Obesity and how to prevent it*

Arsh Lakhanpal

27 April 2022

Abstract

This paper will study to correlation between eating habits and activities that an individual takes part in throughout their day. We observe that the the individuals meals per day, their physical activity frequency, their consumption of food between meals and consumption of alcohol to be the main factors which attributed to higher BMIs. This is significant because it can inform the readers about activities they can do or not do to live a healthy life. The consequences of obesity go past one's physical health, it includes their mental health, sleep and the environment.

Keywords: obesity, body mass index, weight, food, health

1 Introduction

Obesity is defined as the "abnormal or excessive fat accumulation that presents a risk to health (Organization, n.d.)" Although not as large a problem in previous years, as of 2017, over 4 million people have died as a result of obesity. According to the World Health Organization, an individual with a body mass index (referred to as BMI) that is greater than 30 is considered obese and an individual with a BMI greater than 25 is overweight (Organization, n.d.). Similarly a BMI of less than 18.5 suggests that the individual is underweight. This value is calculated by dividing the individuals weight in kilograms by the square of their height which is measured in meters(Canada 2022). The reason that BMI will be used in this paper is because of the fact that it accounts for the height of an individual which provides a standard for us to compare people with different heights. This allows results of this study to be relevant as height also is a determining factor in an individuals weight.

In this report, from data collected from individuals living in Peru, Mexico and Columbia, I plan to examine various factors and determine their significance in a person being overweight or obese. I start out my data section by creating a number of graphs and studying a lot of the variables from the data set. From this, I am able to identify the most significant variables that play a role in an individual having a higher weight. To be precise, this is because from the formula of BMI, a higher weight would indicate a higher BMI and thus resulting in someone being overweight or obese. With the variables that are most significant, I construct a linear model. With the benchmarks set by the BMI, this model could be used in determining what factors and the severity of those factors play a role in someone being overweight or obese.

Section 2 of this paper talks more about where and how the data was collected and if any modifications were made to this data set for the purpose of this study. This includes the construction of new variables such as "BMI" and removal of other variables. In section 3 of this paper, we will make a linear model with the predictors we choose to be relevant and show the results of us testing for the assumptions in section 4 of this paper. Section 4 also presents many graphs to justify the choices of the predictors that were chosen. Section 5 is the discussion section where we talk about possible biases and ethical concerns in the dataset and model. We also mention how problems with the dataset transfer over to the model and some next steps we can take to improve the validity and accuaracy for our model as well.

*Code and data are available at: https://github.com/lakhan99

2 Data

This dataset was obtained from the UCI Machine Learning Repository and was donated on August 27, 2019 (Palechor and Hoz Manotas 2019a). It consists of 17 variables and 2111 observations, all based of individuals aged 14-61 who lived in Mexico, Peru and Colombia. This data set was used for a paper by Fabio Palechor and Alexis Manotas in which they simply presented the data that was collected (Palechor and Hoz Manotas 2019b). The responses were conducted using surveys which can be found in the appendix of this paper. I worked on this data on R (R Core Team 2020) and used readr (Wickham, Hester, and Bryan 2022) to help load the data. Other packages such as tidyverse (Wickham et al. 2019) and dplyr (Wickham et al. 2021) were used to clean the data whereas ggplot2 (Wickham 2016) was used to construct the graphs in this paper. knitr (Xie 2021) and kableExtra (Zhu 2021) were all packages that were used to help make the tables in this paper whereas patchwork (Pedersen 2020) was beneficial in captioning the graphs. Finally, the linear model was created using the stats package, something included in base R and the car package (Fox and Weisberg 2019).

2.1 Data Cleaning and Modifications

The raw data for this dataset did not contain any empty or ambiguous responses however, I did run the code to omit any "N/A's" that may have been in the dataset. The gender variable was also removed since it was insignificant to the study being conducted. Although women are more likely to become obese in comparison to men in general, this statistic does vary for many countries (Kanter and Caballero 2012). Along with that, this paper aims to focus on decisions people make in their day-to-day lives which involve their diet and physical activity, something which is accessible to both genders included in the study.

Variables with an individuals height, weight, age and their family history with obesity were included in the data along with responses about their eating habits. These included their,

- Consumption of high caloric food
- Frequent consumption of vegetables
- Consumption of food between meals
- Consumption of water daily
- Consumption of Alcohol

There were also questions on their physical habits. The habitual questions were based on their:

- Calorie consumption monitoring
- Transportation Method Used
- Physical activity frequency
- Time using technological devices

There were various modifications that needed to be done for the data. The first modification was creating the "BMI" variable as the dataset itself only consisted of the variable in which it stated the weight class for every individual. To do this, as per the formula to calculate BMI, every individuals weight (kilograms) was divided by the square of their height (metres). This variable was important for the modeling part of this section as the response variable for this study had to be a numerical value. Once the BMI variable was created, the age and height variables were removed from the dataset as our model focuses on predicting a specific proportion between these variables and so, neither of these variables could be used as predictors for the model. Other variables that were created were those which pertained to whether or not an individual consumed high caloric food frequently, their consumption of food between meals, their family history with obesity, their consumption of alcohol and their method of transportation. These were all variables that were in the dataset however to use these variables in our linear model, we had to give numerical values to the responses that were recorded. The response for an individuals family history with obesity was either "yes" or "no" and the new variable simply records these as "1" and "0" respectively. However, for variables such as consumption of food between meals had responses as follows:

- "Frequently"
- "Always"

- “Sometimes”
- “no”

and combined the postive responses (frequently and always) and negative responses (sometimes and no) so that we have a binary response instead. This was also done for the alcholol consumption variable. A variable for frequency of physical activity was also created which took responses in the numeric form to responses such as “4 to 5 days” which helped with modelling Figure 4. Finally, an indicator variable was created to aid with splitting our data into training and test data sets for model validation.

2.2 Data Visualization

There were only two numeric variables in this dataset that we could visualize being BMI and Age of the population. From the Figure 1, we see a very strong right skew in terms of the distribution of the variable. This is something important to consider as this suggests that the majority of the population is young meaning that the results of this study would be based on habits of a younger individual. Any conclusions made would have to be made with this in mind. Since we are considering the different classes the BMI would put one in, the weights are also highlighted. The trends we see make sense since as one gets older, they begin to care about their weight more and actively take steps to reduce their weight. This is shown as the proportion of obese and overweight individuals decrease as we consider the higher age ranges. Looking at the relation between age and BMI as shown in Figure 1 , due to the right skew of the age variable, a majority of the points are situated on the left side of the graph however, the higher ages are generally associated with lower levels of BMI. Finally, Figure 2 is a histogram of the BMI variable where we notice a multi-modal distribution with various peaks in all classes, the highest one being in the overweight class.

