Introduction

The increasing prevalence of high blood pressure (HBP) in all age is recognized as a public health concern. Since many people with HBP also smoke, cigarettes is considered to be one of the important factors contributing to HBP. This study examining the relationship between combined systolic blood pressure reading and smoking have shown that people who smoke now were not more likely to have greater combined systolic blood pressure reading.

Methods

Data analysis

There are 743 data in total. 400 data were used in analyses and build models. The remaining 343 data are used as the test set. And there is no missing data in this data set. And after a brief treatment, it was found that the mean and median blood pressure of smokers was slightly lower than that of the others. In addition, education level, depressed level and gender seem to have little effect on blood press, too (see figure 1).

Techniques

This research applied VIF, AIC, BIC, LASSO, ridge penalty, box-cox transformation and so on to study the relationship between blood pressure and other factors and to construct models.

The VIF and correlation plots (figure 2) shows that there is a strong correlation between some predictors which may cause multicollinearity problem. So some variables need to be removed from models. Considering that BMI is a better parameter to measure weight and height, Race3_White has high correlation with other race group, Education_Some College also has high correlation with other education level, and poverty can be used as a measure of total annual gross income, the new rough model will remove Weight, Height, HHIcome group and education group except Education_Some College.

When selecting variables, I used some different methods, like Akaike information criterion (AIC), Bayes Information criterion (BIC), Least Absolute Shrinkage and Selection Operator (lasso). By applying these methods with both direction, AIC have selected 6 variables, BIC selected 3 variables, and lasso only remain 1 variable. In addition, when trying to apply shrinkage methods, just as I expected, ridge penalty does not shrink any coefficients of predictor to 0, so it can not be used for variable selection. Lasso penalty shrink lots of coefficients of variables to 0, except age and BMI.

After fitting new model for each variable selection methods, I used 10 fold cross validation to validate the accuracy of those models and try to avoid overfitting problems. And for each model, I have checked linearity and homoscedasticity with QQ plot and residual plots.

In this process, by normal QQ plot and residual plot, I find that all these model slightly violated homoscedasticity. So I try to apply box cox transformation to deal with this problem but after box cox transforming, the prediction error on test set become dramatically large.

Result

SmokeNow with HBSysAve

All p-value from linear regression model as well as variable selection methods indicated that smoking was not significantly associated with blood pressure reading, which is consistent with what we get from

box plot. The mean and median value of HBSysAve of smoke group is 122 and 119 respectively, compared to 127 and 124 in the non-smoking group.

Other variables with HBSysAve

According to AIC and BIC, Age, BMI and SleepTrouble were found to be associated with blood pressure readings and sleep trouble problems is reported to occur more frequently among people with higher blood pressure readings. Linear regression model reveals that blood pressure rises slightly with age and BMI. In addition, AIC model indicates that gender, physical activity and being married or not may affect blood pressure. However after checking correlation we can find that marital status has some correlation with age, and physical activity has some correlation with BMI.

Model

AIC:

$$HBSysAve = 90.5096 + 0.4478 * Age + 0.3922 * BMI + 2.8923 * Gender_{male} + 4.5113 * MaritalStatus_{NeverMarried} - 4.0423 * SleepTrouble_{Yes} - 2.6893 * PhysActive_{Yes}$$

Prediction error is 247.9896

BIC:

$$HBSysAve = 93.2852 + 0.4267 * Age + 0.3778 * BMI - 4.3624 * SleepTrouble_{Yes}$$

Predication error is 254.8991

LASSO:

$$HBSysAve = 102.4026 + 0.4329 * Age$$

Prediction error is 247.8412

The final model I choose is BIC model. Although the prediction error of BIC model a little higher than AIC's and LASSO's, LASSO model shrink too many variables to 0 and AIC contains some predictors with high correlation, like BMI and PhysActive_Yes, Age and MaritalStatus_NeverMared. In addition, the p-value of PhysActive_Yes is greater than 0.1, which is a little high.

