Factors Associated with Systolic Blood Pressure – A Survey Study with NHANES

Author: Mengyu Lei

Date: 23 Dec 2020

Contents

Abstract

Systolic blood pressure has been identified to be related to smoking status. However, the risk factors associated with systolic blood pressure are not fully identified and evaluated. The primary objective of this project is to identify a set of risk factors for systolic blood pressure given the exposure of smoking status. By utilizing the NHANES data and stratified random sampling procedure, gender, age, diastolic blood pressure, and marital status have been found to be positively correlated with systolic blood pressure. While the household income is negatively associated with systolic blood pressure. The findings could potentially help clinicians to identify high-risk subjects with higher systolic blood pressure and provide personalized treatment suggestion accordingly. The analysis code can be found in https://github.com/Mengyu1201/STA_304_Project.

Key words

NHANES, Survey Data, Systolic Blood Pressure, Smoking, Stratified Random Sampling, Risk Factor

Introduction

In 1971, the National Center for Health Statistics (NCHS) conducted the first National Health and Nutrition Examination Survey (NHANES) to evaluate the health and nutritional status of the United States citizens (adults and children) [1]. Interviews, including demographic, socioeconomic, dietary, and health-related questions, and medical, dental, and physiological examination, and relevant laboratory tests are listed in the NHANES to track the changes of health and nutritional status over time. The NCHS organizes the NHANES annually from 1999 and released the first full report in the year 2001 [2].

By now, the survey data has been widely analyzed in studies that investigated the association between health promotion and disease prevention, and researches that identify the risk factors for diseases. Recently, one study revealed that the leading death in the United States is cardiovascular disease both for men and women [3]. According to its estimation, for every 36 seconds, one person in the United States will die because of cardiovascular disease [3]. Moreover, the total expenses of heart-diseases-related costs for the United Stated account for $219 billion each year from 2014 to 2015 [4].

Two of the most important biomarkers in monitoring heart health, systolic blood pressure, and diastolic blood pressure, have been identified to be related to smoking status [5]. However, the risk factors associated with systolic blood pressure are not fully identified and evaluated. Given all this information, the purpose of this project is to identify a set of risk factors for systolic blood pressure given the exposure of smoking status. The NHANES data consists of demographic, socioeconomic, and health-related factors and an adjusted weighting variable will be utilized to fulfill the objective.

The data description, analysis strategies, and model used in this project will be provided in the **Methodology** section. Results are given in the **Results** section and the inference of this analysis along with conclusions are presented in the **Discussion** section.

Methodology

The raw data is collected by the US National Center for Health Statistics (NCHS) with 76 variables. In this project, only subjects who participated in the NHANES from 2011 to

2012, and age above 17 will be included in the analysis. Moreover, the total number of variables has been reduced to 12, with one ID variable, 7 demographic factors (*Gender, Age, Education, MaritalStatus, HHIncome, HHIncomeMid,* and *HomeRooms*), two physical factors (systolic blood pressure and diastolic blood pressure), one lifestyle variable (*Smoke100*) and one stratum variable (*SDMVSTRA*).

In this project, the outcome of interest is the averaged systolic blood pressure (*BPSysAve*) and the exposure variable is *Smoke100.* The demographic table that summarizes all variables grouped by the exposure variable along with their definition is shown in **Table 1**.

