SUMMARY

The goal of this study is to determine the risk factors of high blood pressure. We used logistic regression to analyze the data of 15643 U.S. adults from the National Health and Nutrition Examination Survey (NHANES III)[1]. The results of the analysis show seven health conditions increase people's risk for high blood pressure. Some risk factors cannot be controlled, like age or race. Changes on the lifestyle are able to significantly lower the risk of high blood pressure, like losing weight or quitting tobacco especially for men.

1. INTRODUCTION

In this study, we studied the possible contributions to the probability of getting high blood pressure with data from the National Health and Nutrition Examination Survey (NHANES III), which was conducted by the National Center for Health Statistics (NCHS) from 1988 to 1994. The survey was performed through both physical examinations and clinical and laboratory test. There are 17030 subjects in the dataset, who were all adults age 20 and older and represented 177.2 million adults living in the U.S. at that time.

The health information from the dataset consists of 16 variables in total: Identification Number, Pseudo-PSU, Pseudo-stratum, Statistical Weight, Age, Sex, Race, Body weight, Standing Height, Average Systolic BP, Average Diastolic BP, Smoking History, Current Smoking, Smoking, Serum Cholesterol and High Blood Pressure. Detailed description of the variables are shown in Table A.1 in the Appendix.

The results of our analysis are important because we figured out several risk factors of high blood pressure. Despite some uncontrollable factors, the results offer some clear and feasible advice to help people decrease the risk of high blood pressure and boost their wellbeing. That is to say, through analyzing the results of this analysis, we could minimize our probability of getting high blood pressure by changing some lifestyle.

The report is structured as following: we performed an exploratory data analysis in Sec. 2, followed by selecting models in Sec. 3. In Sec. 4, we analysed our results from the final model. Conclusion is in Sec. 5. Detailed information about the fitted models are provided in Appendix., which is the final section of our report.

2. EXPLORATORY DATA ANALYSIS

An exploratory data analysis was conducted before modelling the data. To begin with, we performed some data cleaning on the variables. We neglected the statistical weight of subjects and excluded variable 1 - 4. Since variable 14 ("Smoking") is made up by variable 12 ("Has respondent smoked 100 cigarettes in life") and variable 13 ("Does respondent smoke cigarettes now"), we kept variable 14 instead of variable 12 and 13.

[1] https://wwwn.cdc.gov/nchs/nhanes/nhanes3/

The variables Average Systolic BP and Average Diastolic BP are quantities showing blood pressure and hence not considered factors contributing to high blood pressure in our analysis. In particular, the Average Systolic BP is used to decide whether the subject is high blood pressure or not. From Figure 1.1, it is shown that the subject with Average Systolic BP