Discussion

Results from this study indicated that blood pressure and smoking has no significant relation. However, age, BMI and sleep trouble problem may affect the blood pressure conspicuously. And by the regression analysis, we can conclude that with other variables held constant, blood pressure increased by an average of 0.4267 mm Hg for each year of age. Similarly, blood pressure increased by an average of 0.3778 mm Hg for each increase in BMI with other variables held constant. And blood pressure would decrease 4.3624 mm Hg on average when having sleep trouble compared to those without sleeping trouble, all else being equal. Thus this study stressed the potential importance of sleep quality and BMI to health by showing that the strong correlation among blood pressure and age, BMI and sleep trouble. So it is important to control BMI and improve sleep quality in order to keep blood pressure at a healthy level.

Limitation

Despite the generally accepted association between smoking and blood pressure, this study did not find a significant association between them. It may be due to the small sample size and the fact that the number

of smokers in the sample was slightly smaller than the number of non-smokers. In addition, by QQ-plot, it can be easily found that the error is not strictly follow normal distribution. It showed a little left skewed. However after applying box cox transformation, the prediction error will become dramatically large. This will affect the accuracy of this model.

Figure 1

Bpxplot of some catagorical predicators

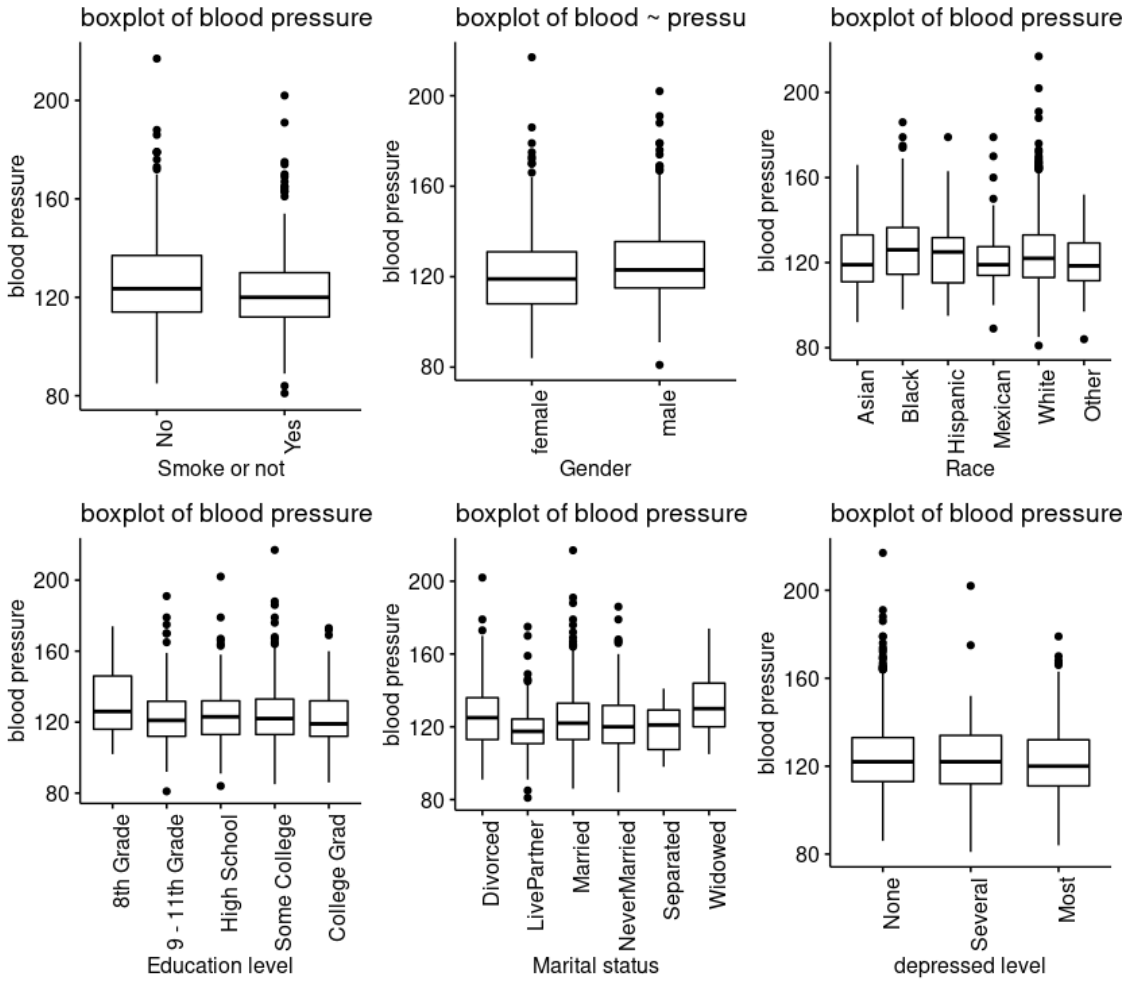

Figure 2

correlation plot of some predictors
Height
Weight
BMI
Race3_White
Race3_Black
Race3_Mexican
Race3_Hispanic
Race3_Other
Education_Some College
Education_High School
Education_9 - 11th Grade
Education_College Grad
Poverty
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

figure 3: distribution of BPSysAve

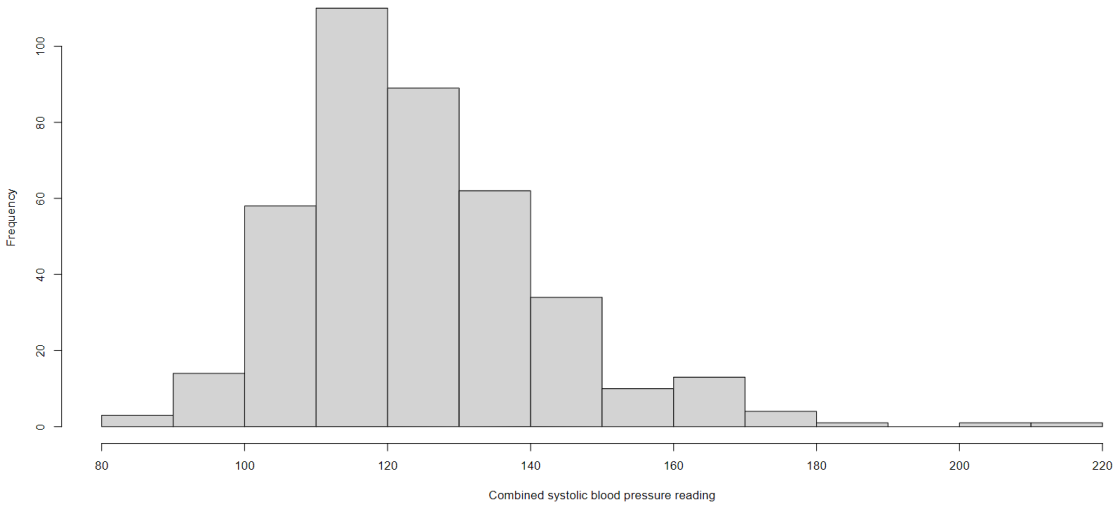

Table1.	Association between BPSysAve and other predictors					
	AIC model		BIC model		LASSO model	
	β	(95%)CI	β	(95%)CI	β	(95%)CI
Intercept	90.509	(80.43, 100.59)	93.2852	(84.47,102.09)	102.4026	(97.39,107.41)
Age	0.4478	(0.35,0.55)	0.4267	(0.34,0.52)	0.4329	(0.34,0.53)
BMI	0.3922	(0.13, 0.65)	0.3778	(0.12,0.64)		
Gender_male	2.8923	(-0.37,6.16)				
MaritalStatus_NeverMarried	4.5113	(0.098,8.92)				
SleepTrouble_Yes	-4.0423	(-7.39,-0.69)	-4.3624	(-7.71,-1.02)		
PhysActive_Yes	-2.6893	(-5.92,0.55)				