Analysis of combined systolic blood pressure reading in the U.S

Ying Xiong
1004795885

- **Introduction:**

1. Background of Study

It is universally acknowledged that blood pressure is a health measure index. High blood pressure indicates a large burden on heart and arteries, which increases the risk of diseases like heart attacks[1]. The National Health and Nutrition Examination Survey (NHANES) is the survey data collected by the US National Center for Health Statistics (NCHS). Approximately 5000 individuals of all ages have participated in the health examination since 1999. The raw data oversampled over the minority subpopulations. To reduce the oversampling effects, the dataset contains 10000 resampled observations and has 75 variables including the combined systolic blood pressure.

2. Statistical Objectives

The primary interest of my analysis is to identify the best factors for prediction and generate a model to predict blood pressure with the least errors for people older than 17. As there has arisen debates about the relationship between smoking and blood pressure, I would also statistically research the effect of smoking on blood pressure for the same group.

- **Methodology:**

For the prediction model, I will use multiple linear regression where the response is the combined systolic blood pressure and the predictors will be the best set of variables for prediction.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon_i ,$$

where Y is the response; X_i are the predictors; β_i stands for the regression parameters; ε_i stands for the random errors and follows $N \sim (0, \sigma^2)$.

1. Variable Selection

To choose the best subset of predictors, I will use two different methods, the stepwise variable selection and the shrinkage method LASSO. For the stepwise approach, I will apply both AIC and BIC as criterion when examining all possible regression models. Thus, there will be three models with the "best" variables.

- Akaike's Information Criterion ($AIC \propto nlog\left(\frac{RSS}{n}\right) + 2p$)

  Preferred model will have the lowest AIC.
- Bayesian Information Criterion ($BIC \propto nlog\left(\frac{RSS}{n}\right) + (p+2)\log(n)$)

  Preferred model will have the lowest BIC.
- LASSO

  The LASSO regression will shrink some of the parameter estimates to 0 and thus result in a set of variables that can produce the least mean squared error.

[1] https://heartinsight.heart.org/February-2014/Why-blood-pressure-matters/

2. Model Validation

I will use cross-validation to check the prediction accuracy because it can identify the ability of models to accurately predict the response. I will prefer a model with less deviation from the 45-degree line in the cross-validation calibration.

3. Model Diagnostics

- MLR Model Assumption Violation Checks:
  - *linearity* of the relationship, *homoscedasticity*, *independence* and *normality* of errors.

I will check the residual plots to see whether there is symmetric pattern in the residuals. The residuals should uniformly scatter around 0. Linearity is violated when there is a curve pattern. Homoscedasticity is violated when there is a fanning pattern. Independence is violated when there are clusters of residuals with obvious separation from the rest. For normality, I will check the Q-Q plot to see whether residuals are in a linear trend matching the qqline plotted with them. If not, then normality is violated. I will do models transformations when the assumptions are violated. When there is only a violation in the normality, I will consider removing outliers/influential observations or ignore it if the violation is small.

- MLR Diagnostics:

I will identify the *leverage points* by calculating the hat values of each observation and threshold will be $h_{ii} > \frac{2(p+1)}{n}$, where p is the number of predictors and n is the number of observations. Then I will identify the *influential observations* using three methods: Cook's Distance ($> \frac{4}{n-p-1}$), DFBETA ($> \frac{2}{\sqrt{n}}$), and DFFITS ($> 2\sqrt{\frac{p+1}{n}}$). I will remove observations that appear under either two methods. For *multicollinearity*, I will check the variance inflation factor ($VIF_j = \frac{1}{1-R_j^2}$) and remove variables with a VIF > 5.

4. Effect of Smoking

I will fit a multiple linear regression model with 15 predictors and check the p-value of variable SmokeNow. With a cutoff of 0.05, I will make a conclusion on the effect of smoking giving the null hypothesis that there is no relationship between smoking and blood pressure.

• Results:

1. Data Description

The dataset I used in my analysis consists of 743 observations and 17 variables, selected from the raw data NHANES. I removed observations with age under 17 and repeated IDs and named the new dataset small.nhanes. Except for the response variable BPSysAve and the ID variable, there are 6 numerical predictors (Age, Poverty, Weight, Height, BMI, SleepHrsNight) and 9 categorical ones (Gender, Race3, Education, MaritalStatus, HHIncome, Depressed, SleepTrouble, PhysActive, SmokeNow). I separated the data into a train set (400 observations) and a test set (343 observations) for further analysis.

