What are the biggest factors that associate smoking and high blood pressure? A population based study based on 17 years or older.

ChanYoung Cho (victorchanyoung.cho@mail.utoronto.ca)
University of Toronto, St George Campus
7 King's College Cir, Toronto, ON M5S

Abstract

Objectives: To describe the characteristic patterns and associations in between gender, age, race, education level, marital status, household income, weight, height, BMI, depression, sleeping hours, physical activity, and smoking with systolic blood pressure reading, and to determine the best model to predict our data.

Methods: The NHANES dataset was given to be used for data analyses. An examination of diagnostic checks, verification of models and selecting variables were used to drive the most plausible, optimal variable for our goal of interest by choosing the best model of data analyses.

Results: Using the best models for evaluating the data analyses, it was shown that age was the most significant variable amongst the other predictors.

Conclusion: Overall, it was conclusive that among the predictors, there was a biggest association in between age systolic blood pressure.

Introduction

For this final project, we are concerned with examining the association of various factors and systolic blood pressure reading. Blood pressure has been an overarching concern regarding health to many adults, and we aim to do analyses to find out the relevant factors that associate with it. The dataset that will be used for this specific project will be the NHANES dataset, which is data collected by the US National Center for Health Statistics (NCHS). NHANES is a programme that is designed to survey health statuses of the United States, each survey on average, contains 5000 responses every year. Similar to the description, the dataset used for this project contains the health and nutritional status of adults and children of the United States. This dataset contains many variables such as smoking, age, height, and race, where these are grouped into each person along with the data that indicates their health and nutritional states.
Assessing these criteria, our main goal is to finding out what is the most significant, relevant factor that highly associates with systolic blood pressure reading, through applying a series of statistical analyses methods to this specified dataset.

Methods

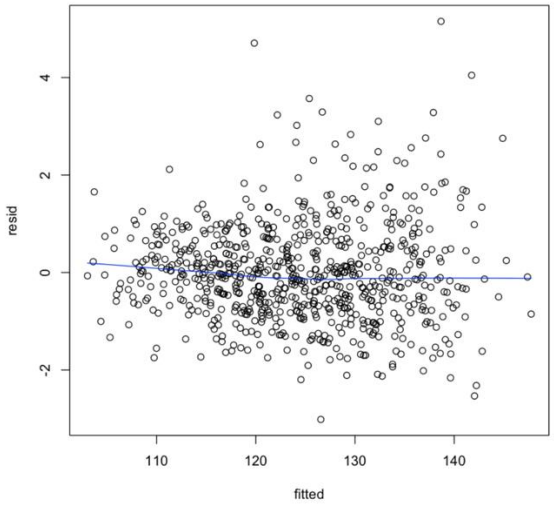
The data analysis methods of this project was done with R and and its necessary libraries to model, graph, and draw conclusions that aim towards our goal. First, I prepare the data by selecting subjects that are older than 17 years old. The very first step taken was fitting the data into a linear regression model, and plotting the studentized residuals versus the systolic blood pressure and fitted values. Then, I compute the leverage points, influential points, the DFFits and DFbetas. Then, I examine the variance inflation factor for this dataset to check for the presence of multicollinearity. When judging for whether not a variable has a high correlation, I used the criterion of a VIF of five to ten is highly correlated, and a VIF above 10 would be a predictor that is poorly estimated. After verifying the estimators, I do a

stepwise variable selection in AIC and BIC to verify to assess the best predictors that associate with systolic blood pressure. For the shrinkage methods, I model different formats of regression models and calculate the prediction error; I do this one with regular regression, in terms of regular regression, elastic net, ridge regression, LASSO. Finally, I perform the model validation, where I select the variables via LASSO selection, and perform cross validation techniques with AIC, BIC, and LASSO. Judging by the graphs comparing the ideal versus actual values and the prediction error outputs, I determine the best model that is suitable for our dataset, as well as the most relevant, significant predictor with the biggest association with systolic blood pressure reading.

Results

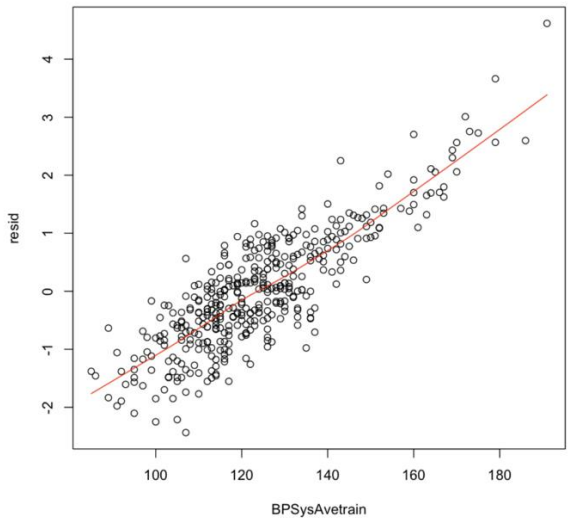
My study sample consisted of males that were 18 years or older. Before jumping straight into the model diagnostics, I first fitted the studentized residuals versus the fitted values, as well as residuals versus the systolic blood pressure readings. From figure 1, we can see that the residuals are spread out from the line of 0. We also see residuals that are way above the line of 0, which means there exist a huge variation on the residuals versus the fitted values.

