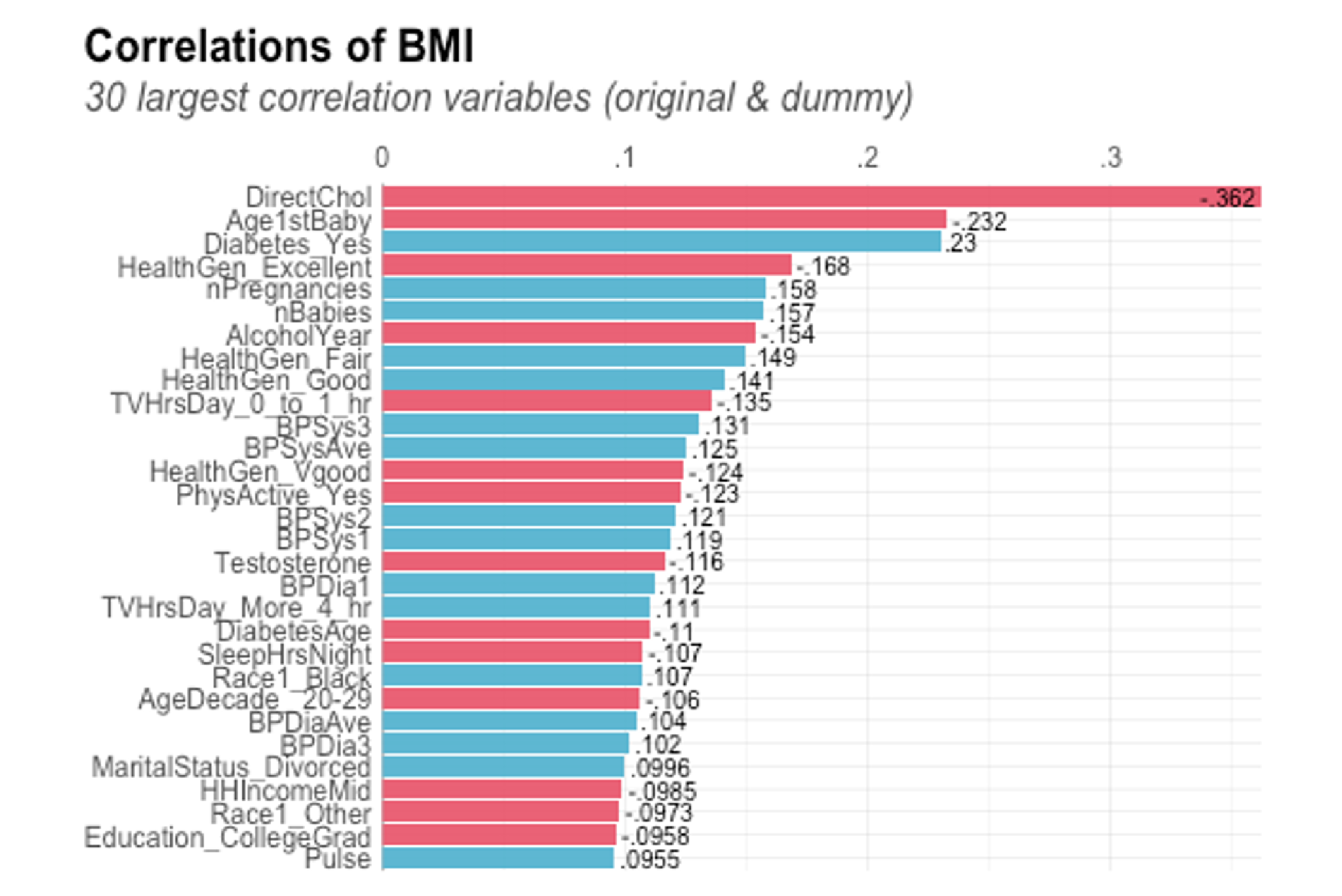
Oct 13, 2022

# “Predicting Body Mass Index using Multiple Regressions and Exploratory Data Analysis: An Analysis of the NHANES Dataset” *by Ian Musumba*

The National Health and Nutrition Examination Survey (NHANES) is a comprehensive, ongoing survey conducted by the US National Center for Health Statistics (NCHS) to assess the health and nutrition status of the American population. The data used in this research project was collected during the years 2009-2010 and 2011-2012 and includes information on approximately 5,000 individuals of all ages. These individuals were interviewed in their homes and completed a health examination as part of the survey.

Our study utilizes a random sample of 1800 adults from the NHANES dataset, with an equal distribution of 907 females and 900 males. Our objective is to develop a model that accurately predicts an individual's body mass index (BMI) using the variables provided in the dataset. We will employ various multiple regression techniques and conduct both graphical and numerical exploratory data analysis (EDA).

To begin our EDA, we will examine the top 30 variables that have the strongest correlation with BMI. The stronger the correlation, the more effective the variable will be in predicting future values. We will use the "lares" package and the "corr\_var" function to correlate the entire dataframe with a single feature. The function automatically applies one-hot-smart-encoding, eliminating the need for inputting only numerical values. After running the function, we will have a plot of the top 30 variables that best fit. From these, we have selected DirectChol, Diabetes, HealthGen, TVHrsDay, BPSys3, Testosterone, BPDial, Race1, AlcoholYear, and DaysMentHlthBad. Blue bars indicate positive correlations, and red bars indicate negative correlations. The correlation significance level is estimated to be around 0.01, with parameters ranging from -1 to 1.

**Figure 1:**

This shows the rankings of top 30 best data points from the NHANES dataset that have the highest correlation to BMI }

Another important package that we used to help with the categorical variable that had the best correlation with BMI was “fastDummies,”. This package was used to create dummy variables for our categorical variables of interest: *Diabetes, HealthGen, TVHrsDay,* and *Race1*. These categorical variables were converted into dummy columns(1’s and 0’s) because of them containing simple answer responses (yes/no, good/bad, etc) so this way we were able to actually include our categorical variables in a numeric sense.