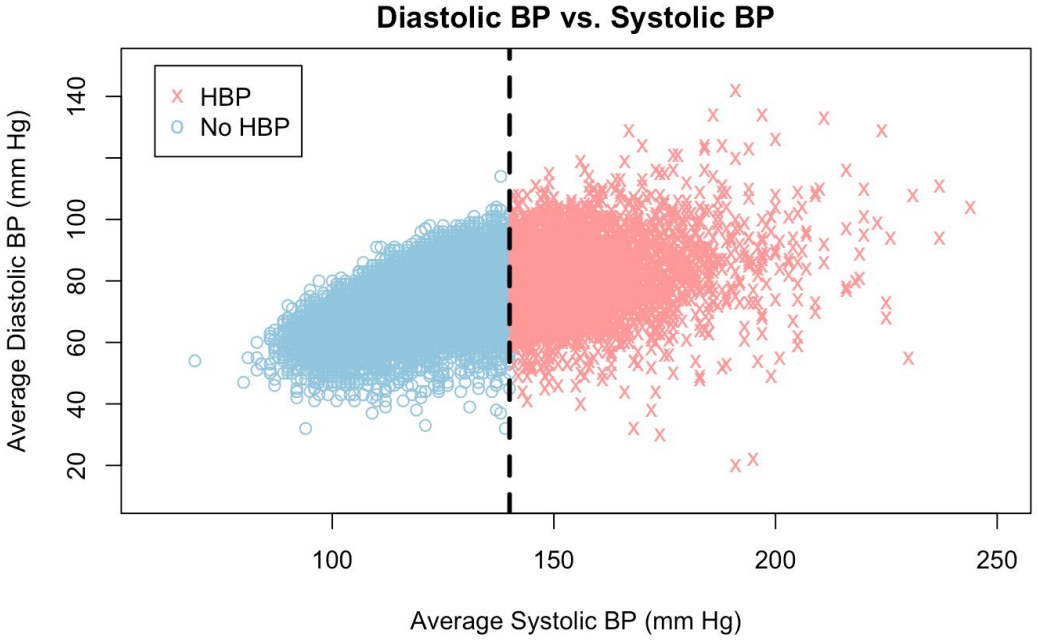

Figure 1.1 Diastolic BP vs. Systolic BP

larger than 140 mm Hg is labeled as high blood pressure. The subject with higher Average Systolic BP tends to have higher Average Diastolic BP, but there is no clear linearity between these two quantities from the plot. The correlation coefficient between these two quantities is around 0.5, which supports our observation.

After deleting features and NA inputs in the previous data cleaning, we have 15643 subjects with 7 variables: age, sex, race, body weight, standing height, smoking, serum cholesterol which may contributing to high blood pressure.To explore the data, we plot the distributions of the response (high blood pressure) and seven independent variables in Fig. 2.1 and Fig 2.2. Skewed distributions of response and variables may influence the generalization ability of our analysis to some extent.

In the top left subplot of Fig. 2.1, we could see that in our dataset, the number of subjects with high blood pressure is quite different from that without high blood pressure. The former is almost four times as many as the latter. The rest three subplots describe the distributions of the three categorical factors in our data: Sex, Race and Smoke. The numbers of females and males are about the same. The number of White people is above twice of black people and there is a small number of people who are neither white nor black. About half of the subjects never smoked while the rest have equal possibility to be still smoking and having quit smoking.

