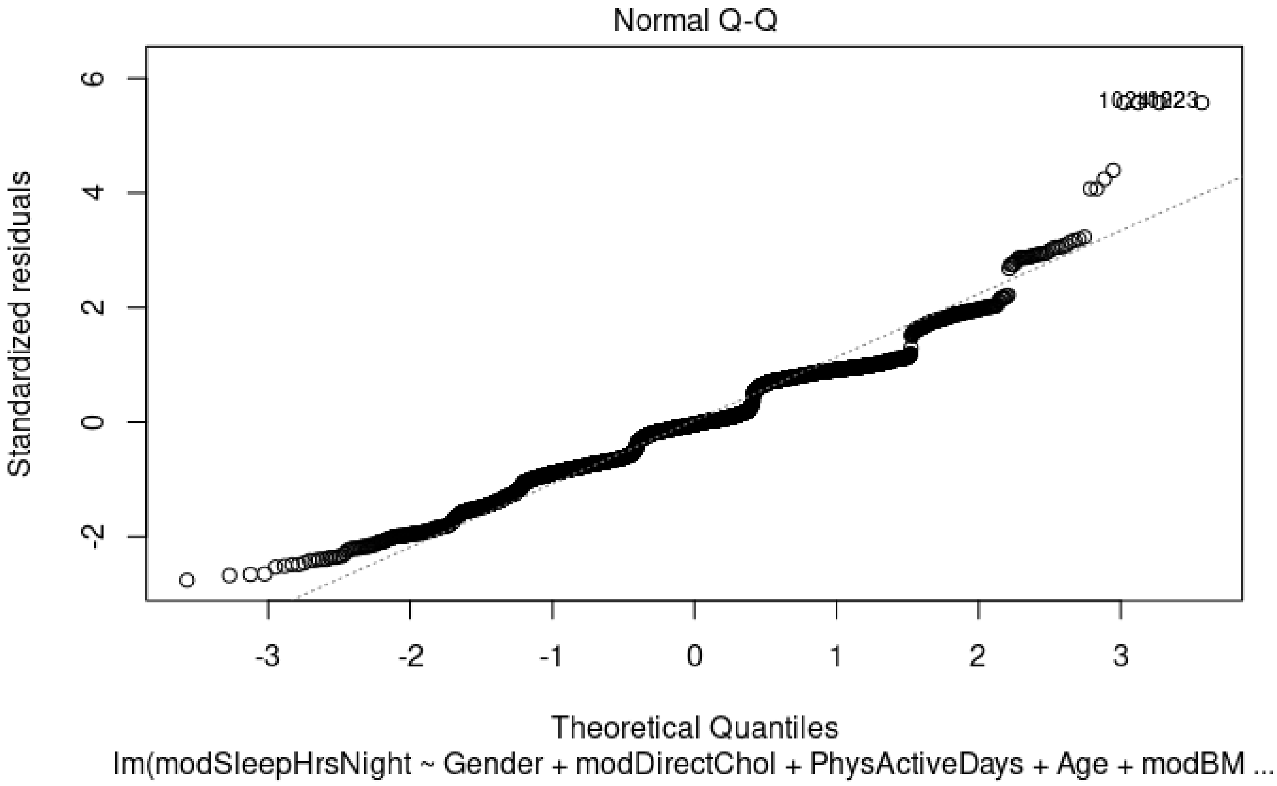

Response vs Predictor
Residual
80
60
40
20
0
-20
-40
0
1
2
3
4
5
6
modAlcoholYear


Residual vs Predictor
Residual
80
60
40
20
0
-20
-40
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
modBMI

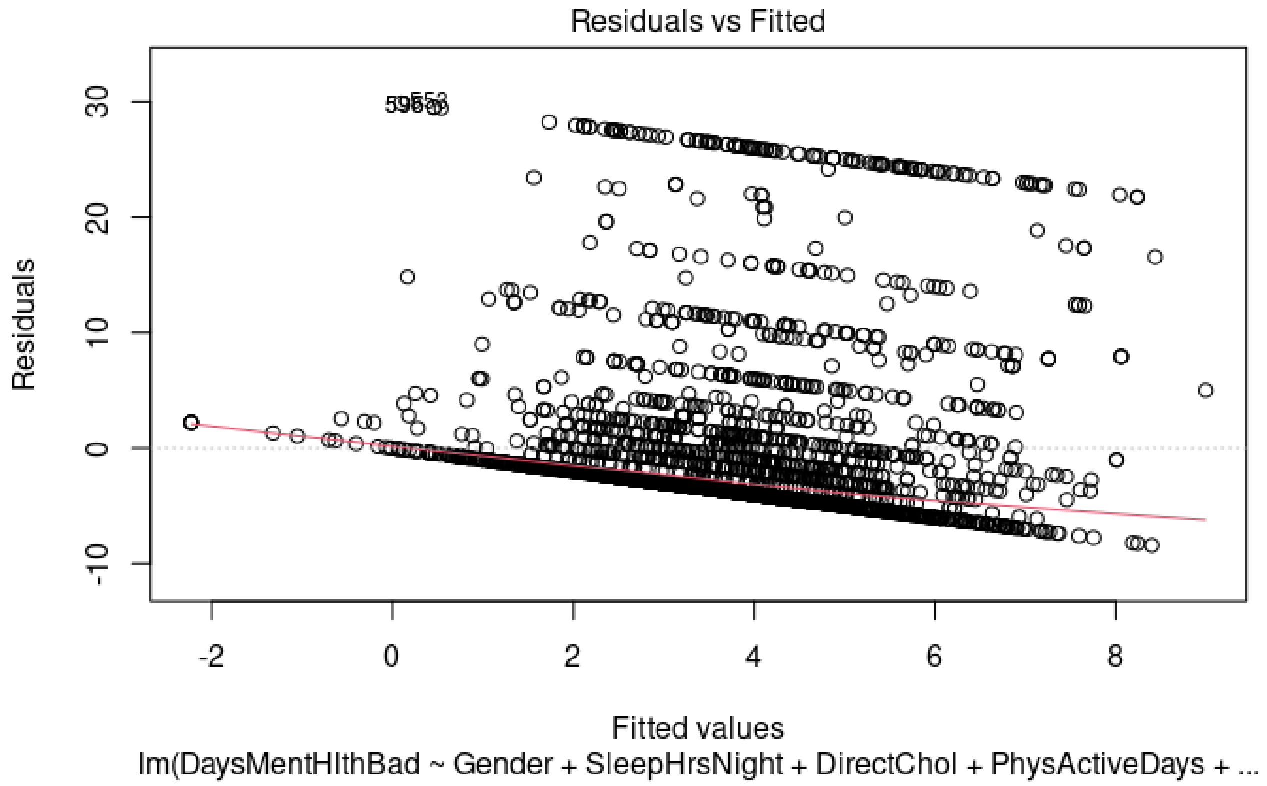

Residual vs Predictor
Residual
80
60
40
20
0
-20
-40
4.4
4.6
4.8
5.0
5.2
5.4
modBPSysAve

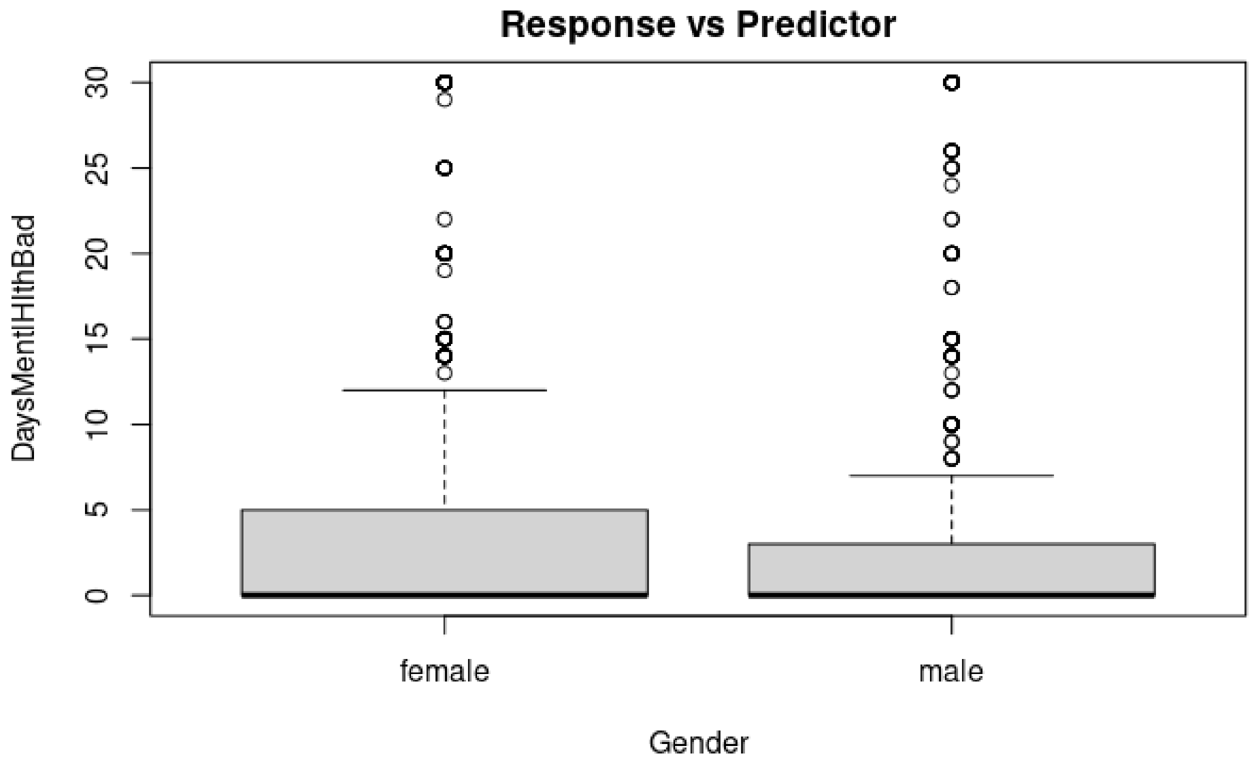


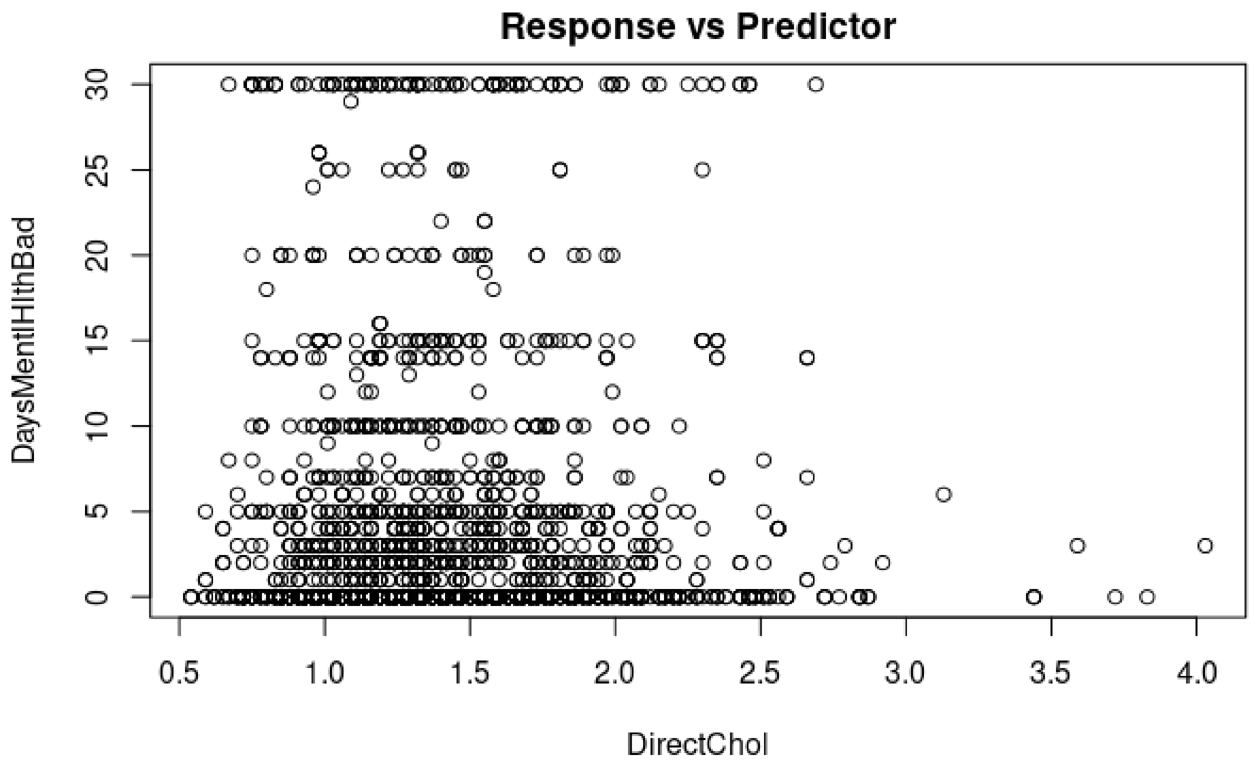
Standardized residuals histogram & Normal QQ plot