**MODEL1**:

𝐵𝑀𝐼 = 𝐷𝑖𝑟𝑒𝑐𝑡𝐶ℎ𝑜𝑙 + 𝐷𝑖𝑎𝑏𝑒𝑡𝑒𝑠\_𝑌𝑒𝑠 + 𝐻𝑒𝑎𝑙𝑡ℎ𝐺𝑒𝑛\_𝐸𝑥𝑐𝑒𝑙𝑙𝑒𝑛𝑡 + 𝑇𝑉𝐻𝑟𝑠𝐷𝑎𝑦\_0\_𝑡𝑜\_1\_ℎ𝑟 + 𝐵𝑃𝑆𝑦𝑠3 +

𝐵𝑃𝐷𝑖𝑎1 + 𝑇𝑒𝑠𝑡𝑜𝑠𝑡𝑒𝑟𝑜𝑛𝑒 + 𝑅𝑎𝑐𝑒1\_𝐵𝑙𝑎𝑐𝑘 + 𝐴𝑙𝑐𝑜ℎ𝑜𝑙𝑌𝑒𝑎𝑟 + 𝐷𝑎𝑦𝑠𝑀𝑒𝑛𝑡𝐻𝑙𝑡ℎ𝐵𝑎𝑑

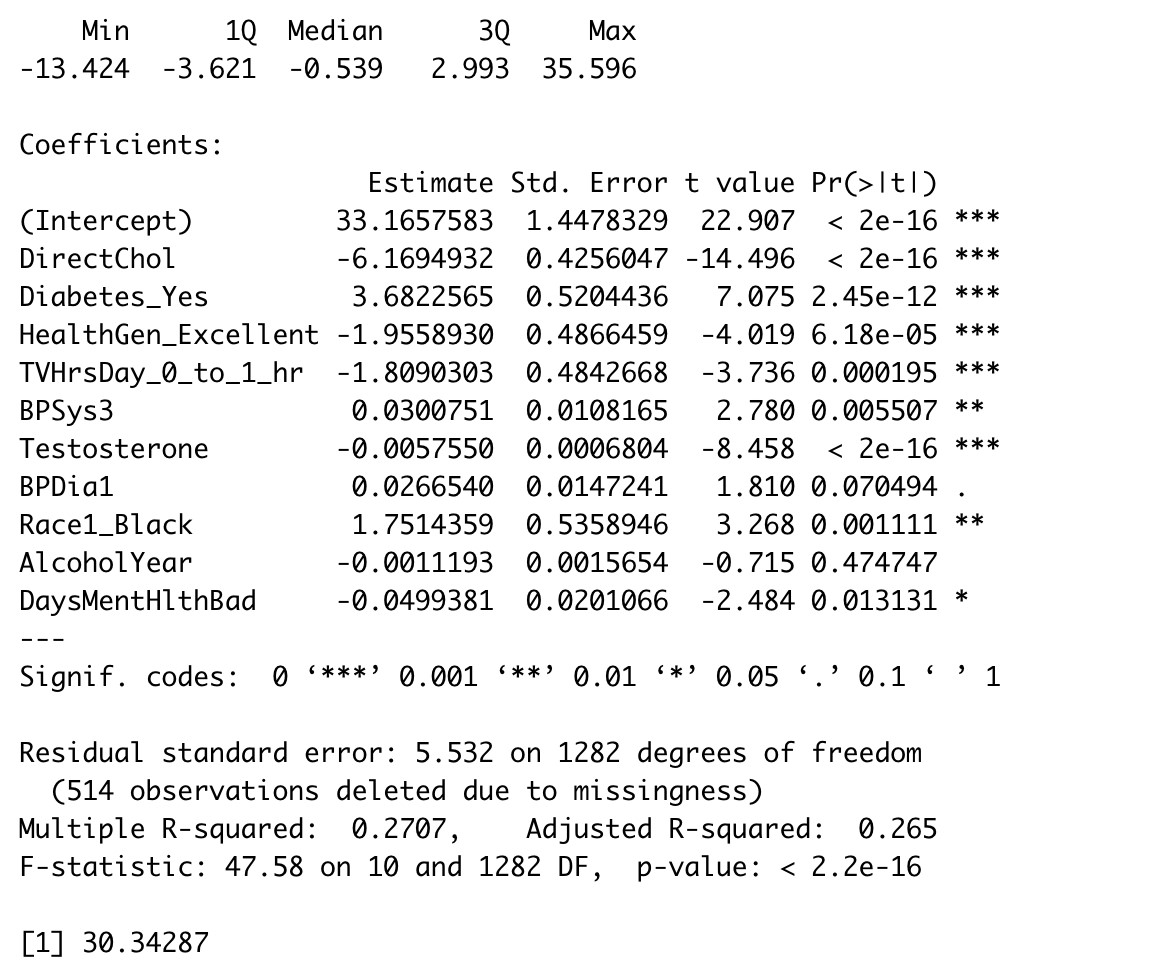
The first model is a multivariate linear model with *BMI* as the quantitative response variable indicating the body mass index of an individual and 10 different predictor variables (6 quantitative and 4 dummy variables for each of categorical variables in this model). Our first predictor variable is *DirectChol,* which is a numerical variable and denotes the direct HDL cholesterol in mmol/L for participants aged 6 years and older. Next, *Diabetes\_Yes,* is a dummy variable created for the “Yes” response for the binary variable *Diabetes,* which reports if the participant aged 1 year or older had diabetes and has two values, “Yes” and “No”. 𝐻𝑒𝑎𝑙𝑡h𝐺𝑒𝑛\_𝐸𝑥𝑐𝑒𝑙𝑙𝑒𝑛𝑡 is a dummy variable for the “Excellent” response of the categorical variable *HealthGen,* which renders a self-reported rating of a participant’s health in general, aged 12 or older, and has five responses: “Excellent”, “VGood”, “Good”, “Fair”, and “Poor”. Next, 𝑇𝑉𝐻𝑟𝑠𝐷𝑎𝑦\_0\_𝑡𝑜\_1\_h𝑟 is a dummy variable created for the “0\_to\_1\_hr” response of the categorical variable *TVHrsDay,* which reports the number of hours per day on average participant, aged 2 years or older, watched TV over the past 30 days and include the responses: “0\_hrs”, “0\_to\_1\_hr”, “1\_hr”, “2\_hr”, “3\_hr”, “4\_hr”, and “More\_4\_hr”. Moreover, 𝐵𝑃𝑆𝑦𝑠3 and 𝐵𝑃𝐷𝑖𝑎1 are both numerical variables indicating the third reading of systolic blood pressure and first reading of diastolic blood pressure, both in mm of Hg, respectively. 𝑇𝑒𝑠𝑡𝑜𝑠𝑡𝑒𝑟𝑜𝑛𝑒 is the next numerical variable denoting the total level of testosterone (ng/dL) in participants aged 6 years and older. Next, 𝑅𝑎𝑐𝑒1\_𝐵𝑙𝑎𝑐𝑘 is a dummy variable created for the “Black” response of the Race1 variable that renders the race of the participant and includes responses: “Mexican”, “Hispanic”, “White”, “Black”, and “Other”. Lastly, 𝐴𝑙𝑐𝑜h𝑜𝑙𝑌𝑒𝑎𝑟 and 𝐷𝑎𝑦𝑠𝑀𝑒𝑛𝑡𝐻𝑙𝑡h𝐵𝑎𝑑 are numerical variables indicating the estimated number of times over the past year that a participant, aged 18 years or older, drank alcohol and the self-reported number of days the participant's mental health was not good out of the past 30 days.