Table 1 Demographic table group by exposure variable (*smoke100)*

	Smoke100 (smoked at least 100 cigarettes in their entire life)			
	No (N=2611)	Yes (N=1970)	Total (N=4581)	P value
Gender	Gender (sex) of study participant coded as male or female			< 0.001
Female	1543 (59.1%)	767 (38.9%)	2310 (50.4%)	
Male	1068 (40.9%)	1203 (61.1%)	2271 (49.6%)	
Age	Age in years at screening of study participant			< 0.001
Mean (SD)	46.646 (17.761)	51.372 (17.176)	48.678 (17.665)	
Education	Educational level of study participant Reported for participants			< 0.001
8th Grade	216 (8.3%)	191 (9.7%)	407 (8.9%)	
9 - 11th Grade	283 (10.8%)	346 (17.6%)	629 (13.7%)	
High School	479 (18.3%)	471 (23.9%)	950 (20.7%)	
Some College	783 (30.0%)	615 (31.2%)	1398 (30.5%)	
College Grad	850 (32.6%)	347 (17.6%)	1197 (26.1%)	
MaritalStatus	Marital status of study participant			< 0.001
Divorced	201 (7.7%)	280 (14.2%)	481 (10.5%)	
LivePartner	163 (6.2%)	187 (9.5%)	350 (7.6%)	
Married	1326 (50.8%)	929 (47.2%)	2255 (49.2%)	
NeverMarried	624 (23.9%)	341 (17.3%)	965 (21.1%)	

Separated	101 (3.9%)	67 (3.4%)	168 (3.7%)	
Widowed	196 (7.5%)	166 (8.4%)	362 (7.9%)	
HHIncome	Total annual gross income for the household in US dollars			< 0.001
0-4999	60 (2.3%)	61 (3.1%)	121 (2.6%)	
5000-9999	122 (4.7%)	125 (6.3%)	247 (5.4%)	
10000-14999	205 (7.9%)	200 (10.2%)	405 (8.8%)	
15000-19999	206 (7.9%)	164 (8.3%)	370 (8.1%)	
20000-24999	180 (6.9%)	197 (10.0%)	377 (8.2%)	
25000-34999	298 (11.4%)	249 (12.6%)	547 (11.9%)	
35000-44999	254 (9.7%)	222 (11.3%)	476 (10.4%)	
45000-54999	188 (7.2%)	131 (6.6%)	319 (7.0%)	
55000-64999	131 (5.0%)	107 (5.4%)	238 (5.2%)	
65000-74999	151 (5.8%)	84 (4.3%)	235 (5.1%)	
75000-99999	277 (10.6%)	145 (7.4%)	422 (9.2%)	
more 99999	539 (20.6%)	285 (14.5%)	824 (18.0%)	
HHIncomeMid	Numerical version of HHIncome derived from the middle income in each category			< 0.001
Mean (SD)	52221.371 (33562.309)	44304.569 (31754.776)	48816.852 (33027.132)	
HomeRooms	How many rooms are in home of study participant			0.025
Mean (SD)	5.755 (2.288)	5.608 (2.081)	5.692 (2.203)	
BPSysAve	Combined systolic blood pressure reading			< 0.001
Mean (SD)	121.529 (18.103)	125.426 (18.569)	123.205 (18.404)	
BPDiaAve	Combined diastolic blood pressure reading			0.163
Mean (SD)	70.403 (12.418)	70.950 (14.011)	70.638 (13.128)	

In the first step, the univariate and bivariate analyses were carried out to check the characteristics of the outcome systolic blood pressure (*BPSysAve*) and the exposure variable smoking (*Smoke100*). To visualize the two variables and their relationships, the density bar plot, and boxplot was used. Since the systolic blood pressure is a continuous variable and the exposure variable smoking is a binary categorical variable, a two-sample

T-test and linear regression model could be used to compare the mean systolic blood pressure difference between the two smoking groups.

An appropriate linear regression model was constructed to better predict the systolic blood pressure given the exposure variable of smoking by unitizing the stepwise regression method. The full model for stepwise regression consists of all variables except for the ID variable. In the meanwhile, a model that serves *BPSysAve* as response and *Smoke100* as the predictor was chosen as the lower model. The final model is listed as follows.

$$BPSysAve = 63.16 + \begin{cases} 0 & Female \\ 2.50 & Male \end{cases} + 0.33 \times Smoke100 + 0.48 \times Age + 0.50 \times BPDiaAve + \begin{cases} 0 & Divorced \\ 2.08 & Live\ Partner \\ 0.81 & Married \\ 4.27 & Never\ Married \\ 1.83 & Separated \\ 7.21 & Widowed \end{cases} + \begin{cases} 0 & 8th\ Grade \\ 0.04 & 9-11th\ Grade \\ -0.24 & High\ School \\ -0.94 & Some\ College \\ -3.91 & College\ Grad \end{cases} - 1.79 \times 10^{-5} \times HHIncomMid$$