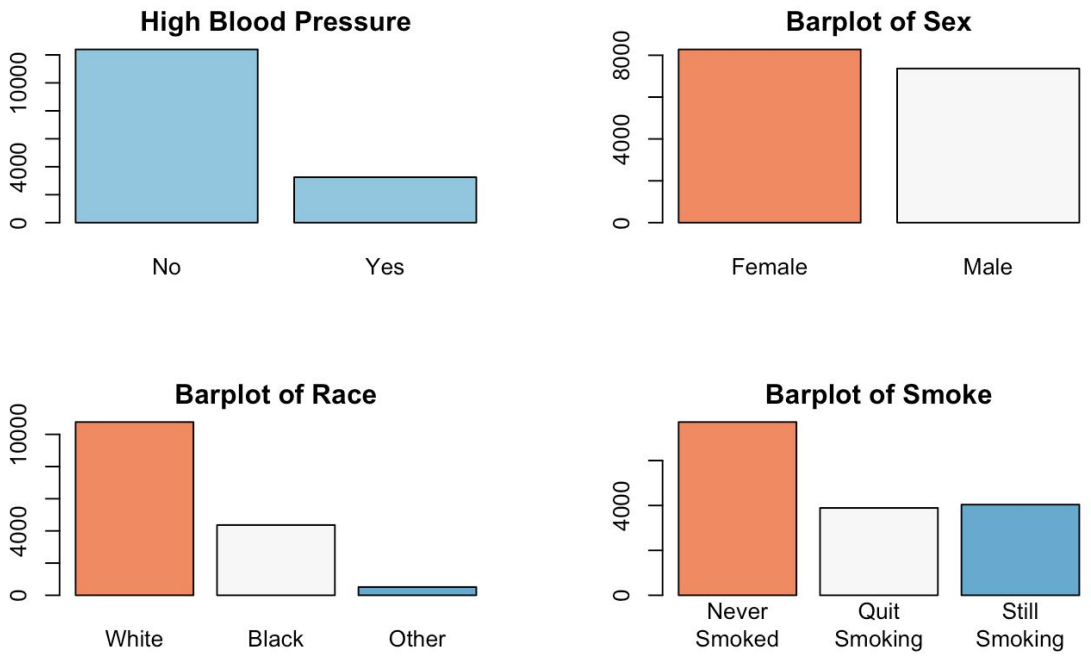

Figure 2.1 Distributions of Variables

In Figure 2.2, we examined the distributions of the rest four variables, which are all numerical. The average age of respondents is around 48 years old while the range is 20 to 90 years old. The average weight is around 166 lbs with a few outliers that are more than 250 lbs. The height with a mean of 66 inches is distributed symmetrically. The mean of TCP is about 206 mg/ml with a few outliers that are more than 350 mg/ml. From the distribution of these four numerical variables, we could see that their distributions are quite symmetric.

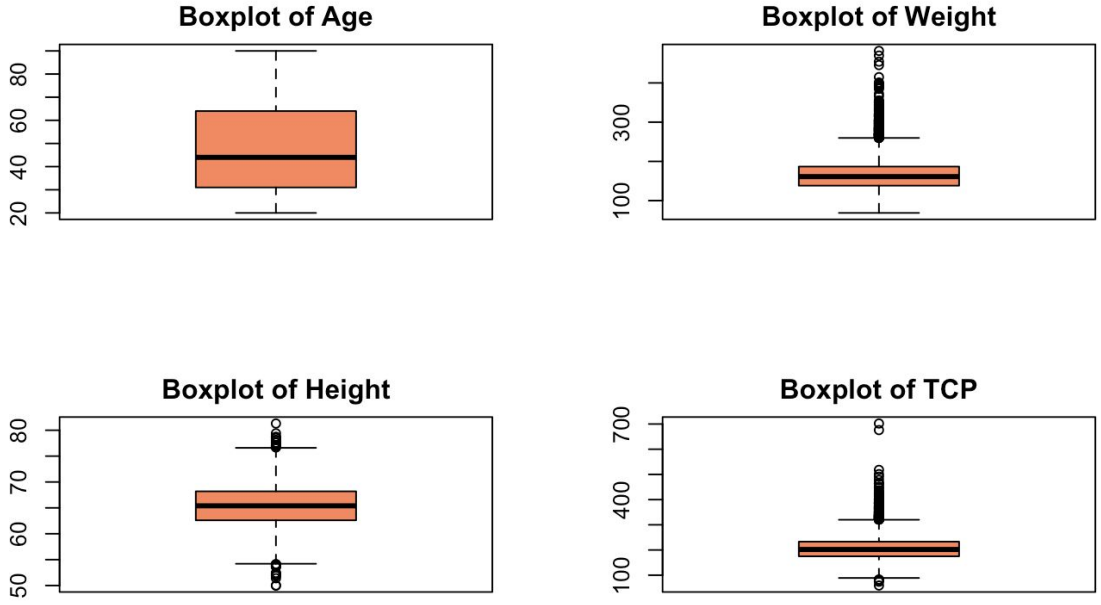

Figure 2.2 Distributions of Variables

3. MODEL SELECTION

In this section, we tried to model whether an individual has high blood pressure with the variables discussed above. Since the response is a categorical variable with two classes, we used logistic regression method. We started with models without interaction and then tested the significance of adding interaction terms between main effects. We provided our fitted model with estimated parameters in the end of this section.

3.1. Models with main effects

We firstly fit a logistic regression model (Model 1) with only main effects. Containing all the seven variables, Model 1 could be expressed as following:

Model 1: HBP ~ Age + Sex + Race + Weight + Height + Smoke + Chole.

in which HBP stands for high blood pressure, Weight for Body Weight, Height for Standing Height and Chole for Serum Cholesterol.

In Model 1, we used 1 dummy variable for Sex: sex1 denotes the effect of being male with female as reference. We used 2 dummy variables for Race: with being white as reference, race2 denotes the effect of being black while race3 denotes that of otherwise. We used 2 dummy variables for Smoke: with never smoked as reference, smoke 2 denotes the effect of having smoked but not smoking now and smoke3 denotes that of having smoked and still smoking.