We chose all our predictor variables in the descending order of the strength of their correlation coefficients, which can be seen in figure 1. *DirectChol* has the strongest correlation coefficient of absolute value 0.362 (figure 1). We decided to not include variables like *Age1stBaby*, *nPregnancies*, and *nBabies* as they had 1361, 1154, and 1219 “NA” values respectively. Our threshold for “NA” values was around 550 for each variable and therefore, we did not include variables that had more “NA” values than the threshold value of 550 in order to make our model more generalizable. Also, we only included one dummy variable for each categorical variable. For example, we only included 𝐻𝑒𝑎𝑙𝑡h𝐺𝑒𝑛\_𝐸𝑥𝑐𝑒𝑙𝑙𝑒𝑛𝑡 for the categorical variable 𝐻𝑒𝑎𝑙𝑡h𝐺𝑒𝑛. Moreover, we did not include variables like *PhysActive\_Yes* and *SleepHrsNight* as they did not increase our adjusted R-squared value by more than 1%, which was our primary criteria to add variables to our model. Also, we only included one variable for measuring each systolic and diastolic blood pressure (one that had the strongest correlation coefficient for both the categories: *BPSys3* and *BPDia1*). Lastly, we restricted the number of predictor variables to 10 and hence came up with this model.

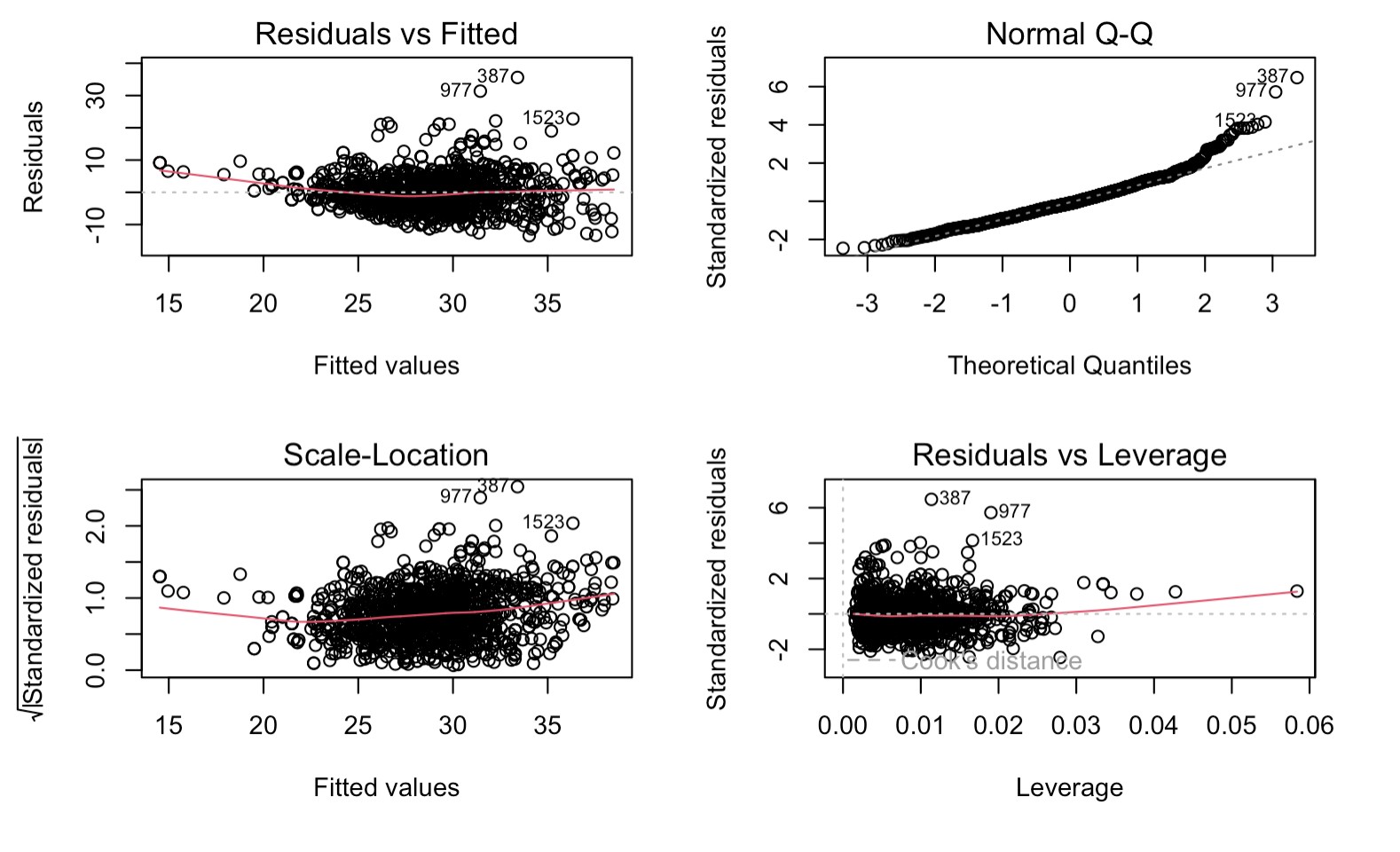
From Table 1, we can see that the R-squared value of this model is 0.2707, which means that 27.07% of the variability in our response variable *BMI* is explained by this model. Also, from Table 1, we can see that all but two variables, *BPDia1* and *AlcoholYear*, have a p-value less than the 5% level of statistical significance and are significant predictors of our response variable *BMI*. Moreover, from Table 1, as the F-statistics is 47.58, which is greater than one and the associated p-value is very small, we have evidence that the model explains a significant amount of variability in the response variable *BMI* and the overall model is effective for predicting it. Moreover, after examining the residuals versus the fitted values plot we conclude that the assumption of linearity is violated as we observed a curved pattern in the distribution of the residuals. Also, the plot of residuals versus fitted values shows a fan shape with the variability in the residuals tending to increase as the fitted values get larger, which violates the constant variance assumption. Furthermore, the assumption of normality is somewhat violated as there is a departure from the straight line of the normal probability plot at the right end and the histogram for residuals is skewed towards right as it has a big tail at the right end. Therefore, we have to undertake some transformations in our model to get a better agreement with these assumptions. Most potential transformation seems to be a logarithmic transformation of either the response variable, the predictor variable, or possibly both in order to solve the issue of non-constant variance.

