Prediction of Cholesterol Level based on its Relating Factors

Jinli Xu

2022-04-07

Abstract

This report is based on data from The National and Nutrition Examination Survey (NHANES). The National Health and Nutrition Examination Survey (NHANES) (Pruim 2015) is a program with a series of studies aimed at determining the health and nutritional status of Americans, including adults and children. The NHANES program started in the early 1960s and was transformed into a continuous program in 1999. Based on NHANES data, after classification, cleaning, analysis, and modelling, this paper aims to study the relationship between hyperlipidemia and a variety of different population labels (eg: age, gender, past medical history); Predictive power of lipid profile.

Introduction

In modern life, hyperlipidemia is a common but serious health threat. Its main hazards include hypertension, cardiovascular and cerebrovascular diseases, cerebral hemorrhage, myocardial infarction and so on. Hyperlipidemia can easily induce other diseases with a high fatality, and itself is very common due to changes in modern life and diet. Hyperlipidemia itself is mainly judged based on the total cholesterol content in human blood, so this paper will mainly study the relationship between cholesterol level and various composite variables.

In a practical sense, the medical and health factors behind hypercholesterolemia have been thoroughly researched. Therefore, the main research direction of this paper is the relationship between some seemingly unrelated social attribute variables and cholesterol levels. These variables include education level, psychological status, average income, and more. The main purpose of the study is to establish a population classification label through these social attribute variables and then by studying the relationship between cholesterol levels and these sociological variables, to obtain the possibility of different populations having high cholesterol health risks. In a practical sense, the research can contribute to individual health risk assessment and customized health advice for different populations, thereby improving the overall living standards of the social population.

The data in this article comes from The National and Nutrition Examination Survey (NHANES) (Pruim 2015). In the article, the process of data cleaning, sorting, and visualization is first carried out, and basic EDA is done on the data itself, and relevant conclusions and research results are drawn. The data and code processed in this project is done with (R Core Team 2020) After that, the process of further modeling is shown in detail in this paper; mainly by studying the relationship between different variables and cholesterol levels, as well as the degree of coincidence between the variables themselves, screening of variables and model simplification, using Steprise Regression and Methods such as Backstep Reduction refine the model. All data and R-code used in this paper are done in a reproducible fashion and can be found at: github link[1].

Data

The data used in this article come from The National Health and Nutrition Examination Survey (NHANES)(Pruim 2015). We have extracted 6,220 different data in NHANES since 1969, including a variety of categorical data that can be used to build an individual's profile, as well as medical data on an individual's cholesterol level. The purpose of extracting data in this paper is mainly to try to use sociological factors

[1] https://github.com/jinlixuokok/Prediction-of-Cholesterol-Level-based-on-its-Relating-Factors.git

