Explore the factors that may be associated with obesity

Xueyan Hu

2024-04-07

# 1. Summary/Abstract

Obesity is a chronic complex disease which may increase the risk of severe health issues and even cancers. And the condition of obesity in real life could be related to physical behaviors. The project is exploring the potential factors that might be related to obesity level using machine learning models. The data analysis result has shown that age, gender, family history, drinking water, physical activity, and commute transportation are all predictors for obesity level, which is represented by individual BMI. As for the outcome, when linear model is fit to data which has BMI as continuous variable, the performance is better than using obesity level as categorical variable.

# 2. Introduction

## 2.1 General Background Information

[Obesity](https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight) is a chronic complex disease defined by excessive fat deposits that can impair health. Obesity can lead to increased risk of type 2 diabetes and heart disease, it can affect bone health and reproduction, it increases the risk of certain cancers.

In the recent decades, obesity has become a more and more common and severe problem across the world.

Worldwide adult obesity has more than doubled since 1990, and adolescent obesity has quadrupled. In 2022, 1 in 8 people in the world were living with obesity.

There is no doubt that obesity is a multifactorial disease due to obesogenic environments, psycho-social factors and genetic variants in most cases, but personal behavior could also has correlation with the result of overweight and obesity.

There is a lot of discussion on the Internet and social media about fitness and weight loss, and people assume that physical activity can lose weight as a fact. Also KOLs ususally encourage people to drink more water for higher metabolic levels. But I have not seen a clear data to prove it. At the same time, the effect of physical activity to lose weight is also different from person to person.

In addition, according to what I learned from nutrition course, alcohol consumption may be related to weight gain because [alcohol metabolism](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6527027/) and carbohydrate metabolism both go through the same pathway/cycle in the middle and late stages.

## 2.2 Description of data and data source

The data is obtained from a website [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition) which have collected the survey results of individuals from the countries of Mexico, Peru and Colombia on their eating habits and physical condition, as well as their estimation of obesity levels.

## 2.3 Questions/Hypotheses to be addressed

I would like to figure out if higher frequency of physical activity and drinking more water could keep people away from obesity and whether heredity is an important factor for obesity. Also, I will focus on the relationship between alcohol consumption frequency and obesity level.

# 3. Methods

Data cleaning including generating necessary variable, renaming categories and variables and so on.

First to take a deeper look of the data, I would like to do some descriptive analysis including bar chart, boxplot, scatterplot etc. to indicate the distribution of the data.

Then simple linear fit will be applied to explore the relationship between variables and outcomes (both categorical and numeric).

Prediction will be made.

## 3.1 Schematic of workflow

## 3.2 Data acquisition

The dataset presents the results of a survey on the topic of obesity including personal physical conditions and answers to behavioral questions.

## 3.3 Data import and cleaning

The detailed raw data and processed data are stored in ‘data’ folder. And the code for data cleaning is in the ‘processing-code’ subfolder under ‘R’ folder.

I made a code book based on the article reporting this dataset (Palechor & Hoz Manotas, 2019) first before doing any cleaning.

Obesity level is a categorical variable defined by BMI range, so BMI as a new variable is created and it could be analyzed as a numeric variable. Also, to make the variable name easy to understand, I changed some abbreviated names to ones with fully spelled key words. Male is replaced by 1 and female by 2 under Gender variable for more convenient further analysis.

## 3.4 Statistical analysis

Bar chart, box plot will show the distribution of obesity level by different variables, as well as the population features of the people under investigation. All figures will be saved in ‘figure’ subfolder under ‘result’ folder and shown in the manuscript.