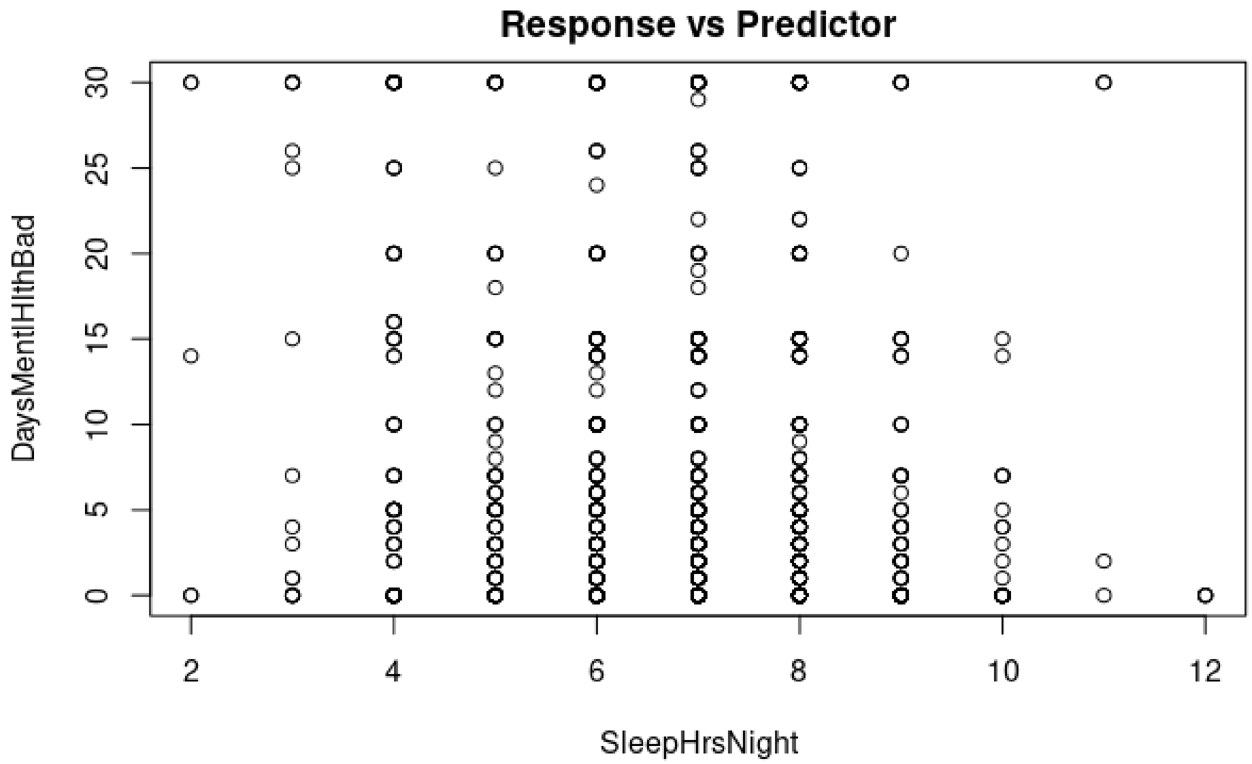


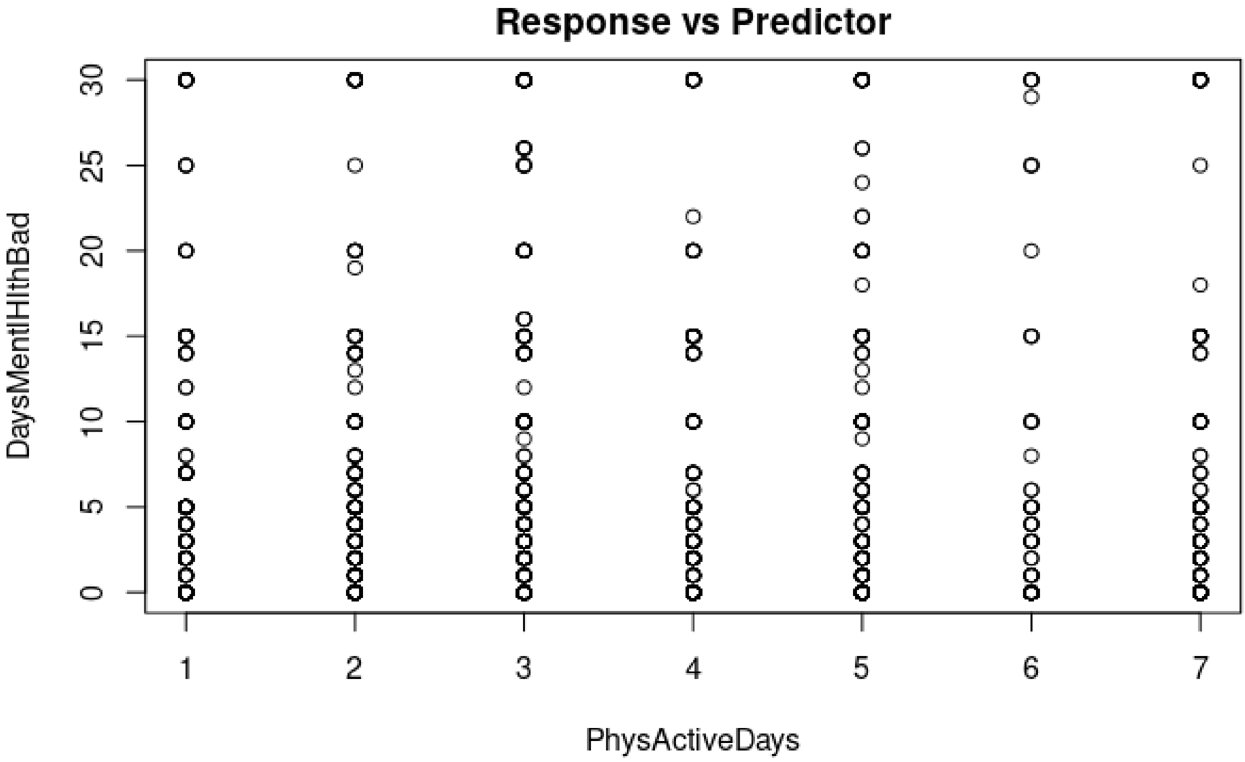
Preliminary Model

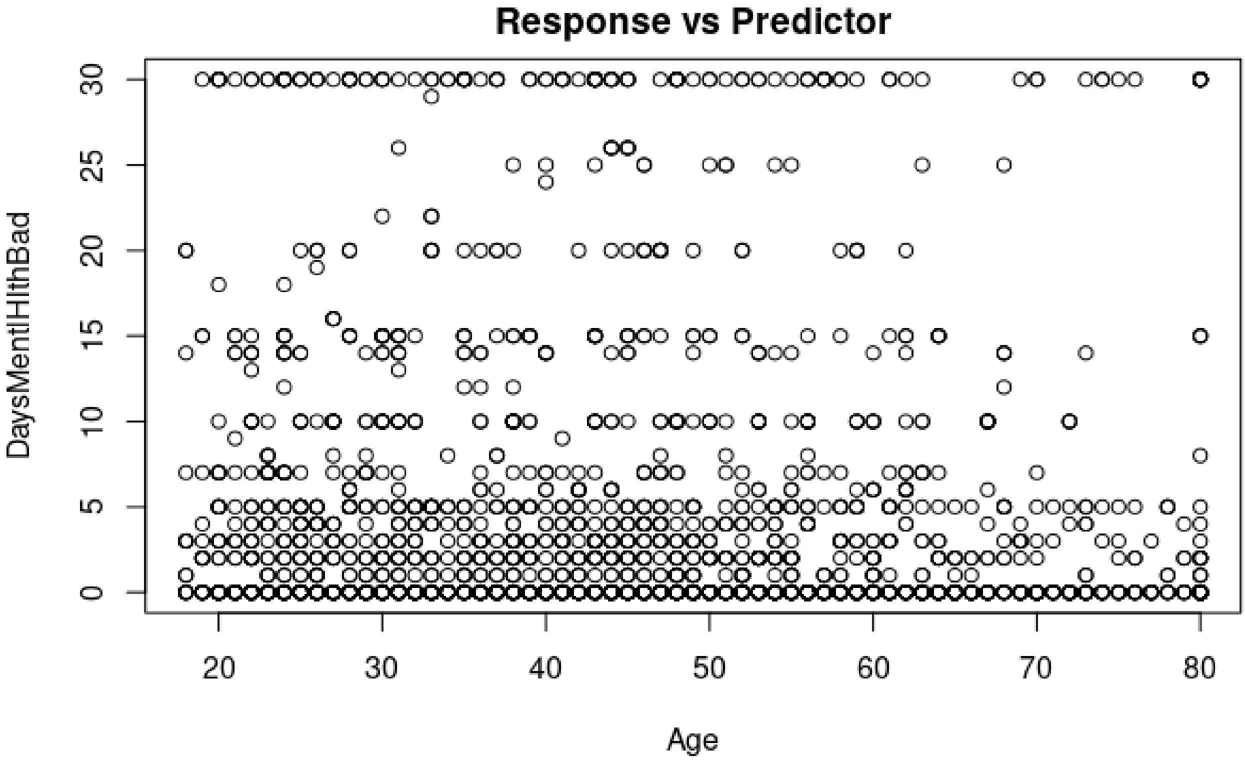


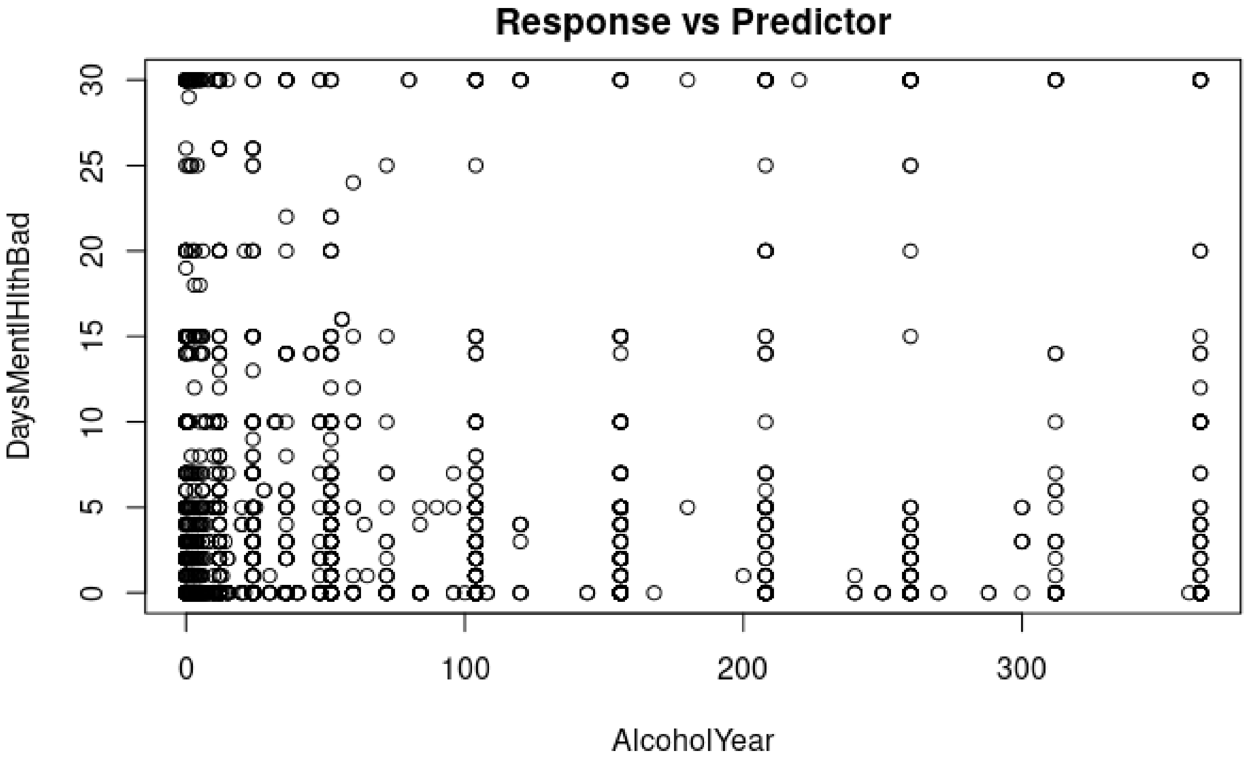


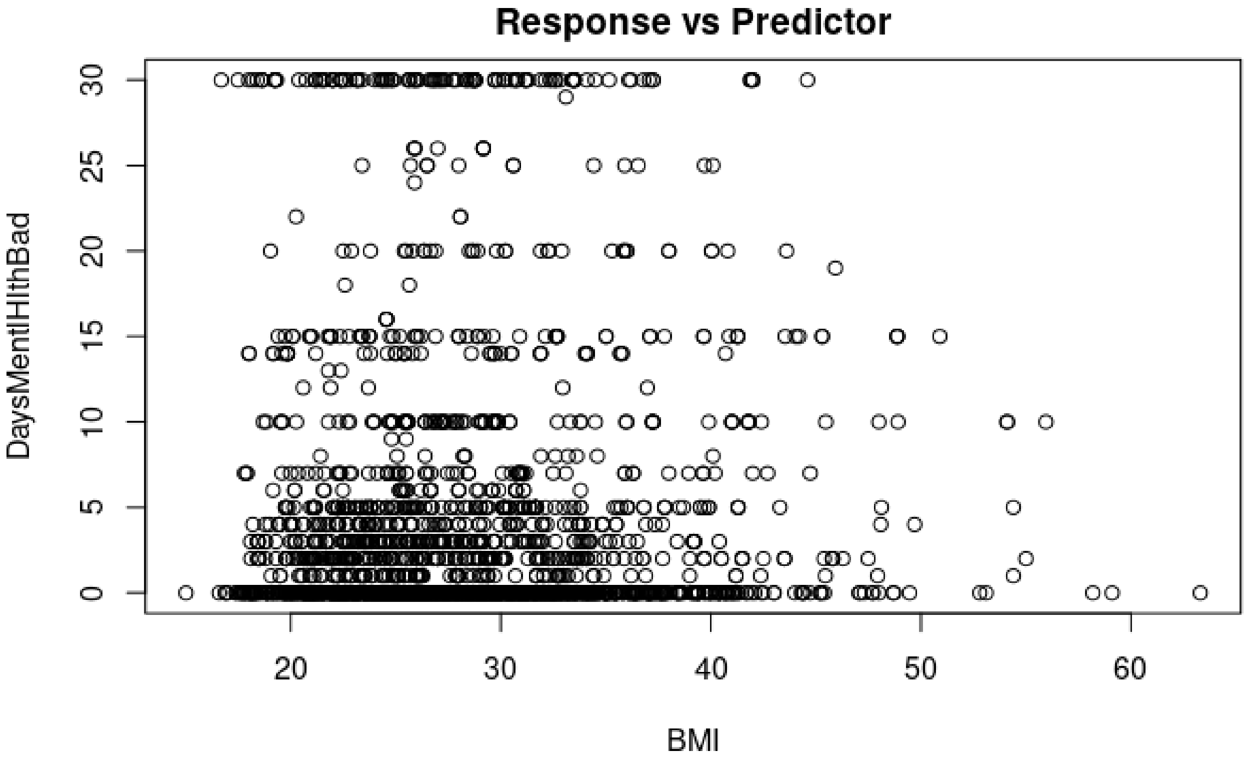

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