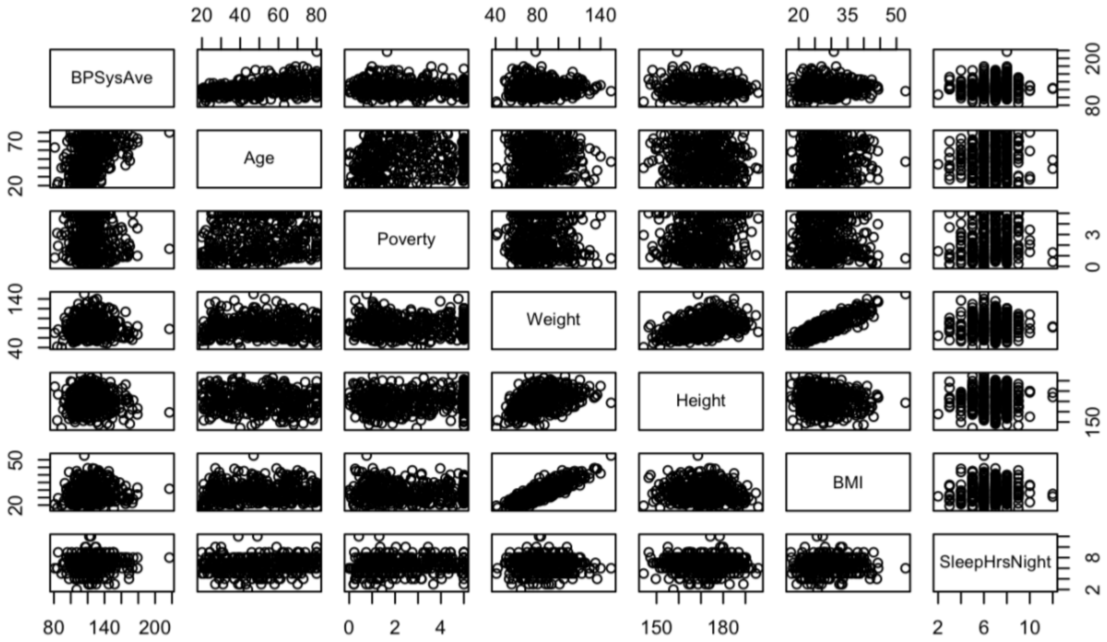


Figure 1. Correlation plots of numerical variables

As shown in the plots, the response BPSysAve has a weak positive linear relationship with the variable Age and Poverty and unclear relationship with the others. Also, Weight and BMI have a strong correlation, which means these two variables will lead to the multicollinearity if we include both in a model.

