STA302 Final Project

Summer 2020, U of T

Ruo Ning Qiu
1004079631

June 25, 2020

Contents

1 Introduction 2

2 Methods 3

2.1 Variable Selection . . . 3

2.2 Variable Validation . . . 3

2.3 Variable Diagnostics/Violations . . . 3

3 Results 4

3.1 Data Description . . . 4

3.2 Process of Obtaining the Final Model . . . 5

3.3 Goodness of the Final Model . . . 7

4 Discussion 8

4.1 Interpretation and Importance . . . 8

4.2 Limitations and Future Direction . . . 8

Bibliography 9

A Appendix 10

Chapter 1

Introduction

In this project, we will examine and fit a multiple linear regression model of a small subset of the NHANES (National Health and Nutrition Examination Surveys) data collected by the US National Center for Health Statistics in 2011-12. The data includes 15 predicting variables to be chosen from for the response variable – combined systolic blood pressure reading from people with age > 17 examined at 30 different study locations with various races [4]. In addition, we are interested in the effects of smoking on blood pressure and constructing an ideal model for prediction, inferences and interpretation by identifying which variables out of the 15 predictors are significant.

Chapter 2

Methods

2.1 Variable Selection

In order to produce an ideal model for the interest of prediction and interpretation of the blood pressure reading, we need to perform several model selections such as stepwise regression of AIC & BIC, and with shrinkage methods such as Elastic-Net & LASSO to eliminate the statistically insignificant predictors, to identify the appropriate predicting variables while preventing the model to have problems like over/under-fitting.

2.2 Variable Validation

To confirm the accuracy of the prediction of a model and detect over/under-fitting, we must validate it, as the model could be a great fit to the data that we used, but not to any other data from the population. A popular method for model validation is k-fold Cross Validation, where the data is randomly split into k-parts. It fits the model with $k-1$ parts, predicts the outcomes for the remaining part, and use all the k-parts as the test set. The predictions can be plotted with the observed values to check the accuracy of the estimates visually by using calibration plots that estimate the consistency of model performance across portions of the data.

2.3 Variable Diagnostics/Violations

We want to check if our assumptions of a multiple linear regression model – linearity, homoscedasticity, and normality – are satisfied, while avoiding problems of multicollinearity and bad leverage points that are problematic for inferences. Thus, we should check Normal Q-Q plot for normality, residuals plot for linearity and homoscedasticity assumptions, to see if we need to perform transformation on the dataset if the linearity assumption failed as patterns appear in these plots. Moreover, we should calculate Cook's distance, DFFITS, and DFBETAS for determining the influential observations, and possible outliers. Interaction between the predictors might be an issue for modelling; however, most of them are categorical variables. We will check the correlation between numerical predictors using the variance inflation factor (VIF) after the model selection.

Chapter 3

Results

3.1 Data Description

The data that we analyze is a small subset of NHANES dataset from people with age > 17 that includes 15 predictor variables – one's gender **(Gender)**, age **(Age)**, race **(Race)**, education level **(Education)**, marital status **(MaritalStatus)**, total income per year **(HHIncome)**, poverty rate **(Poverty)**, weight **(Weight)**, height **(Height)**, body mass index **(BMI)**, depression level **(Depressed)**, average sleep hours per night **(SleepHrsNight)**, insomnia **(SleepTrouble)**, physicality **(PhysActive)**, and smoking status **(SmokeNow)**. The response variable is the combined systolic blood pressure reading **(BPSysAve)**. In total, there are 743 observations each labelled by a unique number **(ID)** where 400 are randomly selected as the training set to fit the model. The rest 300 are used as the testing set to calculate prediction error. The followings are visualization plots that describe the characteristics of the training dataset, with a detailed table attached in appendix A as table A.1.