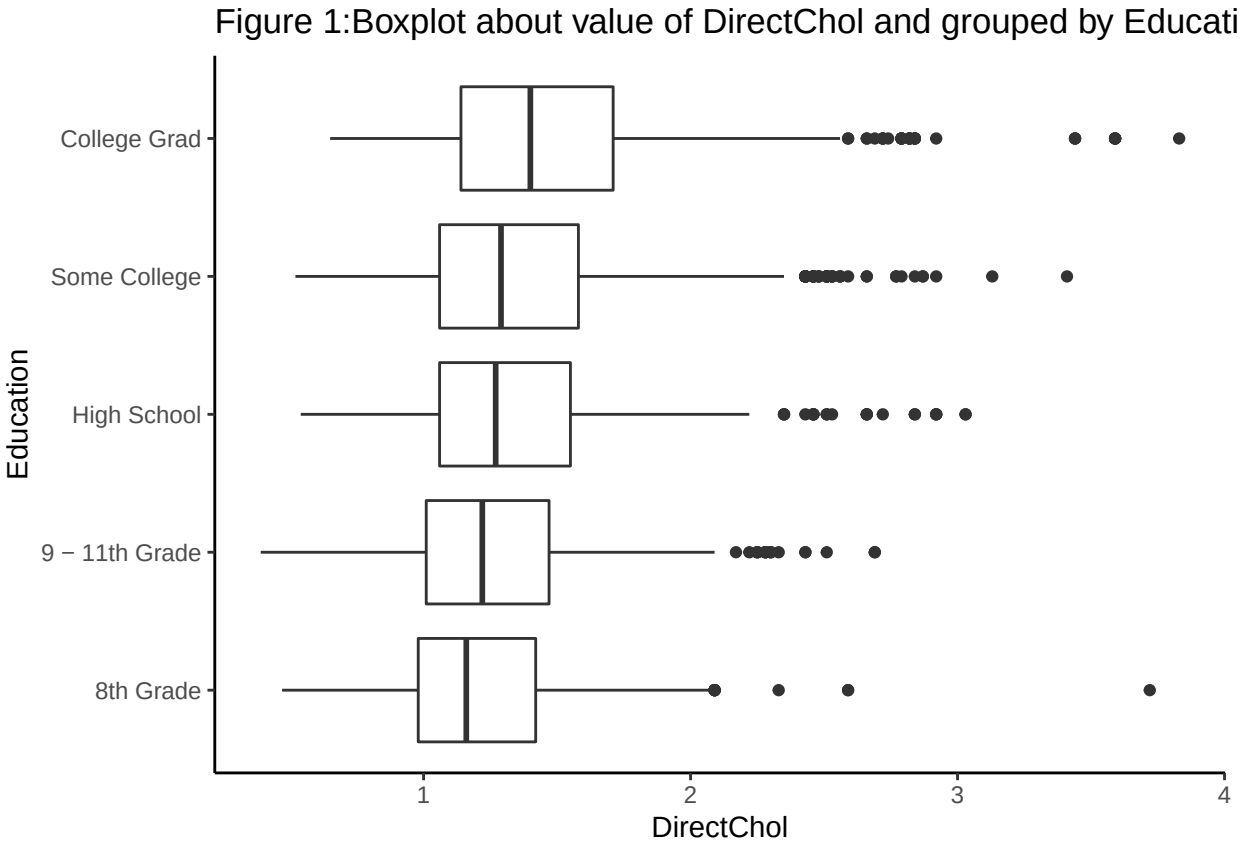
to analyze the cholesterol level of individuals without medical background and medical-related impression factors.

The data of NHANES comes from a completely voluntary and anonymous questionnaire. The data itself covers a wide range, but it mainly focuses on the health and nutritional level of the individual. In this case, wanting to analyze the social influence factors associated with the research object means that we need to focus on cleaning and classifying the data. In the following EDA section, we further visualized the variables related to the study subjects, and made relevant inferences and analyses.

Methodology

The research method of this paper is relatively straightforward. After first finding the data related to this study on HNANES(Pruim 2015), we then entered the data cleaning and preliminary analysis stage. After filtering out the data related to the research direction of this paper, we then visualized the data and built multiple charts to help the analysis. See the Results module next for the visualization results. After data visualization and EDA, we then started the modeling step by using (Harrell Jr 2021); first, we screened out multiple variables that may be related to cholesterol levels through the EDA results, carried out preliminary modeling, and obtained a Model1 containing many variables, after which we Analyze the correlation between different variables in the comparison model, start to filter redundant variables and simplify the model. In this process, we mainly used the Steprise Regression and Backstep Reduction methods to simplify the model, compared the P-value between different variables, and analyzed the Residuals vs fitted, Normal QQ plot and other charts of the model, and determined the validity of the model. The data is simulated and organized by using (Wickham et al. 2019)

