Data Report for Assignment 2 by Naihe Xiao

1. Introduction

Nowadays people are becoming increasingly concerned about high blood pressure. To discover possible relations between health indices and blood pressure, I'll explore a subset of NHANES, a dataset collected by the US National Center for Health Statistics. I'll try to establish the best prediction model with primary interest in the influence of smoking, to comprehend the connection between blood pressure and common health indicators, and predict one's blood pressure based on certain factors.

2. Methods

I'm primarily interested in the relationship between variables BPSysAve and SmokeNow. Specifically, BPSysAve is a continuous variable representing the Combined systolic blood pressure, and SmokeNow is a binary variable where "Yes" or "No" are decided by whether the participants smoke regularly, given that they are at least 20 years old and have smoked at least 100 cigarettes in their lifetimes.

I'll start with exploratory data analysis, including a side-by-side boxplot that illustrates the relationship between the median blood pressures of regular and non-regular smokers, a t-test that checks if the mean blood pressures between regular and non-regular smokers are equal, and a VIF check on the full model followed by deletion of highly correlated predictors.

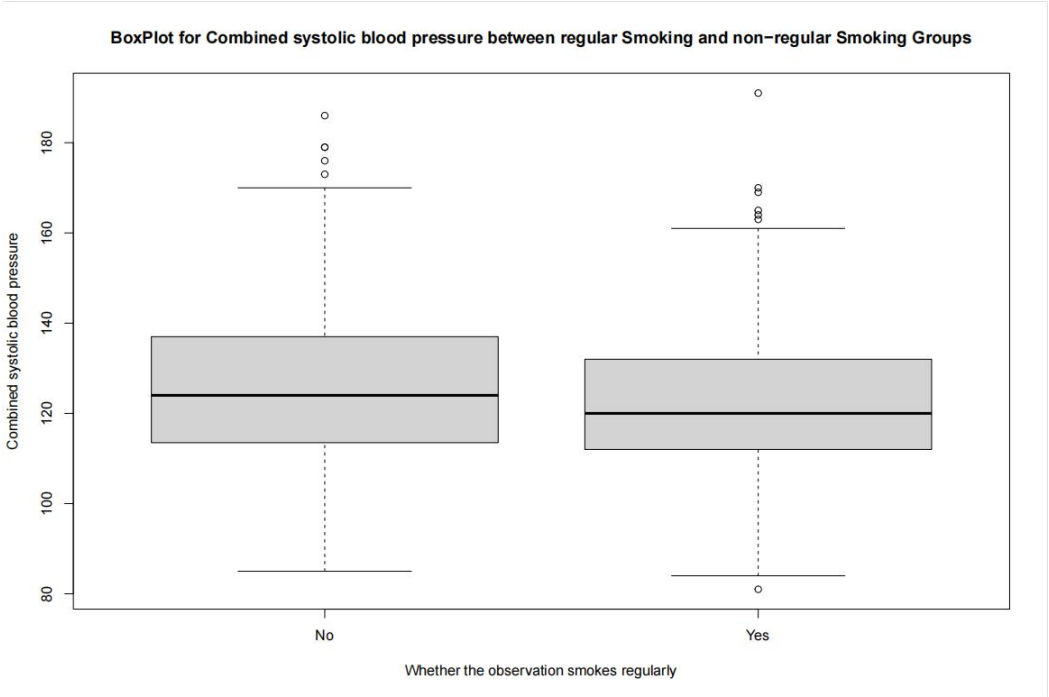
Then I'll generate 3 models based on criteria AIC, BIC, and LASSO. I'll use 5-fold cross-validation to conduct model validation, inspect the corresponding calibration plots, and determine the best prediction model after scrutinizing the prediction errors for the remaining models.

Finally, I'll perform diagnostics on the best model. I'll detect and remove influential points based on the Cook's distance, DFFITS, and DFBETAS. I'll examine the plot between predicted blood pressure and residuals to check linearity and homoscedasticity and parse the Normal Q-Q plot to check normality. If either the linearity or the normality assumption is violated, I'll implement the Box-Cox transformation on blood pressure; if the equal variance assumption is violated, I'll implement a variance stabilizing transformation on blood presure. Thirdly, I'll study the multi-collinearity, and delete predictors with high VIFs.