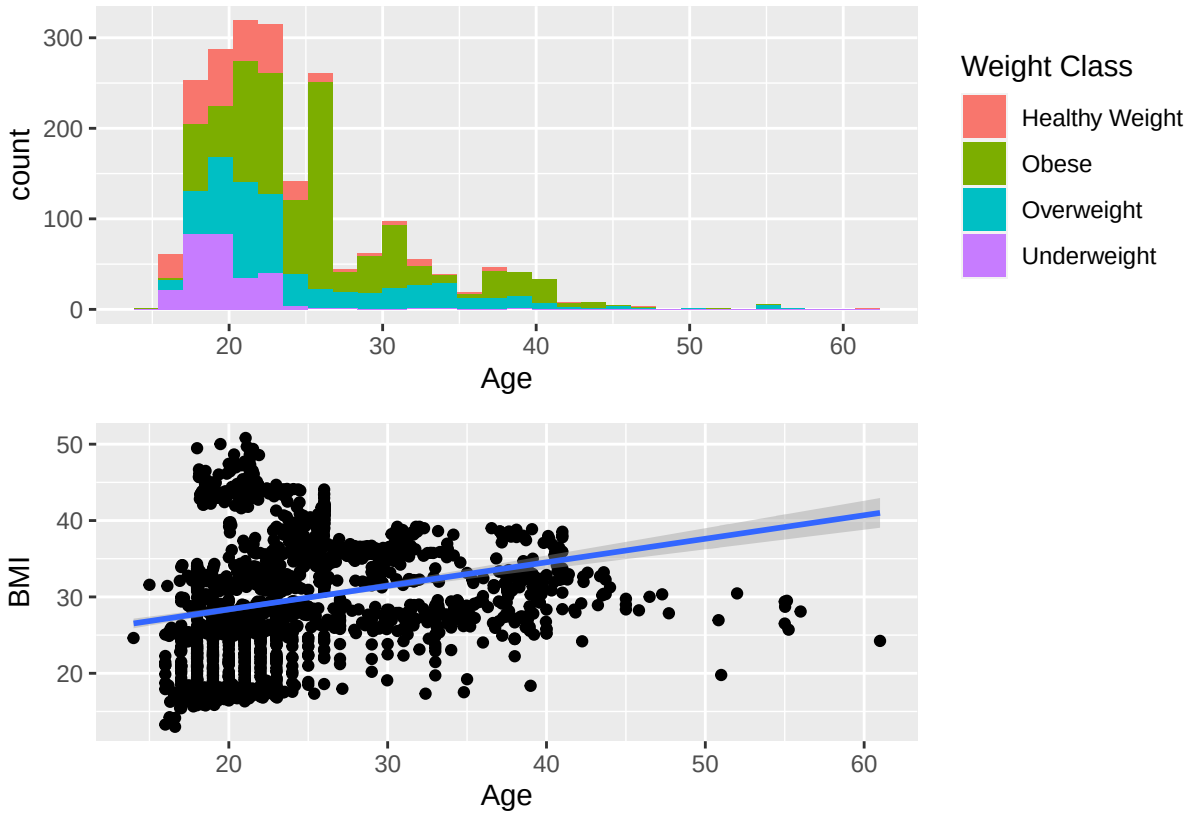


Figure 1: Distribution of Age predictor

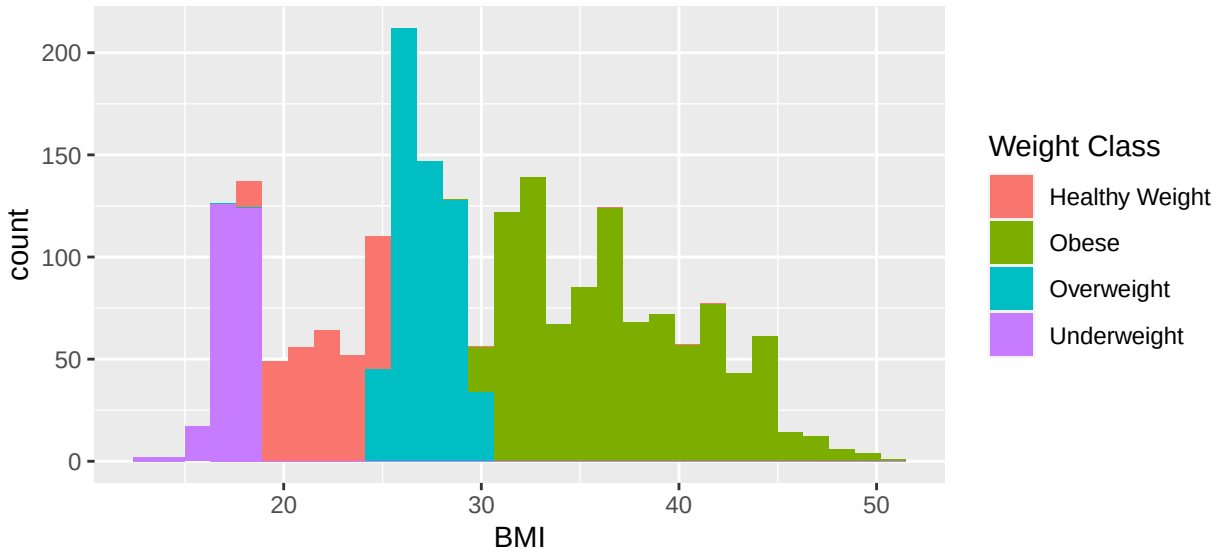


Figure 2: Distribution of BMI

3 Model

Table 1: Number of predictors for every model obtained from backwards selection with its adjusted R squared

	Number of Predictors	Adjusted R Squared
Model 1	11	0.4201
Model 2	10	0.4206
Model 3	9	0.4201
Model Final	7	0.413

The purpose of the model is to be able to find the most relevant predictors that can affect one's BMI. Since BMI is a numeric response variable, linear regression seemed to be the correct choice of model as opposed to a logistic regression which is more beneficial when the response variable is binary. The data was randomly split into the training and test datasets with 1056 obervations in the training dataset and 1054 observations in the test. The training dataset is what we do all of our testing and build our models with while the test dataset serves as a method for us to validate our model. If the results based off our training and test datasets are similar, our model is validated.

As stated in the data section, the new variable BMI was created for this and many of the responses were re-coded to have number values with the understanding that the higher numbers are indicative of actions that are more likely to increase one's BMI. One thing to note is that the method of transportation was excluded from this model as it is difficult to quantify the different methods of transportation and decide which is best for each person. This is highly subjective and can be different for everyone.

To decide which model was best, the backwards selection method method was used with the significance level of 0.01. This means that any predictor with a p-value higher than the significance level of 0.01 was taken out of the model. Each predictor was removed one at a time and this process was repeated until the remaining predictors in the model had a p-value less than the threshold of 0.01.Along with this, research on healthy eating habits was also done to determine which predictors were most important significant to appropriately create the model. Table 1 looks at the different models, the number of predictors each model has and also the adjusted r squared of all these models. The reason we want to look at adjusted r squared is because we

can compare how well our model fits with models consisting of a different number of predictors. From this, we can see that that model 4 as it has the least predictors making it easier to interpret compared to the other models. Along with its simplicity, the margins between the adjusted r squared of that model compared to model 1 (the model with all predictors) is small. This model was made of the the following predictors:

- Family history with Obesity (Yes/No)
- Frequent consumption of High Caloric Food (Yes/No)
- Frequent consumption of vegetables (Always/Sometimes/Never)
- Consumption of food between meals (Frequent or Always / Sometimes or No)
- Calorie Consumption Monitoring (Yes/No)
- Frequency of physical activity (4-5 days/ 2-4 days/ 1-2 days/ Never)
- Age (numerical)

This also means that assumptions for linear regression were also checked as this is a vital step so that we can make inferences off the results of our model. We aren't able to check for multi-collinearity since the model that we use only have one numerical variable (Age). The results of the model will be presented in the next section of this report. This section will also present various tables and graphs which helped show which predictors were relevant in making this model.

4 Results

4.1 Choice of predictors

Table _ looks takes a look at the individuals who count their calories. We notice that between both the overweight and healthy weight groups, only 2% of people keep track of their calories whereas, in the healthy population, about 12% of the population monitors their calories. when looking specifically at the obese population, of the 971, only 3 of them track their calories.

Table 2: Monitoring of Calories for Obese and Overweight vs Healthy Weight individuals

	Obese or Overweight	Healthy Weight
Yes	37	37
No	1503	262
Total	1540	299

When considering calories, one must also consider the diet of these individuals. Table 3 looks at specific diet choices both groups of people make. It shows that 515 people who are either obese or overweight have a high consumption of calories and drink less than 2L of water. 492 of them consumes high calories and either sometimes or never eat vegetables. On the contrary, these statistics are 71 and 133 for those who are have a healthy weight showing a clear correlation between the diet statistics in this dataset and number of obese individuals.

