Hypertension high risky groups estimation and proprtion of hypertension people prediction in the US

Mengyuan Wang (1005239341)

December 20, 2020

GitHub link: https://github.com/Mengyuandaisy/304-Final

Abstract

High blood pressure (hypertension) is a serious medical condition that will cause many complications and also impacts a large proportion of people in the world. In this paper, it's concluded that age, gender, weight and poverty level are four important causes for high blood pressure via multilevel regression modelling. Moreover, it shows that there are about 33.51% adults are potentially having high blood pressure in the USA.

Keywords: High blood pressure, hypertension, Post-stratification, Logistics model, USA, Multilevel Regression Model

Introduction

High blood pressure (hypertension) is a serious medical condition that is even more dangerous than heart disease. It can quickly damage organs and bodies even before symptoms appear, which cause difficulties for patients to realize. (Mayo Clinic Staff. 2019) Hence, it is very important to figure out what reasons will potentially cause high blood pressure so that people will be able to realize to have a test if it's necessary. Moreover, many complications of hypertension will cause damage to the heart, the brain, the kidneys, the eyes and many other organs all over the body. (Mayo linic Staff. 2019) According to WHO report, 'In 2015, 1 in 4 men and 1 in 5 women had hypertension. (WHO, 2019)' Also, 'Hypertension is a major cause of premature death worldwide. (WHO, 2019)' Hypertension is very dangerous, which impacts a large proportion of people in the world. Overall, it is crucial to predict the probability of having high blood pressure nationally so that the government can have better control.

In this paper, the prediction is specifically focus on population and surveys of USA, since data sources are feasible and reliable. However, according to data from American Heart Association's, the newest update was in 2013 and used data from 2007 to 2010. Hence, it is meaningful to make new prediction with newest data for update. In order to achieve this goal, the paper will use meaningful variables to build a model so that we can estimate the probability of having high blood pressure for American individuals. In that case, it will be easy to see which groups of people are potentially with high risk, and finally predict the proportion of people with high blood pressure in the US.

According to Lancet report, there are many reasons which will potentially cause the high blood pressure including age, gender, stress weight and other aspects. According to WHO report, income is also an important cause, because almost "two-thirds of patients living in low and mid- income countries". (WHO, 2019) These aspects will be considered in the process of data resources searching.

Overall, the NHANES and IPUMS will be used for later analysis. In the Methodology section (Section 2) will talk about the model which is used to show the relationship between blood pressure and other aspects. Moreover, the results section (Section 3) will show a detailed results about each aspect, and predict the proportion of people with high blood pressure. Conclusion section (Section 4) will be about overall analysis and weakness discussion.

Methodology

Data

The survey data used is the National Health and Nutrition Examinations from Centers for Disease Control (CDC) and Prevention's National Center for Health Statistics (NCHS). This survey is designed to interview many aspects for individual health includes demographic and other health-related questions. It has long history since 1970s. The target population for this survey is the total noninstitutionalized civilian U.S. population residing in the 50 states and District of Columbia. The sample design is a clustered design. The data we used in this paper is the data from 2011 to 2012 specifically, since it has the important outcome (Combined systolic blood pressure reading). This data its self is easily accessible in R package ('NHANES'), and we select 16 highly relative variables according to WHO guidelines from total 32 variables. All variables will be listed in Table 1. This data is picked because US is a multi-culture country with a large population. Generally, samples are sufficient with various backgrounds. This data is reliable and representative even for worldwide using. This data will be used to build a model in later steps.

Table1: Survey variables description list

Variable	Description
ID	Participant identifier.
Gender	Gender (sex) of study participant
Age	Age in years
Race3	Reported race of study participant, including non-Hispanic Asian category
Education	Educational level of study participant
MaritalStatus	Marital status of study participant.
HHIncome	Total annual gross income for the household in US dollars.
Poverty	A ratio of family income to poverty guidelines.
Weight	Weight in kg
Height	Standing height in cm.
BMI	Body mass index (weight/height2 in kg/m2).
BPSysAve	Combined systolic blood pressure reading,
Depressed	Self-reported number of days where participant felt down, depressed or hopeless.
SleepHrsNight	Self-reported number of hours study participant usually gets at night on weekdays
SleepTrouble	Participant has told a doctor or other health professional that they had trouble sleeping.
PhysActive	Participant does moderate or vigorous-intensity sports, fitness or recreational activities.

The census data used is U.S. census data for Social Economic and Health research from IPUM's collection, and samples are cluster samples. All the samples are also stratified by key characteristics. They are samples of households or dwellings, and we picked individual information from these large units. It used post-stratification, which divided groups according to household unit. In order to match variables in the model, I selected variable which are highly matching to the survey data, like age, gender, weight, household income etc. This data was rearranged so that the format for each variable exactly matches the survey data. This data will be used to estimate the proportion of people with high blood pressure, since the census data can widely represent the whole country.