DirectChol

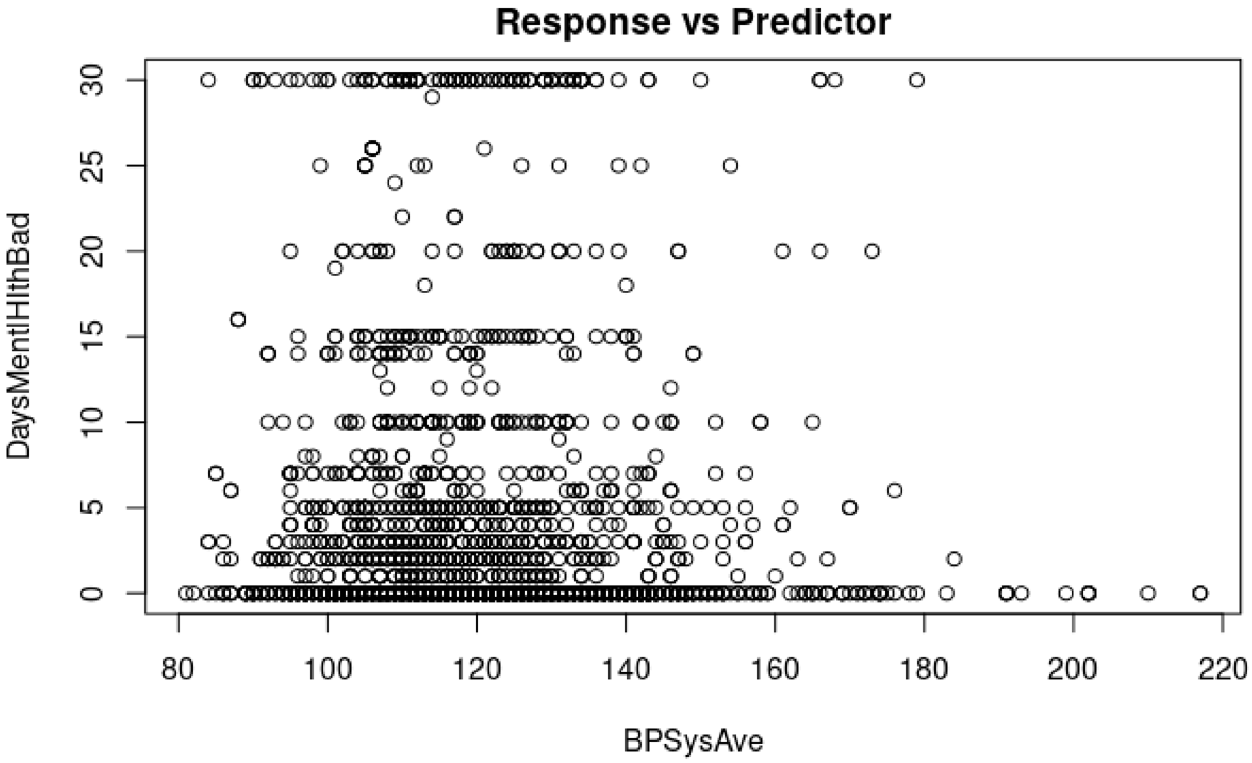

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
2
4
6
8
10
12
SleepHrsNight

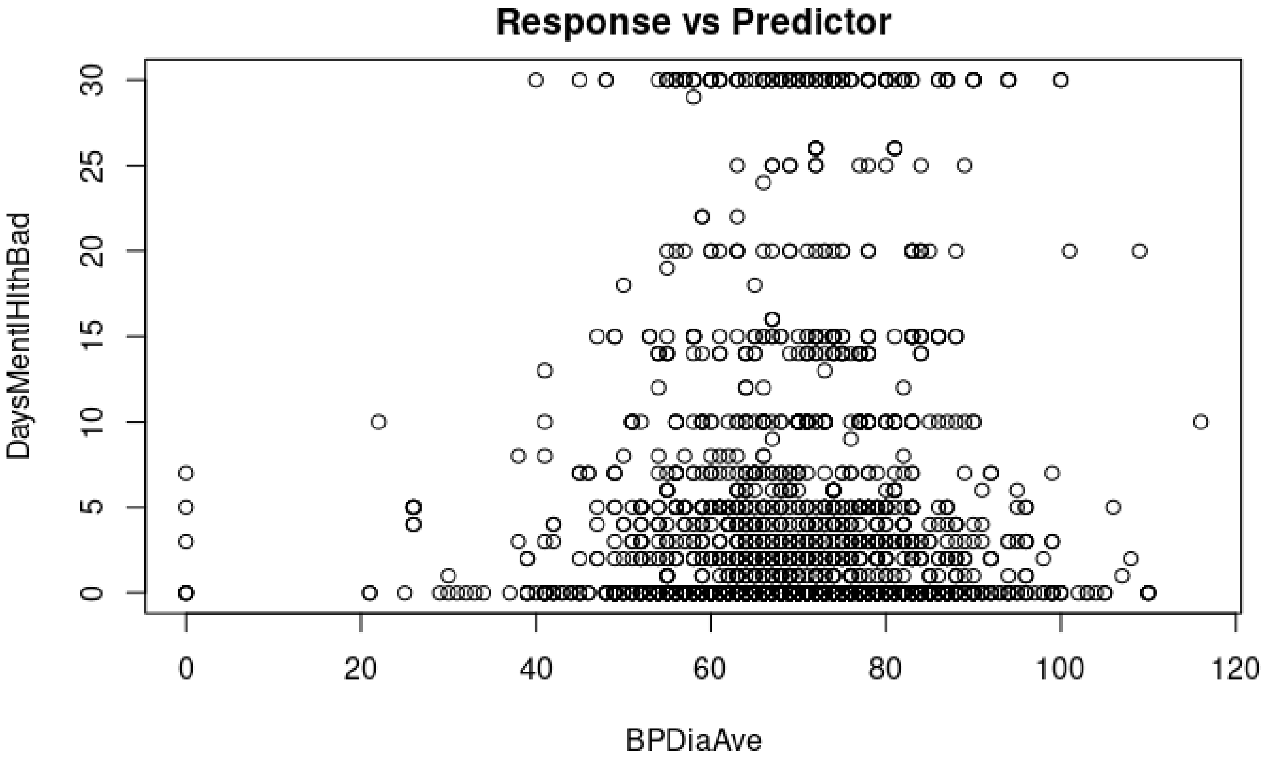

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
1
2
3
4
5
6
7
PhysActiveDays