Table 3: Diet choices for Overweight and OBese versus Healthy Weight individuals

	Obese or Overweight	Healthy Weight
High Consumption of Calories and drinking less than 2L of water	515	71
High Consumption of Calories and Sometimes/Never having Vegetables	492	133

However, their diet isn't to be confused with amount of meals they eat in a day because Figure 3 shows how people who work out in a healthy weight range eat between meals and have a relatively less amount of people that are obese provided that they workout atleast three times a week. The main conclusion taken from Figure 3 is that people who have a lower frequency of physical activity per week and eat less frequently between meals are either overweight or obese. Of the overweight and obese population, a strong majority of

the population “sometimes” eats between meals compared to the health weight population where more of them eat more frequently.

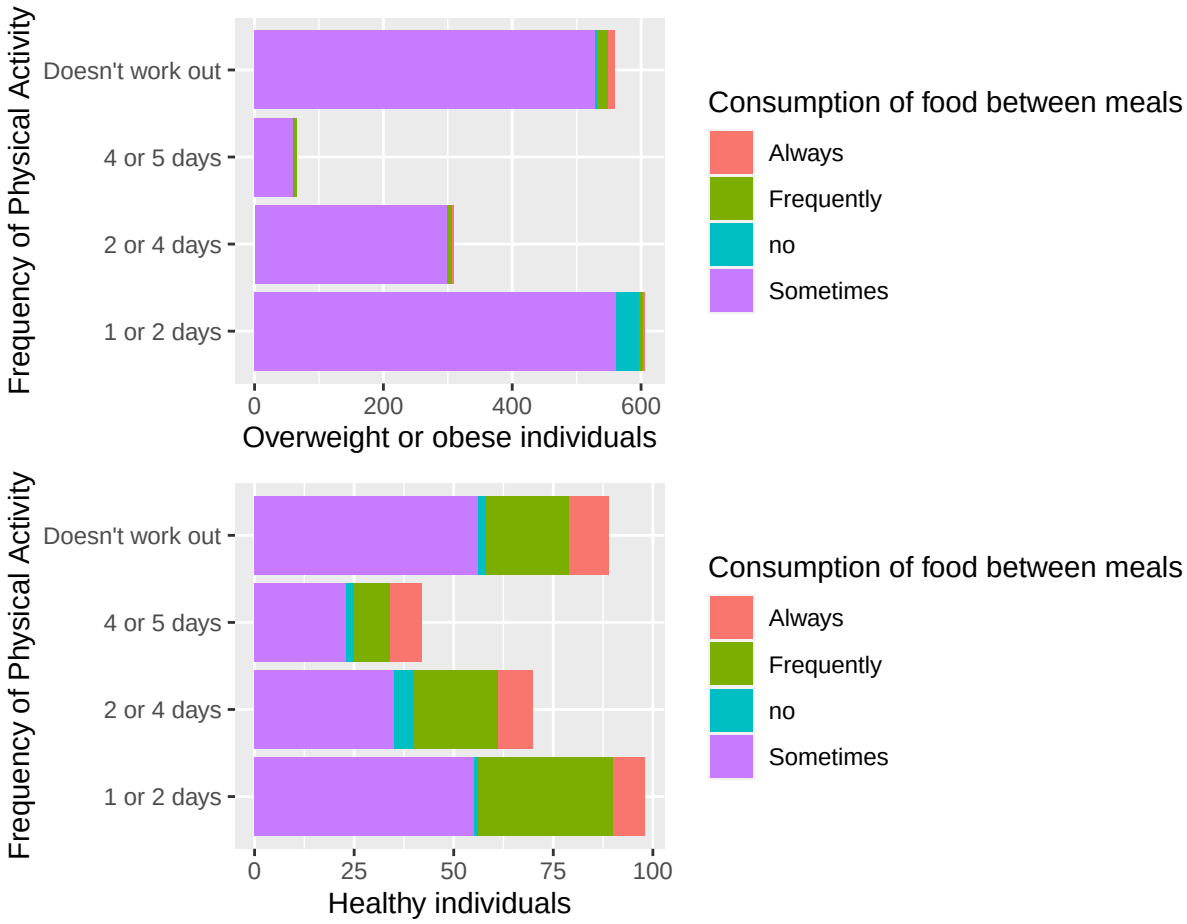


Figure 3: Affect of physical activity and frequency of meals on overweight and healthy indivudals

Figure 4 is relevant because it shows that the overweight population in general works out less compared to the healthy weight population while also consuming more calories making both of these variables significant predictors. The first graph shows that more than 1100 people workout less than twice a week and have a frequent consumption of high caloric food whereas the second graph shows that less than 200 people in the healthy weight workout less than twice a week. This graph also has a lower proportion of people who are eating high caloric food.

Finally, Figure 5 outlines the impact of having family history with obesity. We see that of overwight and obese population, those who have a family history with obesity are much more likely to be obese and overweight compared to those without obesity in their family. This is informative of the role of genetics in determining one’s BMI.

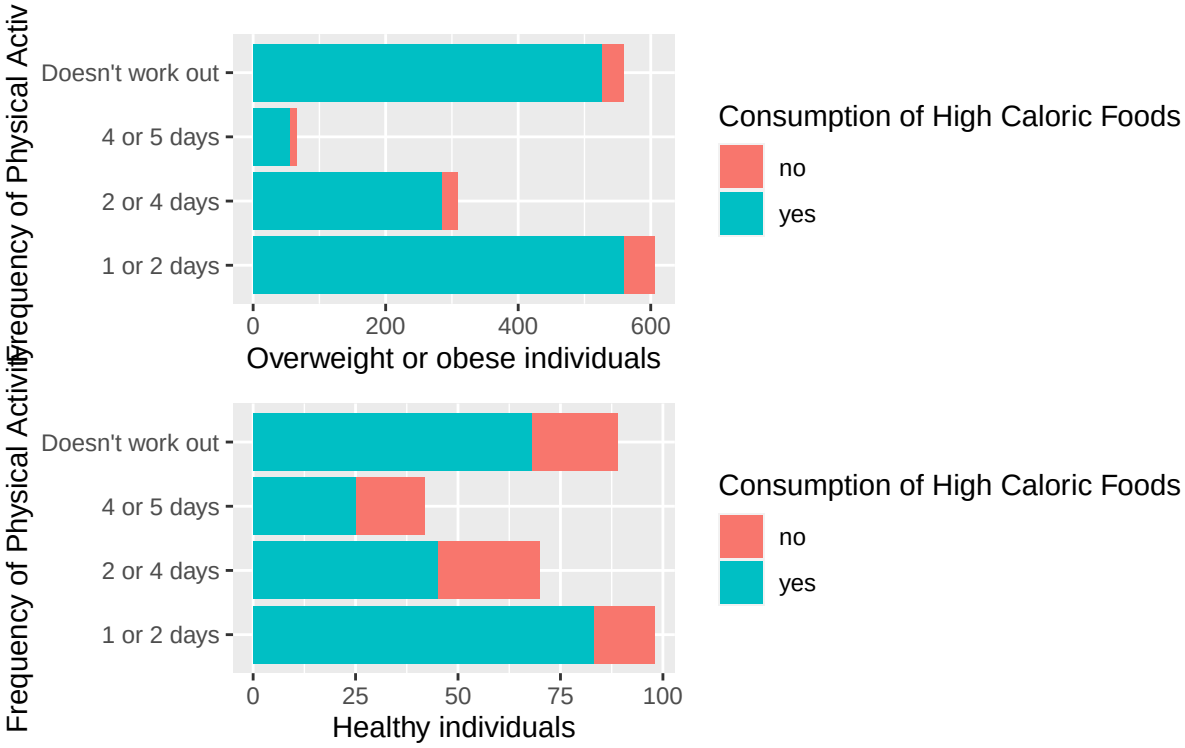
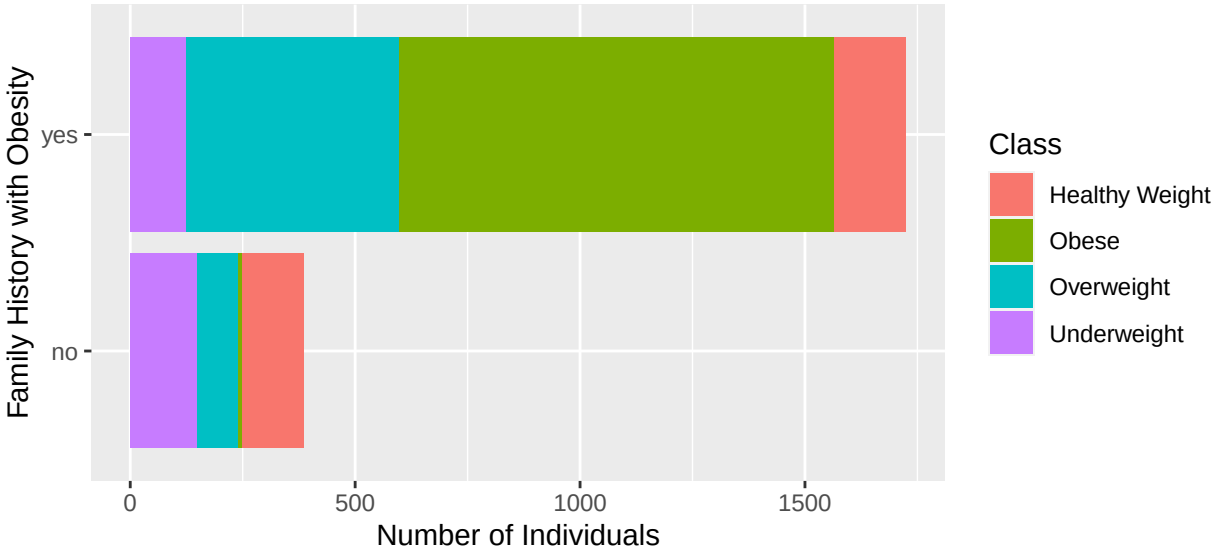