Table 3.1 Siginificance of variables in Logistic Regression Model 1

Variable	age	sex1	race2	race3	weight	height	smoke2	smoke3	chole
Significant	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes

We put the detailed fitted results of logistic regression Model 1 in Table A.2. Here we show whether each variable is significant by comparing its p-value with 0.05 significance level in Table 3.1. In later analysis, we all used this 0.05 significance level. If p-value of one variable is smaller than 0.05, we conclude it as significant. Otherwise, it is not significant and we could delete it from the model. Through the p-values in Table 3.1, we could see that the following main effects are significant: age, sex1, race2, weigth, height, smoke3 and chole. However, some dummy variables for race and smoke have p-values greater than 0.05, so they might not be significant at level of significance 0.05, which triggered us to examine the model.

In order to test whether race and smoke can be deleted from the model, we performed analysis of deviance by fitting a logistic regression model without Race and Smoking (Model 2) and comparing with Model 1.

Model 2: HBP ~ Age + Sex + Weight + Height + Chole.

From the result from analysis of deviance between Model 2 and Model 1 in Table 3.1, we could see that the deviance between these two models is greater than the critical value

9.488. So the null hypothesis is rejected and we conclude that race and smoke can not be deleted from the model at the same time.

Table 3.2 Analysis of Deviance between Model 2 and Model 1

	Residual. Df	Residual. Dev	Df	Deviance
Model 2	15637	11774		
Model 1	15633	11679	4	94.974

Next, we test whether we could delete either smoke or race. Starting with the former, we fit a logistic regression model without Smoking (Model 3).

Model 3: HBP ~ Age + Sex + Race + Weight + Height + Chole.

The results are shown in Table 3.3. Since the deviance is greater than the critical value 5.991, we reject the null hypothesis and conclude that smoke cannot be deleted from the model.

Table 3.3 Analysis of Deviance between Model 3 and Model 1

	Residual.Df	Residual.Dev	Df	Deviance
Model 3	15635	11691		
Model 1	15633	11679	2	11.902

Then we test whether race is significant. Similar with the above anaylsis, we fit a logistic regression model without Race (Model 4).

Model 4: HBP ~ Age + Sex + Weight + Height + Smoke + Chole.

The results are shown in Table 3.4. Since the deviance is greater than the critical value 5.991, we conclude that race cannot be deleted from the model.

Table 3.4 Analysis of Deviance between Model 4 and Model 1

	Residual.Df	Residual.Dev	Df	Deviance
Model 4	15635	11755		
Model 1	15633	11679	2	75.492

In conclusion, for the models consisting of only main effects, we have to keep all seven variables since each of them is significant to high blood pressure. In other words, we keep Model 1 after this subsection discussion.

3.2. Model with interaction terms

Now, we consider the interaction effects based on Model 1. There would be a lot of terms if we blindly compose interaction terms between two terms, three terms or even all seven terms. The analysis would be messy and the results may be statistically significant but meaningless and hence fail to provide valuable insights into the problem.

Trying to keep our model as simple but strong as possible, we tested interaction terms added based on common sense. As mentioned above, there are three categorical main effects: sex, race and smoke. Would it be possible that interaction exists among these effects? According to the report from American Lung Organization, there is no significant difference between the adult smoking rate of the African-Americans and that of Non-Hispanic Whites[2]. So we ignore the interaction between race and smoke considering the fact that white people and black people made up most of our data set, as mentioned in the exploratory data analysis in Sec 2. In this subsection, we examined two possible interaction terms: the one between race and sex and the one between sex and smoke.

To begin we, we fit a logistic regression model containing interactions between Sex and Race, between Sex and Smoking (Model 5):

Model 5: HBP ~ Age + Sex + Race + Weight + Height + Smoke + Chole

+ Sex*Race + Sex*Smoke.

The detailed fitted results are shown in Table A.3. From Table 3.4, we could see that some p-values for interactions terms are greater than 0.05, so they might not be significant at level of significance 0.05. We also observed that when adding the interaction terms Sex*Race, Sex*Smoke, Sex as a main effect kind of lost its significance.