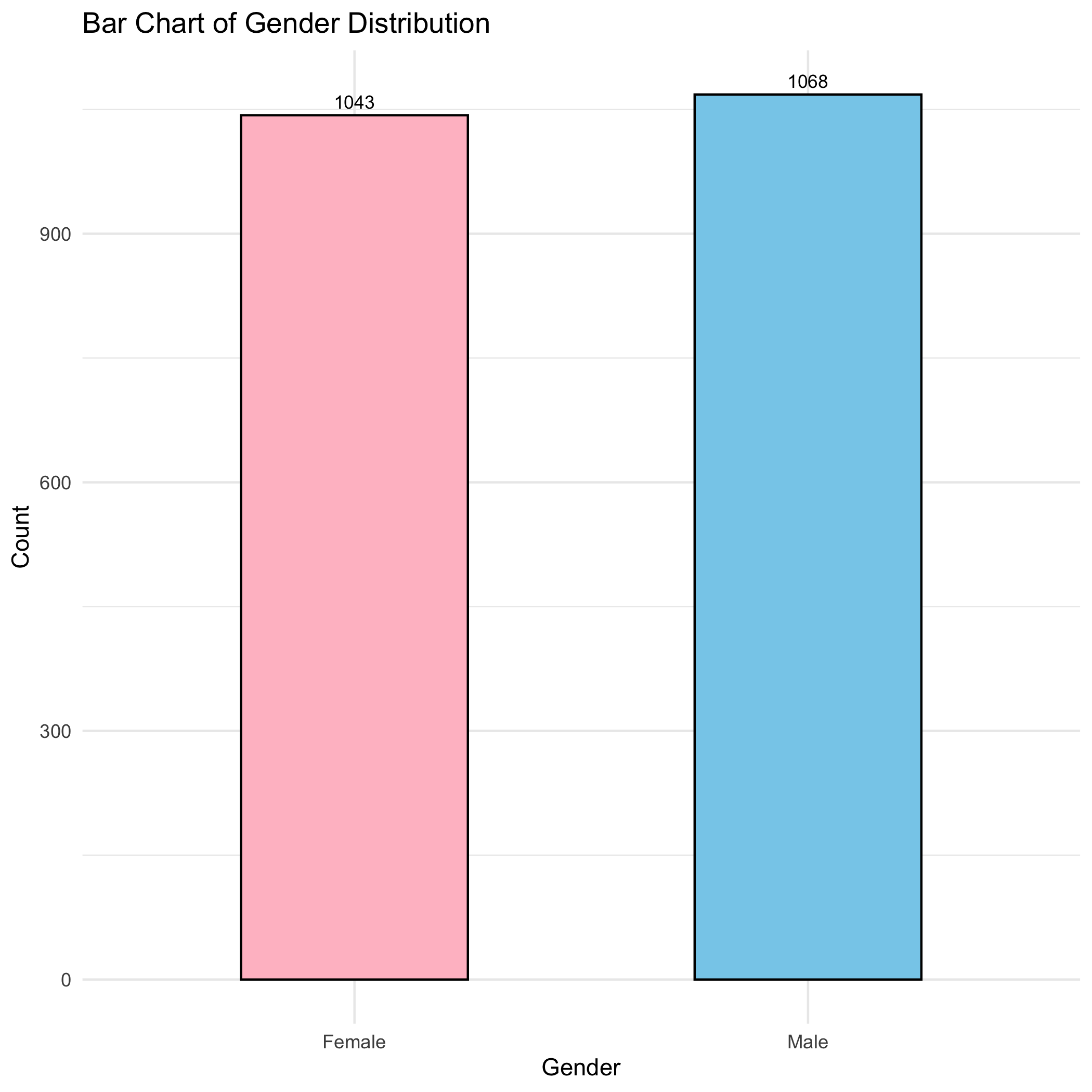
By fitting the data with simple linear model, the result will show whether each independent variable has main effect to the dependent variable through the P value. And the summary table is stored in ‘table’ subfolder under ‘result’ folder.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Bar chart shows that the genders of the cohort under investigation are balanced.

Figure 1. Gender distribution.



It can be seen that almost two thirds of the people are identified with over weight.

