A Rudimentary Analysis of Smoking on Blood Pressure

Johnson Vo

ABSTRACT

PURPOSE: Smoking has historically and scientifically been detrimental to healthy individuals, however the association of smoking with individuals with uncontrolled blood pressure (systolic) is unknown. We also seek in the prediction of systolic blood pressure using smoking as a predictor.

METHODS: Seven hundred and forty three subjects aged 18 or over were selected from the National Health and Nutritional Examination Survey for the 2011-2012 sample years collected with a clustered sample design. Multiple weighted regression models were used to examine the association of smoking with levels of blood pressure using stepwise selection models and ridge regression.

RESULTS: The mean age of the sample was 50.67 years and approximately 58.54 percent of the sample were male. Prediction error for smokers was nearly identical for non-smokers (259.5114 vs. 259.5056).

CONCLUSION: There is a lack of evidence that correlates smoking with uncontrolled blood pressure. An explanation for their relation is unknown. It is also unclear whether smokers are more likely to have higher systolic blood pressure.

1. INTRODUCTION

Hypertension is the leading cause of cardiovascular diseases in adults, where 46% of all US adults have uncontrolled blood pressure, which amounts to less than 140/90 mm Hg. The sample data was derived from the National Health and Nutritional Survey (NHANES) which is an American national series of health and nutritional surveys conducted on US citizens. There are clear documented studies on the dangers of smoking, increasing risks of heart disease, cancer, and kidney failure. Despite these claims, there is a lack of evidence that suggests smoking as a risk factor for high blood pressure. The goal of this brief study is to derive any association between smoking and uncontrolled blood pressure, and attribute this to smoking in lieu of other factors such as age.

For this study I used data from the NHANES dataset for the 2011-2012 cycle. Participants included in this sample were those eighteen or older and had completed all relevant testing. Individuals under eighteen were not included due to smoking not being very relevant in this subgroup. Since the NHANES target population is "the non-institutionalized civilian resident population of the United States", the dataset can be treated as a simple random sample from the American population. A total of seven hundred and forty three individuals were included in the sample.

2. METHODS

2.1 Study Population

NHANES is survey data collected by the US National Center for Health Statistics (NCHS) which has conducted health studies since the 1960s. They were a series of cross-sectional studies conducted continuously in 2-year cycles. Participants undergo a combination of home interviews as well as physiological measurements and laboratory tests. The study is described to be a collection of "health examination data for a nationally representative sample of the resident, civilian noninstitutionalized U.S. population". Certain subpopulations are intentionally over-sampled to increase precision, such as racial minorities, and so a naive analysis of the data can lead to misleading conclusions. Note the data is done under a weighted sampling scheme to reduce sampling bias.

2.2 Blood Pressure Measurement

For the purpose of this analysis, only the combined systolic blood pressure readings were taken into account. Systolic blood pressure was measured by trained physicians following the BPXSAR standard procedures. Up to four total measurements were taken, where only the last three measurements (if possible) were taken into account.

2.3 Other Notable Measurements

A number of other factors were taken into interest in developing the regression models. Physical activity was categorized as a binary variable, classified on whether the individual was physically active or not. Participants were also questioned on the the number of hours they normally slept on weekdays or workdays, and whether they had trouble sleeping. Other factors such as gender, age, height, weight and race were taken into account. Subjects aged 80 or older were categorized as 80. Education was reported by individuals aged 20 or older, with categories being one of 8th-grade, 9-11th grade, completed highschool, some college or being a college grad. Income was recorded by a subjects total annual gross income for the household, and poverty as a ration of family income to poverty guidelines, where smaller numbers indicated more poverty.

2.4 Data Cleaning

In order to prepare the dataset, a large number of potential predictors were removed for a variety of reasons. BMI is derived from a calculation of one's height and weight, and so to reduce multicollinearity in the data, BMI is removed. Income and poverty are two very similar factors and so to reduce multicollinearity, income is removed due to reducing VIF values the greatest. Other predictors such as marital status, and whether the individual suffered from depression or had trouble sleeping were removed.

2.5 Model Selection

A number of regression models were derived from a number of statistical techniques (Model I, Model II, Model III, Model IV and Model V). These models differ by the types of predictors employed in order to test estimation and prediction. Model I is the full model fitted with gender, age, race, education, poverty, weight, height, number of hours slept per night, whether the subject is physically active and is currently smoking. Models II - V are smaller in number of predictors in order to avoid multicollinearity and increase interpretability and simplicity. Thus Model II is fitted by gender, age, height, weight, and if the subject is physically active and currently smokes. I also employed a number of model selection tools in hopes of deriving models that are proficient in describing the NHANES data. Model III is derived under Akaike's information criterion (AIC) using stepwise selection, and Model IV is derived under Bayesian Information Criterion (BIC) using stepwise selection. Due to the large dataset, I did not consider using corrected AIC to avoid overfitting. Model III is then fitted with gender, age, poverty, weight and height, and Model IV is fitted with gender, age, weight and height. The final model is developed usiong ridge regression, a type of shrinkage method in order to minimize prediction variance and hope to develop a model that is optimized for predicting a subject's blood pressure. Model V is fitted with only one predictor - age.

2.6 Model Diagnostics

I used a standard Z-test to determine the significance of each model's associations and to test whether a linear relationship is held. The 5 models went under a number of diagnostics to test whether they violate any model assumptions. This includes a plot of fitted values against standardized residuals to test homoscedasticity (random error variance). I also employed Q-Q plots in order to test normality of errors for each model. Finally I used Cook's distance plot which is used to test any influential points and in conjunction with DFFITS and DFBETAs, I could determine if the data was negatively influenced by any leverage points. Due to the large dataset and the results from these methods, I found that there were not any valid influential points that needed to be removed.