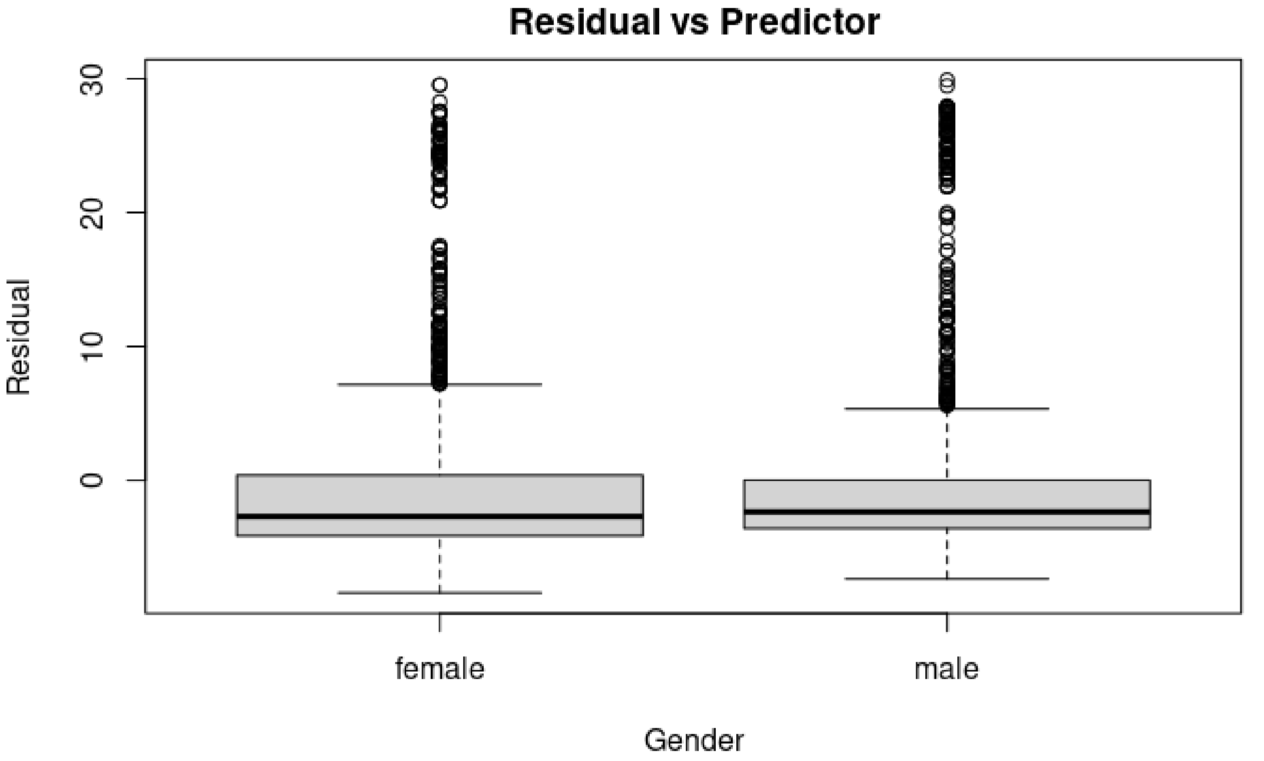

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
20
30
40
50
60
70
80
Age

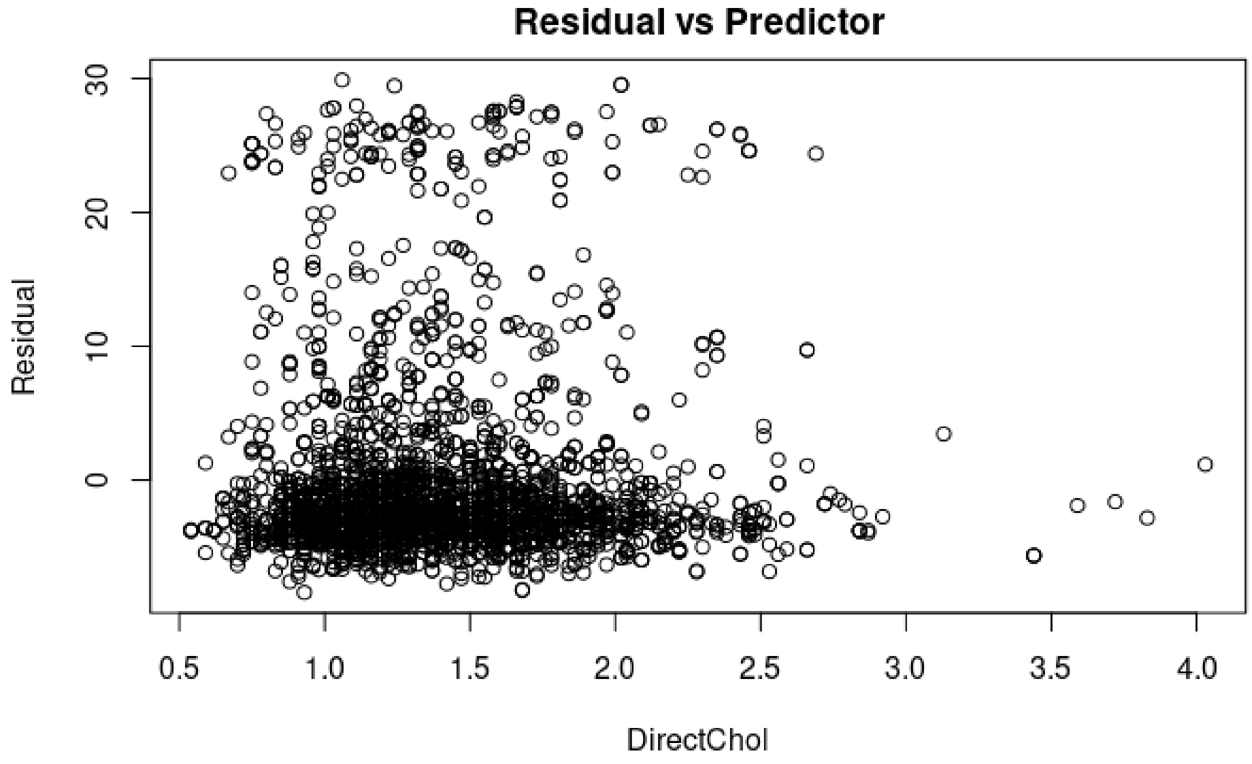

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
AlcoholYear
0
100
200
300

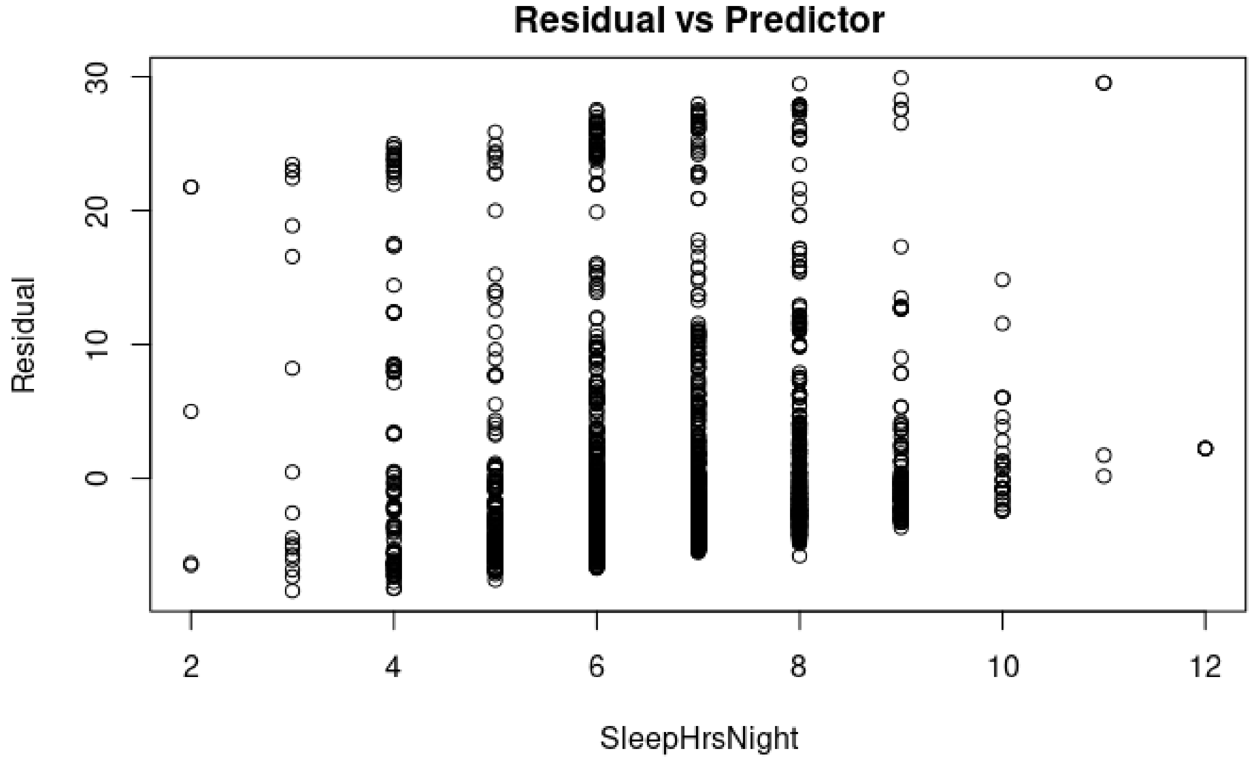

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
20
30
40
50
60
BMI

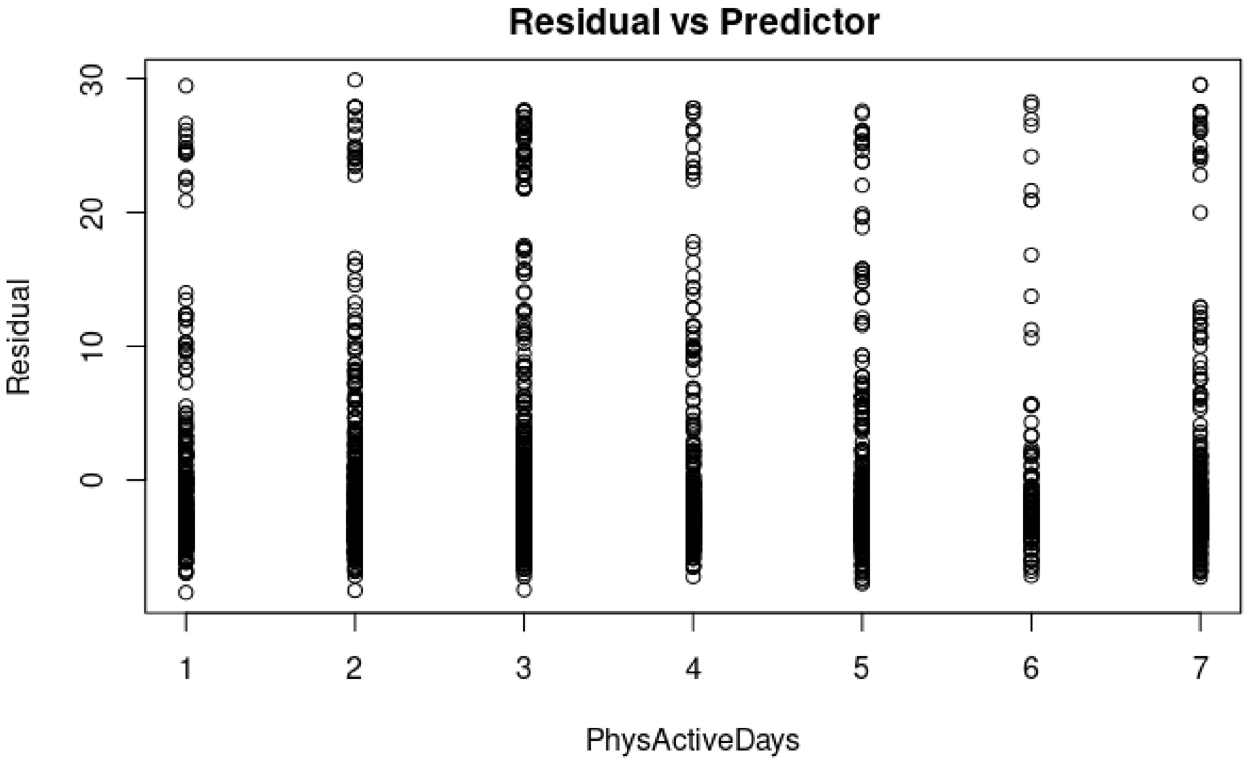

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
80
100
120
140
160
180
200
220
BPSysAve

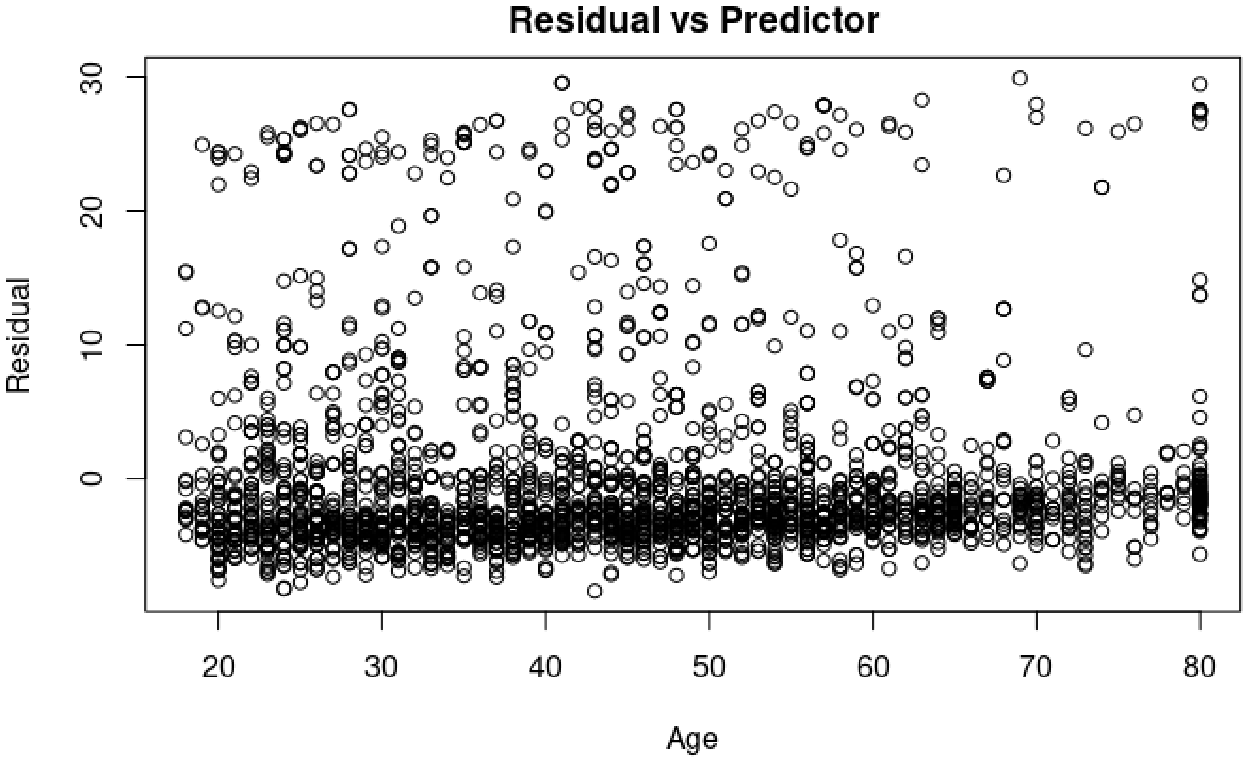

Response vs Predictor
DaysMentHlthBad
0
5
10
15
20
25
30
0
20
40
60
80
100
120
BPDiaAve