The above model did not account for the stratum variable *SDMVSTRA,* which potentially give more error in the estimation and lead to undesired results. Therefore, stratified random sampling was adopted to handle such issues. It is assumed that the margin of error equals 4 and the level of significance $\alpha = 0.01$. The stratum specific variance is estimated by $\sigma_j = Range\ of\ BPSysAve) / 4$. The previous model derived was refitted with a new subsample derived from each stratum. The comparison of the results for the two models is provided in the **Results** section.

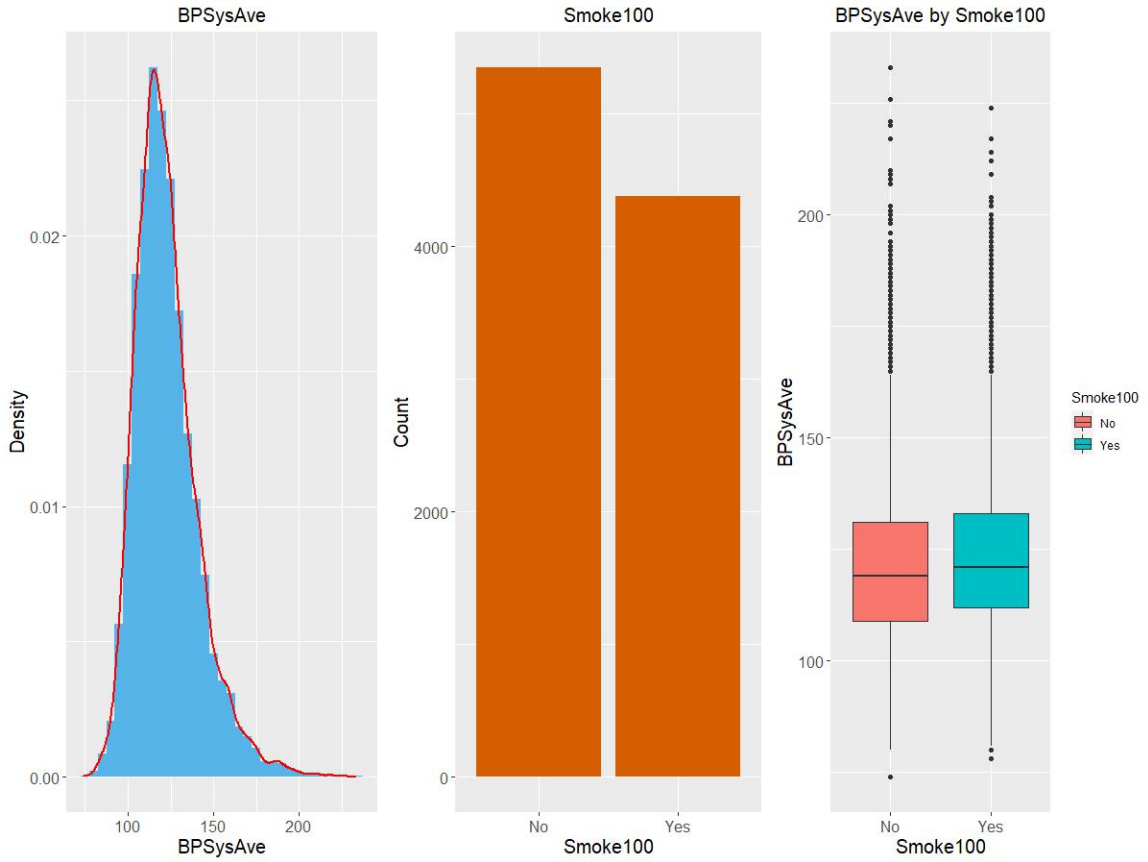
The same strategy described before for the stepwise regression method was applied to the new subsample. The final stratified model is provided as follows.