Figure 2. Age distribution

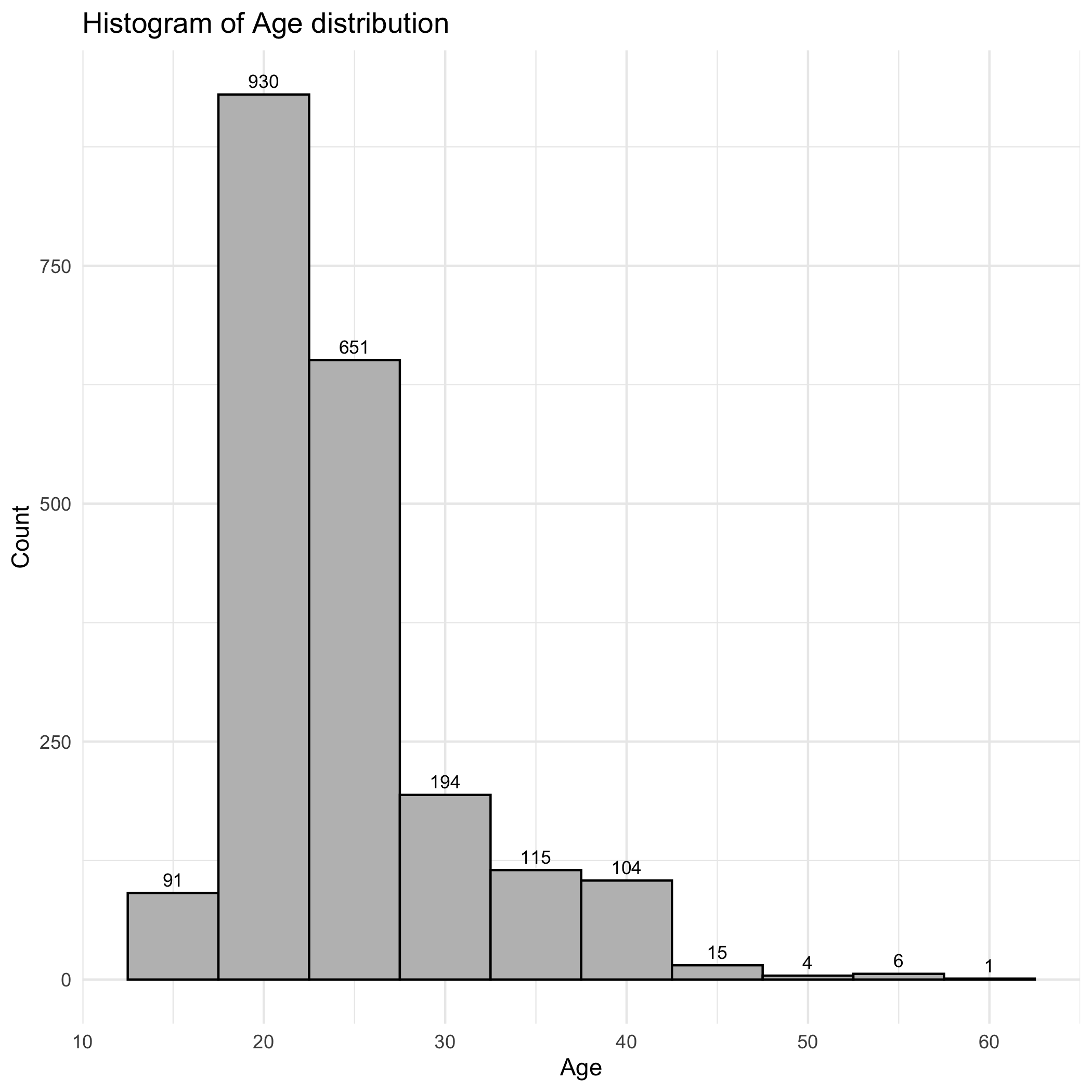