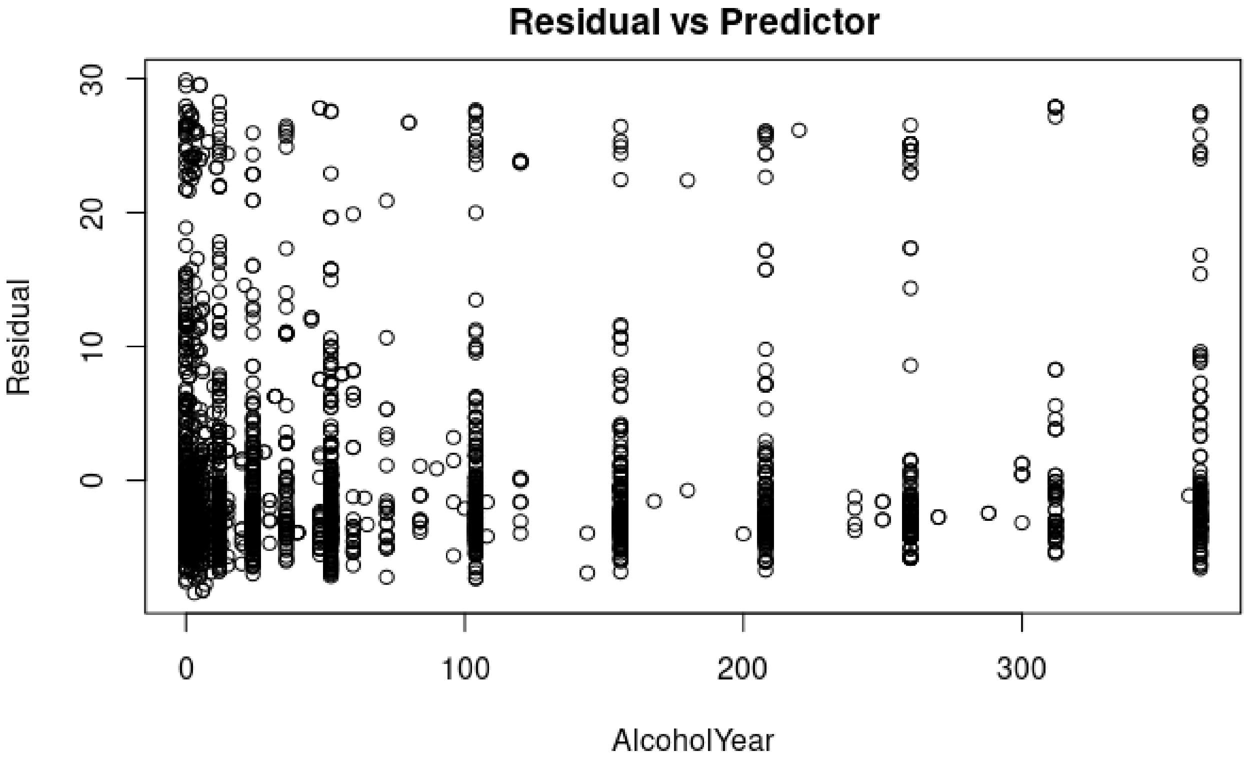

Residual vs Predictor
Residual
0
10
20
30
female
male
Gender

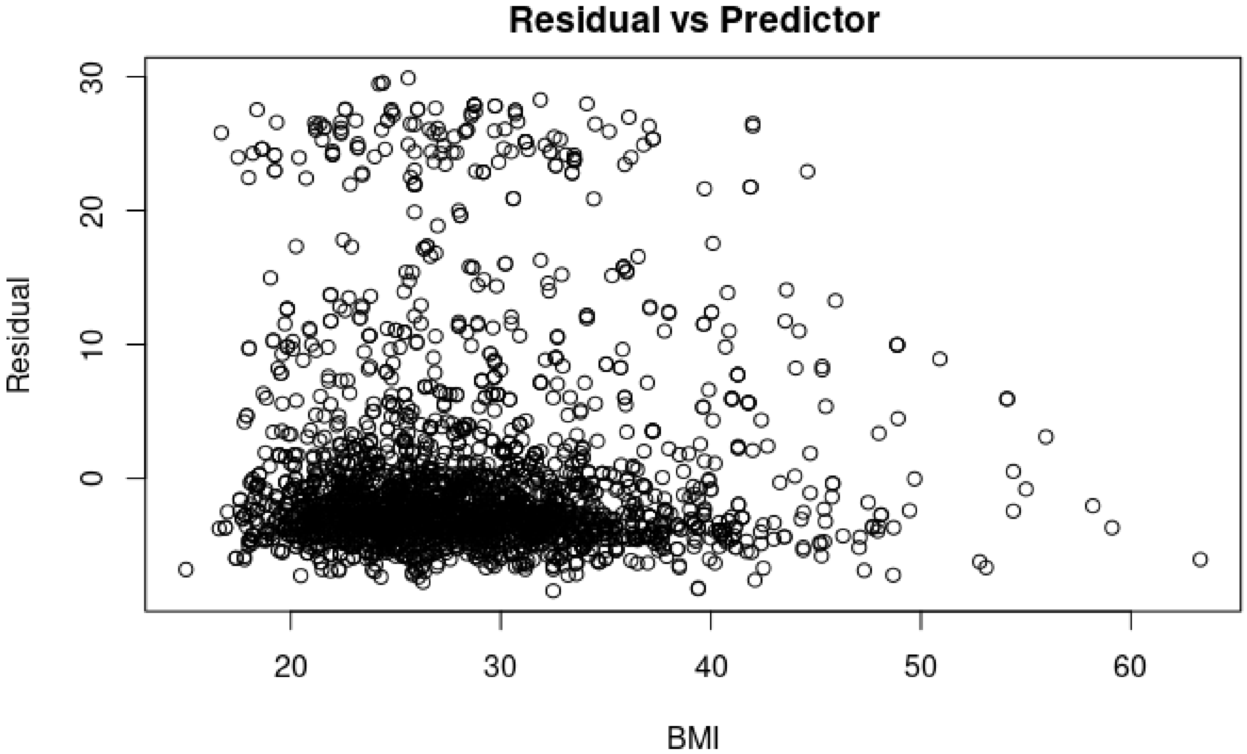

Residual vs Predictor
Residual
30
20
10
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
DirectChol