$$BPSysAve = 52.00 + \begin{cases} 0 & Female \\ 3.68 & Male \end{cases} - 0.08 \times Smoke100 + 0.42 \times Age + 0.69 \times BPDiaAve + \begin{cases} 0 & Divorced \\ 2.99 & Live\ Partner \\ 0.34 & Married \\ 4.72 & Never\ Married \\ 0.35 & Separated \\ 10.10 & Widowed \end{cases} - 7.82 \times 10^{-5} \times HHIncomMid$$

Results

There is a total of 4581 participants involved in the analysis. The systolic blood pressure for all participants has a range from 74 to 233, with mean and median equals to 123.2 and 120, respectively. 2611 (57%) subjects smoked less than 100 cigarettes at the time of the survey. **Figure 1** provides the density plot for the outcome systolic blood pressure (*BPSysAve*), the barplot of exposure variable smoking, and the boxplot of *BPSysAve* by *Smoke100*.

Figure 1 Visualization of Outcome BPSysAve and Exposure Smoke100

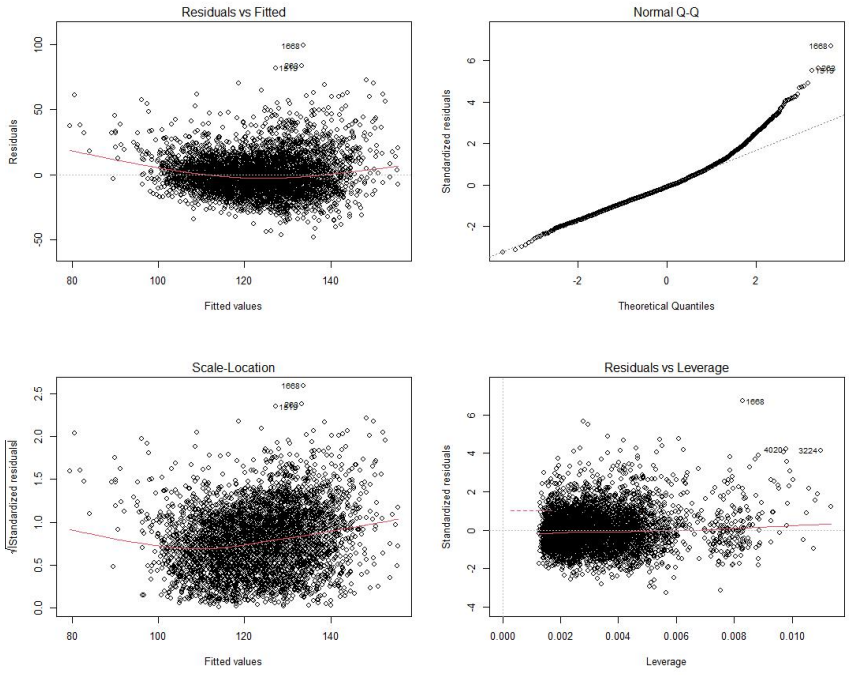


It is obviously observed that the systolic blood pressure was slightly right-skewed from the density plot of *BPSysAve*. The boxplot of *BPSysAve* by *Smoke100* revealed that people who smoke more than 100 cigarettes tended to have higher systolic blood pressure. The mean (median) value of *BPSysAve* for the two groups is 122 (119) and 125(123), respectively. An F-test is employed to evaluate the variances of systolic blood pressure across *Smoke100* groups. The corresponding F-statistic and P-value is 0.9504 and 0.2269,

respectively. This suggested that the variances of systolic blood pressure are the same between *Smoke100* groups. According to this result, the two-sample t-test with equal variance is used to check the group mean systolic blood pressure difference. The P-value is less than 0.0001, which suggested that the mean systolic blood pressure between the *Smoke100* groups is different. The results generated from the linear regression model are identical to the result of the t-test. The 95% confidence interval for the mean difference between the two groups ($BPSysAve_{Smoke100Yes} - BPSysAve_{Smoke100No}$) is (2.83,4.97)