# Table 1: Regression Output of Model 1

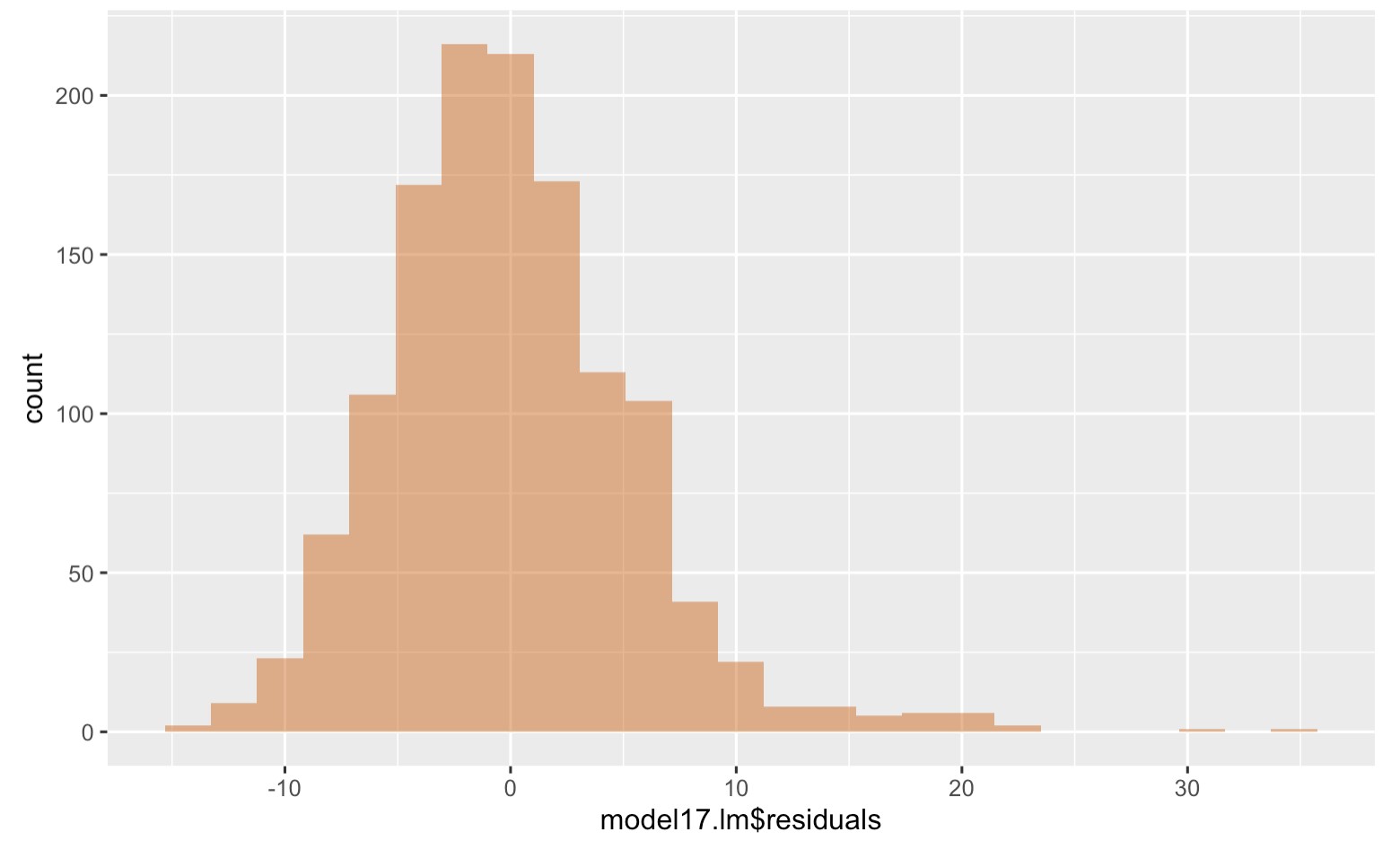
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | 33.1658 | 1.4478 | 22.91 | 0.0000 |
| DirectChol | -6.1695 | 0.4256 | -14.50 | 0.0000 |
| Diabetes\_Yes | 3.6823 | 0.5204 | 7.08 | 0.0000 |
| HealthGen\_Excellent | -1.9559 | 0.4866 | -4.02 | 0.0001 |
| TVHrsDay\_0\_to\_1\_hr | -1.8090 | 0.4843 | -3.74 | 0.0002 |
| BPSys3 | 0.0301 | 0.0108 | 2.78 | 0.0055 |
| Testosterone | -0.0058 | 0.0007 | -8.46 | 0.0000 |
| BPDia1 | 0.0267 | 0.0147 | 1.81 | 0.0705 |
| Race1\_Black | 1.7514 | 0.5359 | 3.27 | 0.0011 |
| AlcoholYear | -0.0011 | 0.0016 | -0.71 | 0.4747 |
| DaysMentHlthBad | -0.0499 | 0.0201 | -2.48 | 0.0131 |
| R-squared: 0.2707  Adjusted R-squared: 0.265  F-statistic: 47.58 on 10 and 1282 DF, p-value: < 2.2e-16 | | | | |



**{ Summary statistics of model 1. F-statistic is 47.58 and P-value is <2.2e-16. MPSE is 30.34 }**



**{Plot summary of model 1}**



**{Histogram of model 1. Visible right skewness. A transformation is necessary in order to improve the above diagram}**

**MODEL 2:**

𝑙𝑜𝑔(𝐵𝑀𝐼) = 𝐷𝑖𝑟𝑒𝑐𝑡𝐶ℎ𝑜𝑙 + 𝐷𝑖𝑎𝑏𝑒𝑡𝑒𝑠\_𝑌𝑒𝑠 + 𝐻𝑒𝑎𝑙𝑡ℎ𝐺𝑒𝑛\_𝐸𝑥𝑐𝑒𝑙𝑙𝑒𝑛𝑡 + 𝑇𝑉𝐻𝑟𝑠𝐷𝑎𝑦\_0\_𝑡𝑜\_1\_ℎ𝑟 + 𝐵𝑃𝑆𝑦𝑠3