2.7 Model Validation

In order to test how well the models predict, I validated every model through cross validation. By using 10 as an upper limit on the number of resamples on which variables were selected, I was able to test the predictive ability of each model. This was done using a calibration plot that plotted predictions against observed values to test the accuracy of the models.

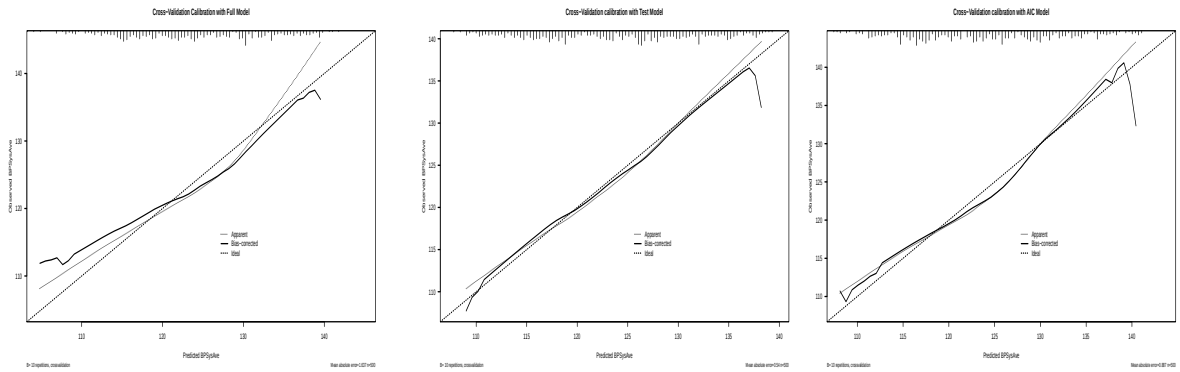

Figure 1. Calibration plots of Model I, II and III

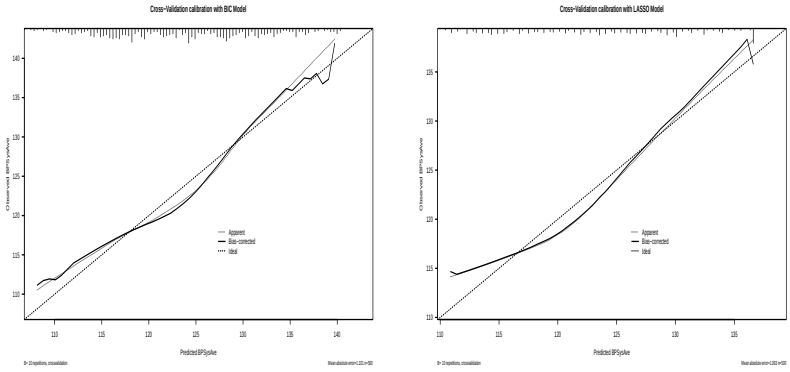

Figure 2. Calibration plots of Model IV and V

3. RESULTS

A total of 743 subjects were included in this analysis, where the mean age of the sample was 50.67 years old and approximately 58.54% of the sample were male. To test the predictive ability of the five models, I can take a look at the calibration plots for each model.

The above calibration plots applies a loess smoother curve plotting the predictive probabilities (x-axis) to the empirical probability (y-axis) in order to test line-of-fit. That is, the closer the plot lines on alone the diagonal line, the model is more calibrated and thus is a considerable better model for prediction. It becomes evident from the plots that Model II is the closest to the diagonal line. Next I split the dataset into a test dataset and training dataset (n=500) and computed the prediction error of each model.

Table 1. Mean Squared Prediction Error

Model	Prediction Error
Model I	264.7980
Model II	259.5114
Model III	265.8969
Model IV	266.2635
Model V	272.7590

Prediction error, specifically mean square prediction error, is the expected value of the squared differences between the fitted values derived from the models against the actual dataset values themselves. Therefore a smaller prediction error is desirable. Clearly from the table that Model II is the best model in terms of predictive

ability. However I wanted to test if smoking did have a strong influence in predicting blood pressure. Model II is fitted by gender, age, height, weight, physical activity and smoking as predictors, and so the predictive ability of the model may be explained by the other predictors rather than smoking.

Table 2. Smoking as a Predictor

Model	Prediction Error
Model II With Smoking	259.5114
Model II Without Smoking	259.5056

By removing smoking as a predictor from Model II, the prediction error actually decreased (meaning it became a stronger predictive model). This means that smoking did not actually help predict blood pressure.

3.1 Conclusion

In this study of 743 participants I found there is no statistical significant relationship between smoking and blood pressure through the use of linear regression models and calibration plots. However there were a number of caveats made during this analysis that may have played a role in its conclusion. For every model, there were a small amount of non-constant variance that were present in subjects with very high blood pressure, which could have lead to inconsistency in the analysis. In that case, variance-stabilizing transformations may have been employed in order to combat this. Another caveat was the large number of points identified by DFFITS and DFBETAs. Due to the complex nature of the dataset, it is difficult to ascertain which points may be considered bad outliers, and so I abstained from removing any points from the analysis, however a closer inspection may have been helpful in this ordeal.

I conclude from this study that there is no relationship to be found between individuals who smoked and with blood pressure.