3. Results

i. Exploratory Data Analysis

There are 743 observations and 17 variables in this dataset, while the training set contains 400 observations and testing set contains 343 observations.

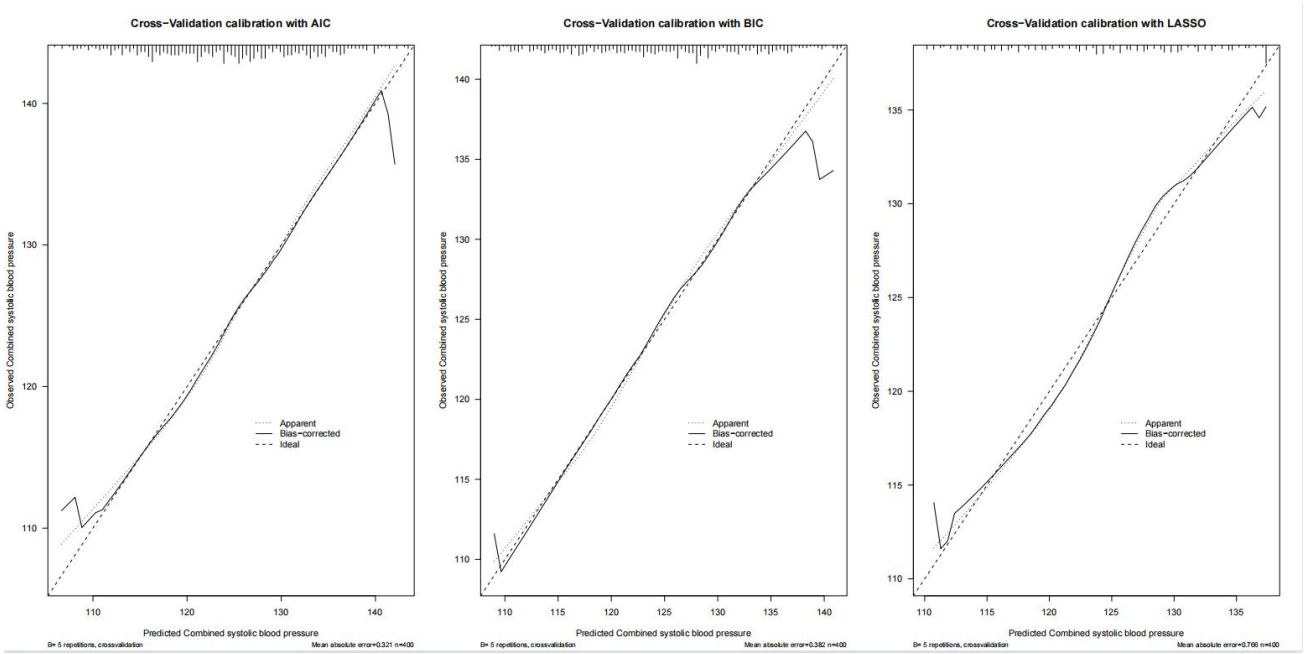


The boxplot shows that the median blood pressure of the regular smoking group is slightly lower than that of the non-regular smoking group.
The p-value for the t-test with null hypothesis stating the equivalence between the mean blood pressure of the two groups is 0.009314, indicating an unequality between the mean blood pressure of the regular and non-regular smoking groups.
Variables whose VIFs are greater than 5 include HHIncome, Poverty, BMI, Height, and Weight. After deleting variables BMI and HHIncome, I found that the VIF of all variables are smaller than 5, hence there is no more multi-collinearity issues in the new full model.

ii. Finding Final Model

The corresponding models generated based on AIC, BIC, and LASSO are Model A, Model B, and Model L, respectively. The AIC model includes predictors Gender, Age, Poverty, Height, and Depressed; the BIC model includes predictors Gender, Age, and Height; the LASSO model includes only predictor Age.