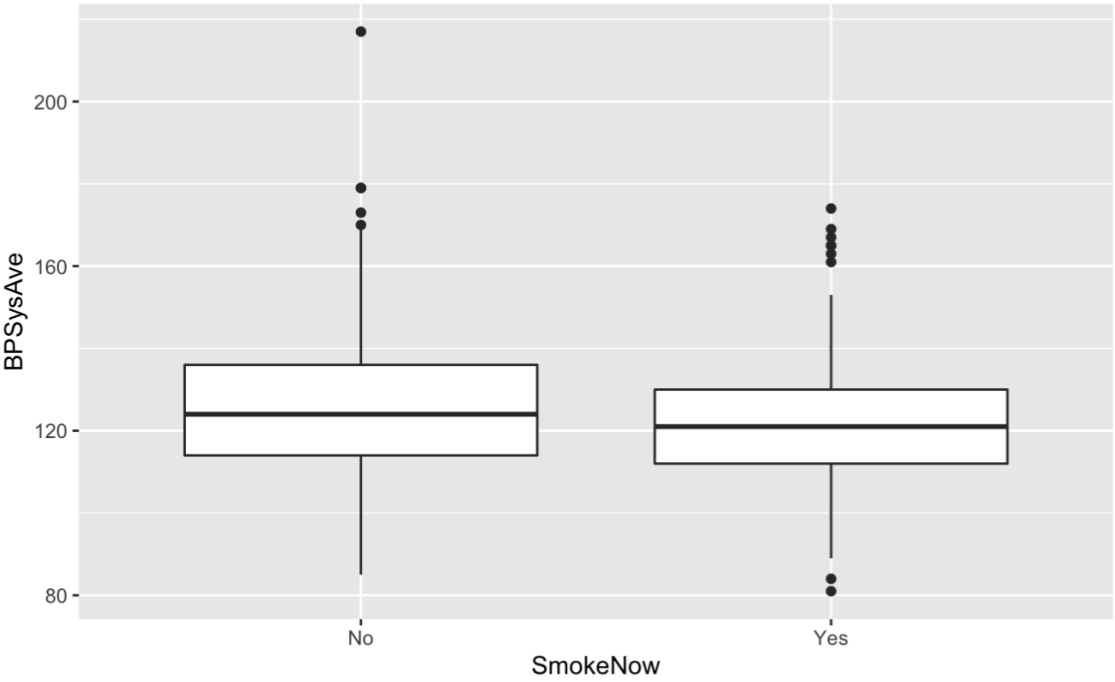


Figure 2. Boxplots of SmokeNow and BPSysAve

For categorical variables, I will focus on the discussion about SmokeNow. As exhibited, the shapes of two levels are similar and have close means while the range of no is wider. This may imply that the blood pressure is not much affected by smoking.

Before the analysis, I checked the model diagnostics with the residual plots and Q-Q plots. With all other assumptions satisfied, there is a small violation in normality (Appendix 1). I decided to remove one outlier observation because I believed it will deviate my line of prediction from the true one. That observation is considered as influential in DFFITS and DFBETA. There are no influential observations identified by all three methods. Thus, I will use 399 observations in the train set for further modeling.

2. Effect of Smoking

$$H_0:\ \beta_{SmokeNow} = 0, H_A:\ \beta_{SmokeNow} \neq 0$$

I fitted a multiple linear regression model with all variables and found that the p-value for variable SmokeNow is 0.49, which is much larger than the cutoff the 0.05. With the hypothesis test, we can conclude that there is no evidence against the null hypothesis of smoking does not affect the blood pressure reading.

3. Model Selection

I applied the stepwise variable selection and shrinkage method LASSO and got 3 models. For AIC and BIC, I used stepwise selection and checked the VIF. There are no variables with VIF >5 in either model. The variables chosen in each model are listed below:

- AIC: Gender, Age, Poverty, Height
- BIC: Gender, Age
- LASSO: Age

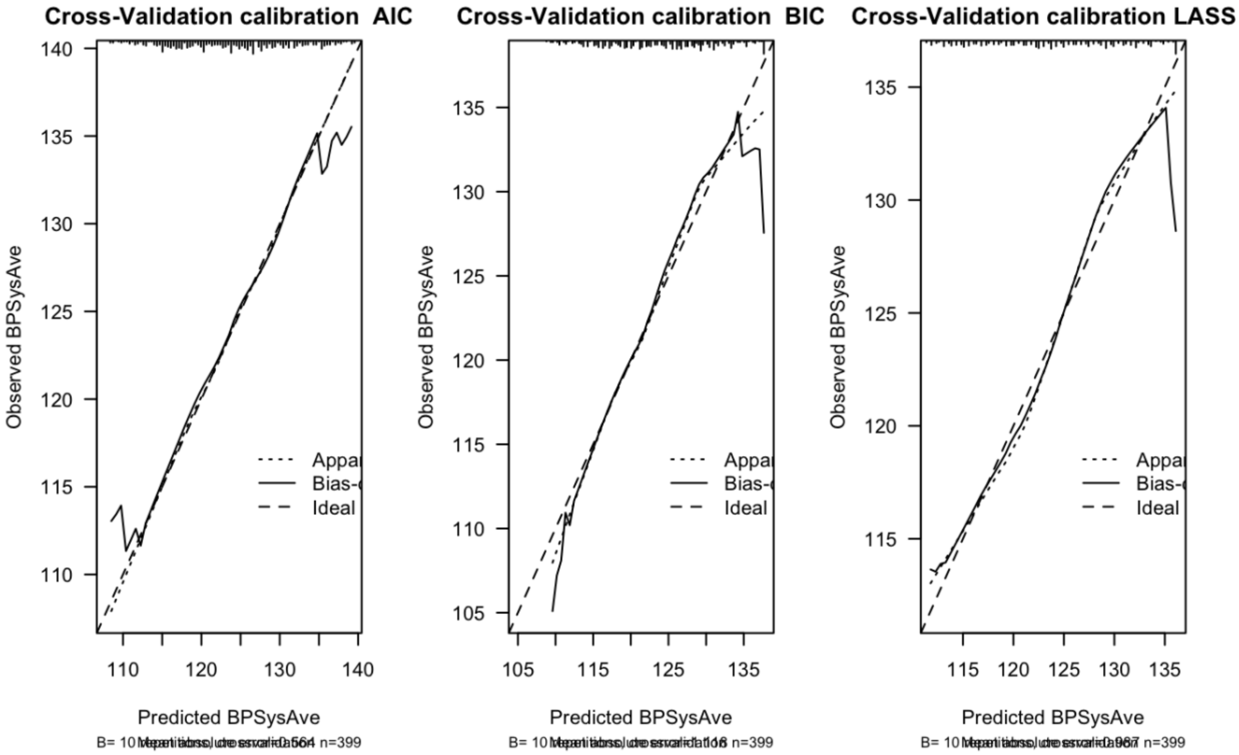


Figure 3. Cross-Validation Calibration Comparison between models

As shown, the AIC model has the least variation from the 45-degree line as most part perfectly matches the line with small differences at the ends.

	Full Model	AIC	BIC	LASSO
Prediction Error	276.577	267.043	269.233	273.950

Figure 4. Table of Comparison between models

As displayed, the stepwise model with AIC criteria has the least prediction error. Since our aim is to predict accurately, I chose AIC to be the final model.

4. Final Model Diagnostics

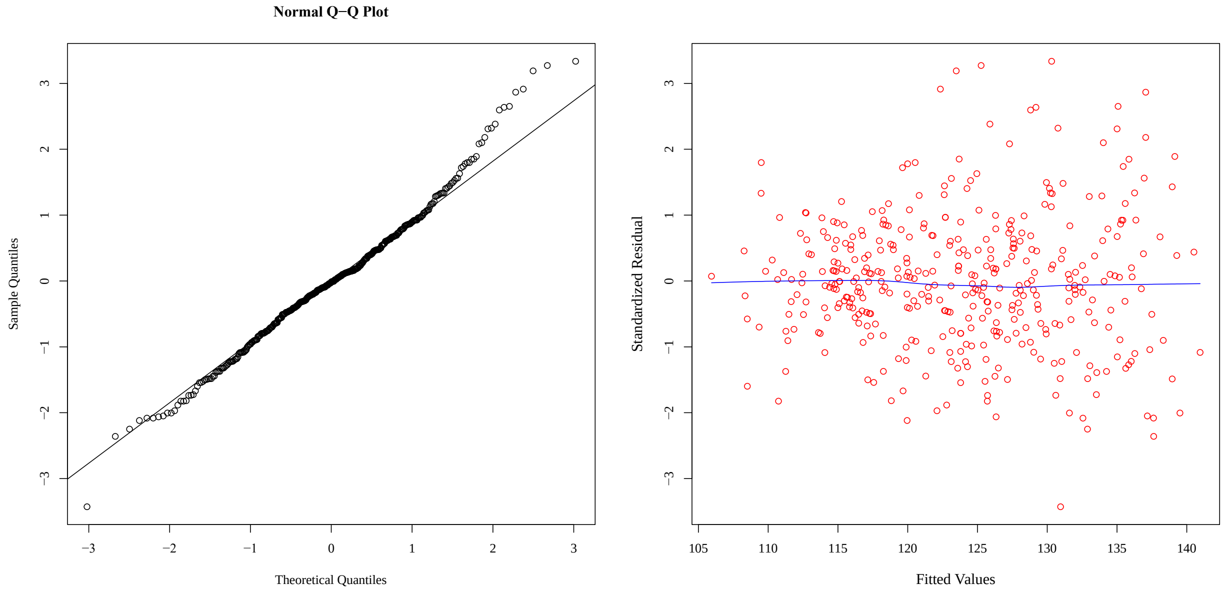


Figure 5. Q-Q plot and Residual plot of AIC model

As shown, residuals scattered uniformly around 0. There is no curve pattern, fanning pattern, and clusters. Hence, the assumptions of linearity, homoscedasticity, and independence of errors are met. From the Q-Q plot, we can see a slight variation at the upper right corner while all the other residuals match the qqline. We can ignore the small violation here as others are well satisfied. For model diagnostics, there are no influential points identified by Cook's distance and DFFITS, while there are a lot of leverage points. Multicollinearity does not exist here as no predictor has VIF > 5.