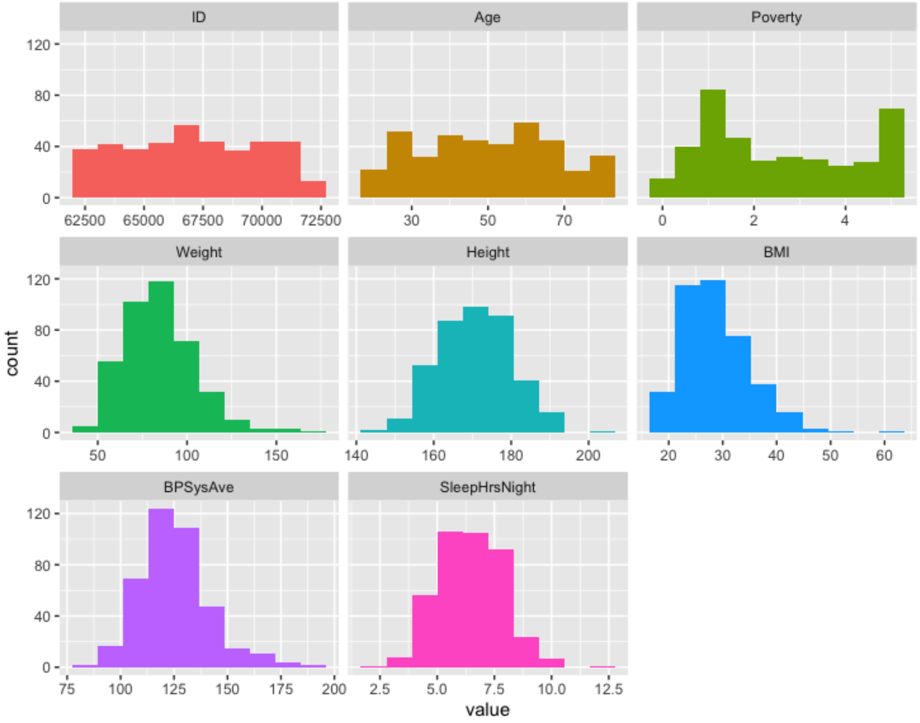

Figure 3.1: The Distribution of Numerical Variables in the Training Set

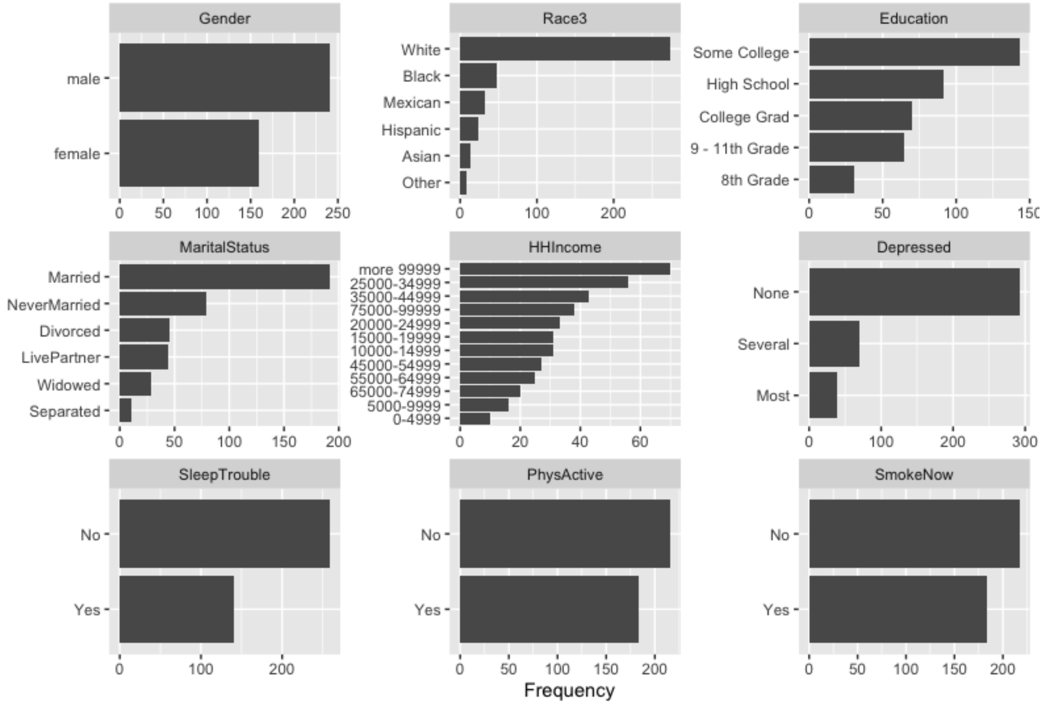

Figure 3.2: The Distribution of Categorical Variables in the Training Set

3.2 Process of Obtaining the Final Model

As mentioned in 2.1, performing several model selection methods that are validated with cross validation produces 4 candidate models. There aren't any influential observation that falls under Cook's distance, DFFITS, and DFBETAS for the multiple linear regression with all 15 predictors, so we would not remove any data points in the training set before applying model selection. No patterns are found in the Normal Q-Q plot and residual plots, so we would not apply any transformation before the model selection as well. Below is a summary chart of the criteria values from the candidate models.

Table 3.1: Summary of Model's Criteria

	R^2	$R^2_{Adjusted}$	AIC	AICc	BIC	Prediction Error
AIC Model	0.249	0.228	2167.69	2168.62	2223.57	289.18
BIC Model	0.194	0.190	2177.78	2177.88	2197.74	283.90
Elastic-Net Model	0.175	0.173	2185.10	2185.16	2201.08	286.82
LASSO Model	0.175	0.173	2185.10	2185.16	2201.08	286.82

[1] Best option is coloured in blue.

Notice that the Elastic-Net and LASSO both select the same variable, age **(Age)** only, which is why they have exact same values in table 3.1. At this stage, AIC model that has the greatest

$R^2_{Adjusted}$ and smallest AIC, AICc values, and BIC model that has the smallest prediction error and BIC value, are the more competitive models. The followings are the calibration plots of candidate models.