The five-fold cross validations of each model generate the following calibration plots.



The calibration plots reveal the poor performance of Model L, where the bias-corrected line is mostly far from the perfect prediction line; both bias-corrected lines for Model A and Model B are close to their ideal lines, but the performance of the line for Model B is better at two ends. Hence the cross-validation results suggest the supremacy of Model B.

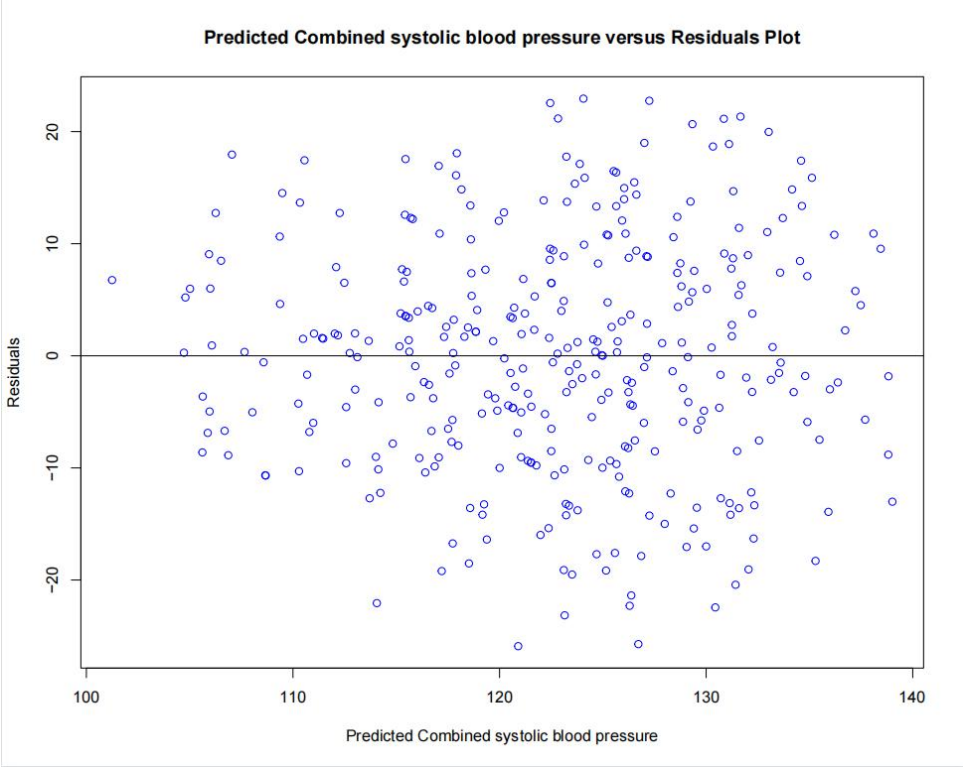
The prediction errors of the models are shown below, where Model B again performs the best.

Prediction Error Table		
Model A	Model B	Model L
269.5131	263.5091	264.0076

Since Model B has the smallest prediction error and the best cross-validation performance, I determined that Model B is the best model for predicting blood pressure. Since my interest is the relationship between smoking and blood pressure, I decide to add SmokeNow as a predictor to Model B to form the final model.

iii. Model Diagnostics

There are 13 leverage points and 72 influential points according to DFFITS check.



The predicted blood pressure versus residual plot shows that the residuals are normally distributed around the 0-line, and the residuals have approximately equal variances across the x-axis, justifying the linearity and the homoscedasticity assumptions. In addition, in the Normal Q-Q plot the points follow a straight line except for the tails, so the normality assumption holds.

The VIF of predictors Gender, Age, Height, and SmokeNow in the final model are 1.836886,1.136310,1.857938, and 1.090217, all much smaller than 5, hence there's no multi-collinearity.