Figure 4: Affect of physical activity and consumption of high caloric foods on overweight and healthy indivudals


Family History with Obesity
yes
no
0
500
1000
1500
Number of Individuals
Class
Healthy Weight
Obese
Overweight
Underweight

4.2 Model Assumptions and Validation

From Figure _, the true relationship between BMI and the fitted values of the model seem to be linear while the residual plots show randomness which is ideal. This means that the uncorrelated errors and constant variance assumption is also satisfied. One thing to note is that the residual plot for the Age variable is also skewed right, something that we noticed when visualizing our data. Finally, since a majority of the points align very closely with the line, the normality distribution is also satisfied.

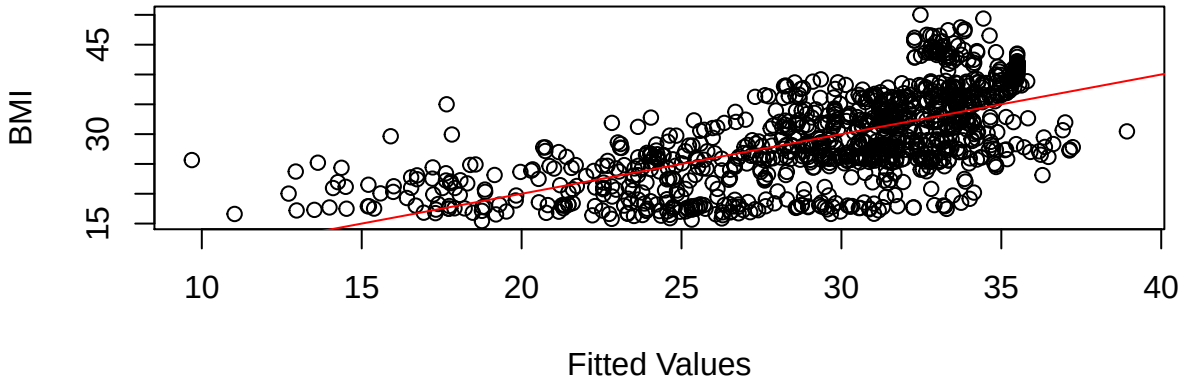
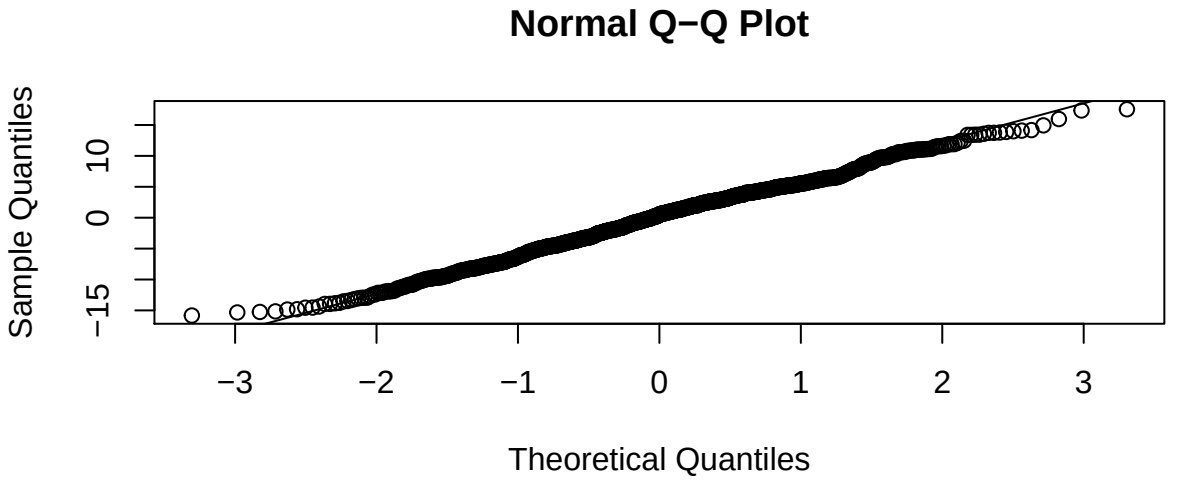


Figure 5: Response vs Fitted Values

```
## integer(0)
```



A boxcox transformation was also applied to this model in hopes of finding a model with a better fit however, the differences in results were minimal so the original model will be used so that it can be interpreted and understood clearly. The fitted values versus BMI plot and residual plots remained similar to the untransformed model in shape. However, the QQ plot for this model is worse as less of the points are on the line. The graphical results for the transformed model can be found in the appendix of this paper.