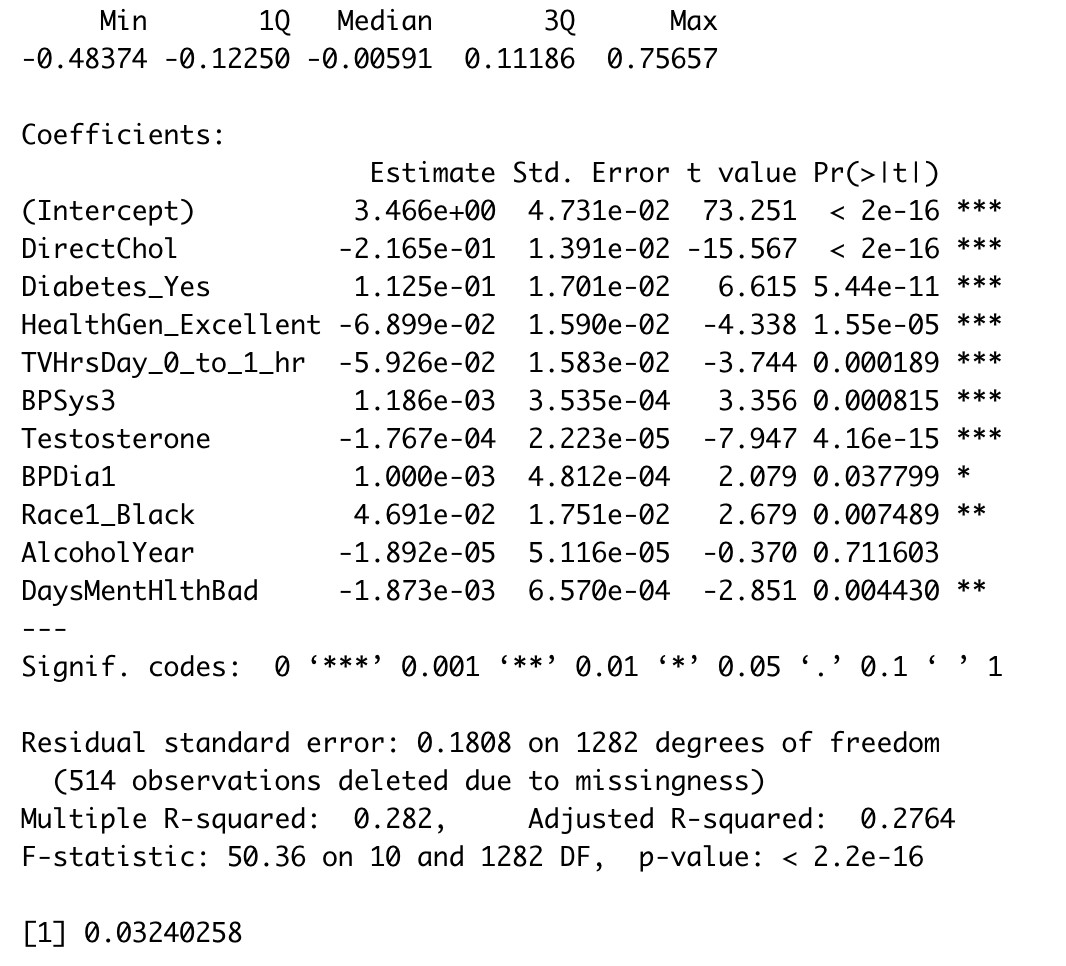
From Table 2, we can see that the R-squared value of this model is 0.2820, which means 𝐵𝑃𝐷𝑖𝑎1 + 𝑇𝑒𝑠𝑡𝑜𝑠𝑡𝑒𝑟𝑜𝑛𝑒 + 𝑅𝑎𝑐𝑒1\_𝐵𝑙𝑎𝑐𝑘 + 𝐴𝑙𝑐𝑜ℎ𝑜𝑙𝑌𝑒𝑎𝑟 + 𝐷𝑎𝑦𝑠𝑀𝑒𝑛𝑡𝐻𝑙𝑡ℎ𝐵𝑎𝑑

that 28.2% of the variability in our response variable log(*BMI)* is explained by this model. Also, from Table 2, we can see that all but one variable, *AlcoholYear*, have a p-value less than the 5% level of statistical significance and are significant predictors of our response variable log(*BMI)*. Moreover, from Table 2, as the F-statistics is 50.36, which is greater than one and the associated p-value is very small, we have evidence that the model explains a significant amount of variability in the response variable log(*BMI)* and the overall model is effective for predicting it.

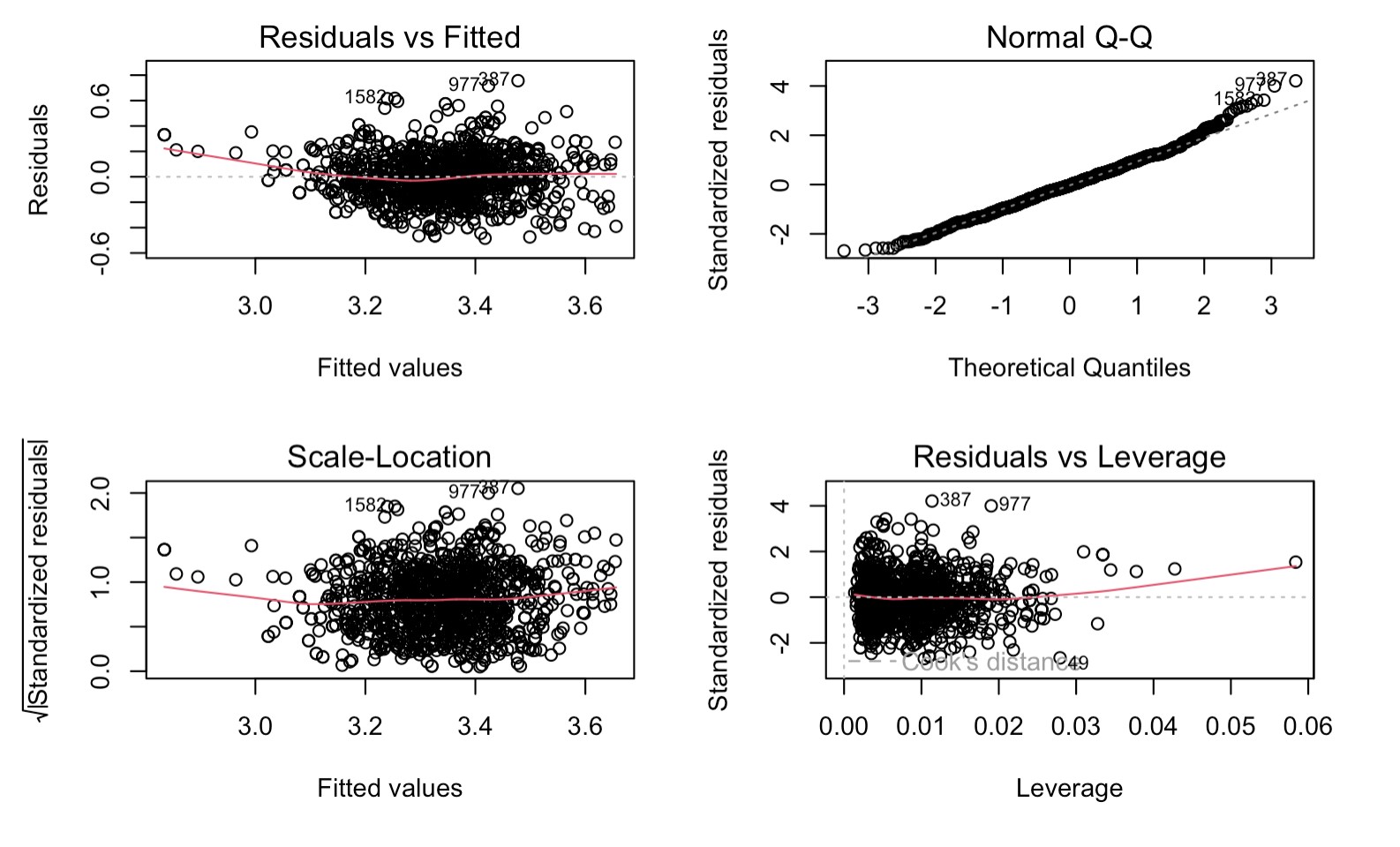
Furthermore, in this model we have logged the response variable, *BMI* in order to fix the issue of non-constant variance. This transformation shows considerable improvement in meeting the normality assumption and constant variance assumption.The normal probability plot shows less departure from the straight line and the histogram of residuals is considerably less skewed towards right. However, the assumption of linearity is still violated. This compels us to consider a logarithmic transformation in both response and predictor variable.