4. Discussion

The summary table of my final model is exhibited below:

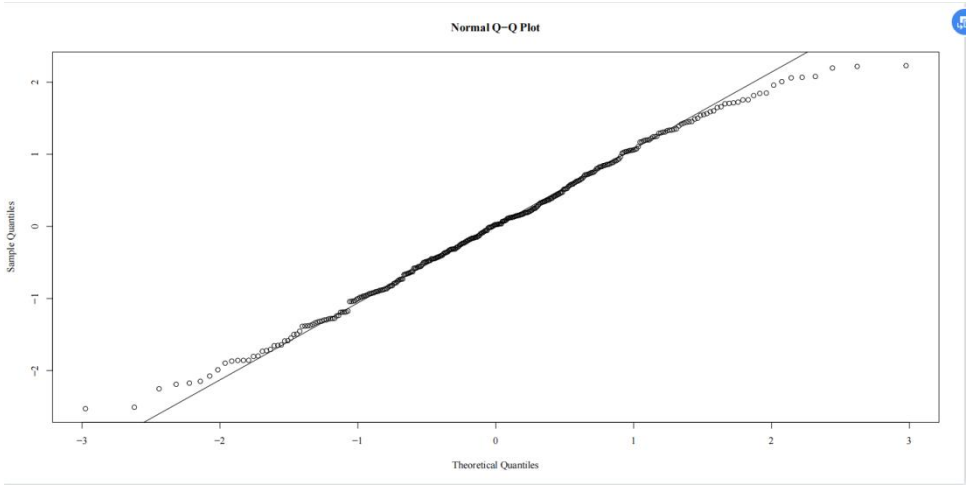
Final Model Summary Table

	Estimates	P-value of t-test
Intercept	144.37451	2e^-16
Gender|Male	7.40007	1.73e^-6
Age	0.41606	2e^-16
Height	-0.27456	0.000876
SmokeNow|Yes	0.41356	0.715899
Adjusted R-squared		0.3841

According to the table, the average blood pressure of a 0-years-old and 0-cm tall female who doesn't smoke regularly is 144.37451. Fixing other predictors at a constant value, the average blood pressure increases by 0.41606 when age increases by 1, increases by 7.40007 when gender is changed from female to male, decreases by 0.27456 when height increases by 1cm and increases by 0.41356 when the smoking status is changed from No to Yes.
The p-values for the t-tests suggest that all predictors are statistically significant except for SmokeNow, whose p-value is 0.715899, indicating strong evidence that the coefficient for SmokeNow is 0; that is, there are no linear association between SmokeNow and Combined systolic blood pressure. In other words, smoking has no effect on blood pressure.
The adjusted R-squared is 0.3841, meaning that 38.41% of the variability observed in blood pressure is explained by this model.
In general, the model has an outstanding performance, therefore is reliable for predicting the Combined systolic blood pressure of a future patient whose gender, age, height, and smoking status are known.
Limitations of this model include the small sample size to number of variables ratio, leading to high risk of Type I error during t-tests. Secondly, researches show a positive relationship between smoking and blood pressure, which my model fails to detect. Possible reasons include small sample size or dishonest participants who probably disavowed the fact of smoking regularly.
The applicability of the model can be enhanced with larger samples, more objective techniques for collecting data, and more comprehentive variable selection methodology like elastic net application.

Appendix

1. Specifically, the mean blood pressure for the smoking group is 121.7401, while the blood pressure for the non-smoking group is 126.3798 in the training set.
2. The mean absolute error for Model A, B, and L is 0.532,0.382,0.766, respectively.
3. The normal QQ plot



4. The simple linear regression model fitted with combined systolic blood pressure

against Smoke Now tells me that when SmokeNow status changes from No to Yes, the average combined systolic blood pressure decreases by 4.621 units, and the average combined systolic blood pressure for the non-smoking group is 131.083. The p-value is 0.00931, thus I reject the null hypothesis which states the coefficient is 0. However, I also observed that the adjusted R-squared is 0.0144, meaning that only 1.44 percent of the variability observed in the response variable is explained by this simple model, a very bad performance.