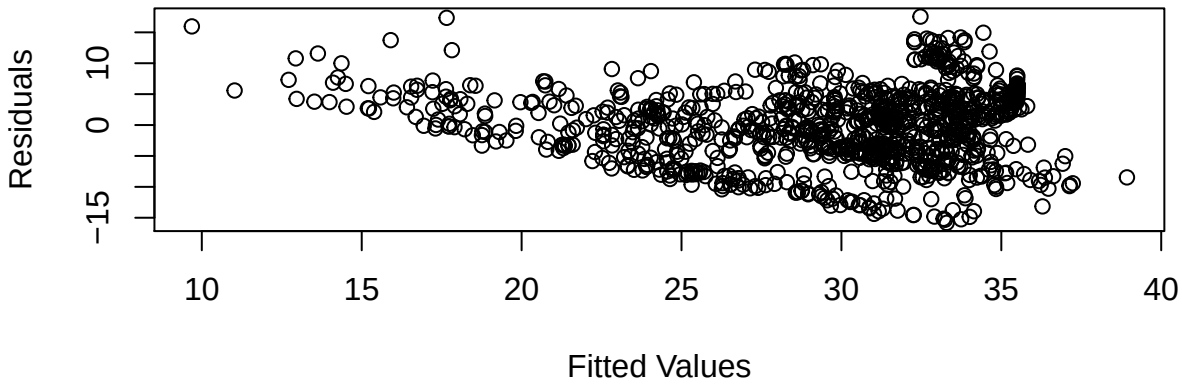


Figure 6: Residuals versus Fitted Values to check uncorrelated errors and constant variance assumption

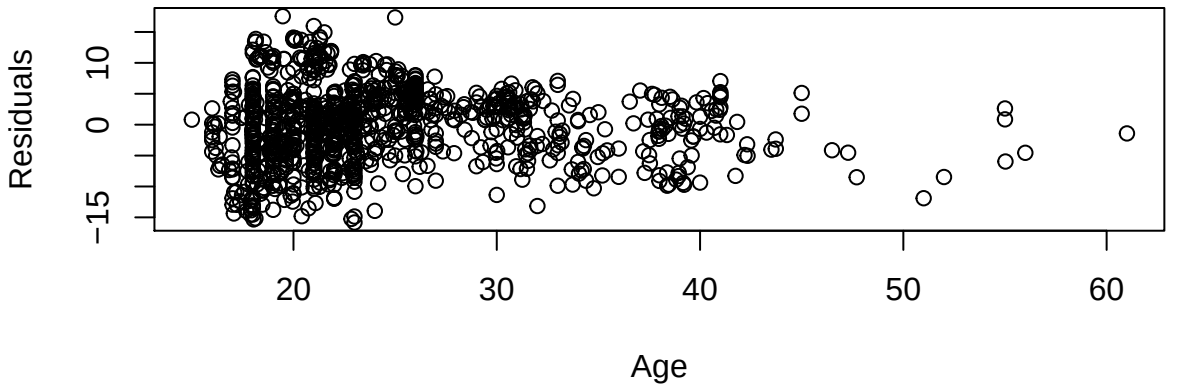


Figure 7: Residuals versus Age to check uncorrelated errors and constant variance assumption

The final step was validating the model by doing the exact same steps with the test dataset. When comparing the same model with both the training and test dataset, we notice that both have similar model statistics. Table 2 compares key statistics for both models and due to the similarities between both these models, we are able to validate our model.

Table 4: Key model statistics between training and test datasets

	Adjusted R Squared	Multiple R Squared	RSS
Model Train	0.413	0.4213	5.974
Model Test	0.446	0.4496	6.103

4.3 Interpreting the model

The way to interpret this model is to look at every individual variable individually and hold other variables fixed. From this, the model tells us that:

- If the individual has obesity in their family, they can expect their BMI to increase by approximately 6.04 with a standard error of 0.53
- If they eat high caloric food frequently or always, the individual can expect their BMI to increase by 2.15 with a standard error of 0.60
- Consumption of vegetables atleast sometimes would increase BMI by 3.52 with a standard error of 0.35
- Consumption of atleast 1-2 meals will increase BMI by 0.67 with a standard error of 0.24.
- Frequently or always consuming food between meals will decrease BMI by 7.43 with a standard error of 0.55
- Monitoring your calories will decrease BMI by 2.73 with a standard error of 0.95
- Working out atleast once per week will decrease BMI by 1.17 with a standard error 0.23
- For an increase of age by one year, BMI would increase by 0.13 with a standard error of 0.03.

5 Discussion

The purpose of this model was to find build something in which we can identify the most prominent factors that can lead to one being overweight or even obese. After the completion of the model there are many things that one can learn but also many things that can be done to improve this model and get more accurate results for this study. The model is slightly successful in terms of detecting a linear relationship between the predictors and response variable. However, there are some results from the model which contradict prior research and must be considered.

5.1 Obesity in family and its effects

From the model, the predictor which plays the largest role in increasing one's BMI significantly is the predictor which considers if obesity is already something common in your family. Someone with obesity in their family has a 70% risk for the disease(Golden and Kessler 2020). Figure 5 of this paper shows the effect that genetics has on the obese population since the largest class of body weight for the obese population is the obese class. This could be closely related to those with diabetes as well as it is proven that those with diabetes also have higher BMI's and higher concentrations of cholesterol and glucose than the regular person(M A van der Sande 1 2001)However, lifestyle choices are equally as important for these families with history with these diseases. A child can be born at a normal weight range but because of the choices made by parents, that child can end up higher on the BMI scale to the point where they also become obese. A study conducted by the 1958 British Cohort shows that if a parent has a high BMI during childhood and adulthood, the risk for their child to become overweight or obese is also heightened(Nielsen L.A 2015).

This can happen for a number of reasons, one of the main ones being that "junk food" is widely available and also relatively cheap compared to the foods that are healthier for one's body (See 2020). Another reason includes the socio-economic status of these families. Due to the cheap prices of these foods, sometimes these foods are all a family can afford. Kids with parents that have a lower level of education are more likely to become obese than children with parents having higher levels of education which is an important consideration to consider (Anke Hüls 2021). This study focusing on residents in Mexico, Peru and Columbia, countries with poverty rates of 44%, 20% ("Poverty & Equity Brief, Peru, Latin America & the Caribbean" 2020) and 40% ("Colombia Poverty Declined in 2021, but Still Above Pre-Pandemic Levels" 2021) respectively would explain a large proportion of the respondents having family history with poverty.

5.2 Consumption of food between meals

From this model, we are able to conclude that consuming food between your main meals can actually help lower your BMI by 7.43 which is slightly counter intuitive at first glance. One would assume that eating food along with the main meals would be something that would increase the calorie intake however, it actually works in reverse. The benefit of consuming more between meals is that it kills your hunger for the next meal ("The Science of Snacking," n.d.). This means that when you do eat your next meal, because of the snack you ate earlier, there is a lower probability that one eats more than necessary thus keeping their calorie intake lower. The problem with eating less between meals is that you feel you should compensate so that you have enough till the next time you eat. This becomes risky from a diet point of view because easier to attain junk or fast foods become more attractive as they will satisfy you the fastest ("How Often Should You Eat?" 2013).

However, it is important to make note of what you eat between meals. Regularly eating junk food between meals will increase your BMI because junk food is high in calories and from our model, a frequent consumption of high caloric food is associated with a BMI increase of 2.15. When looking at Figure 3 of this paper, we notice that most of the overweight and obese population only sometimes eat food between their meals. When looking at the healthy weight population, we see the proportion of individuals who "Always" or "Frequently" snack increase, likely because this population understands the benefits of snacking.

5.3 Bias and Ethical Concerns

Any data based on people carries a level of bias and ethical concern which limits the extent to which we can rely on this model. One thing that limits the reliability of this model is the population that was surveyed. As stated in the data section, the respondents were citizens from Mexico, Peru and Columbia so we cannot claim that this model will work for any population across the world. There are different factors that we must consider when using this model for populations in different countries, one of them being socio-economic class of the population. Although the socio-economic class of a person is something the dataset didn't collect, we know that Mexico, Peru and Columbia are countries with higher rates of poverty than countries like United States or Canada. Along with this, culture plays a large role the types of foods people in Latin America may eat and the practices they may follow. For example, in Latin American culture, food is generally eaten at a slower pace ("10 Differences in Latin Culture Compared to U.s. Culture" 2020) while chatting whereas in North American culture, with the high availability of fast food, we generally eat faster. The benefit to eating slower is that it allows for food to digest and can prevent people from overeating compared to North American culture where we are often left wanting more ("All About Eating Slowly," n.d.). This means that if the same model was built of a North American population, perhaps some of the statistics we see regarding weight or number of meals in a day could be entirely different.This makes generalizing this model for other populations difficult to do as it could produce results which are inaccurate.