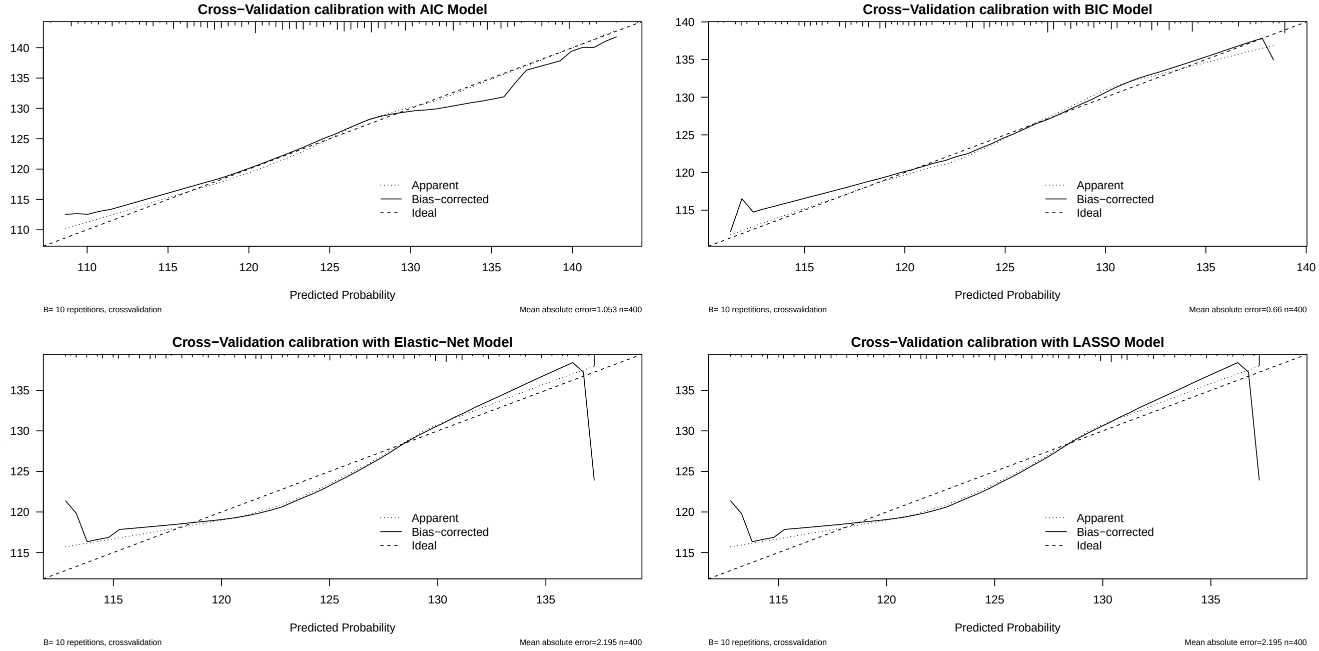

Figure 3.3: The Cross-Validation Calibration Plots

There are no large disparities between the calibration plots for each model, while the Elastic-Net and LASSO models are getting further away from the ideal line at the two ends. They demonstrate an under-forecasting (above the line) pattern in the beginning with small probabilities and over-forecasting (below the line) pattern at the end with large probabilities. AIC and BIC models are closer to the ideal line relatively with a lower mean absolute error (1.05 and 0.66 respectively, 2.195 for LASSO model).

Evaluating from all the criteria (table 3.1 & fig. 3.3), we will choose the final model from AIC or BIC model by inspecting the correlation (VIF values) between the chosen predictors.

For AIC model, 3 of the predictors, weight, height, and BMI have VIF value 9, 10, and 11 respectively, which is greater than the common cutoff. In comparison, predictors from BIC model do not have any VIF greater than 5. This implies that AIC model might have multicollinearity issue where the 3 predictors are correlated that hinders the inferences of the model.

As a result, BIC model is chosen as the final model. We are also interested in how the smoking status of the participants affects blood pressure. Although none of the variable selection methods includes the smoking status as a predictor variable, the final model will include it due to its scientific importance outlined in these research papers [3, 5, 6]. In addition, the p-value of the t-test on the blood pressure with smoking status is statistically significant under the 95% confidence interval (table A.2).

3.3 Goodness of the Final Model

The final model is the multiple linear regression with categorical predictors gender **(Gender)**, smoking status **(SmokeNow)**, and numerical predictor age **(Age)** for response variable combined systolic blood pressure reading **(BPSysAve)**. A summary of the model is given in table A.3.

A research paper [1] supported that men would have a higher blood pressure than women and blood pressure increased with older ages. This is consistent with our findings as to the estimated coefficients for the predictors – age and gender – are positive (table A.3). The p-values for gender and age are statistically significant under the 95% confidence interval. Although the p-value for the smoking status predictor is insignificant, the model includes it for scientific interest.

For inference and diagnostic checks of the final model, there aren't any influential observation that falls under Cook's distance, DFFITS, and DFBETAS, so we would not remove any data points in the training set. There are also no correlated predictors under VIF cutoff.