# Table 2: Regression Output of Model 2

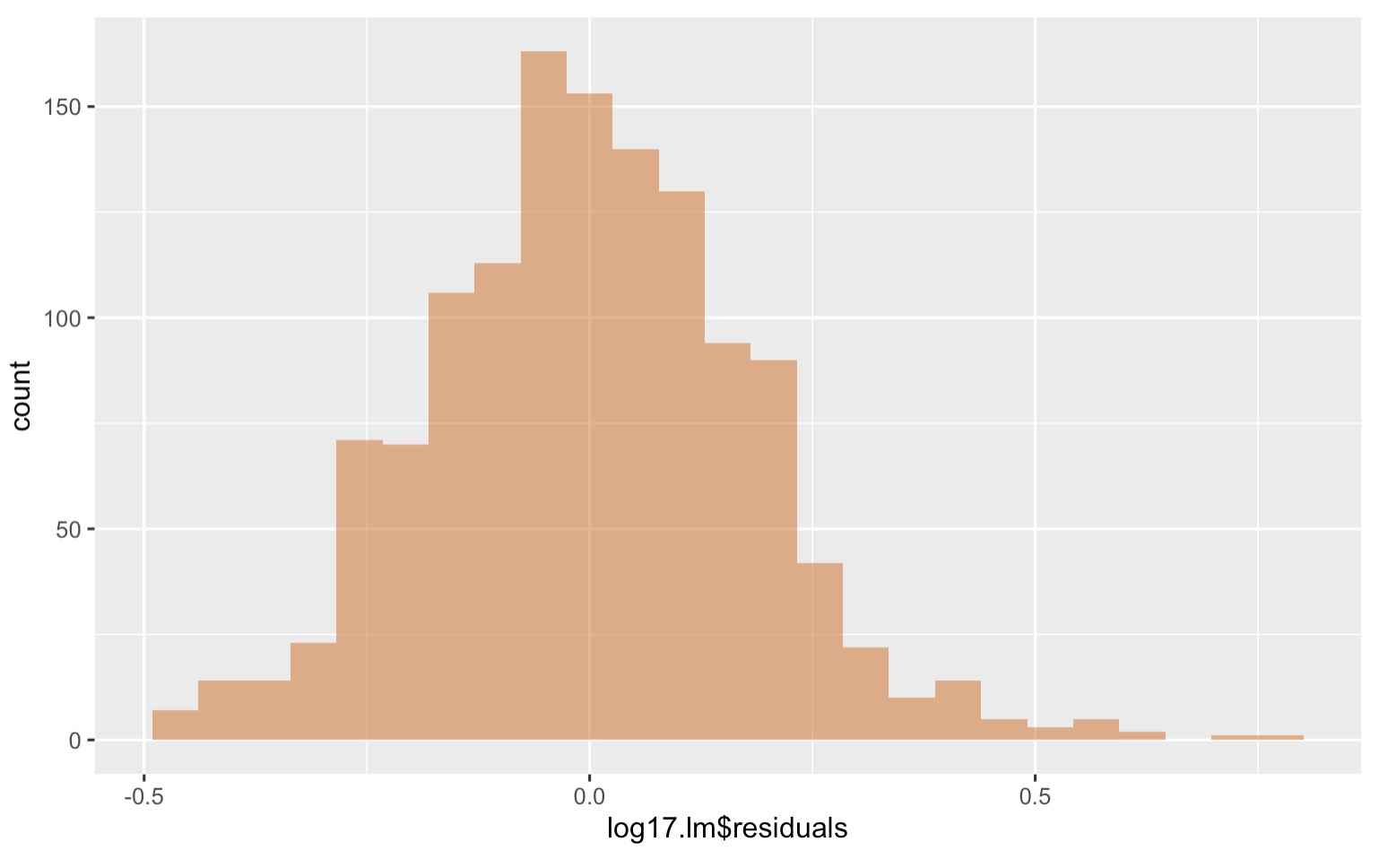
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| (Intercept) | 33.1658 | 1.4478 | 22.91 | 0.0000 |
| DirectChol | -6.1695 | 0.4256 | -14.50 | 0.0000 |
| Diabetes\_Yes | 3.6823 | 0.5204 | 7.08 | 0.0000 |
| HealthGen\_Excellent | -1.9559 | 0.4866 | -4.02 | 0.0001 |
| TVHrsDay\_0\_to\_1\_hr | -1.8090 | 0.4843 | -3.74 | 0.0002 |
| BPSys3 | 0.0301 | 0.0108 | 2.78 | 0.0055 |
| Testosterone | -0.0058 | 0.0007 | -8.46 | 0.0000 |
| BPDia1 | 0.0267 | 0.0147 | 1.81 | 0.0705 |
| Race1\_Black | 1.7514 | 0.5359 | 3.27 | 0.0011 |
| AlcoholYear | -0.0011 | 0.0016 | -0.71 | 0.4747 |
| DaysMentHlthBad | -0.0499 | 0.0201 | -2.48 | 0.0131 |
| R-squared: 0.282  Adjusted R-squared: 0.2764  F-statistic: 50.36 on 10 and 1282 DF, p-value: < 2.2e | | | | -16 |