The model obtained with the completed data is summarized in **Table 2**. In general, Gender (Male), Age, and diastolic blood pressure are statistically significantly positively associated with systolic blood pressure. Household income and education level are negatively associated with systolic blood pressure. However, the exposure variable smoking is not statistically significantly associated with the outcome, with a P-value of 0.4763. The adjusted R-squared for the model is 0.3473. **Figure 2** is the visualization of the model diagnostics for the assumptions of a linear regression model. Basically, the assumption of homoscedasticity and normality is slightly violated, but it is still predictive in terms of R-Squared.

Figure 2 Model Diagnostics



SDMVSTRA is the stratum variable in this dataset with 14 different strata. The cost of sampling for each stratum is C = 52, 50, 46, 53, 48, 48, 47, 57, 53, 47, 54, 40, 43, and 44. It is assumed that the margin of error equals 4 and the level of significance α = 0.01. The stratum specific variance is estimated by $\sigma_j = Range\ of\ BPSysAve)/4$. The optimal required sample size for stratified random sampling is 302 (It will become 310 if set the stratum-specific sample sizes to the smallest integer). The same model was applied to a subset of the original dataset by using stratified sampling, and the results are listed in **Table 2**. Gender (Male), Age, and diastolic blood pressure are still statistically significantly positively associated with systolic blood pressure. However, education is no longer statistically significant with all P-value greater than 0.05. The adjusted R-squared for the stratified model is 0.4184. Generally, the stratified model is preferred as the stratified random sampling could better reflect the entire population that was investigated and has a higher adjusted R-squared.

	Final Model (N=4581)		**Stratified Model (N=310)**	
	Estimate (SE)	P-Value	Estimate (SE)	P-Value
(Intercept)	63.16 (1.86)	<0.001	55.13 (6.58)	<0.001
Smoke100				
No	-	-	-	-
Yes	0.33 (0.47)	0.476	-0.20 (1.65)	0.903
Gender				
Female	-	-	-	-
Male	2.50 (0.46)		3.82 (1.59)	0.017
Education				
8th Grade	-	-	-	-
9 - 11th Grade	0.04 (0.96)	0.039	-3.32 (3.51)	0.345
High School	-0.24 (0.90)	0.791	-2.10 (3.22)	0.514
Some College	-0.94 (0.88)	0.286	-2.33 (3.12)	0.456
College Grad	-3.91 (0.94)	<0.001	-5.44 (3.37)	0.108
MaritalStatus				
Divorced	-	-	-	-
LivePartner	2.08 (1.09)	0.055	2.68 (3.49)	0.443
Married	0.81 (0.77)	0.297	0.39 (2.50)	0.877
NeverMarried	4.27 (0.89)	<0.001	5.01 (2.80)	0.075
Separated	1.83 (1.35)	0.174	0.29 (3.92)	0.941
Widowed	7.21 (1.08)	<0.001	10.22 (4.33)	0.019
Age	0.48 (0.02)	<0.001	0.41 (0.06)	<0.001
BPDiaAve	0.50 (0.02)	<0.001	0.68 (0.06)	<0.001
HHIncomeMid	-1.80e-5 (7.8e-6)	0.023	-5.68e-5 (2.79e-5)	0.043
Adjusted R-Squared	0.3473	<0.001	0.4184	<0.001

Table 2 Final Model and Stratified Model Summary

Since the education level is not statistically in the stratified model, remove this variable could potentially have a better prediction of systolic blood pressure. The same strategy described for the stepwise regression method was then applied to the stratified subsample. The results final stratified model is provided in **Table 3**. The variable eduction has been dropped out in the final stratified model. Gender (Male), Age, and diastolic blood pressure are still statistically significantly positively associated with systolic blood pressure. Household income is negatively associated with systolic blood pressure. The marital status trend to increase systolic blood pressure to some extent. And the exposure variable smoking is not statistically significantly associated with the outcome in this model.