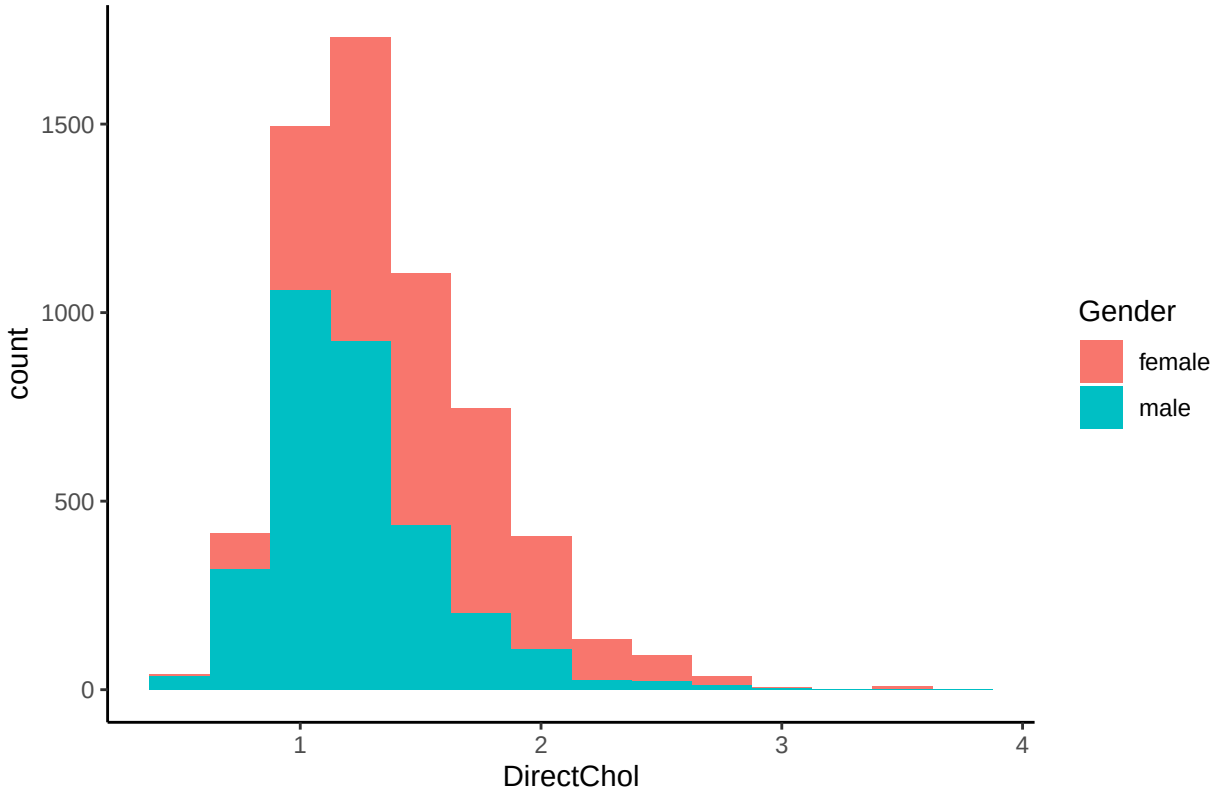
EDA and Results



This graph shows the link between educational levels and cholesterol levels via Bar Plot. As shown, we

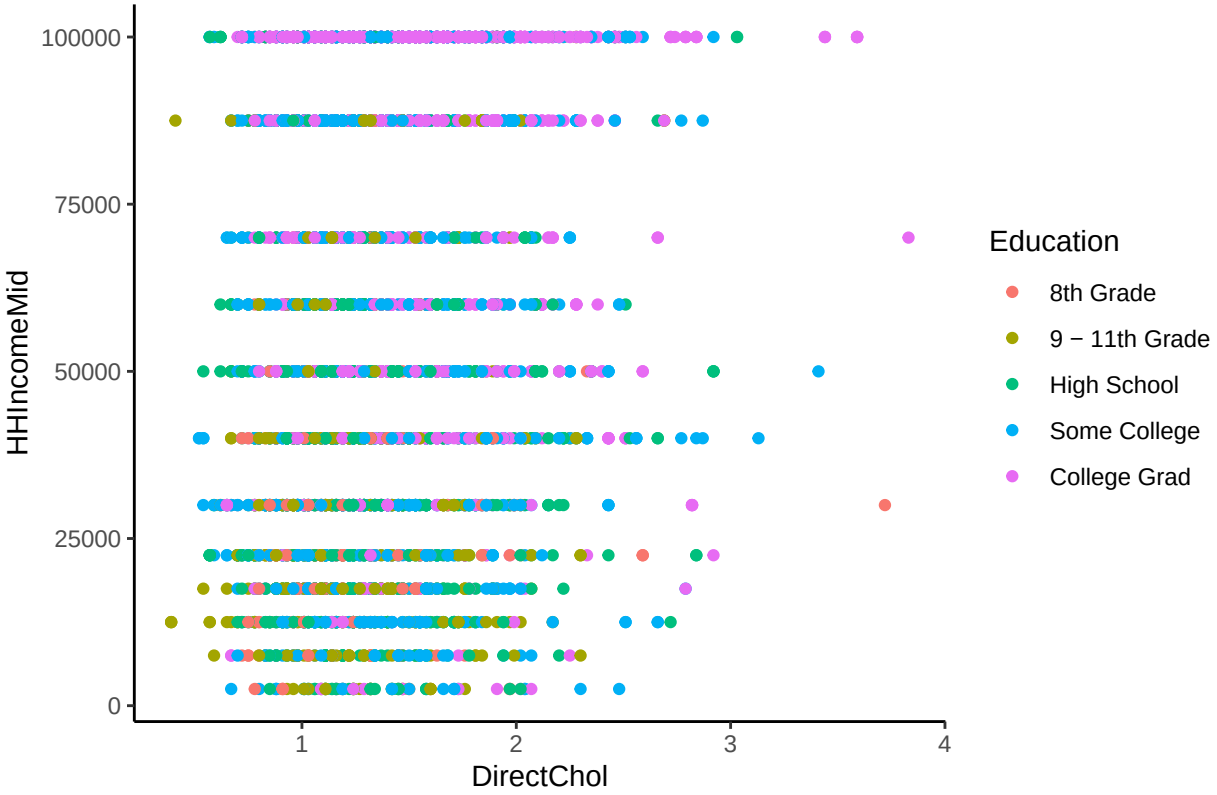
found that relatively higher educational levels were more likely to have higher cholesterol levels. The reason for this correlation may be due to the positive correlation between education level and income level, and high cholesterol level is often caused by usual eating habits; higher education level means higher income level, which means that high calorie level is usually caused by Food intake is more. Of course, the logical relationship behind the derivation of this correlation is mainly based on empiricism. To prove its own correctness, we still need to further study the relationship between more variables in the data.

Figure 2: Plot about value of DirectChol and colored by Gender



This graph shows the average cholesterol levels between genders in a succinct way. It is not difficult to see that the average cholesterol level of women is much higher than that of men; the possible reasons for this phenomenon include innate physiological differences between men and women; differences in eating habits between women and men; women's exercise and life compared with men Get used to the difference. However, the fact that women have a higher risk of high cholesterol than men cannot be ignored, and health recommendations based on this report will take this into account.

Figure 3:Plot of HHIncomeMid Vs. DirectChol group by Education level



This graph visualizes variables such as income level, education level, and cholesterol level. It is not difficult to see in the figure that the higher the income level, the higher the average cholesterol level. This is in line with the inferences from the analysis we did under the previous chart, on the other hand, the higher the income level, the higher the average education level. In this case, we found that there is a considerable correlation between income level and education level, so in the subsequent modeling process, both education level and income level are used as influencing factors, and we will only consider keeping one of them.

Model and Results

After cleaning and visualizing the data itself and doing preliminary analysis, we started to clean up the model itself. In this paper, the selected model is based on the Linear Regression model, which can effectively integrate many different types of variables (numerical and categorical) and conduct a comprehensive analysis. The model is generated based on :(Simon et al. 2011)

We first linear model as model_1

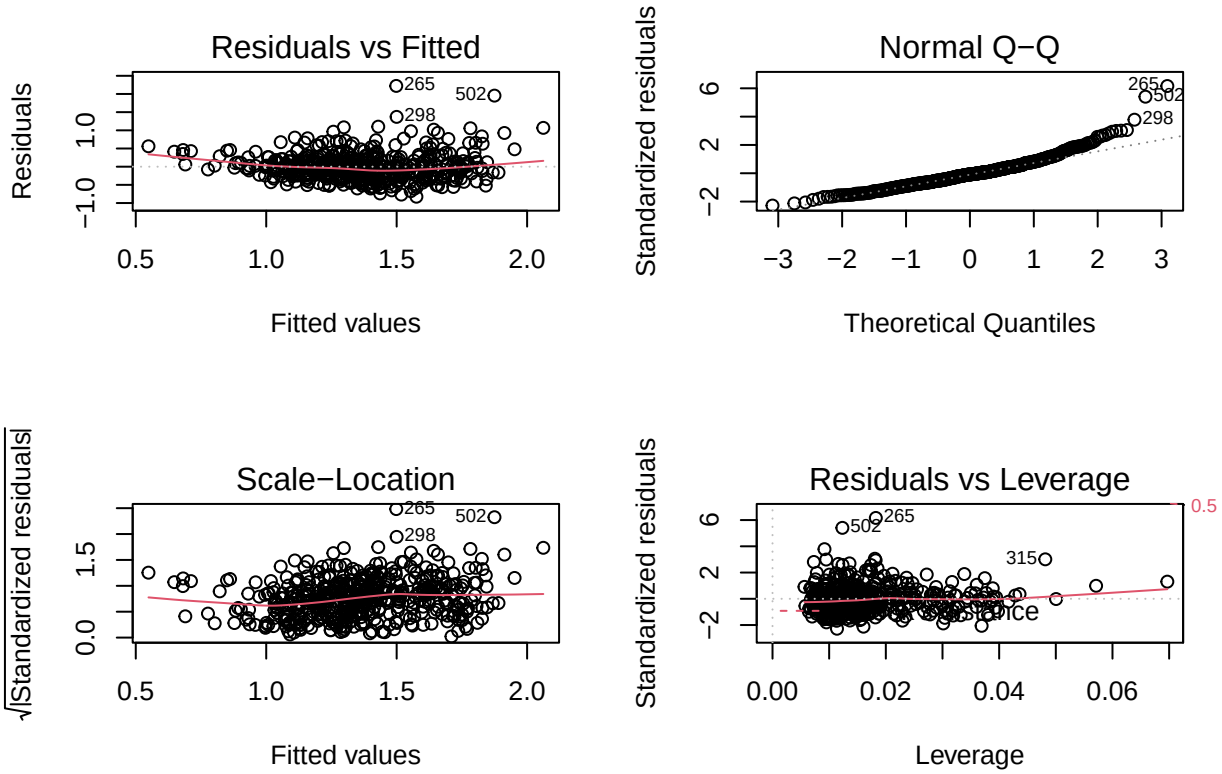
$$DirectChol = \beta_1 Gender + \beta_2 Age + \beta_3 BMI + \beta_4 Weight + \beta_5 HHIncomeMid + \beta_6 Education + \beta_7 MaritalStatus + \beta_8 Depressed + \beta_9 SleepHrsNight + \beta_a PhysActiveDays + \beta_b SmokeNow + \beta_c BPSysAve + \beta_d HardDrugs + \beta_e Alcohol12PlusYr + \beta_f Diabetes + \beta$$

By checking the variance inflation factor, we analyze the model and find that there is a very high correlation between BMI and Weight, as well as between income level and education level, so we will delete Weight and education level in subsequent models' variables. And then cut it down to:

$$DirectChol = \beta_1 Gender + \beta_2 Age + \beta_3 BMI + \beta_4 HHIncomeMid + \beta_5 MaritalStatus + \beta_6 Depressed + \beta_7 SleepHrsNight + \beta_8 PhysActiveDays + \beta_9 SmokeNow + \beta_a BPSysAve + \beta_b HardDrugs + \beta_c Alcohol12PlusYr + \beta_d Diabetes + \beta$$

Next use backward steprise regression to Reduce model as its final form:

$$DirectChol = \beta_1 Gender + \beta_2 Age + \beta_3 BMI + \beta_4 HHIncomeMid + \beta_5 SmokeNow + \beta_6 BPSysAve + \beta_7 Diabetes + \beta_8$$



Next, by plotting the residual analysis charts of this model, including its Residual vs Fitted, Normal QQ Plot, etc., we found that the Normal QQ plot line of the model itself is stable and in line with its baseline. There is no obvious linear relationship in the other figures, so this model can be regarded as valid.

Finally we get the linear regression model for predicting the value of DirectChol:

$$DirectChol = -0.266 Gender + 0.00252 Age - 0.0284 BMI + (8.52e-07) HHIncomeMid - 0.137 SmokeNow + (3.73e-03) BPSysAve - 0.138 Diabetes + 1.784$$

Gender: female - 0, male - 1

Age: Age of years

BMI: Body Mass Index.

HHIncomeMid: The total income of the unit id US Dollar. Higher income will usually yield higher amount of food consumption.

SmokeNow: If the person is having the habit of smoke. Shown by 1/0 as yes/no.

BPSysAve: Average bloodpressure.

Diabetes: If the person is having diabete. Shown by 1/0 as yes/no.

Discussion

This paper studies the social factors related to high cholesterol by cleaning, classifying, and studying relevant data from NHANES in detail; and through modeling and analyzing models, constructs a basic method for predicting personal cholesterol levels through relevant social factors. In general, we found that in modern society, cholesterol levels are highly influenced by an individual's diet, lifestyle, and exercise habits, and on an individual basis, high-income women have a higher associated risk. The main cause is that high cholesterol is caused by high carbohydrate intake, and people with similar eating habits tend to have higher incomes. On the other hand, this report further found the relationship between cholesterol level and BMI. Studies have shown that BMI level and cholesterol level are positively correlated, and the cholesterol level of people with BMI higher than 30 is significantly higher than the average. Smoking, age, and diabetes also greatly affect an individual's cholesterol levels.

In conclusion, this report examines the relevant data in NHANES in detail through data visualization methods, and further studies the related causes of individual hypercholesterolemia. The model established in this study can help individuals predict their own cholesterol levels through non-medical relevant information, identify the risk associated with hypercholesterolemia, and adjust their living status and habits in a timely manner. From a social point of view, this study also helps to analyze the risk of cholesterol level of a group through personal information profiling, and timely issue medical dietary advice to solve the risk when it is found, so as to avoid a large amount of medical resource consumption in the future.

However, it cannot be ignored that cholesterol levels are also affected by a large number of genetic factors. This study only considered the relevant social factors of individuals but did not consider the genetic factors related to their parents. Therefore, from this perspective, the establishment of this study in this study the model may have some inaccuracies in its predictions. On the other hand, the data source of this study is essentially a personal questionnaire. When similar questionnaires are used to collect and study medical-related questions, they are often inaccurate and biased in their answers due to the lack of professional knowledge of the respondents.

Reference

Harrell Jr, Frank E. 2021. *Rms: Regression Modeling Strategies.*

Pruim, Randall. 2015. *NHANES: Data from the US National Health and Nutrition Examination Study.*

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

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

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.