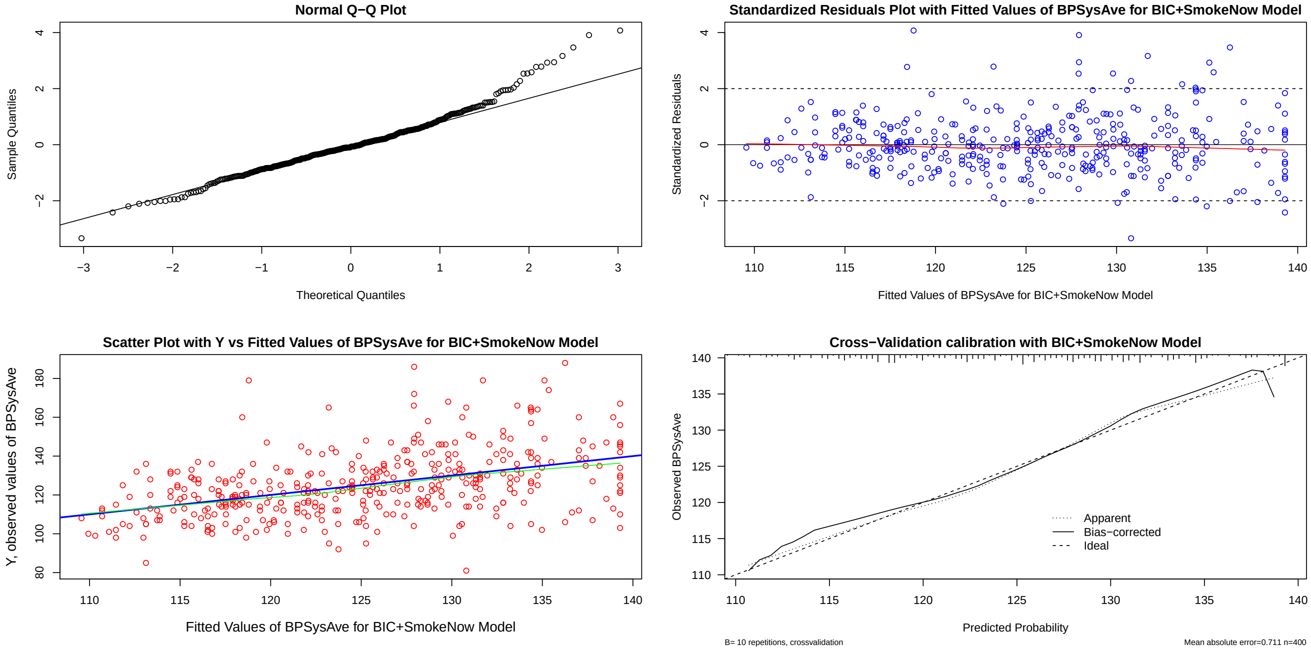

Figure 3.4: Diagnostic and Calibration Plots of the Final Model

In fig. 3.4, observe that the normality assumption holds as the middle component stays close to the line for the Normal Q-Q plot. There are no obvious patterns in the residuals and scatter plots, so linearity and homoscedasticity hold and we would not apply any transformation. The new cross validation calibration plot has a pretty similar pattern as BIC model, with a slight increase in mean absolute error from 0.66 to 0.71. The prediction error shows a small increase from 283.90 to 287.12 (table A.3), but is still lower than AIC model's and very similar to LASSO model's (table 3.1). The $R^2_{Adjusted}$ increases from 0.1903 to 0.1914, while AIC, AICc values increase by around 0.44.

Chapter 4

Discussion

4.1 Interpretation and Importance

The final model states that if there is a unit increase in the age of an individual, then one's average blood pressure will increase by 0.38 times more. If the individual is male, the average blood pressure of this individual is nearly 5 times more than a female individual's blood pressure. If the individual smokes, the blood pressure of this individual is approximately twice lower than a non-smoker on average. Although the estimated coefficient for the smoking status predictor is negative (table A.3), which is very counter-intuitive, this result is consistent with this study paper [2]. It supported that smokers have a lower blood pressure than non-smokers.

The project is important since it examines the factors that might affect the blood pressure of U.S. citizens older than 17-year-old and identifies which of the factors are confounding. As discussed in section 3.3, the final model maximizes the accuracy of prediction and inferences among all the models with statistical and scientific significance. The model reveals how blood pressure is affected and can be used to extrapolate blood pressure data in order to provide health suggestions based on one's age, gender and smoking status.

4.2 Limitations and Future Direction

A strength of this study is it uses a sample from a large dataset as 9338 U.S citizens were examined [4]. However, this study is also limited by the usage of NHANES dataset that is self-reported information with potential response bias. The adjusted coefficient of determination is very low ($R^2_{Adjusted} \approx 0.191$) for the final model; in fact, all the candidate models have very low $R^2_{Adjusted}$ (table 3.1). This might imply a weak correlation between the given predictors and blood pressure. Future studies can be targeted at investigating other factors that also affect blood pressure. One could repeat this study with other samples to conclude more consistent information regarding the effects of smoking on blood pressure.

Bibliography

[1] Hiroyuki Daida et al. “Peak Exercise Blood Pressure Stratified by Age and Gender in Apparently Healthy Subjects”. In: *Mayo Clinic Proceedings* 71 (May 1996), pp. 445–52. DOI: 10.4065/71.5.445.

[2] Manfred S. Green, Eliezer Jucha, and Yair Luz. “Blood pressure in smokers and nonsmokers: Epidemiologic findings”. In: *American Heart Journal* 111.5 (1986), pp. 932–40. ISSN: 0002-8703. DOI: https://doi.org/10.1016/0002-8703(86)90645-9. URL: http://www.sciencedirect.com/science/article/pii/0002870386906459.