**{ Summary statistics of model 2. F-statistic is 47.58 and P-value is <2.2e-16. MPSE is 0.032}**



# { Plot summary of model 2. Visible improvement when applied the log transformation}



**{Histogram for model 2. The diagram has the intended Bell-Shape}**

**MODEL 3:**

𝑙𝑜𝑔(𝐵𝑀𝐼) = 𝑙𝑜𝑔(𝐷𝑖𝑟𝑒𝑐𝑡𝐶ℎ𝑜𝑙) + 𝐷𝑖𝑎𝑏𝑒𝑡𝑒𝑠\_𝑌𝑒𝑠 + 𝐻𝑒𝑎𝑙𝑡ℎ𝐺𝑒𝑛\_𝐸𝑥𝑐𝑒𝑙𝑙𝑒𝑛𝑡 + 𝑇𝑉𝐻𝑟𝑠𝐷𝑎𝑦\_0\_𝑡𝑜\_1\_ℎ𝑟 +

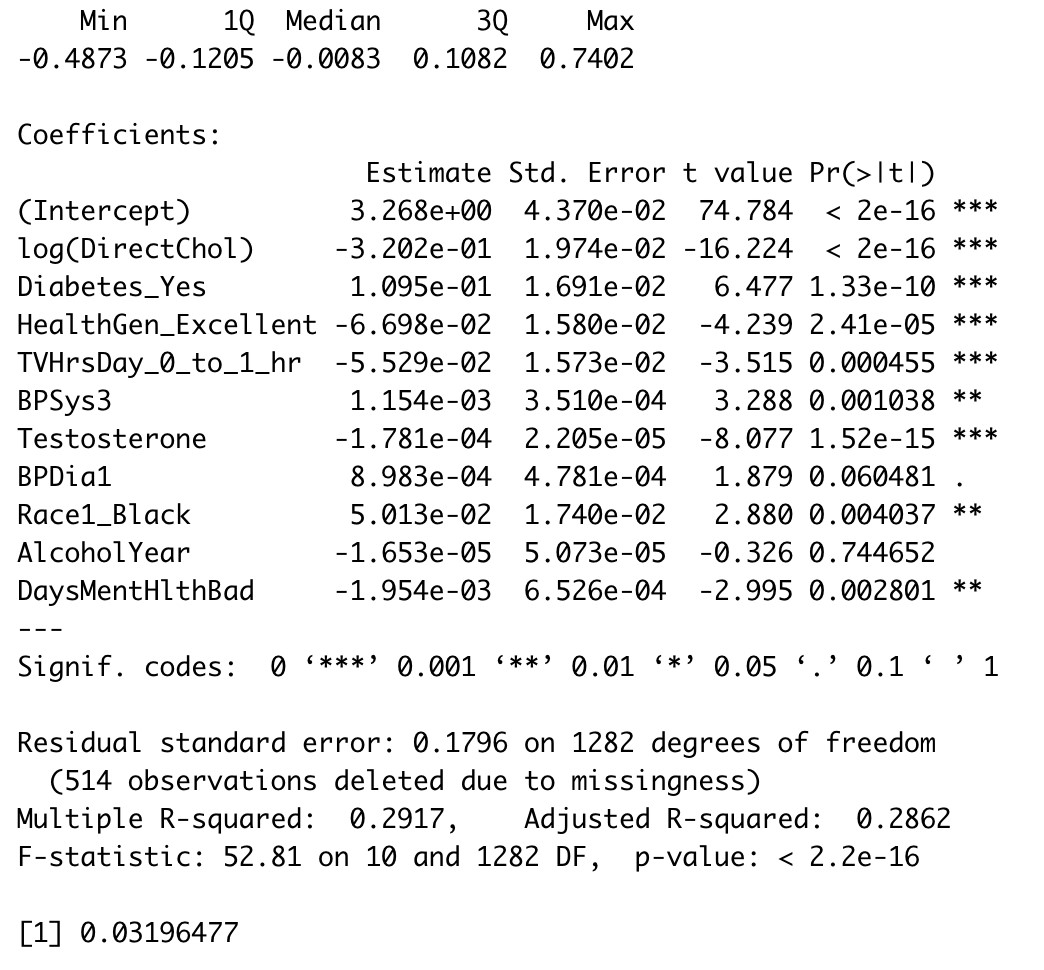
𝐵𝑃𝑆𝑦𝑠3 + 𝐵𝑃𝐷𝑖𝑎1 + 𝑇𝑒𝑠𝑡𝑜𝑠𝑡𝑒𝑟𝑜𝑛𝑒 + 𝑅𝑎𝑐𝑒1\_𝐵𝑙𝑎𝑐𝑘 + 𝐴𝑙𝑐𝑜ℎ𝑜𝑙𝑌𝑒𝑎𝑟 + 𝐷𝑎𝑦𝑠𝑀𝑒𝑛𝑡𝐻𝑙𝑡ℎ𝐵𝑎𝑑

In this third model we have logged both the response variable, *BMI* and the predictor variable *DirectChol.* The residuals vs fitted values plot does not show a clear pattern or trend, meaning the assumption of linearity is valid in this case. Also, as the distribution of residuals seems to follow a horizontal band above and below the zero line, the assumption of constant variance is valid too. Lastly, as the distribution follows the straight line of the normal probability plot for the most part, except for the right end, the assumption of normality holds as well. Same can be seen from the histogram of the residuals as it seems symmetric, unimodal, and centered at zero. Therefore, this model seems the best at predicting our response variable *BMI.*