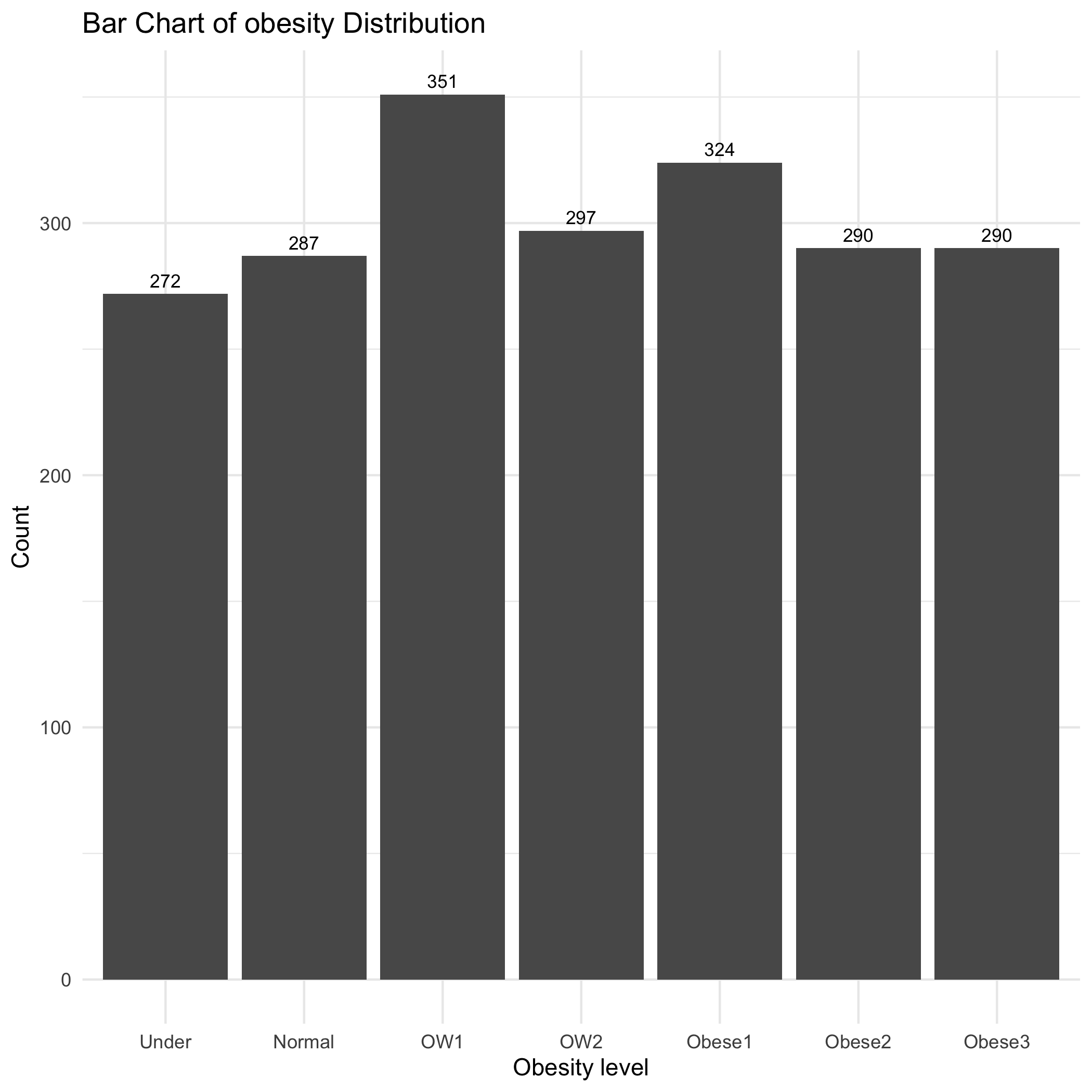
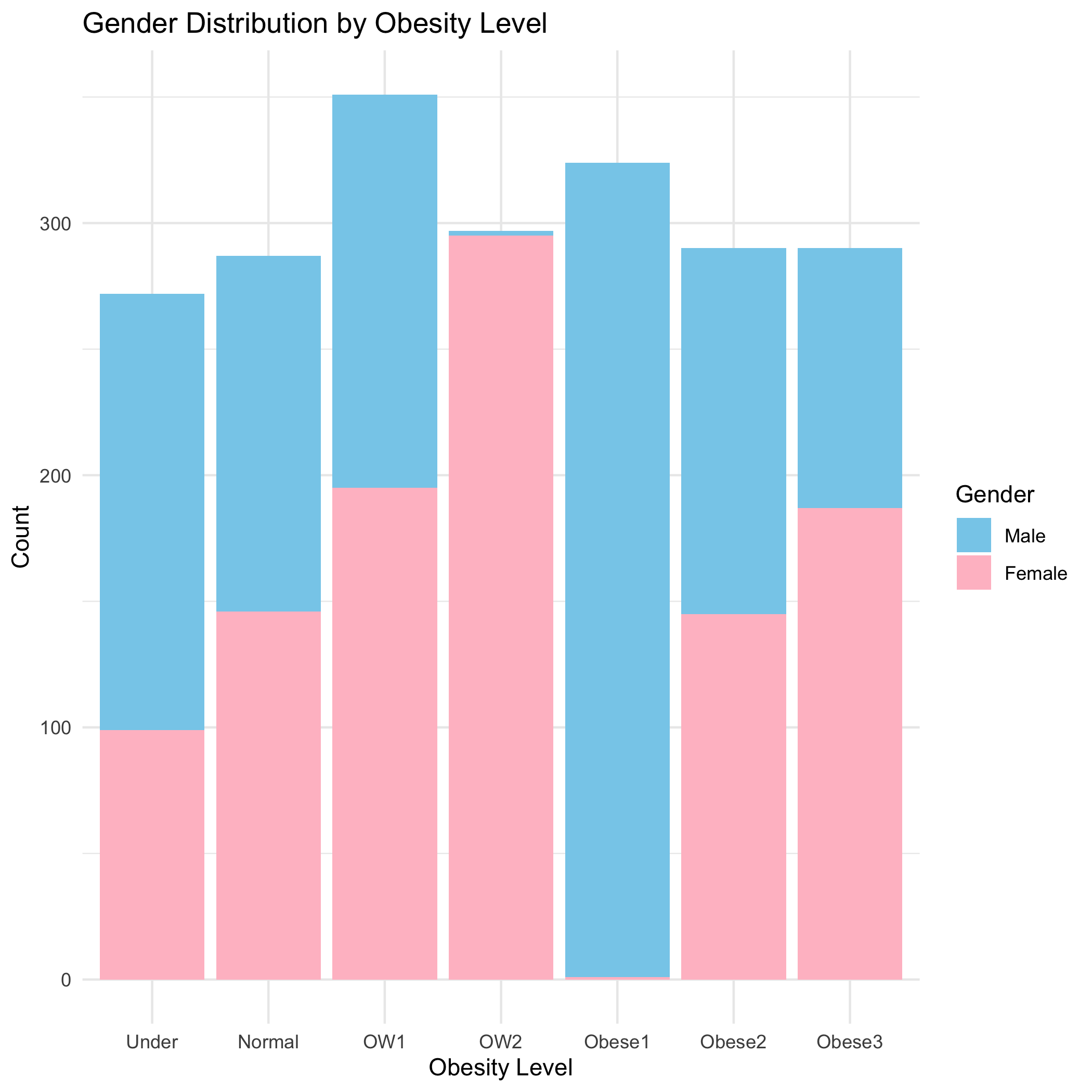