Figure 1.



Fitting a line with on the systolic blood pressure readings on figure 2, we see that this shows a linear pattern.

Figure 2.



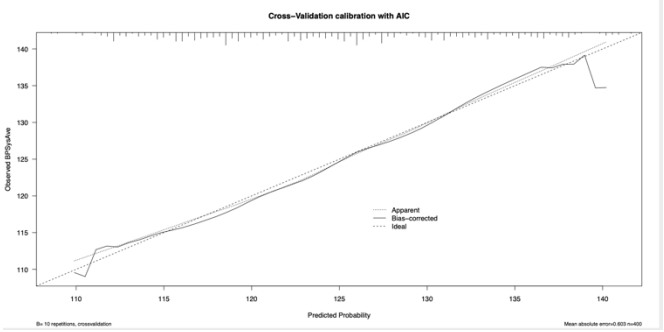
Further diagnostics reveal that our model for this data contains no influential points with many leverage points, DFFits, and DFFbetas.

Identifying the degree of how well the predictors are predicted through the VIF analysis, I performed variable selection methods in BIC, AIC, and LASSO , which all yielded different results. Each outputted different selected variables.

Applying the shrinkage methods, I assessed all my models through model selection, which yielded that Age was the most significant variable in association to systolic blood pressure,

Plotting an AIC cross validation graph as you see in figure 3, it was observable that my apparent values were nearly consistent with ideal values.

Figure 3.



Performing the identical task with BIC (figure 4) and LASSO (figure 5) selection models, we see that there are some variations with the model in between the apparent values and the ideal values.

Figure 4

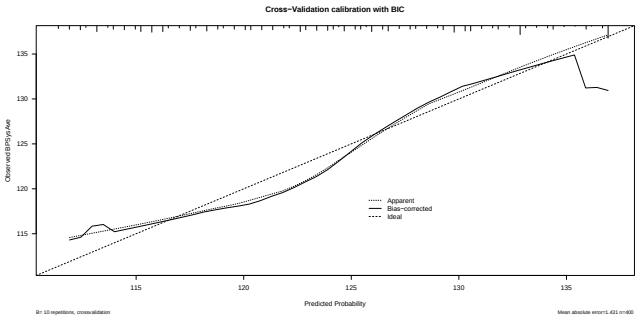
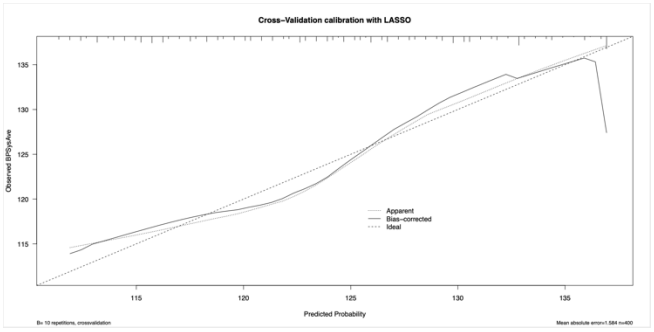


Figure 5.



From these results, it was shown that AIC was the best model to predict my data when modelled into a linear regression. Further calculation the prediction errors of all AIC, BIC, and LASSO also reflected a similarly referable result, such that AIC had the lowest prediction error compared to the two other models.

Discussion

According to our best model which resulted in the AIC, it was shown that Age variable was the most significant variable that had the biggest association with systolic blood pressure reading. While there were many predictors to consider as blood pressure contains many factors to consider as reflected in the NHANES dataset, it was observable that Age amongst all of them was a crucial factor. Despite these results, there are limitations for this data analyses methods. Although the analysis gives us an overall conclusion of what we observe, we are not able to draw conclusions merely from the associations. Moreover, through the VIF analysis, there were some variables that were poorly predicted, which may result in potential multicollinearity. This could have affected my result such that it may as not be as optimal.

References

NHANES Questionnaires, Datasets, and Related Documentation. (2020, February 21). Retrieved June 25, 2020, https://wwwn.cdc.gov/nchs/nhanes/