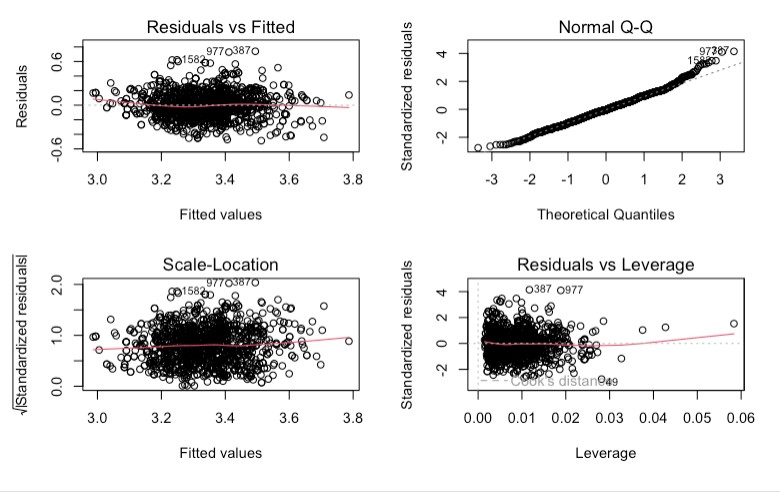
From Table 3, we can see that the R-squared value of this model is 0.2917, which means that 29.17% of the variability in our response variable log(*BMI)* is explained by this model. Also, from Table 3, we can see that all but two variables, *AlcoholYear* and *BPDia1*, have a p-value less than the 5% level of statistical significance and are significant predictors of our response variable log(*BMI)*. Moreover, from Table 3, as the F-statistics is 52.81%, which is greater than one and the associated p-value is very small, we have evidence that the model explains a significant amount of variability in the response variable log(*BMI)* and the overall model is effective for predicting it.

# Table 3: Regression Output of Model 3

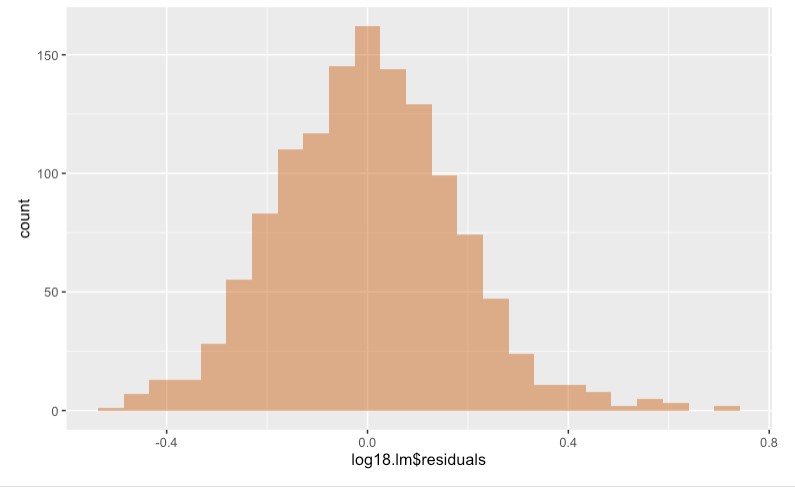
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Estimate** | **Std. Error** | **t value** |  | **Pr(>|t|)** |
| (Intercept) | 33.1658 | 1.4478 | 22.91 | 0.0000 |  |
| DirectChol | -6.1695 | 0.4256 | -14.50 | 0.0000 |  |
| Diabetes\_Yes | 3.6823 | 0.5204 | 7.08 | 0.0000 |  |
| HealthGen\_Excellent | -1.9559 | 0.4866 | -4.02 | 0.0001 |  |
| TVHrsDay\_0\_to\_1\_hr | -1.8090 | 0.4843 | -3.74 | 0.0002 |  |
| BPSys3 | 0.0301 | 0.0108 | 2.78 | 0.0055 |  |
| Testosterone | -0.0058 | 0.0007 | -8.46 | 0.0000 |  |
| BPDia1 | 0.0267 | 0.0147 | 1.81 | 0.0705 |  |
| Race1\_Black | 1.7514 | 0.5359 | 3.27 | 0.0011 |  |
| AlcoholYear | -0.0011 | 0.0016 | -0.71 | 0.4747 |  |
| DaysMentHlthBad | -0.0499 | 0.0201 | -2.48 | 0.0131 |  |
| R-squared: 0.2917  Adjusted R-squared: 0.2862  F-statistic: 52.81 on 10 and 1282 DF, p-value: < 2.2e-16 | | | | |  |



**{ Summary statistics of model 3. F-statistic is 52.81 and P-value is <2.2e-16 MPSE is 0.031}**



**{ Plot summary of model 3. Visible improvement with linearity in the QQ plot}**



**{ Histogram of model 3. By applying log transformation to *DirectChol*, we achieved**

# Bell-Shaped curvature}

**FINAL MODEL**:

𝑙𝑜𝑔(𝐵𝑀𝐼) = 3. 268000 − 0. 320200[𝑙𝑜𝑔(𝐷𝑖𝑟𝑒𝑐𝑡𝐶ℎ𝑜𝑙)] + 0. 109500[𝐷𝑖𝑎𝑏𝑒𝑡𝑒𝑠\_𝑌𝑒𝑠]

0. 069800[𝐻𝑒𝑎𝑙𝑡ℎ𝐺𝑒𝑛\_𝐸𝑥𝑐𝑒𝑙𝑙𝑒𝑛𝑡] − 0. 055290[𝑇𝑉𝐻𝑟𝑠𝐷𝑎𝑦\_0\_𝑡𝑜\_1\_ℎ𝑟] + 0. 001154- [𝐵𝑃𝑆𝑦𝑠3] +

0. 000898[𝐵𝑃𝐷𝑖𝑎1] −0. 000178[𝑇𝑒𝑠𝑡𝑜𝑠𝑡𝑒𝑟𝑜𝑛𝑒] + 0.050130[𝑅𝑎𝑐𝑒1\_𝐵𝑙𝑎𝑐𝑘] - 0.000017[AlcoholYear] -0.001954[DaysMenHlthBad]

Therefore, considering *DirectChol* as the primary predictor to predict our response variable of interest, *BMI*, we can conclude that a 1% increase in the direct HDL cholesterol of an individual, aged 6 years and above, is associated with a 0.32% decrease in the body mass index of that individual, on average, holding other factors constant.

**Conclusion/Results:**

In conclusion, our research on the NHANES dataset identified a combination of highly correlated variables as the best predictors of BMI, including DirectChol, Diabetes, HealthGen, TVHrsDay, BPSys3, Testosterone, BPDial, Race1, AlcoholYear, and DaysMentHlthBad. By conducting numerous regression tests and taking the log of some of these values, we were able to improve the accuracy of our predictions.

According to our findings, the variables that had the highest correlation with BMI were the most effective in accurately predicting BMI. Furthermore, taking the log of some of these variables helped to better fit the regression model and address the skewness of the data.

In the future, we plan to replicate our research by searching for similar data in previous and potential future NCHS datasets. Additionally, we hope to explore the potential of a single value that may be more closely related to the collection of values that make up BMI, in order to further support the validity of our conclusions.