This dataset also restricted its gender variable to only men or woman which disregards a lot of other communities. The problem with confining the gender variable to either men or woman is that it hinders with the accuracy of our data. There could be cases where someone who is neither male nor female chooses one of the options presented to them leaving us analyzing data which isn't actually true. The other problem with this would be the fact that if individuals that aren't male or female feel uncomfortable with selecting male or female, they may not provide us with their data at which point we are losing data.

There are also many areas in the data where a categorical response such as "Always" or "Frequently" were given numerical values such as "3" and "2" and the frequent consumption of vegetables variable is an example of that. It is evident that the intention was to quantify the response to use it for their results however, assigning these values arbitrarily adds a level of bias to the dataset. We have to question why a response such as "Frequent" got assigned a value of "2" as there is not any explanation justifying this choice. There is also a lot of ambiguity with responses such as "Frequently" and "sometimes" because it does not specify how "frequent" and when working with statistics, you want to know little details like this to get the most accurate results possible.

5.4 Weaknesses and Limitations of Model

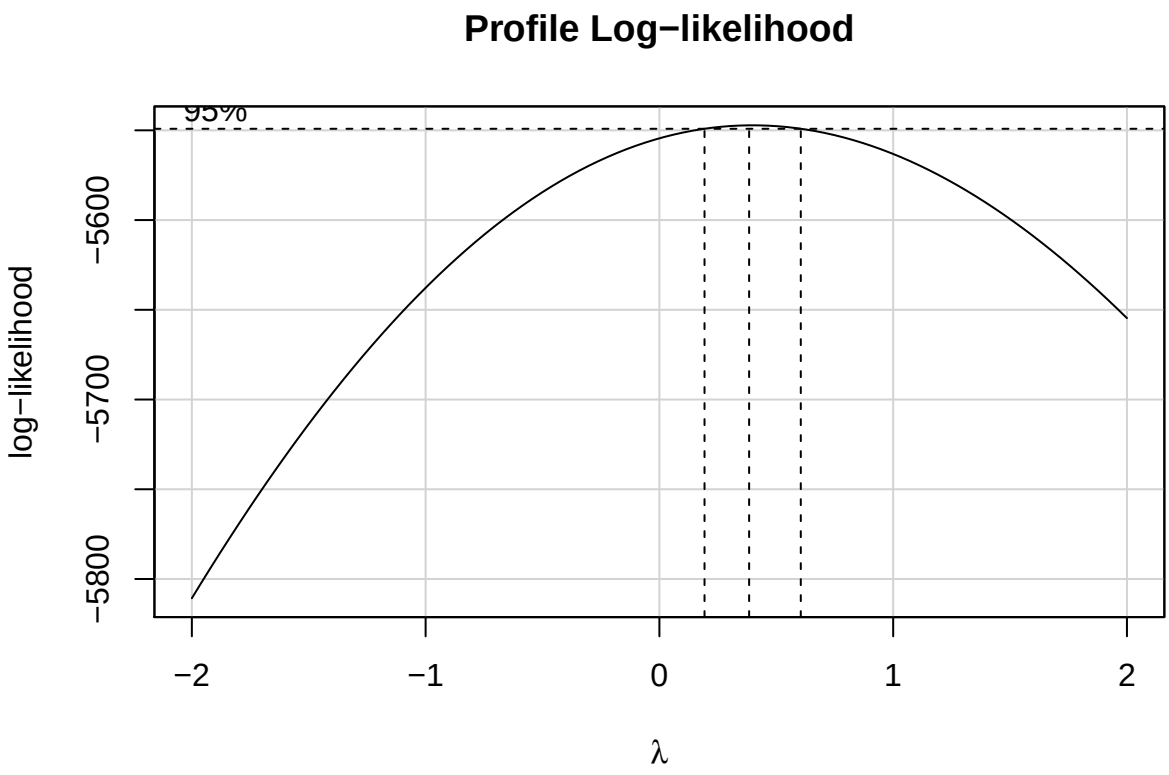
Due to the weaknesses in the data highlighted in the previous section, there are also some weaknesses in the model which are important to highlight. The first weakness is that BMI is not a perfect measure of an individuals weight class. This is because BMI takes the weight that is written on a scale but it doesn't specify if the weight is muscle or fat. The problem with this is that muscle weighs more than fat because it is more dense ("Why Bmi Is Inaccurate and Misleading" 2022). This becomes a problem because BMI isn't able to distinguish that difference which is misleading in terms of which weight class an individual may be in.

Another limitation of this model is that it doesn't take stress and sleep into account as it is something which was not measured for these individuals. Both of these are factors which play a role in an individual being overweight ("What Causes Obesity & Overweight?" 2021) so the fact that the model does not consider these factors is a limitation of this model. In a similar manner, the model states that a frequent consumption of vegetables leads to an increase in BMI by 3.52. From an article titled "The Relationship between Vegetable Intake and Weight Outcomes: A Systematic Review of Cohort Studies" by Monica Nour and Sarah Alice Lutze, they find that there is a negative correlation with consumption of vegetables and weight gain ("The Relationship Between Vegetable Intake and Weight Outcomes: A Systematic Review of Cohort Studies" 2018). The reason that vegetables are encouraged when try to lose weight is because they are often low in calories but high in fiber which keeps you full throughout the day. Them being low in calories allow for a greater volume of vegetables to be consumed while also keeping the calorie intake low ("The Relationship Between Vegetable Intake and Weight Outcomes: A Systematic Review of Cohort Studies" 2018). This is likely due to the arbitrary values given to the responses in that observation.

The last limitation of this model is a result of data modification. When considering for non-binary and non-numerical responses that did not already have a numerical value associated with them, the four potential responses were grouped into binary responses. These were transformed into binary responses so that "Always and Frequently" were together and "Sometimes and no" were together. This was done so that we could work with binary responses instead of non-binary ones for ease of interpretation. Although this was done to help interpret the model, this eliminates accuracy from the model because it generalizes the positive and negative responses.

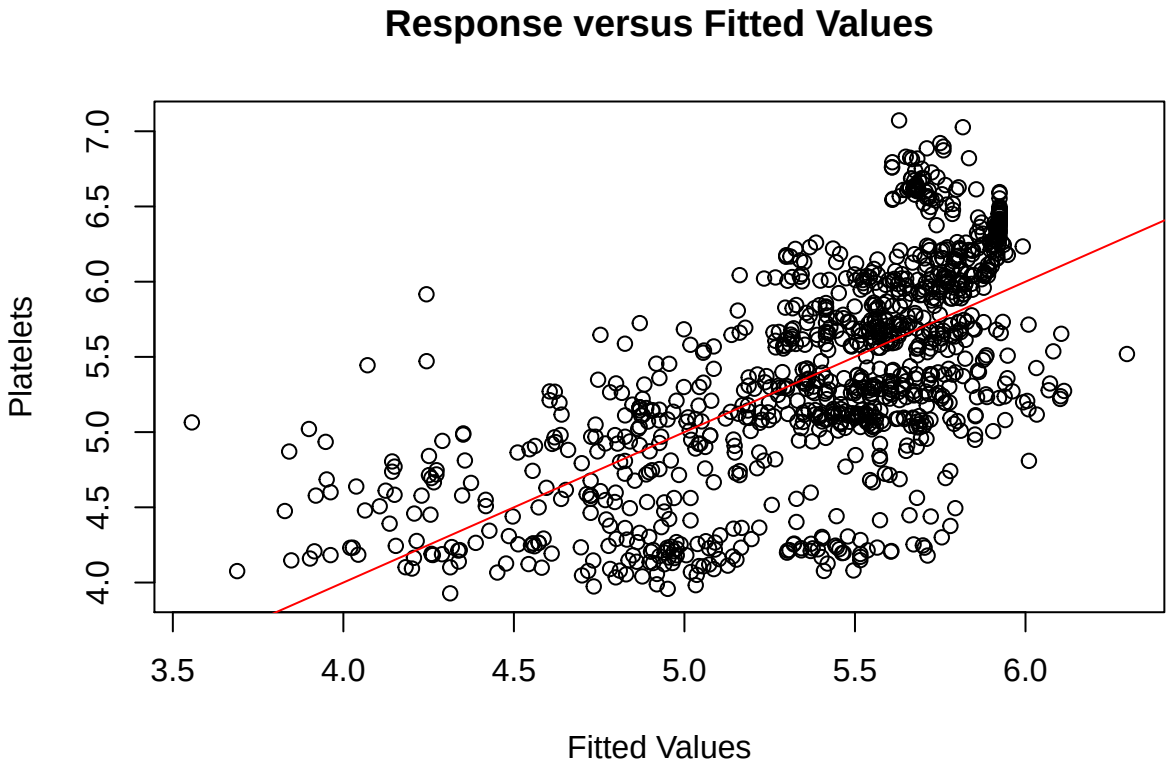
The many weaknesses in the model are apparent in the R squared value of this model. Although a linear trend does show, it isn't strong as outlined by the R squared value of 0.4213. In the future, it would be better to find a dataset that had binary and numerical variables as interpretations of any other type of variable in linear modeling becomes complicated. It would also be beneficial to consider factors such as sleep and stress as causes to one becoming obese. To get a more accurate measurement of one's weight category, waist-to-height ratio may be more beneficial but also more difficult to obtain. These types improvements are all likely to not only improve the accuracy of the model but the ability for the model to be used for prediction purposes.