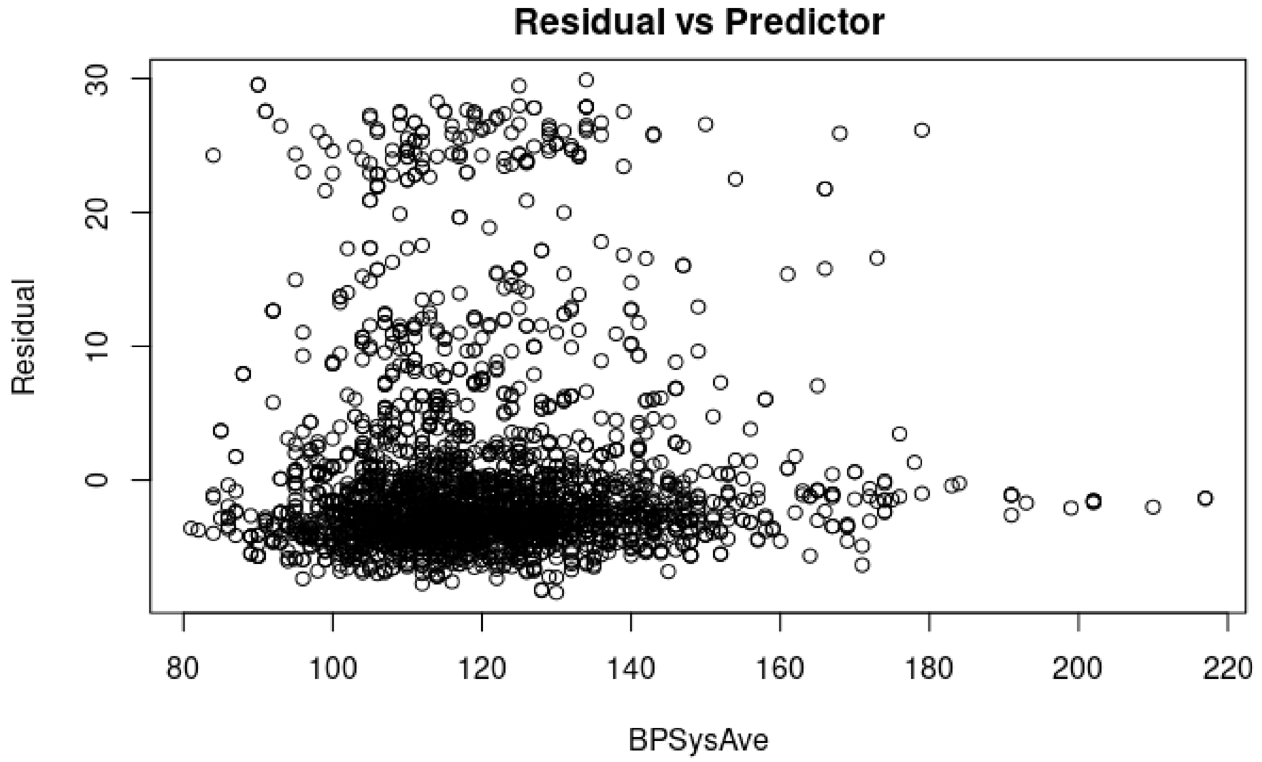

Residual vs Predictor
Residual
30
20
10
0
2
4
6
8
10
12
SleepHrsNight

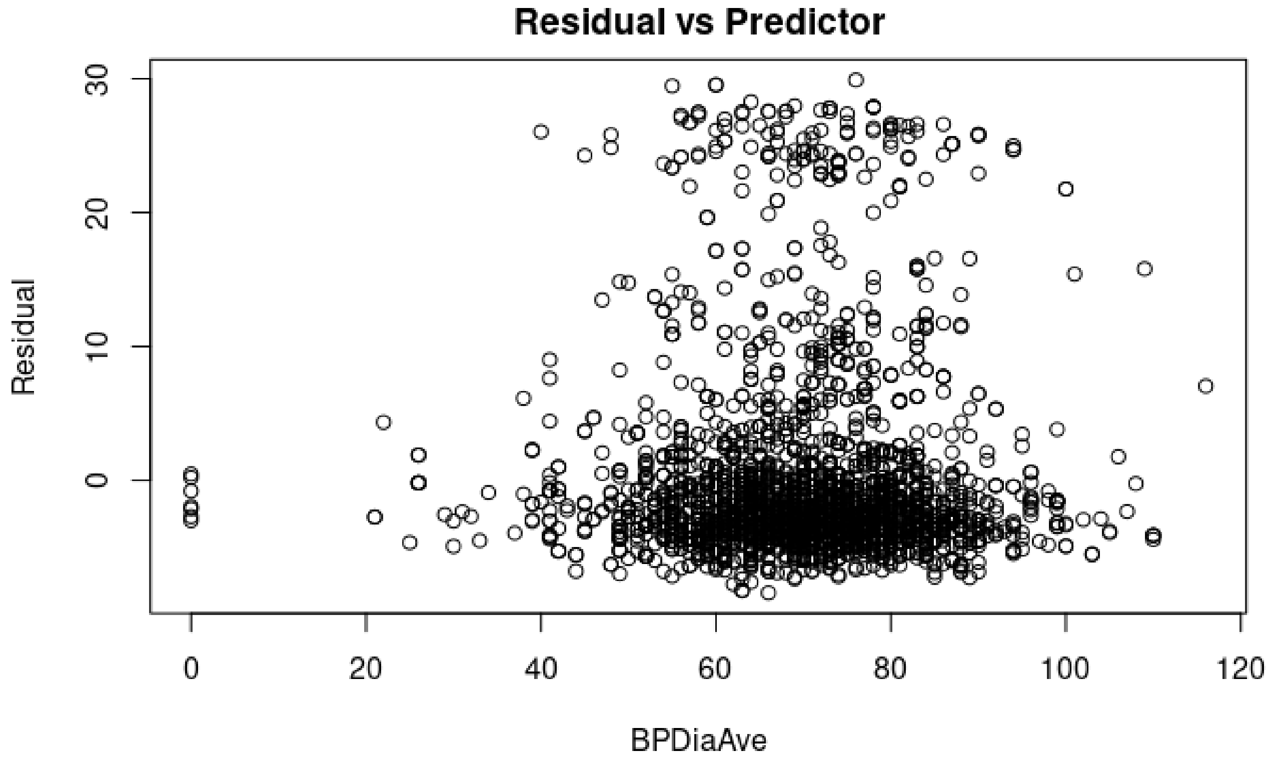

Residual vs Predictor
Residual
30
20
10
0
1
2
3
4
5
6
7
PhysActiveDays

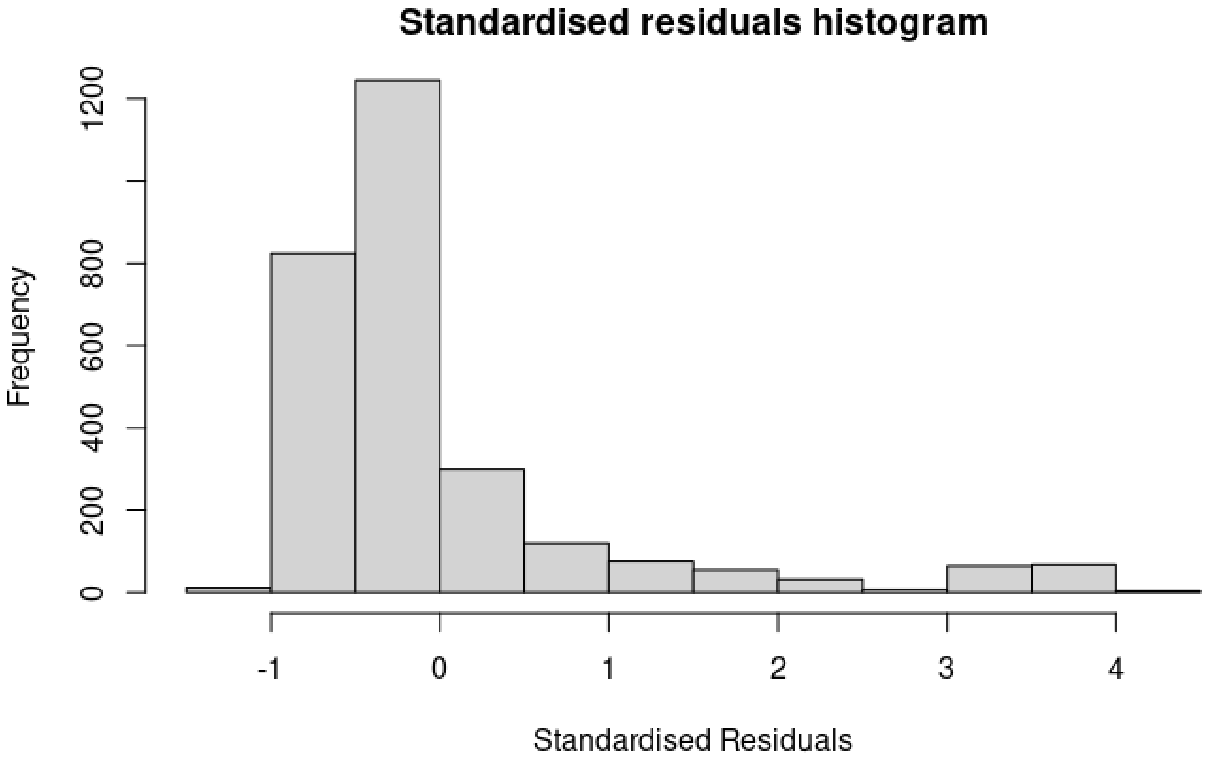

Residual vs Predictor
Residual
30
20
10
0
20
30
40
50
60
70
80
Age

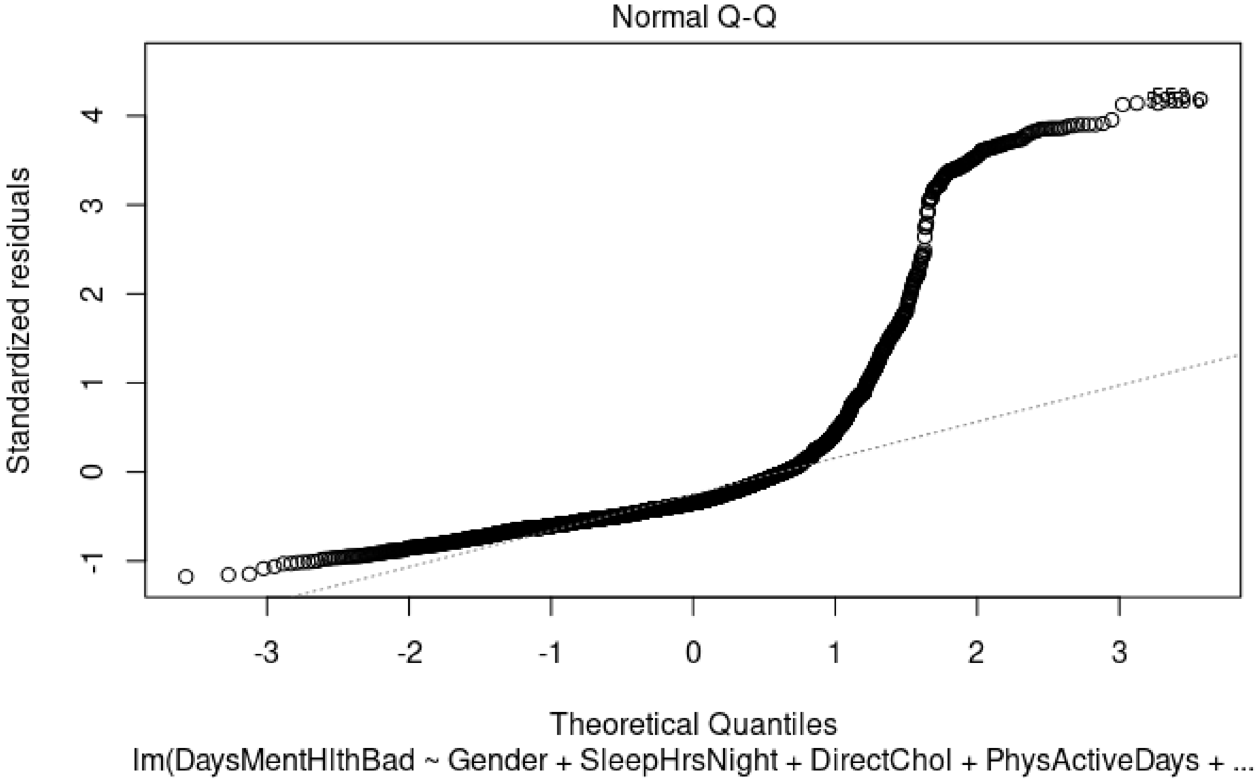

Residual vs Predictor
Residual
30
20
10
0
0
100
200
300
AlcoholYear


Residual vs Predictor
Residual
30
20
10
0
20
30
40
50
60
BMI

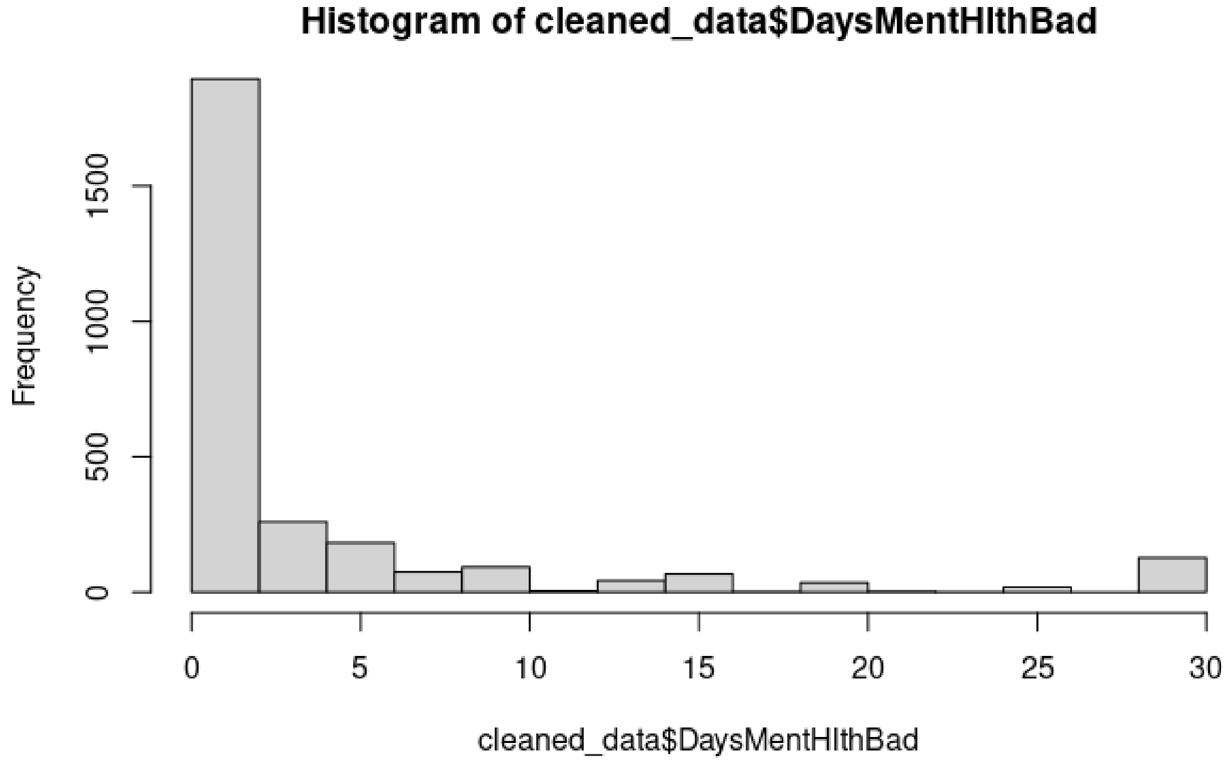

Residual vs Predictor
Residual
30
20
10
0
80
100
120
140
160
180
200
220
BPSysAve