[3] Antonella Groppelli et al. “Persistent blood pressure increase induced by heavy smoking”. In: *Journal of hypertension* 10.5 (June 1992), pp. 495–9. DOI: 10.1097/00004872-199205000-00014.

[4] National Center for Health Statistics. *NHANES 2011-2012 Overview*. Feb. 2020. URL: https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Overview.aspx?BeginYear=2011.

[5] James D. Neaton and Deborah Wentworth. “Serum Cholesterol, Blood Pressure, Cigarette Smoking, and Death From Coronary Heart Disease Overall Findings and Differences by Age for 316099 White Men”. In: *Archives of Internal Medicine* 152.1 (Jan. 1992), pp. 56–64. DOI: 10.1001/archinte.1992.00400130082009.

[6] Per Omvik. *How smoking affects blood pressure*. 1996. URL: https://www.tandfonline.com/doi/abs/10.3109/08037059609062111.

Appendix A

Appendix

Table A.1: Characteristics of the training set ($n = 400$)

Gender	Age	Race3	Education	MaritalStatus
female: 159	Min.: 20.00	Asian: 14	8th Grade: 31	Divorced: 45
male: 241	1st Qu.: 35.75	Black: 47	9 - 11th Grade: 65	LivePartner: 44
	Median: 50.50	Hispanic: 24	High School: 91	Married: 192
	Mean: 50.14	Mexican: 32	Some College: 143	NeverMarried: 79
	3rd Qu.: 63.00	White :274	College Grad : 70	Separated: 11
	Max.: 80.00	Other: 9		Widowed: 29

HHIncome	Poverty	Weight	Height	BMI
more 99999: 70	Min.: 0.000	Min.: 44.40	Min.: 141.3	Min.: 16.70
75000-99999: 38	1st Qu.: 1.117	1st Qu.: 69.83	1st Qu.: 164.3	1st Qu.:23.90
35000-44999: 43	Median: 2.185	Median: 82.70	Median: 171.2	Median :28.20
25000-34999: 56	Mean: 2.550	Mean: 84.33	Mean: 170.8	Mean: 28.85
20000-24999: 33	3rd Qu.: 4.072	3rd Qu.: 95.30	3rd Qu.: 177.3	3rd Qu.: 32.60
10000-14999: 31	Max.: 5.000	Max.: 172.50	Max.: 200.4	Max.: 59.10
(Other): 129				

Depressed	SleepHrsNight	SleepTrouble	PhysActive	SmokeNow
None: 292	Min.: 2.00	No: 259	No: 216	No: 217
Several: 70	1st Qu.: 6.00	Yes: 141	Yes: 184	Yes: 183
Most: 38	Median: 7.00			
	Mean: 6.74			
	3rd Qu.: 8.00			
	Max.: 12.00			

ID	BPSysAve			
Min. :62172	Min. : 81.0			
1st Qu.: 64513	1st Qu.:114.0			
Median: 67116	Median :123.0			
Mean: 67032	Mean :125.1			
3rd Qu.: 69506	3rd Qu.:133.0			
Max.: 71868	Max. :188.0			

Table A.2: Summary of the Two Sided Sample T-test of BPsysAve and SmokeNow

	Estimate	Statistic	P-value	95% Confidence Interval
Sample T-test	6.50	3.99	0.0000789	(3.29, 9.70)

Table A.3: Summary of the Final Model

	Estimate	SE	T-value	$Pr(> \|t\|)$	
(Intercept)	104.002	2.903	35.830	$< 2e^{-16}$	***
GenderMale	4.916	1.559	3.154	0.00173	**
Age	0.380	0.047	8.067	$8.67e^{-15}$	***
SmokeNowYes	-2.040	1.638	-1.245	0.21371	
Residual SE = 15.18	DF = 396	$R^2_{Adj} = 0.191$	F-stat. = 32.49	P-value = $< 2.2e^{-16}$	
AIC = 2178.21	AICc = 2178.4	BIC = 2202.2	PE = 287.12		

[1] Significant Codes: 0 ‘$***$’, 0.001 ‘$**$’, 0.01 ‘$*$’, 0.05 ‘.’