Table 3.5 Siginificance of variables in Logistic Regression Model 5

Variable	age	sex1	race2	race3	weight	height	smoke2
Significant	Yes	No	Yes	No	Yes	Yes	No
Variable	smoke3	chole	sex1*race2	sex1*race2	sex1*smoke2	sex1*smoke3	
Significant	No	Yes	No	No	No	Yes	

In order to test whether the interaction terms can be deleted from the model, we performed the analysis of deviance between Model 1 (containing only main effects) and Model 5 (containing the interaction terms Sex*Race, Sex*Smoke) and the test results are shown in Table 3.5.

Table 3.6 Analysis of Deviance between Model 1 and Model 5

	Residual.Df	Residual.Dev	Df	Deviance
Model 1	15633	11679		
Model 5	15629	11668	4	11.235

[2] https://www.lung.org/stop-smoking/smoking-facts/tobacco-use-racial-and-ethnic.html

Since the deviance is greater than the critical value 9.488, we conclude that interactions between Sex and Race, between Sex and Smoking cannot be deleted from the model at the same time.

Next, we test whether interaction between Sex and Race is significant in Model 5. We fit a logistic regression model without the interaction between Sex and Race (Model 6) and the test results are shown in Table 3.7. Since the deviance is smaller than the critical value 5.991, we conclude that the interaction between Sex and Race can be deleted from the model.

Model 6: HBP ~ Age + Sex + Race + Weight + Height + Smoke + Chole + Sex*Smoke.

Table 3.7 Analysis of Deviance between Model 6 and Model 5

	Residual.Df	Residual.Dev	Df	Deviance
Model 6	15631	11673		
Model 5	15629	11668	2	4.5845

Next, we test whether interaction between Sex and Smoking is significant in Model 6. We performed the analysis of deviance between Model 1 (containing only main effects) and Model 6 (containing the interaction Sex*Smoke) and the test results are shown in Table 3.8. Since the deviance is greater than the critical value 5.991, we conclude that the interaction between Sex and Smoking cannot be deleted from the model.

Table 3.8 Analysis of Deviance Table between Model 1 and Model 6

	Residual.Df	Residual.Dev	Df	Deviance
Model 1	15633	11679		
Model 6	15631	11673	2	6.6504

3.3. Final Model

From subsection 3.2, we found that the interaction between Sex and Smoking should be included in the model due to its significance. So we treat the interaction between Sex and Smoking as a factor with sex levels and fit a logistic regression model Model 7:

Model 7: HBP ~ Age + Race + Weight + Height + Chole + Sex*Smoke.

in which we used 5 dummy variables to denote the interaction term Sex*Smoke while kept the same form of other variables: f2 for females having smoked but quit, f3 for current smoking females, m1 for males never smoked, m2 for males having smoked but quit and m3 for current smoking females with females never smoked as reference. The detailed fitted results are shown in Table A.4.

Then, in order to test whether this factor is significant, we fit a logistic regression model without this factor Sex*Smoke (Model 8) and the test results are shown in Table 3.9. Since the deviance is greater than the critical value 11.071, we conclude that this factor can not be deleted from the model.

Table 3.9 Analysis of Deviance between Model 10 and Model 9

	Residual.Df	Residual.Dev	Df	Deviance
Model 10	15636	11698		
Model 9	15631	11673	5	25.762

From the above analysis, we finally get our model:
Let π denote the probability that a randomly selected individual has high blood pressure, then logit (π) = 0.078*Age + 0.481*Race2 + 0.145*Race3 + 0.008*Body Weight - 0.058*Standing Height + 0.005*Serum Cholesterol - 0.018*Female*Smoking2 + 0.026*Female*Smoking3 + 0.094*Male*Smoking1 + 0.086*Male*Smoking2 + 0.422*Male*Smoking3.

4 ANALYSIS OF THE RESULTS

1). If we increase Age by 1, the estimated odds of a randomly selected individual having high blood pressure will change by a multiplicative factor 1.081, holding Race, Body Weight, Standing Height, Serum Cholesterol, Sex and Smoking fixed. We are 95% confident that if we increase Age by 1, the odds of a randomly selected individual having high blood pressure will change by a multiplicative factor between 1.078 and 1.085, holding Race, Body Weight, Standing Height, Serum Cholesterol, Sex and Smoking fixed.