	Final Stratified Model (N=310)		
	Estimate	Standard Error	P-Value
(Intercept)	52.00	5.87	<0.001
Smoke100			
No	-	-	-
Yes	-0.08	1.64	0.96
Gender			
Female	-	-	-
Male	3.68	1.59	0.021
MaritalStatus			
Divorced	-	-	-
LivePartner	2.99	3.47	0.390
Married	0.34	2.49	0.893
NeverMarried	4.72	2.79	0.091
Separated	0.35	3.92	0.929
Widowed	10.10	4.28	0.019
Age	0.42	0.05	<0.001
BPDiaAve	0.69	0.06	<0.001
HHIncomeMid	-7.82e-5	2.47e-5	0.002
Adjusted R-Squared	0.4190		<0.001

Table 3 Final Stratified Model Summary

Discussion

Summary

In this project, a set of demographic, socioeconomic, and health-related factors associated with systolic blood pressure has been identified by leveraging the NHANES data with an adjusted weighting variable and stratified random sampling procedure. Gender, age, diastolic blood pressure, and marital status have been found to be positively correlated

with systolic blood pressure. While the household income is negatively associated with systolic blood pressure. The findings could potentially help clinicians to identify high-risk subjects with higher systolic blood pressure and provide personalized treatment suggestion accordingly.

Conclusions

The final stratified model obtained with the stratified random sampling is preferred in the analysis. The results are generally agreed upon in reality. Basically, when holding other variables as constant, one unit increase in age will lead to a 0.42 unit increase in systolic blood pressure, and one unit increase in diastolic blood pressure will result in a 0.69 unit increase in systolic blood pressure. Even it is widely known that smoking is a risk factor for cardiovascular disease, but it is not statistically associated with systolic blood pressure from the findings. The marital status was identified to be positively significantly associated with systolic blood pressure, but the underlying mechanism is still unknown. The potential reason could be that marital status is related to age. However, this may need further investigation.

Household income is negatively associated with systolic blood pressure, one potential reason is that people with higher household income may care more about their physical health and have better living conditional and better life habits as well. These factors may be associated with lower systolic blood pressure. More investigation of the hidden factors is required.

Weakness & Future work

Given the analysis performed and findings in this project, several weaknesses should be acknowledged as well.

- The optimal stratified sample size for each stratum is highly correlated to the prior cost information. Inappropriate cost prior may lead to different optimal stratified sample size. The impact of this is unknown.
- The model constructed from stratified random sampling could be totally different if we rerun it.

- The linear regression obtained with the completed data set did not account for the stratum information.

Considering these limitations, future work could be done to address the issues.

- Perform a sensitivity analysis for the impact of cost on the optimal stratified sample size calculation.
- Construct a linear regression model by unitizing the stepwise regression method. The upper model for stepwise regression consists of all variables except for the ID variable and the lower model consists of the exposure variable smoking and the stratum variable *SDMVSTRA* (categorical variable). This could potentially address issues related to stratified populations.
- A reduced model with smoking, age, and diastolic blood pressure may be proposed to predict systolic blood pressure by adjusting the stratum variable. This model may forecast systolic blood pressure in a more simple manner.

Reference

[1] "About the National Health and Nutrition Examination Survey". CDC/National Center for Health Statistics. U.S. Department of Health & Human Services. Retrieved 18 March 2020.

[2] Ozonoff, David (2014). "Biomonitoring". In Rogers, Kara (ed.). Encyclopædia Britannica Online. Encyclopædia Britannica, Inc.

[3] Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.

[4] Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.

[5] Linneberg A, Jacobsen RK, Skaaby T, et al. Effect of Smoking on Blood Pressure and Resting Heart Rate: A Mendelian Randomization Meta-Analysis in the CARTA Consortium. Circ Cardiovasc Genet. 2015 8(6):832-841.