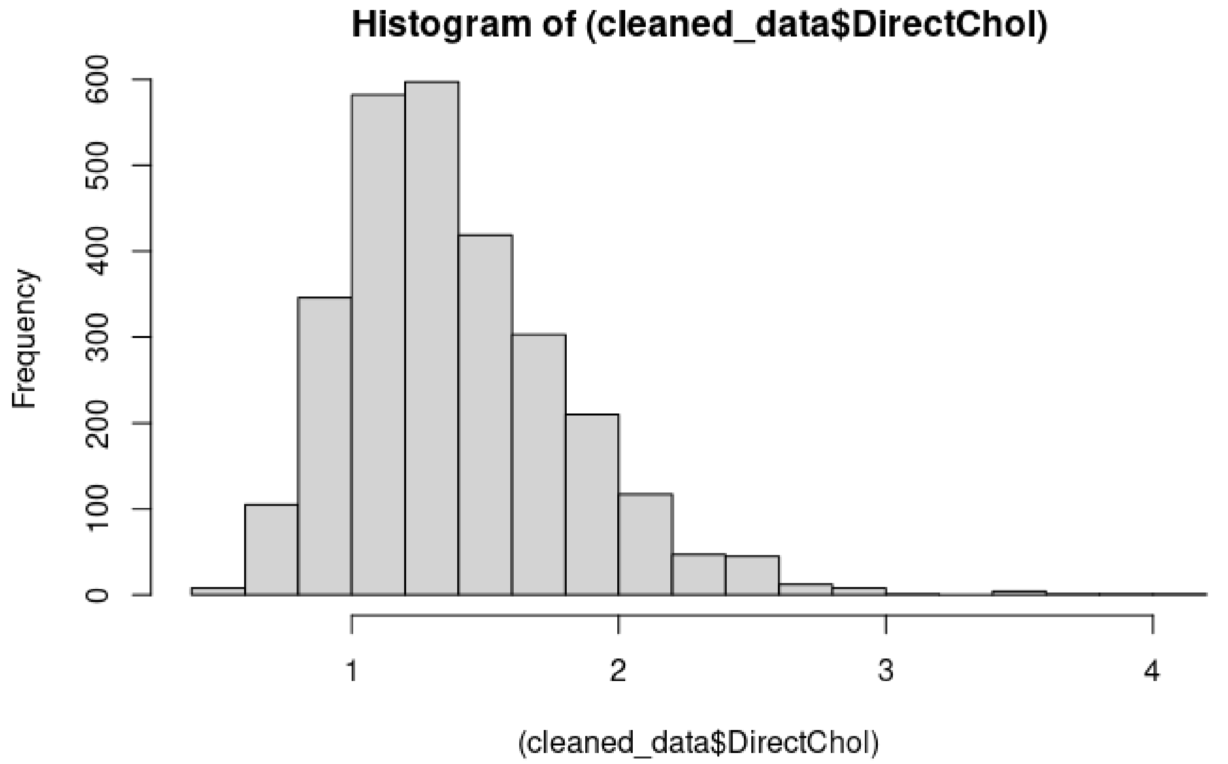

Residual vs Predictor
Residual
30
20
10
0
0
20
40
60
80
100
120
BPDiaAve

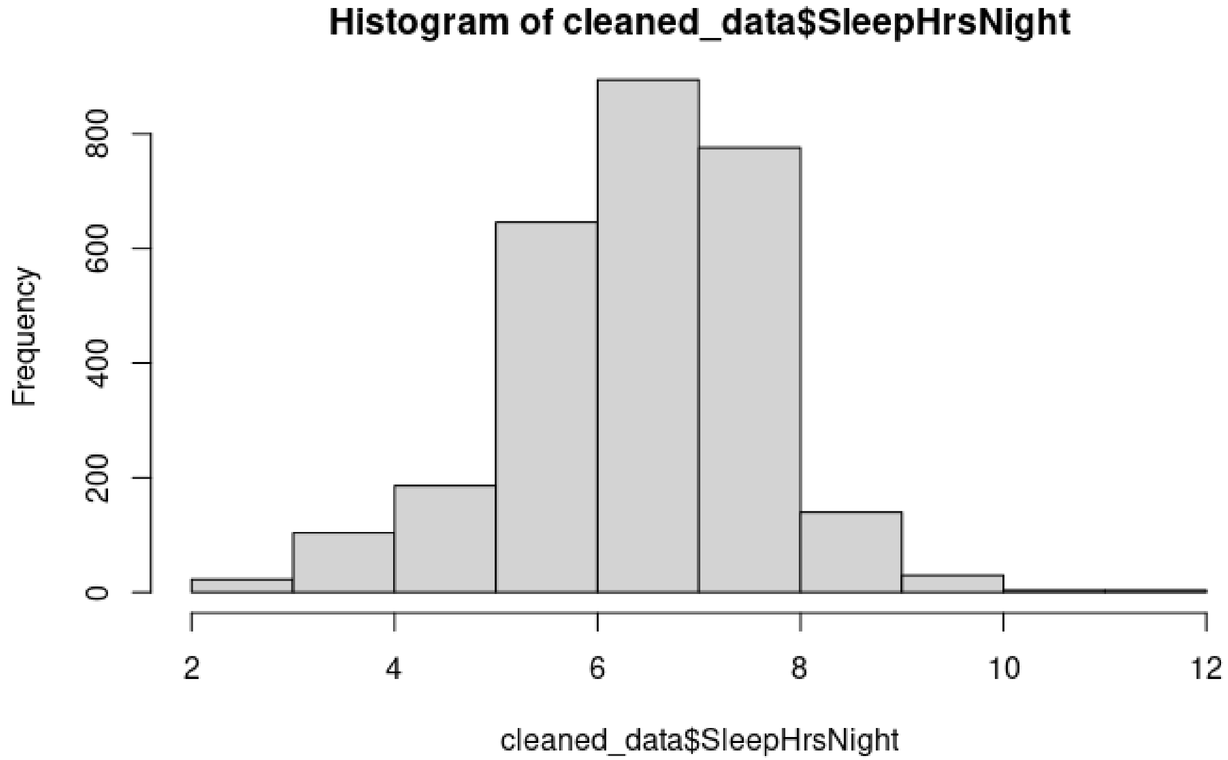

Standardised residuals histogram
Frequency
0
200
400
600
800
1200
-1
0
1
2
3
4
Standardised Residuals

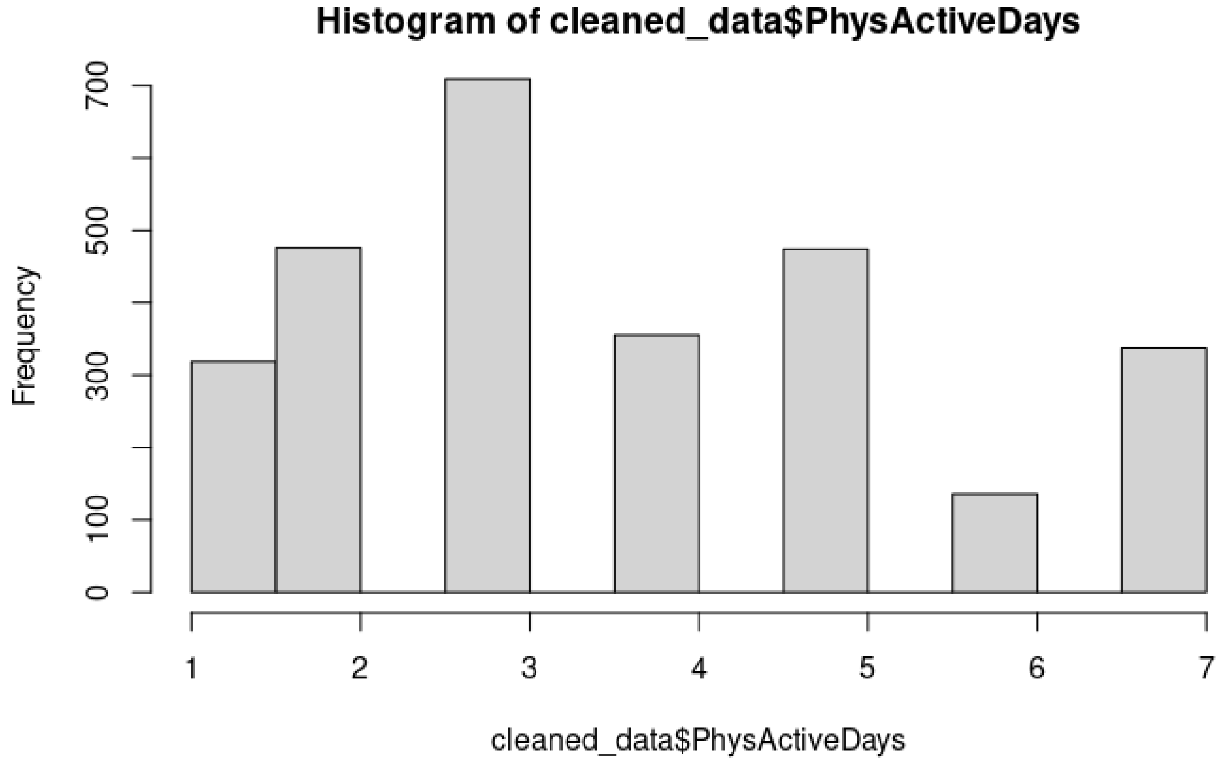

Normal Q-Q
Standardized residuals
-1
0
1
2
3
4
553
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(DaysMentHlthBad ~ Gender + SleepHrsNight + DirectChol + PhysActiveDays + ...

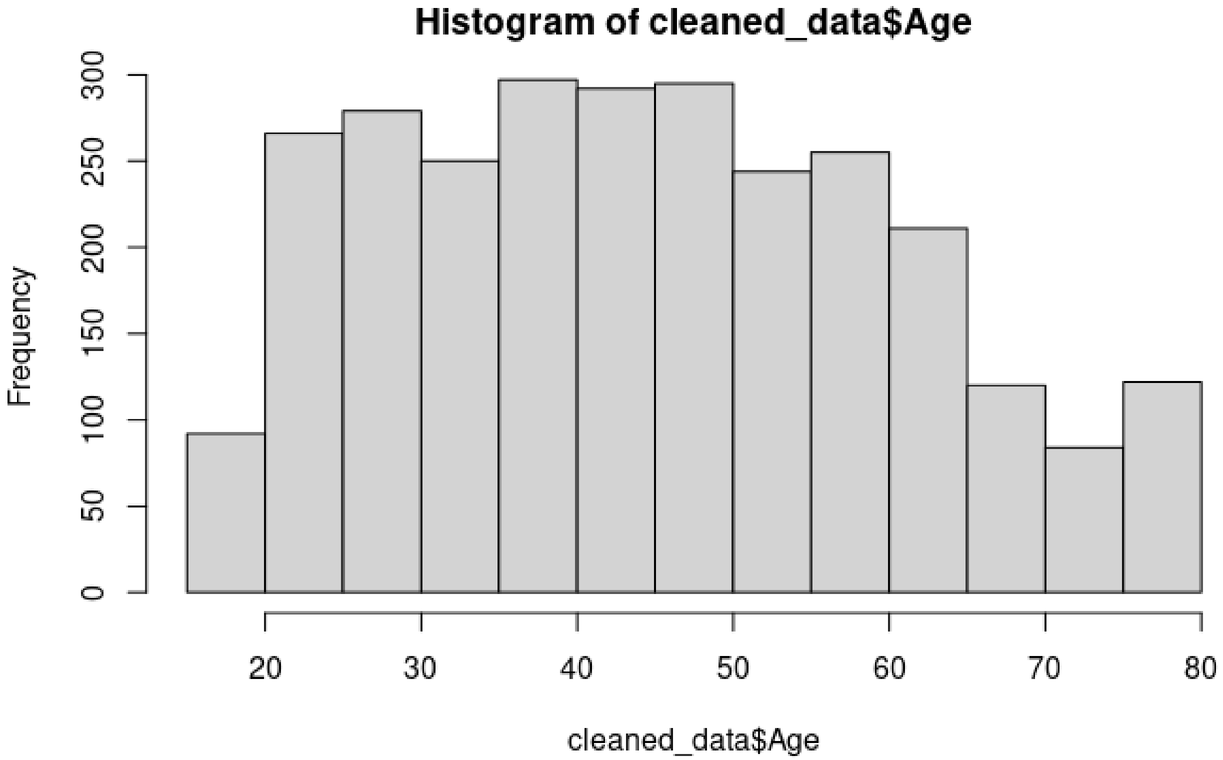
Distribution of Predictor and Response Variables