2). The estimated odds of a black person having high blood pressure is 1.618 times the estimated odds of a white person having high blood pressure, holding Age, Body Weight, Standing Height, Serum Cholesterol, Sex and Smoking fixed. We are 95% confident that the odds of a black person having high blood pressure is a number between 1.451 and 1.804 times the odds of a white person having high blood pressure, holding Age, Body Weight, Standing Height, Serum Cholesterol, Sex and Smoking fixed.

3). The estimated odds of a person with other race having high blood pressure is 1.156 times the estimated odds of a white person having high blood pressure, holding Age, Body Weight, Standing Height, Serum Cholesterol, Sex and Smoking fixed. We are 95% confident that the odds of a person with other race having high blood pressure is a number between 0.855 and 1.546 times the odds of a white person having high blood pressure, holding Age, Body Weight, Standing Height, Serum Cholesterol, Sex and Smoking fixed.

4). If we increase Body Weight by 1 unit, the estimated odds of a randomly selected individual having high blood pressure will change by a multiplicative factor 1.008, holding

Age, Race, Standing Height, Serum Cholesterol, Sex and Smoking fixed. We are 95% confident that if we increase Body Weight by 1 unit, the odds of a randomly selected individual having high blood pressure will change by a multiplicative factor between 1.007 and 1.010, holding Age, Race, Standing Height, Serum Cholesterol, Sex and Smoking fixed.

5). If we increase Standing Height by 1 unit, the estimated odds of a randomly selected individual having high blood pressure will change by a multiplicative factor 0.944, holding Age, Race, Body Weight, Serum Cholesterol, Sex and Smoking fixed. We are 95% confident that if we increase Standing Height by 1 unit, the odds of a randomly selected individual having high blood pressure will change by a multiplicative factor between 0.927 and 0.961, holding Age, Race, Body Weight, Serum Cholesterol, Sex and Smoking fixed.

6). If we increase Serum Cholesterol by 1 unit, the estimated odds of a randomly selected individual having high blood pressure will change by a multiplicative factor 1.005, holding Age, Race, Body Weight, Standing Height, Sex and Smoking fixed. We are 95% confident that if we increase Serum Cholesterol by 1 unit, the odds of a randomly selected individual having high blood pressure will change by a multiplicative factor between 1.004 and 1.006, holding Age, Race, Body Weight, Standing Height, Sex and Smoking fixed.

7). The estimated odds of a female who has smoked more than 100 cigarettes in life but doesn't smoke now having high blood pressure is 0.983 times the estimated odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed. We are 95% confident that the odds of a female who has smoked more than 100 cigarettes in life but doesn't smoke now having high blood pressure is a number between 0.837 and 1.152 times the odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed.

8).The estimated odds of a female who has smoked more than 100 cigarettes in life and still smoke now having high blood pressure is 1.026 times the estimated odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed. We are 95% confident that the odds of a female who has smoked more than 100 cigarettes in life and still smoke now having high blood pressure is a number between 0.859 and 1.223 times the odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed.

9).The estimated odds of a male who doesn't smoke more than 100 cigarettes in life having high blood pressure is 1.099 times the estimated odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed. We are 95% confident that the odds of a male who doesn't smoke more than 100 cigarettes in life having high blood pressure is a number between 0.924 and 1.307 times the odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed.

10).The estimated odds of a male who has smoked more than 100 cigarettes in life but doesn't smoke now having high blood pressure is 1.090 times the estimated odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed. We are 95% confident that the odds of a male who has smoked more than 100 cigarettes in life but doesn't smoke now having high blood pressure is a number between 0.928 and 1.280 times the odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed.

11).The estimated odds of a male who has smoked more than 100 cigarettes in life and still smoke now having high blood pressure is 1.525 times the estimated odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed. We are 95% confident that the odds of a male who has smoked more than 100 cigarettes in life and still smoke now having high blood pressure is a number between 1.275 and 1.824 times the odds of a female who doesn't smoke more than 100 cigarettes in life having high blood pressure, holding Age, Race, Body Weight, Standing Height, Serum Cholesterol fixed.