Appendix

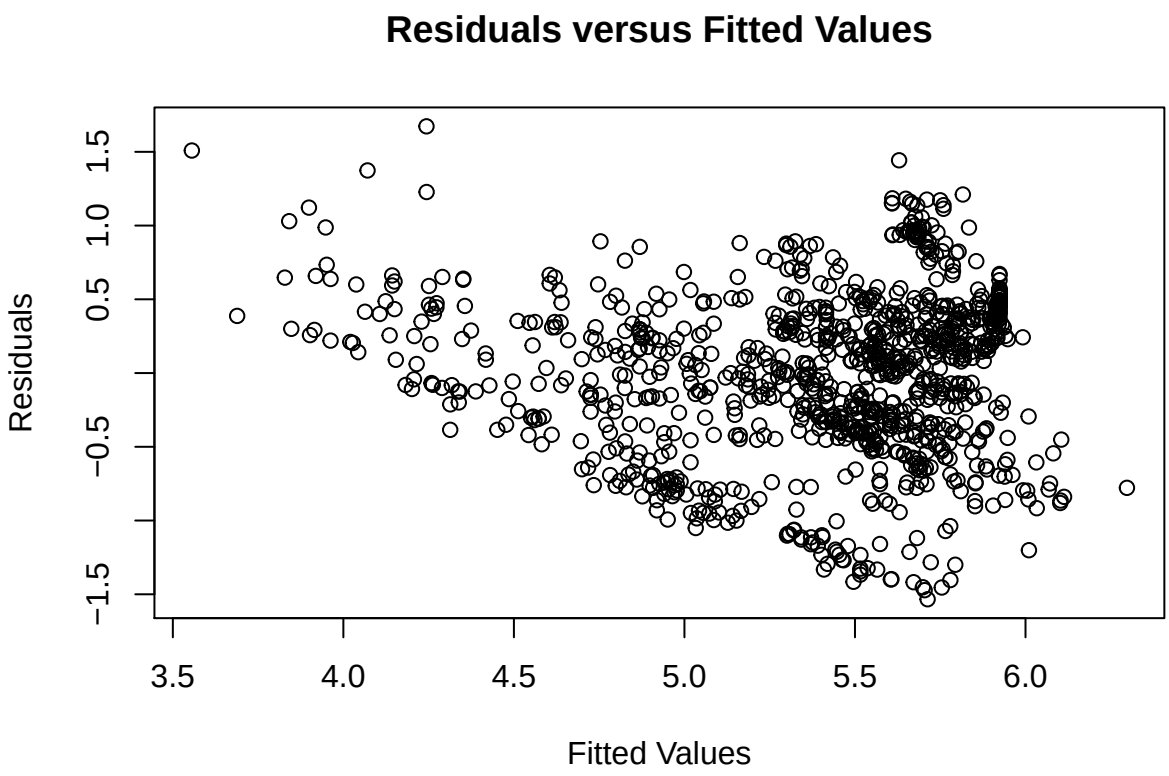


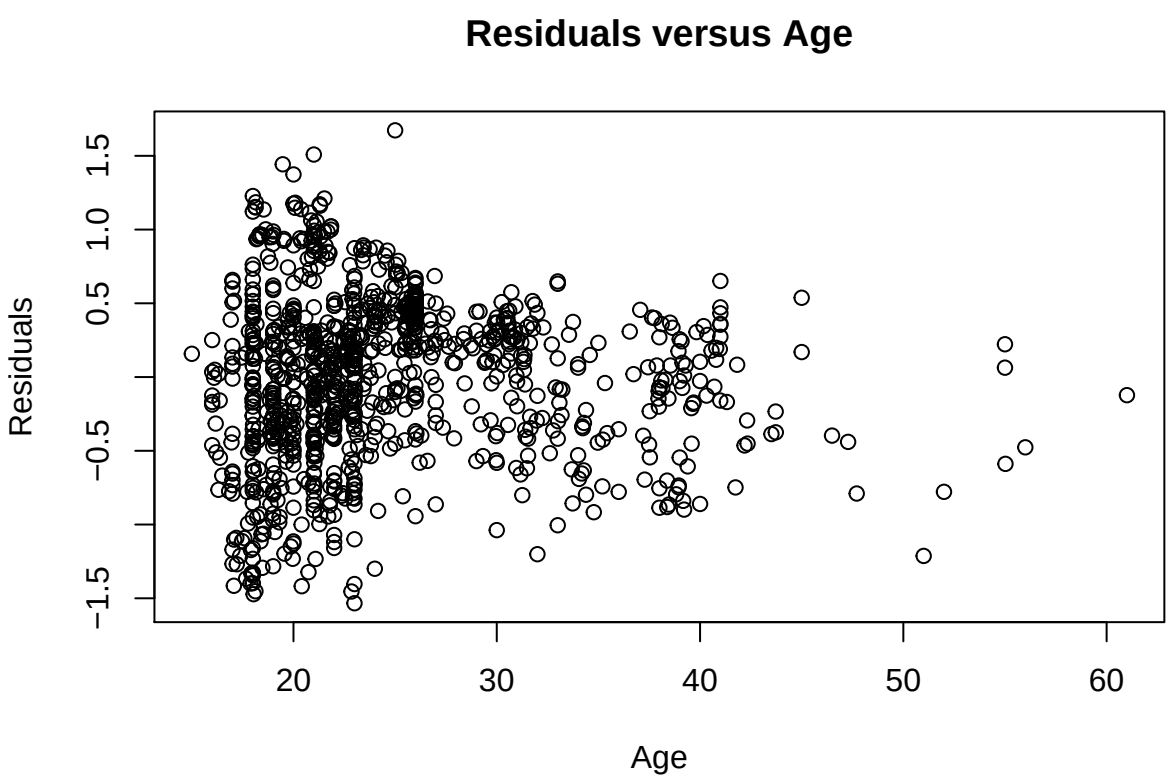

```
##
## Call:
## lm(formula = BMI ~ new_history + new_FAVC + FCVC + new_CAEC +
##     new_SCC + FAF + Age, data = train_modbox)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.53366 -0.37059  0.06917  0.37927  1.67271
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.897554   0.121152  32.171  < 2e-16 ***
## new_history  0.589407   0.048922  12.048  < 2e-16 ***
## new_FAVC     0.184514   0.055600   3.319 0.000936 ***
## FCVC         0.292911   0.032584   8.989  < 2e-16 ***
## new_CAEC    -0.689938   0.050400 -13.689  < 2e-16 ***
## new_SCC     -0.236585   0.088607  -2.670 0.007701 **
## FAF         -0.100974   0.020712  -4.875 1.26e-06 ***
## Age          0.014370   0.002631   5.461 5.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5523 on 1048 degrees of freedom
## Multiple R-squared:  0.4271, Adjusted R-squared:  0.4233
```