Figure 3. Obesity level distribution.



There is no obvious pattern of obesity levels by gender.

Figure 4. Gender distribution in each obesity levels.



It is apparent that people with family obesity history may tend to be over weight according to figure 4 below.

Figure 5. Age distribution in each obesity levels.

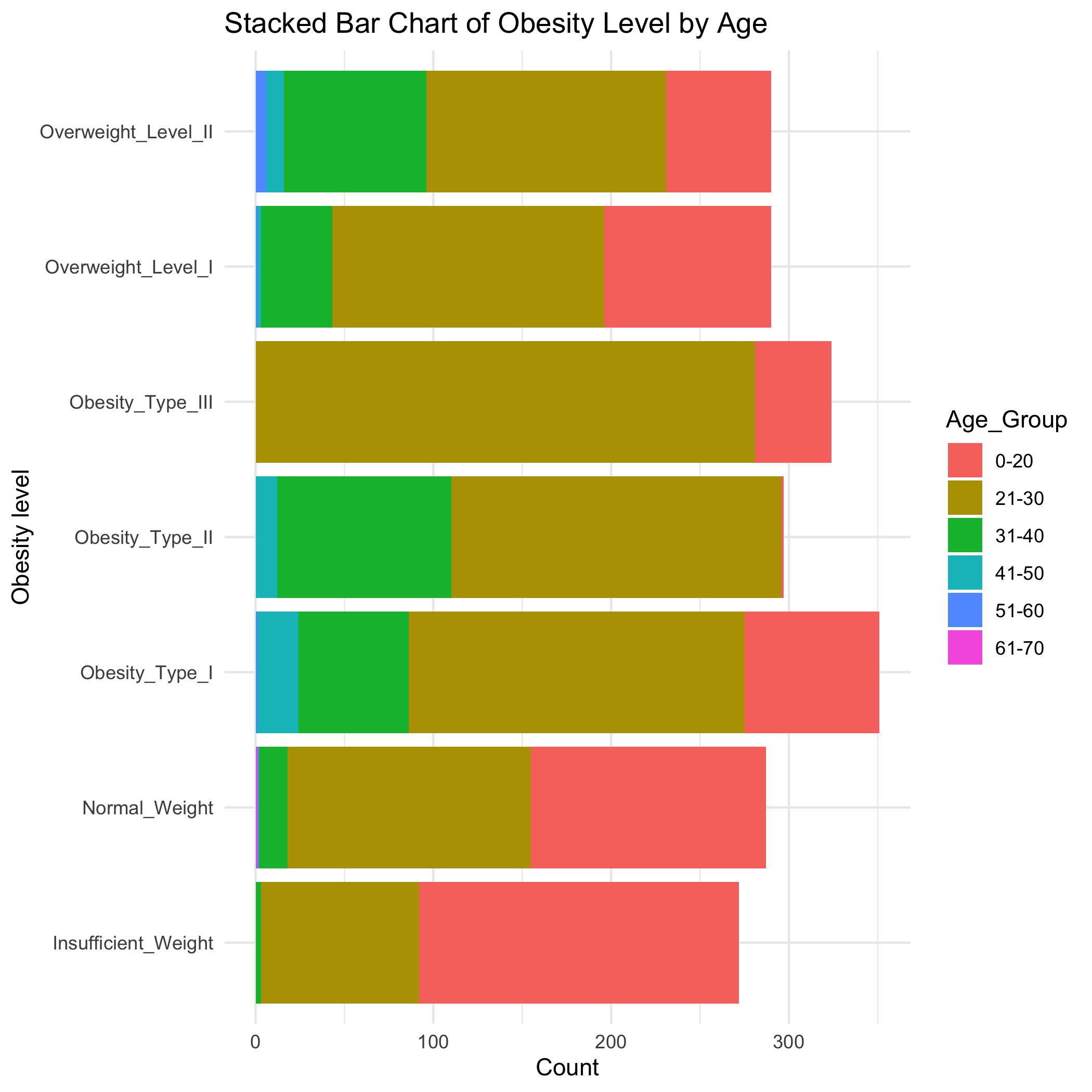
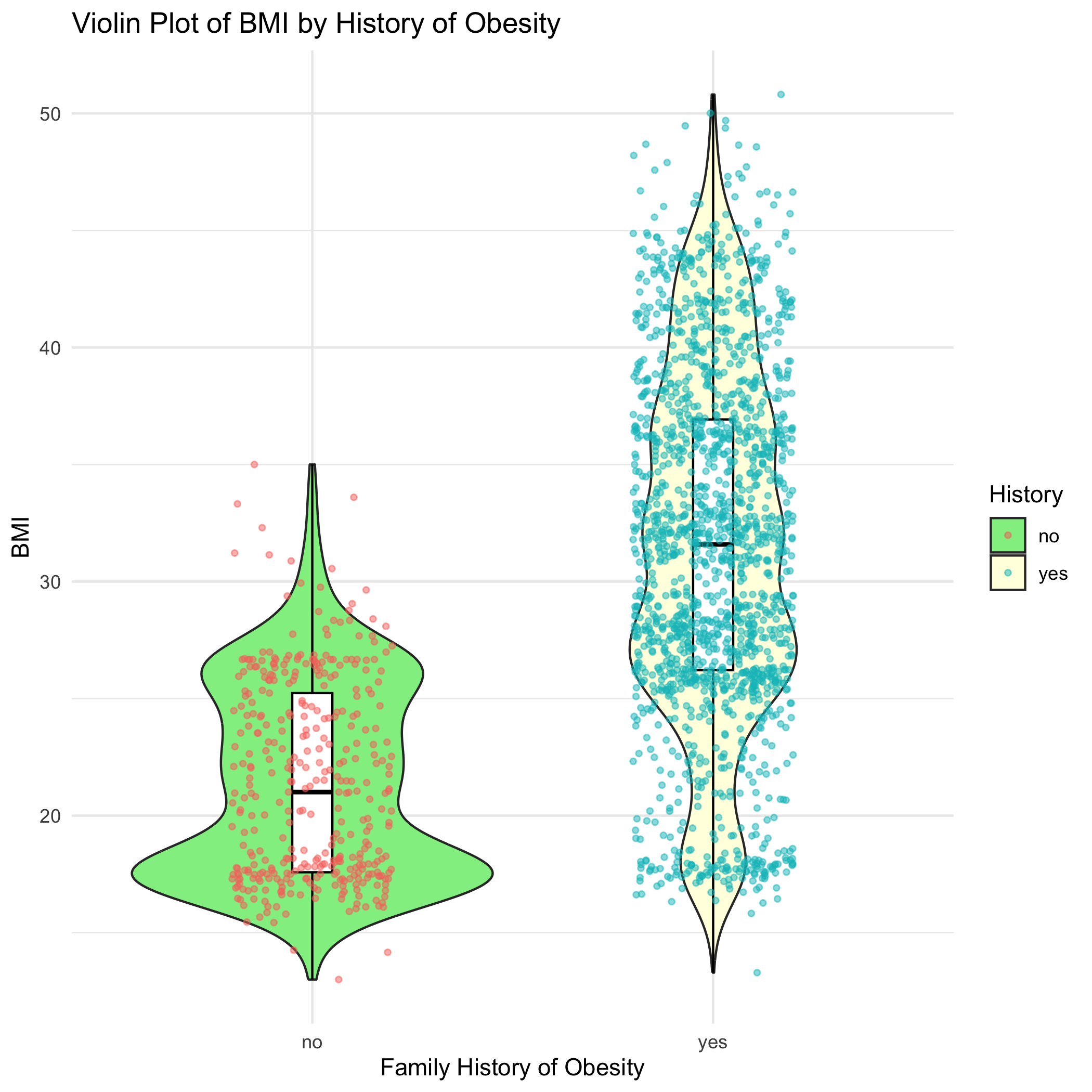