5 CONCLUSION

In this analysis, we studied the effects on high blood pressure based on information of a dataset from NHANES that consists 16 significant variables in total. We used logistic regression as a tool to look for and analyze the factors that may affect the probability of an individual having high blood pressure.

Our analysis showed some results as following: Elderly people have a higher risk than younger people to have high blood pressure. Black people are more likely to get high blood pressure than white people and people of other races. Overweight people have a high probability than others to have high blood pressure. Shorter people are more likely to get high blood pressure than higher people. People with higher chole have a higher risk to get high blood pressure. Currently smoking males are endangered by the risk of high blood pressure than others.

The results of our analysis suggest that excluding the uncontrollable risk factors, it is better for people to adopt healthy lifestyle to minimize the risk of getting high blood pressure, like performing weight control, quit smoking and cholesterol control.

Even though our results are consistent with people's common sense, since the classes in response have distinct difference in size: the proportion of records of HBP(have high blood pressure) is much smaller than proportion of non-HBP(not having high blood pressure), our model might predict towards the more common class(lower probability of having high blood pressure in our case). To get a more generalized predictive model, we may collect more and diverse data.

APPENDIX

Table A.1 Description of variables in the NHANES III Data Set

Variable	Description	Codes/Values
1	Respondent Identification Number	
2	Pseudo-PSU	1, 2
3	Pseudo-stratum	01 - 49
4	Statistical weight	225.93 - 139744.9
5	Age	(in years)
6	Sex	0 = Female, 1 = Male
7	Race	1 = White, 2 = Black, 3 = Other
8	Body Weight	(in pounds)
9	Standing Height	(inches)
10	Average Systolic BP	(mm Hg)
11	Average Diastolic BP	(mm Hg)
12	Has respondent smoked 100 cigarettes in life	1 = Yes, 2 = No
13	Does respondent smoke cigarettes now?	1 = Yes, 2 = No
14	Smoking	1 = if var.12=2 2 = if var.12=1&var.13=2 3 = if var.12=1&var.13=1
15	Serum Cholesterol	mg/100ml
16	High Blood Pressure	0 if var.10 ≤ 140 1 if var.10 > 140

Table A.2 Logistic Regression Model 1

Variable	Estimate	Std.Error	z value	Pr(>z)
age	0.078	0.002	48.102	0.000
sex1	0.174	0.067	2.592	0.010
race2	0.484	0.056	8.712	0.000

race3	0.147	0.151	0.971	0.332
weight	0.008	0.001	12.037	0.000
height	-0.059	0.009	-6.360	0.000
smoke2	-0.038	0.057	-0.674	0.500
smoke3	0.184	0.063	2.896	0.004
chole	0.005	0.001	8.864	0.000

Table A.3 Logistic Regression Model 5

Variable	Estimate	Std.Error	z value	Pr(>z)
age	0.078	0.002	47.964	0.000
sex1	0.039	0.094	0.418	0.676
race2	0.410	0.076	5.391	0.000
race3	-0.146	0.232	-0.632	0.528
weight	0.008	0.001	12.134	0.000
height	-0.058	0.009	-6.300	0.000
smoke2	-0.024	0.082	-0.293	0.770
smoke3	0.034	0.090	0.374	0.708
chole	0.005	0.001	8.732	0.000
sex1*race2	0.148	0.107	1.373	0.170
sex1*race3	0.537	0.305	1.763	0.078
sex1*smoke2	0.025	0.115	0.214	0.831
sex1*smoke3	0.284	0.126	2.244	0.025

Table A.4 Logistic Regression Model 7

Variable	Estimate	Std.Error	z value	Pr(>z)
age	0.078	0.002	47.989	0.000
race2	0.481	0.056	8.665	0.000

race3	0.145	0.151	0.962	0.336
weight	0.008	0.001	12.062	0.000
height	-0.058	0.009	-6.290	0.000
chole	0.005	0.001	8.789	0.000
f2	-0.018	0.082	-0.216	0.829
f3	0.026	0.090	0.288	0.773
m1	0.094	0.088	1.067	0.286
m2	0.086	0.082	1.048	0.294
m3	0.422	0.091	4.623	0.000