- **Discussion:**

1. Final Model Interpretation

$$\widehat{BPSysAve} = 132.89 + 6.95 * Gendermale_i + 0.4 * Age_i - 0.96 * Poverty_i - 0.18 * Height_i$$

The model above implies that the combined systolic blood pressure reading will increase by 0.4 as age increases, decreases by 0.18 as the poverty index decreases, and decreases by 0.96 as height decreases. A decrease in the poverty index indicates more poverty. On average, males' blood pressure reading is 6.95 higher than females. However, the p-values for variable poverty (0.045) and height (0.07) warn us to consider more carefully when including these two variables because there is little evidence against the null hypothesis that either of them is non-related to the response.

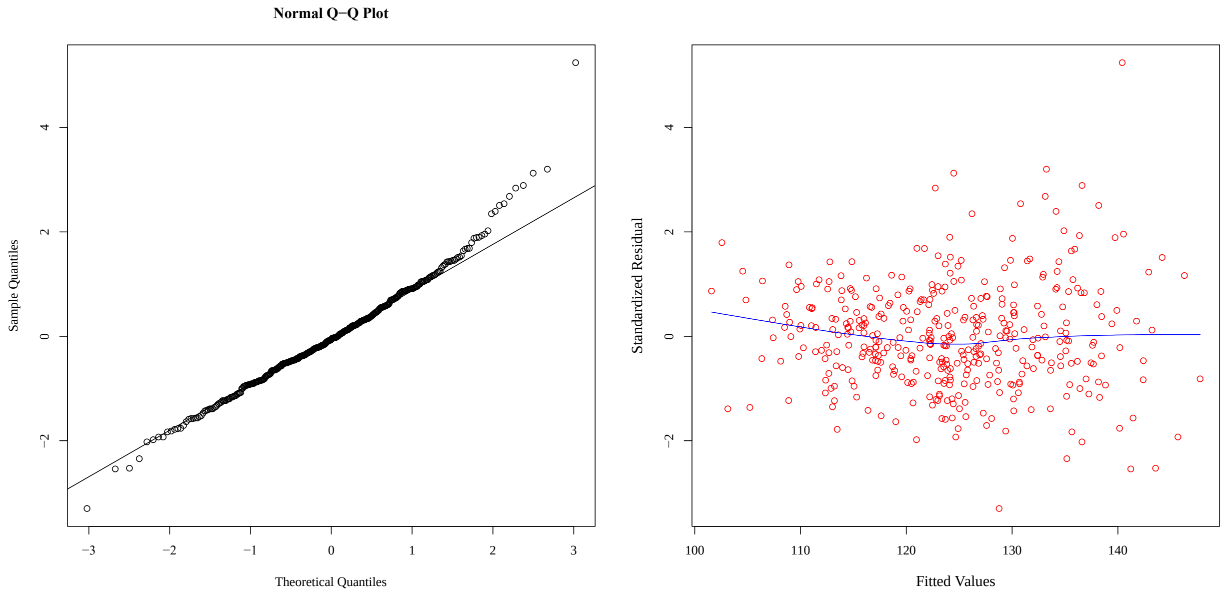
Thus, with this model, we can predict the combined systolic blood pressure of a person if we know his/her gender, age, poverty index, and height. We can reduce the risk of diseases like strokes by keeping attention on changes in these factors. This is the best model we can apply in the future medical crisis prevention system.

2. Limitations

There are two limitations to the final model. The first is that its prediction error (267.043) would still be considered high though it is the least. This implies discounts on the usefulness of the model as it cannot precisely predict the blood pressure. The second limitation is the variation in variable selection. During experiments, I found removing observations has a large impact on the variable selected in different models. Problems like overfitting may take place. Thus, the AIC model could be enhanced if we have use different training sets or including more observations.

Appendix

Appendix 1. Q-Q plot and Residual plot of Full model



Work Cited

American Heart Association. “Why Blood Pressure Matters.” *Heart Insight Mag*, Feb. 2014, heartinsight.heart.org/February-2014/Why-blood-pressure-matters/.