Figure 6. Violin plot of BMI by family obesity history.



## 4.2 Basic statistical analysis

Since some obvious relationship could be shown in figures, I choose other variables as predictors in exploratory analysis.

First, I was curious that whether there is a negative correlation between transportation type and physical activity frequency, I converted transportation types into numbers indicating the energy consuming levels from 1 (lowest automobile) to 4 (highest walking). Then I used correlation function to calculate the coefficient for these two variable but it was pretty low (0.0059).

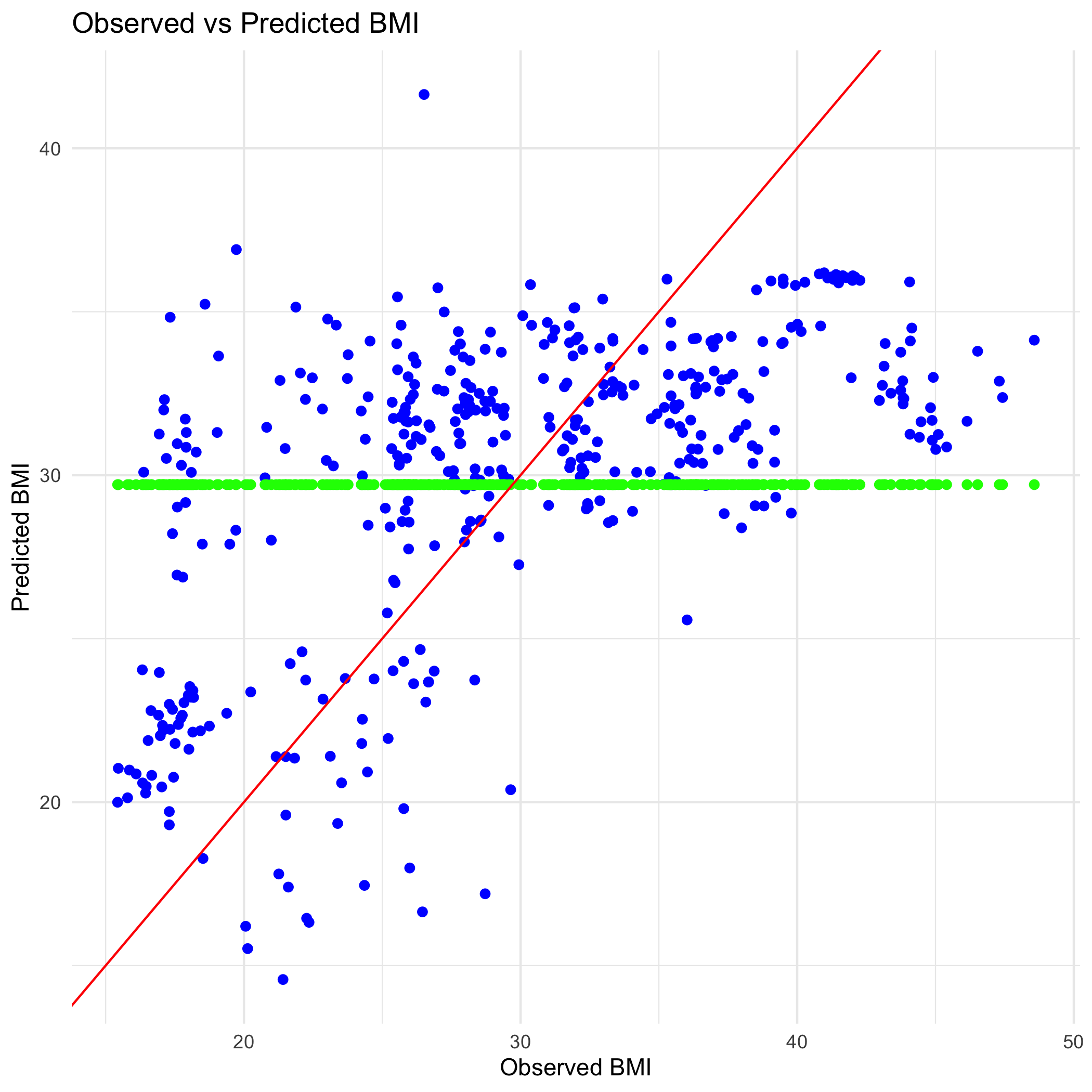
Then, I would like to fit linear model for BMI by both water consumption and physical activity. And the result indicates that there is a connection effect between them two, which means that both of them contribution to obesity control.

Also, I use single indicator, alcohol consumption and BMI as an outcome to fit linear model.

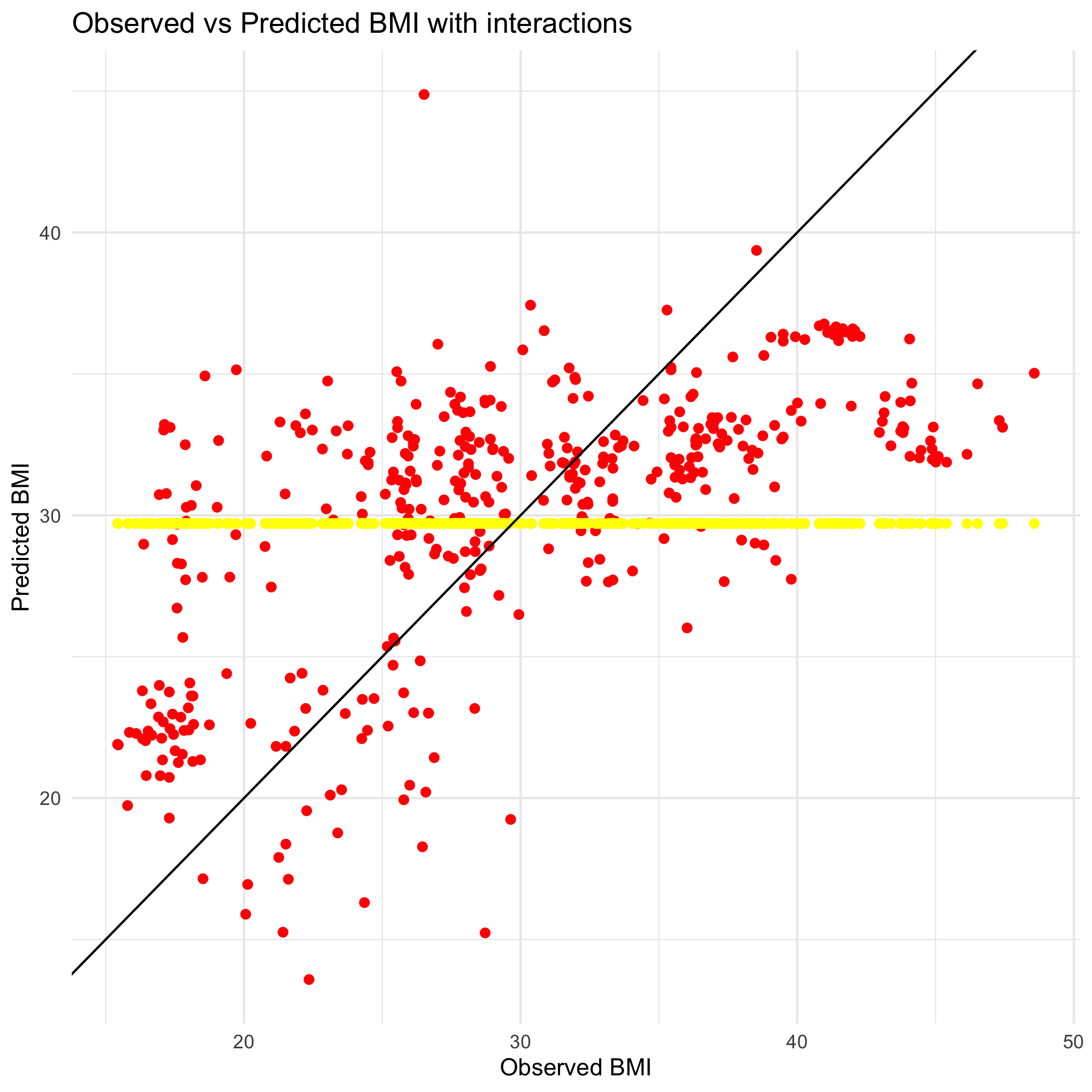
In the last model I select all the variables that I think may have effect on BMI as predictors and BMI as outcome.