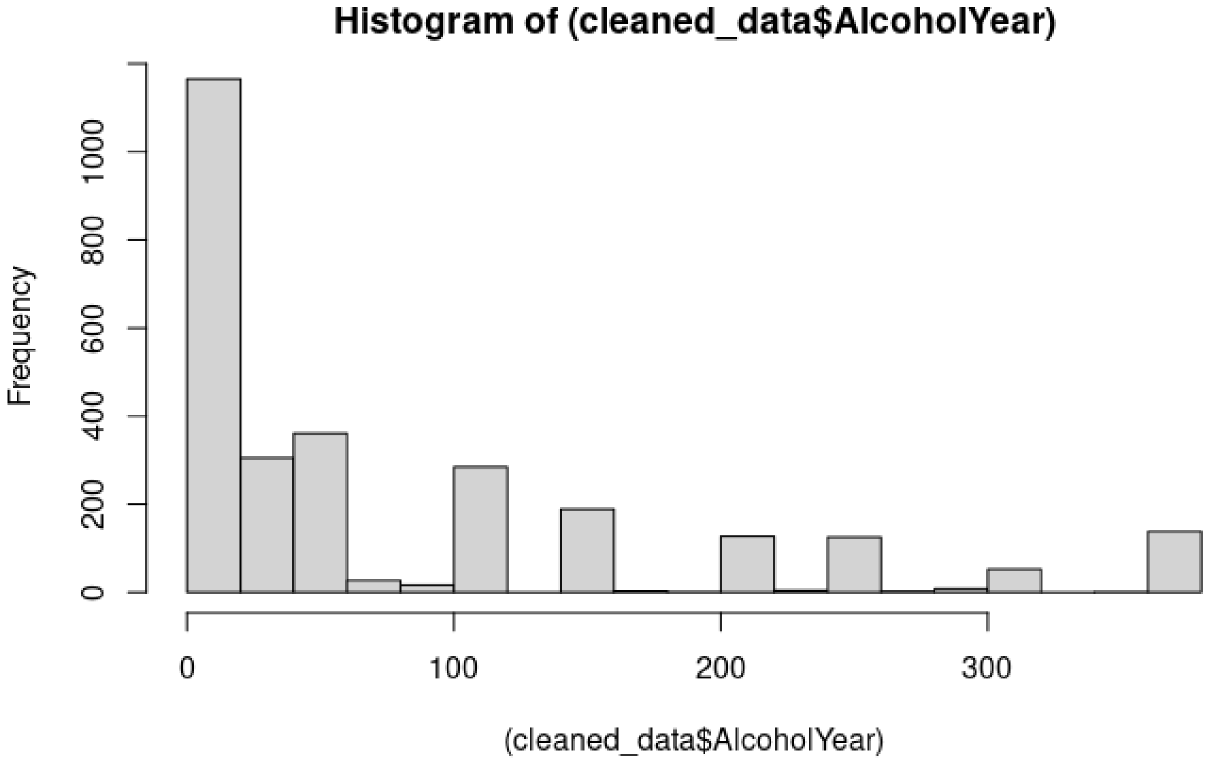


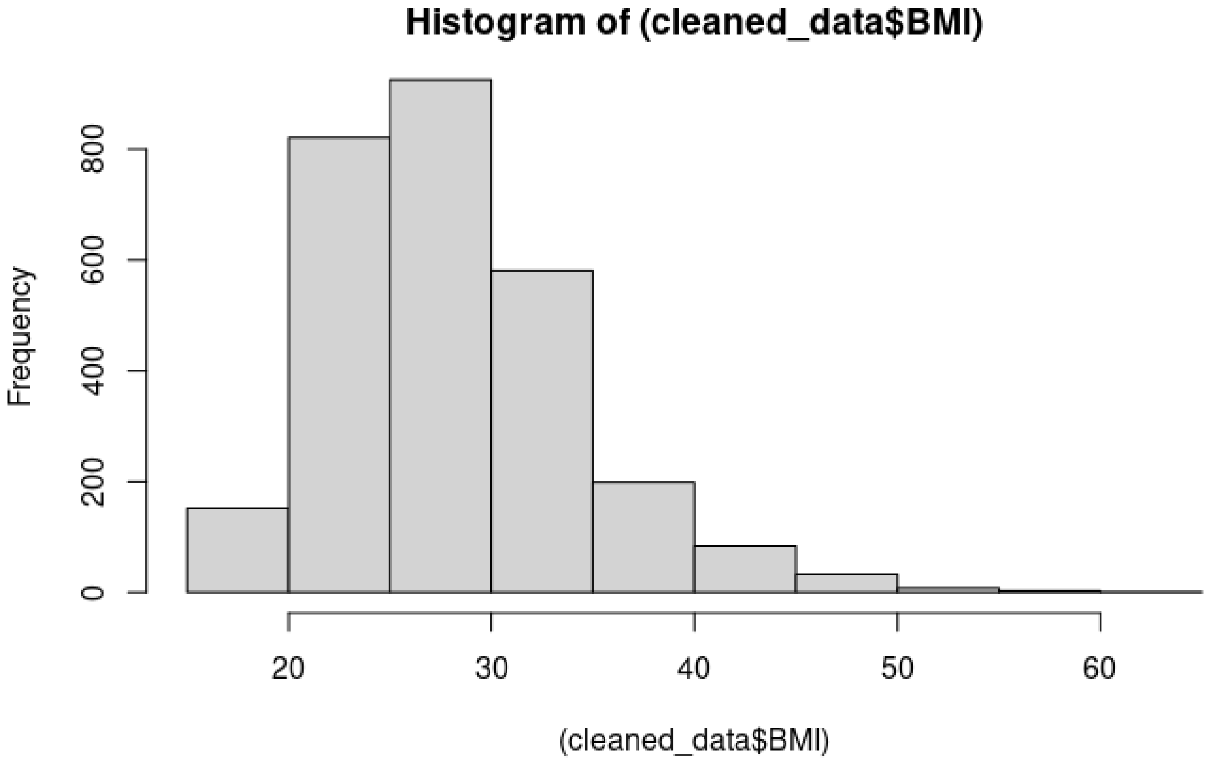


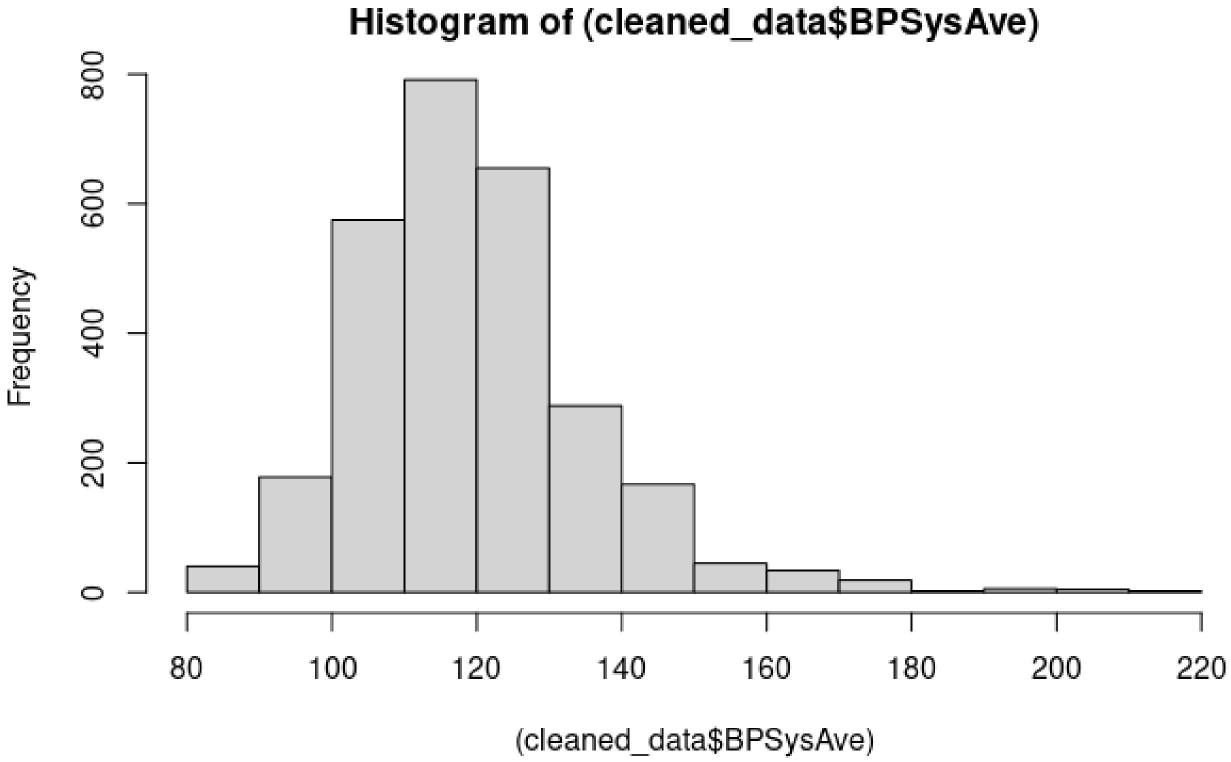

Histogram of cleaned_data$SleepHrsNight
Frequency
0
200
400
600
800
2
4
6
8
10
12
cleaned_data$SleepHrsNight

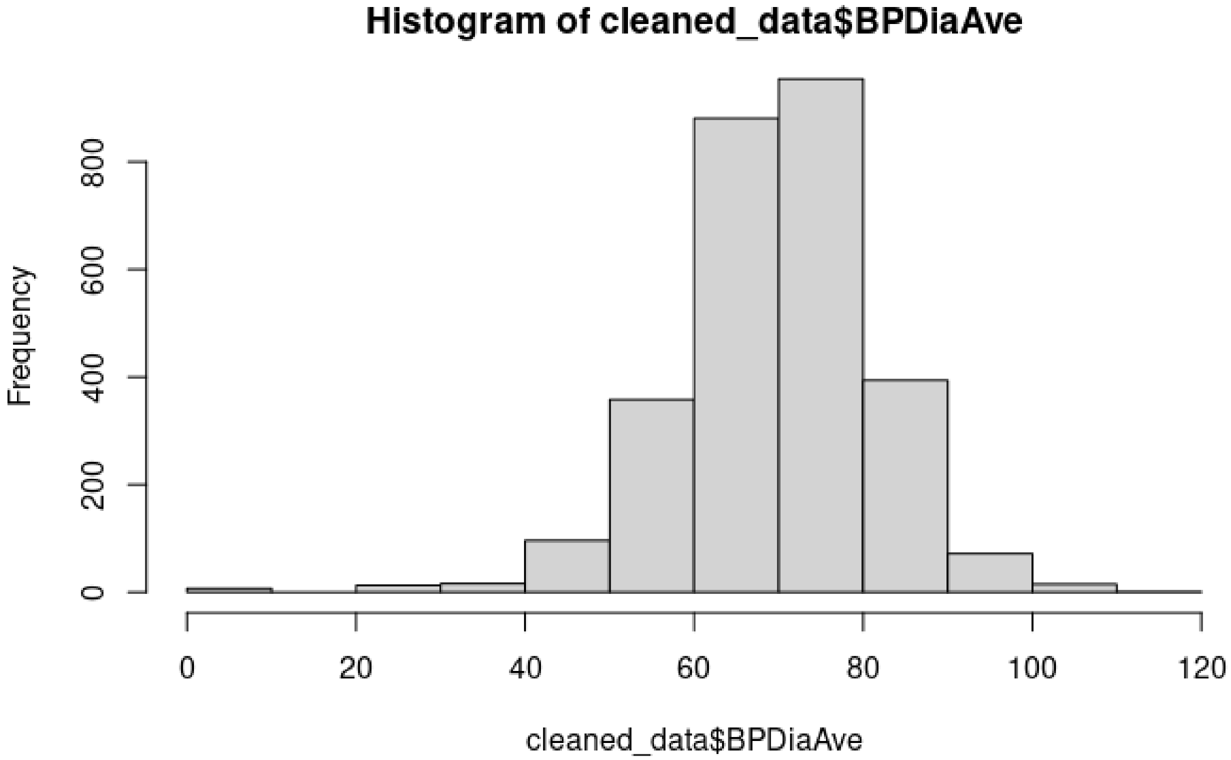

Histogram of cleaned_data$PhysActiveDays
Frequency
0
100
300
500
700
1
2
3
4
5
6
7
cleaned_data$PhysActiveDays


Histogram of cleaned_data$Age
Frequency
0
50
100
150
200
250
300
20
30
40
50
60
70
80
cleaned_data$Age


Histogram of (cleaned_data$AlcoholYear)
Frequency
0
200
400
600
800
1000
0
100
200
300
(cleaned_data$AlcoholYear)


Histogram of (cleaned_data$BMI)
Frequency
0
200
400
600
800
20
30
40
50
60
(cleaned_data$BMI)


Histogram of (cleaned_data$BPSysAve)
Frequency
0
200
400
600
800
80
100
120
140
160
180
200
220
(cleaned_data$BPSysAve)


Histogram of cleaned_data$BPDiaAve
Frequency
0
200
400
600
800
0
20
40
60
80
100
120
cleaned_data$BPDiaAve

Works Cited

De Mello, Marco T., et al. “Relationship between Physical Activity and Depression and Anxiety Symptoms: A Population Study.” *Journal of Affective Disorders*, Mar. 2013, https://doi.org/10.1016/j.jad.2013.01.035.

Dong, Lu, et al. “Association between Sleep Duration and Depression in US Adults: A Cross-Sectional Study.” *Journal of Affective Disorders*, Sept. 2021, https://doi.org/10.1016/j.jad.2021.09.075.

Knox, S., Barnes, A., Kiefe, C., Lewis, C. E., Iribarren, C., Matthews, K. A., Wong, N. D., & Whooley, M. (2006). History of depression, race, and cardiovascular risk in CARDIA. International Journal of Behavioral Medicine, 13(1), 44–50. https://doi.org/10.1207/s15327558ijbm1301_6.