In order to build Multilevel regression model (MRP), I filtered data roughly before the modelling. For example, we just consider adults' high blood pressure issue here, thus samples below 21 years old will be deducted. Also, according to Health line report, we set 130 mm Hg as the boundary. In other words, for individuals with above 130 mm Hg combined systolic blood pressure reading will be considered as high blood pressure group. In this data, we set 1 as high blood pressure results, and 0 for otherwise. Please note, different institutions have different ideas about the standard of high blood pressure. In this paper, we used the newest guideline standard.

Model

In order to find the possibility for an individual to have high blood pressure. We will build a logistics regression Model to find the relationship between probability and other aspects. In this paper, we used R language to achieve this logistics model. As mentioned before, we set numerical number '1' to represent the person has high blood pressure, and '0' for other cases.

Model Check Firstly, we will use all 16 variables which we pre-selected before to build a logistics regression. However, according to model summary, we only found few variables are very significant. In this case, we used step forward method for variable selection according to their P-values (a value to test weather variables are significant). Overall, we obtained 4 significant variables which are Age, Gender, Poverty and BMI. All these variables with around 850 AIC value, which are much smaller than original model. However, we do not have height data in census data, so we used variable weight instead. (BMI is Body mass index, weight/height in kg/m2)

Model Specific Afterwards, we divided variables into with different levels, and built the multilevel regression model (MRP) which shows below. (The process of level division will be explained in Post-stratification section)

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{agegroup31to40} + \beta_2 x_{agegroup41to50} + \beta_3 x_{agegroup51to60} + \dots + \beta_{15} x_{Povertylevel5}$$

P in the left hand of the equation stands for the probability of a random person who potentially have high blood pressure. We broke numerical variables into several levels, thus for all 15 Xs (variables with notations) here are dummy variable. For example, the variable Xage_group31 to 40 represents that if a random individual from samples is from 31 to 40 years old, then the value of the term "X age_group31 to 40" would be 1. Otherwise, it would be zero. The same logic applies to all dummy variables. B0(0.29005) is the intercept here, it represents the log odds if hold all other variables be 0. In other words, if a random female who is between 21 to 30 years old, with 50 to 60 kg weight and whose household income is below the poverty in the US, then the odds for her to have high blood pressure is 33.6% B1 to B15 (detailed results will be listed in Table2) represent coefficients in our equation, which shows how much different aspects will impact to the probability of having high blood pressure. For instance, if a female with 50 to 60 kg weight is aged from 31 to 40, the odds for her to have high blood pressure will increase by 1.28 times. Same idea for all B values here.

Post-stratification As mentioned before, we selected 5 variables according to their p-values, and they would be built a logistics regression. However, in order to make census data fit in the model for prediction, we should divide these numerical variables into different level and build a multilevel regression model (MRP). Gender is categorical variable for two levels already (male and female). For age group division, we referred the age analysis by American Heart Association, and slip them into 6 groups. Wight and poverty were also divided accordingly. The poverty level is the ratio of household income dividing by poverty guideline (use the standard in 2019). 1 represents the lowest income category which means household income is below the poverty line. Then, we selected the same 4 variables from American census data and did the same level split. Overall, we obtained 360 cells in census data, and we used them to predicted the probability for each cell (group of people). Finally, we would summarize for an overall probability prediction.

Result

After variable selection, we used variable age groups, weight, poverty level, gender and built the multilevel regression model with 16 B values. Each B value shows different impact level towards the probability. All B value were listed in Table 2.

Table2: Beta values for each level

Beta	Levels	Value
B0	(Intercept)	-1.8977
B1	age_group31 to 40	0.253
B2	age_group41 to 50	0.8847
B3	age_group51 to 60	1.7816
B4	age_group61 to 70	2.3855
B5	age_groupabove 70	2.869
B6	Gendermale	0.477
B7	weight60 to 70	-0.1607
B8	weight70 to 80	-0.8709
B9	weight80 to 100	-0.4397
B10	weightabove 100	-0.1811
B11	weightLess than 50	-0.7721
B12	Poverty_level2	-0.243
B13	Poverty_level3	-0.1838
B14	Poverty_level4	-0.617
B15	Poverty_level5	-0.6428

After doing post-stratification analysis, we estimated the probability for each cell. In other words, we estimated the probability of having high blood pressure for each distinct group of people. In Table 3, we selected 5 groups with highest probability.

Table3: 5 highest probability groups

Age group	Poverty level	Weight	Genger	Probability
above 70	Poverty_level1	50 to 60	male	0.8097379
above 70	Poverty_level1	60 to 70	male	0.7837372
above 70	Poverty_level1	above 100	male	0.7802626
above 70	Poverty_level3	50 to 60	male	0.7797998
above 70	Poverty_level2	50 to 60	male	0.7694713

Overall, we summarized all estimates together, and obtained the result 0.3351. It means that we predicted that there are about 33.51% adults are potentially having high blood pressure in the USA.