## 4.3 Full analysis

Since I found there were interactions between physical activity-drinking water and age-gender, then I chose BMI as outcome and fitted the data with linear regression model. I asked chatGPT for help to write code and used the code to find the best combination of all the predictors included in the processed dataset. By comparing RMSE values of each combination, all predictors were selected. After that, I calculated RMSE of the null model (8.04) and it was a little bit higher than all predictor model (6.53).



According to the suggestions I received from module 11’s discussion, I also included 2 interactions as predictors and used linear model to fit data again, it was under my expectation that it lowered the RMSE (6.39), but didn’t make much difference in fact.



Then I used cross validation to check if the all predictors model overfits the data and the average RMSE I got was 6.32, which was not so different from data splitting.

And I also tried to fit LASSO model with all predictors inspired by the reply in discord, but I couldn’t interpret the result well because it selected specific categories rather than the whole variables as the predictors as the best combination, which I couldn’t figure out the meaning of that. So I only considered the linear regression result.

Because I had a question for whether using BMI as an outcome is preferable than obesity level, I used discriminant analysis model to fit the data with obesity level as outcome inspired by module 12 discussion since there were 7 categories in total under obesity level. I got the accuracy for all predictors without interactions was 0.45, and 0.48 with interactions. The average accuracy of cross validation result was 0.48, almost as same as the result with interactions included.

# 5. Discussion

## 5.1 Summary and Interpretation

I chose a dataset that summarize an investigation result done in South American countries about individual physical behaviors and the classification of obesity level based on the information of height and weight obtained from the investigation. I had some questions to explore and did data analysis on this dataset.

I found simple linear model fitting can answer my questions, which are that alcohol consumption is correlated to obesity, as well as there was interactions between physical activity frequency and drinking water. Besides, I found there was nearly no correlation between transportation choice and physical activity, which could answer another question I had.

For machine learning model fitting, I found the best predictors for BMI index based on the value of some performance metrics after trying both linear regression and classification model. It is hard to compare the metrics between two models, but I tend to choose linear regression model and BMI as outcome since RMSE at value 6 means relatively low and the model fits well for me compared to accuracy at 0.48 means the model can predict 48% of the test data correctly. And the result of cross validation shows a a sign of not overfitting.

So I think I would choose linear regression as the better prediction model and all variables

## 5.2 Strengths and Limitations

I think my analysis has include two different types of model according to the type of outcome, which can provide more information and the result can be shown in different aspects.

But there is still a concern that whether my model is overfitting not not since the RMSE I have got is much lower than exercise 8 which is several hundreds even if I have applied cross validation. But I still cannot find a good way to figure it out based on my knowledge so far.

## 5.3 Conclusions

The take-home message for this project is that obesity is correlated to individual physical behaviors. Based on the result of data analysis, age, gender, family history, drinking water, physical activity, and commute transportation are all predictors for obesity level which is represented by BMI index.

# 6. References

Palechor, F. M., & Hoz Manotas, A. de la. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, *25*, 104344. https://doi.org/<https://doi.org/10.1016/j.dib.2019.104344>