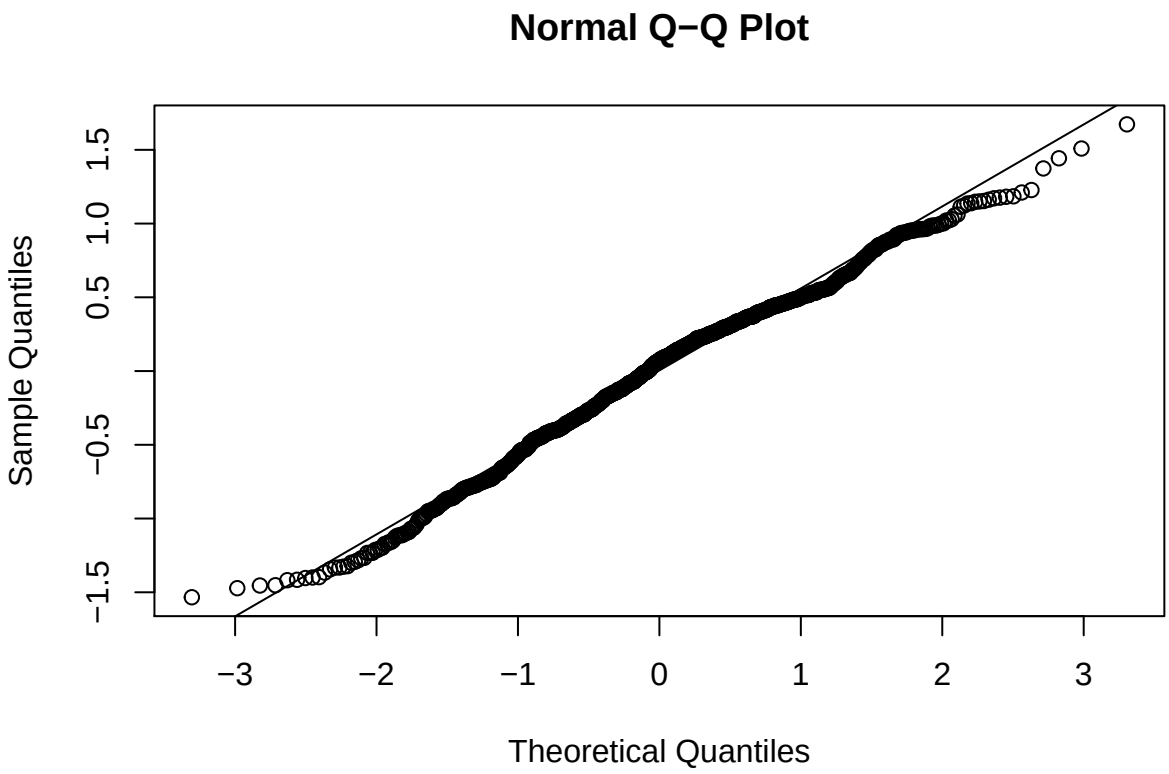
```
## F-statistic: 111.6 on 7 and 1048 DF,  p-value: < 2.2e-16
```




```
## integer(0)
```


Residuals versus Fitted Values
Residuals
1.5
1.0
0.5
−0.5
−1.5
3.5
4.0
4.5
5.0
5.5
6.0
Fitted Values


Residuals versus Age
Residuals
1.5
1.0
0.5
−0.5
−1.5
20
30
40
50
60
Age



A Additional details

References

"10 Differences in Latin Culture Compared to U.s. Culture." 2020. https://www.spanish.academy/blog/10-differences-in-latin-culture-compared-to-u-s-culture/.

"All About Eating Slowly." n.d. https://www.precisionnutrition.com/all-about-slow-eating#:~:text=The%20benefits%20of%20slow%20eating,weight%20gain%2C%20and%20lower%20satisfaction.

Anke Hüls, Leonie H. Bogl, Marvin N. Wright. 2021. "Polygenic Risk for Obesity and Its Interaction with Lifestyle and Sociodemographic Factors in European Children and Adolescents." https://doi.org/10.1038/s41366-021-00795-5.

Canada, Diabetes. 2022. *Body Mass Index (Bmi) Calculator*. https://www.diabetes.ca/managing-my-diabetes/tools---resources/body-mass-index-(bmi)-calculator.

"Colombia Poverty Declined in 2021, but Still Above Pre-Pandemic Levels." 2021. https://www.usnews.com/news/world/articles/2022-04-26/colombia-poverty-declined-in-2021-but-still-above-pre-pandemic-levels#:~:text=The%20share%20of%20Colombians%20living,the%20figure%20stood%20at%2035.7%25.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Golden, Angela, and Christine Kessler. 2020. "Obesity and Genetics." *J. Am. Assoc. Nurse Pract.* 32 (7): 493–96.

"How Often Should You Eat?" 2013. https://www.foodnetwork.com/healthyeats/diets/2013/07/how-often-should-you-eat.

Kanter, Rebecca, and Benjamin Caballero. 2012. "Global Gender Disparities in Obesity: A Review." *Adv. Nutr.* 3 (4): 491–98.

M A van der Sande 1, P J Milligan, G E Walraven. 2001. "Family History: An Opportunity for Early Interventions and Improved Control of Hypertension, Obesity and Diabetes." https://pubmed.ncbi.nlm.nih.gov/11357211/#:~:text=Those%20with%20a%20family%20history,obesity%20and%20diabetes%20was%20increased.

Nielsen L.A, Holm J, Nielsen T. R.H. 2015. "The Impact of Familial Predisposition to Obesity and Cardiovascular Disease on Childhood Obesity." https://doi.org/10.1159/000441375.

Organization, World Health. n.d. *Obesity*. https://www.who.int/health-topics/obesity#tab=tab_1.

Palechor, Fabio Mendoza, and Alexis de la Hoz Manotas. 2019a. *Estimation of Obesity Levels Based on Eating Habits and Physical Condition Data Set*. https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+.

———. 2019b. "Dataset for Estimation of Obesity Levels Based on Eating Habits and Physical Condition in Individuals from Colombia, Peru and Mexico." *Data in Brief* 25 (August): 104344. https://doi.org/10.1016/j.dib.2019.104344.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*.

"Poverty & Equity Brief, Peru, Latin America & the Caribbean." 2020. https://databank.worldbank.org/data/download/poverty/987B9C90-CB9F-4D93-AE8C-750588BF00QA/SM2020/Global_POVEQ_PER.pdf.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

See, Caitlin. 2020. "The Cost of Healthy Eating Vs Unhealthy Eating." https://plutusfoundation.org/2020/healthy-eating-budget/.

"The Relationship Between Vegetable Intake and Weight Outcomes: A Systematic Review of Cohort Studies." 2018. https://doi.org/10.3390/nu10111626.

"The Science of Snacking." n.d. https://www.hsph.harvard.edu/nutritionsource/snacking/.

"What Causes Obesity & Overweight?" 2021. https://www.nichd.nih.gov/health/topics/obesity/conditioninfo/cause.

"Why Bmi Is Inaccurate and Misleading." 2022. https://www.medicalnewstoday.com/articles/265215.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.