Discussion

Overall, we used survey data from NHANES 2012 to build a multilevel regression model. After we set the high blood pressure standard which is over 130 mm Hg combined systolic blood pressure reading, we select 4 variables according to their P-values and AIC values. In other words, we found that age, gender, weight and poverty level are four important causes for high blood pressure. Then we divided each variable to several levels according to health report suggestions, and built multilevel regression model. After post-stratification analysis, we split the USA 2019 census data into 360 cells according to levels in prior survey data. We estimated the probability for each cell, and found 5 groups with highest probability. Finally, we summarized the overall probability according to estimation in each cell. It would be able to conclude that there are about 33.51% adults are potentially having high blood pressure in the USA in 2019.

Conclusions

According to Post-stratification above, causes of high blood pressure are related to age, gender, weight and poverty. If a male is over 70 whose household income is below the poverty line, the odds for him to have high blood pressure will increase by 4.22 times. Overall, according to the B value in Table 2, we can find that elder men will have higher probability to have high blood pressure. For people whose weight is from 50 to 60 kg and household income is below the poverty line (the meaning of poverty level 1), they also have higher probability to have high blood pressure. Hence, it is important for elder people specially for those who are below the poverty line to take body examinations regularly, since they have high risk to have high blood pressure. For the government, it is also important to spend more expense to care more about elder people and people below the poverty line since they are highly risky to have high blood pressure or even other complications.

According to the Table 3, all top 5 highest risk groups are aged 70 above male, and all the top 3 are below the poverty line. It is a strong evidence to declare that the risk of having high blood pressure will significantly increase as people get older. It is also meaningful to explore what reasons might cause male have higher risk than women. It might cause by physical difference between genders or other personal reasons like smoking. It is also important to notice that poverty will bring physical complications like high blood pressure. This phenomenon also confirms that most of high blood pressure patients are from low and mid-income counties (the statement by WHO). Overall, we can conclude that there are about 33.51% adults are potentially having high blood pressure in the USA in 2019. This data is very close to the prediction by American Heart Association in 2010. It means the proportion of high blood pressure patients does not decline, instead shows a slight increasing in recent years. There is still a long way to go to achieve the decreasing goal by WHO.

Weakness

- Even though the census data in the USA is relatively representative for worldwide application, since America is a multicultural country with people from different races background, it is still biased to represent universally, especially because the US is a developed country.
- Even though this is the newest survey data compared to American Heart Association's report, this the survey data is from 2012, which is still a bit far away to predict results with census data 2019. According to WHO report, genetic problems is also a main cause, but it's not included in the survey data.
- According to P-values, we found that BMI value is more significant than weight. However, the height is not included in census data so we cannot do BMI analysis.
- The survey data does not include diet summary (like high salty diet), which might also impact on the final prediction.
- High blood pressure will consider both systolic blood pressure reading and Combined diastolic blood pressure reading in real clinic condition, however, we only considered the systolic in the paper.

Next steps

- This paper is designed for the US prediction, if it's necessary to predict the universal probability, then more data will be needed for more countries.
- In order to improve results of this paper, it is better to obtained survey data with more variables like family disease history or diet. (But NHANES is the most widely used analysis data for high blood pressure discussion.)

- If the newest NHANES will be updated, then use the new survey data to build the model so that the time slot will match the census data time slot (2019).
- Moreover, it is better to consider both Combined systolic blood pressure reading and Combined diastolic blood pressure reading in modeling process in following steps.

References

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., Fox, C. S., Franco, S., Fullerton, H. J., Gillespie, C., Hailpern, S. M., Heit, J. A., Howard, V. J., Huffman, M. D., Kissela, B. M., Kittner, S. J., … Turner, M. B. (2013). Heart disease and stroke statistics—2013 update. *Circulation*, *127*(1). https://doi.org/10.1161/cir.0b013e31828124ad

Johnson, C. L., Dohrmann, S. M., Burt, V. L., & Mohadjer, L. K. (2014). National health and nutrition examination survey: sample design, 2011-2014. *Vital Health Statistics*, (162), 1–33.

Madell, R. (2018, January 26). *Blood pressure readings explained*. https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained

Mayo Clinic Staff. (2019, November 19). *High blood pressure dangers: Hypertension's effects on your body*. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868

National Center for Health Statistics. (2017, September 15). *About the national health and nutrition examination survey*. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Pruim, R. (2015, July 2). *Package 'NHANES': Data from the US national health and nutrition examination study* [R package version 2.1.0]. https://cran.r-project.org/web/packages/NHANES/NHANES.pdf

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). *IPUMS USA: Version 10.0*. Minneapolis, MN, IPUMS. https://doi.org/10.18128/D010.V10.0

WebMD. (2004, November 4). *Causes of high blood pressure*. https://www.webmd.com/hypertension-high-blood-pressure/guide/blood-pressure-causes#1

WHO. (2019, September 13). *Hypertension*. https://www.who.int/news-room/fact-sheets/detail/hypertension

Wickham, H., & RStudio. (2019, November 21). *Package 'tidyverse': Easily install and load the 'tidyverse'* [R package version 1.3.0]. https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf