The link between hypertension and smoking and other factors

Jiekai Yin

December 19, 2020

1. Abstract

In modern society, hypertension is a common issue that troubles many people. Using American National Health and Nutrition Examination surveys (NHANES 2009-2012), we explore the effects of smoking and other factors, in the aspects of physical traits, health and lifestyle, on hypertension (defined as the systolic blood pressure that is higher than 130 mmHg). The propensity score matching is used for the smoking propensity, to eliminate the bias of the estimates of the effects of smoking on hypertension. The logistic regressions are fitted with the factors that are correlated with high blood pressure as the predictors and whether the participants have hypertension as the outcome variable. The estimated coefficients of the model shows that the individuals with the gender of male, a higher age, a larger weight, a higher frequency of depression, having more number of alcoholic drinks in the past year, having a habit of cigarette smoking are more likely to have hypertension.

Key words: Propensity Score, Blood Pressure, Health, Smoking, Hypertension

2. Introduction

Research Inspiration

In modern society, hypertension is a pretty common issue that troubles a large group of population, especially for senior citizens aged over 50 (Pinto).It is a very risky factor for other diseases of heart and blood vessel. The previous studies have shown that cigarette smoking also has an effect of inducing cardiovascular and other types of diseases (NH, and IS). And when the effects of hypertension and smoking are associated, the risk of getting cardiovascular disease has increased dramatically (Leone). In this paper, the effects of cigarette smoking and other factors on hypertension are explored. By checking the effects of the key factors on high blood pressure (defined as the combined systolic blood pressure that is larger than 130 mmHg), the hypertension may be prevented by reducing the influence of these factors such as quit cigarette smoking.

The research analysis is based on the dataset National Health and Nutrition Examination Survey (NHANES 2009-2012). It is collected by the US National Center for Health Statistics (NCHS). The collection of the survey data started from the early 1960's.The raw dataset of NHANES contains 78 columns and 20293 observations from the year between 2009-2010 and 2011- 2012.

Research objective

Given a wide set of features of individuals who completed the survey of NCHS, we want to investigate the characteristics (including smoking) of the individuals that have effects on whether he or she has a blood pressure that is higher than 130 mmHg. The logistic regression is employed to investigate the magnitude and direction of the effects of the key factors on the high blood pressure of the participants. The propensity score matching a quasi-experimental method used in the observational studies to reduce bias. A treatment is defined, and the treated units are matched with the non-treated units with very similar characteristics. In the paper, this method is used to make more accurate estimates of the effects of smoking.

3. Data

Target Population, Sampling Approach and Dataset

NHANES is a survey data. Its target population contains the civilian resident population who are non-institutionalized. The individuals were interviewed at home. They were also required to complete the health examination component of the survey, which was taken in a mobile examination center (MEC). The data we used in the analysis is NHANESraw from the R package NHANES. It contains 78 columns, including the aspects of demography, physical measurements, health and lifestyle, and plus 4 variables that describe the sample weighting adjustment.

Characteristics and Chosen Predictors

The table below displays all selected variables that are considered affecting the outcomes of the observations. The 13 variables, out of 75 of the total, are chosen to inform the relationship between the predictors and the outcomes of the participants. The selected potential factors are used to fit the base model, which will be discussed in the model section. Selecting the potential influential variables rather than using all the predictors can be easier to construct models and might be able to make more accurate estimates. The predictors contain 7 categorical variables and 6 numerical variables.

List of Chosen Variables and Descriptions

VARIABLE.NAME	TYPE	FEATURES	CATEGORIES
Gender	categorical	Gender of the individual	male, female
Age	Numerical	Age of the individual	Numeric values between 0 and 80
Poverty	Numerical	A proportion of household income to poverty guidelines	Numeric values between 0 and 5
Weight	Numerical	Weight of the individual in kg	Numeric values in unit kg
BMI	Numerical	Body mass index of the individual	Numeric values in unit kg/m2

HealthGen	Categorical	Rateing of general health reported by the individual	Excellent, Vgood, Good, Fair, Poor
Depressed	Categorical	The frequency of the days that the participant felt depressed reported by themselves	None, Several, Majority (more than half the days), AlmostAll
SleepHrsNight	Numerical	The number of hours the participant usually sleep reported by themselves	Numeric values in unit hour
SleepTrouble	Categorical	Whether the participant has told a doctor that they have trouble falling asleep	Yes, No
PhysActive	Categorical	Whether the participant does sports regularly	Yes, No
AlcoholYear	Numerical	Estimated number of days that the participant drank alcoholic beverages	Numeric values in unit day
SmokeNow	Categorical	Whether the participant smokes cigarettes regularly	Yes, No
RegularMarij	Categorical	Whether the participant use marijuana regularly	Yes, No
High_bp	Categorical	Whether the participant has hypertension	1, 0

Data Cleaning and Propensity Score Matching

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

The final dataset contains 3600 observations with 13 characteristics of the participants. The observations with missing values are dropped from the dataset. The variable High_bp is created to indicate if the individual has hypertension. The criterion of having hypertension is defined as the combined systolic blood pressure larger than 130 mmHg. The method of propensity score matching is used to finalize the observations of the final dataset. To accurately estimate the effects of cigarette smoking on hypertension, the variable SmokeNow is considered as the treatment and the dummy variable High_bp is considered as the outcome of the interest.

We match the participants who are smokers and non-smokers with similar characteristics in the dataset. In this way, the estimates of the effects of smoking can be less bias by eliminating the effects of other factors that can influence the treatment.

4. Model

The models are fitted by using logistic regression to interpret the relationship between the characteristics of the participants including whether smoking and whether they have hypertension. The binary logistic regression is appropriate to estimate the probabilities of the dependent variables with only two possible values.

The base model is firstly constructed with all 13 predictors included. There might be many insignificant variables in the regression. Therefore, the model selection is performed by using forward stepwise and backward stepwise selection using AIC and BIC. The final model is chosen based on the value of AIC to ensure that the model has a better fit than other candidate models. Considering smoking may have multicollinearity with other predictors, VIF is also checked to make sure that the predictors do not have high correlations with each other. RStudio is used for propensity score matching and model development.

5. Results

The forward stepwise and backward stepwise AIC/BIC selections are performed for model selection. The model selected from the backward stepwise BIC selection is chosen as the final model, since it has the smallest value of AIC, which is 1368.377. The model contains 7 predictors. The variance inflation factors of the variables all have the values around 1, indicating that there is no strong multicollinearity occurs in the final model.

6. Discussion

Summary

Hypertension and cigarette smoking are two risk factors that can induce other cardiovascular diseases. Combing the effects of smoking and high blood pressure, the risk of getting cardiovascular diseases grows exponentially. The research objective is to investigate if smoking and other characteristics, in the aspects of physical traits, health and life style, have effects on high blood pressure. The analysis in based on the dataset NHANES 2009-2012. It is an observational data including a set of health and nutrition surveys in the United State. By employing propensity score matching, we match the participants who are smokers and non-smokers with similar characteristics, in order to make the estimates of the effects of smoking less bias. The forward stepwise and backward stepwise AIC/BIC selections are used for model selection. The final model is selected based on the value of AIC. It contains the variables Gender, Age, Weight, Depressed, SleepTrouble, AlcoholYear and SmokeNow, indicating that these factors have effects on the high blood pressure.

The coefficients of the summary table of the final model

	Estimate	Std..Error	z.value	Pr…z..
(Intercept)	-5.659	0.4576	-12.37	3.938e-35
Gendermale	0.4301	0.1517	2.835	0.004578
Age	0.05235	0.00693	7.554	4.23e-14
Weight	0.01656	0.003006	5.509	3.618e-08
as.factor(Depressed)Several	0.5205	0.1646	3.162	0.001569
as.factor(Depressed)Most	0.1905	0.2551	0.7468	0.4552
as.factor(SleepTrouble)Yes	-0.32	0.1567	-2.043	0.0411
AlcoholYear	0.001573	0.0006191	2.541	0.01104
as.factor(SmokeNow)Yes	0.3233	0.1381	2.34	0.01928

Conclusions

The final model build by using logistic regression is fitted well and all predictors are statistically significant in the model (all p-values are smaller than 0.05). The characteristics that have important relationships with high blood pressure are gender, age, weight, the frequency that the participants felt depressed, whether the participants have trouble sleeping, the days the participants drank alcoholic beverages last year and if they smoke regularly. The estimated coefficients indicate both the magnitude and direction of the effects of each predictor. Based on the summary table, age and gender of the participants are the two most important factors correlated with the high blood pressure, since these variables have the smallest p-values in the model. The estimates shows that the individuals with the gender of male, a higher age, a larger weight, a higher frequency of depression, having more number of alcoholic drinks in the past year, having a habit of cigarette smoking are more likely to have hypertension.

The propensity score analysis indicates that the individuals who smoke regularly are 0.3233 times more likely to develop hypertension (having the systolic blood pressure higher than 130 mmHg) than the people who do not have a habit of smoking. The p-value of the estimated coefficient is 0.01, meaning that the value is statistically significant at the 99% confidence interval. This effect shows that quitting smoking can reduce about 30% possibility of developing hypertension. Therefore, giving up smoking is very essential in preventing hypertension. Combined the elimination of the effect of smoking and the reduction of high blood pressure, the risk of developing other cardiovascular diseases can also be reduced.

Weakness

The dataset NHANES 2009-2012 is constructed under a clustered design that the samples are assigned different weight. Failure to use the variables of the sampling weight in the analysis for weight adjustment may lead to biased estimates.

Propensity score matching is used for the smoking propensity in this research. This method mimics some

features of the randomized controlled trial. Whether the participants smoke is defined as the treatment. Based on the propensity score matching, the other characteristics of the individuals who are smokers and non-smokers can be similar. It helps to reduce the influence of the treatment bias. While, this method has some limitations. There still might be unmeasured or unobserved confounding variables, that can cause bias of the estimates. Therefore, the estimated effect of smoking may not be considered as a causal effect. Additionally, the method can cause data reduction. After employing the method, the sample size decreases by 420 observations. This can decrease the power of the statistical tests.

Next Steps

The dataset NHANES 2009-2012 contains weighted selections. The failure to adjust the sample weights may cause the inaccurate estimates. Accounting for sampling parameters in the analysis can reduce the bias of the estimates.

Logistic regression is the only model used in this research. The model diagnostics can be performed to verify the assumptions of the model. Additionally, the model validation can be performed to check its prediction accuracy.

7. References

1.Pinto, Elisabete. Blood Pressure And Ageing. 2007, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805932/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805932/). Accessed 20 Dec 2020.

2.NH, Miller, and Ockene IS. Cigarette Smoking, Cardiovascular Disease, And Stroke: A Statement For Healthcare Professionals From The American Heart Association. American Heart Association Task Force On Risk Reduction. 1997, https://pubmed.ncbi.nlm.nih.gov/9386200/ (https://pubmed.ncbi.nlm.nih.gov/9386200/). Accessed 20 Dec 2020.

3.Leone, Aurelio. Smoking And Hypertension. 2015, https://medcraveonline.com/JCCR/smoking-and-hypertension.html (https://medcraveonline.com/JCCR/smoking-and-hypertension.html). Accessed 20 Dec 2020.

4.Pruim, R., 2015. Data From The US National Health And Nutrition Examination Study. 2nd ed. [ebook] Available at: <https://cran.r- (https://cran.r-) project.org/web/packages/NHANES/NHANES.pdf> [Accessed 25 June 2020].

5.“Propensity Score Matching”. THE WORLD BANK, https://dimewiki.worldbank.org/wiki/Propensity_Score_Matching (https://dimewiki.worldbank.org/wiki/Propensity_Score_Matching). Accessed 20 Dec 2020.

6.Pruim, Randall. Package ‘NHANES’. 2016, https://cran.r-project.org/web/packages/NHANES/NHANES.pdf (https://cran.r-project.org/web/packages/NHANES/NHANES.pdf). Accessed 21 Dec 2020.

7.“Propensity Score Matching”. En.Wikipedia.Org, 2020, https://en.wikipedia.org/wiki/Propensity_score_matching (https://en.wikipedia.org/wiki/Propensity_score_matching).

8.Randall Pruim (2015). NHANES: Data from the US National Health and Nutrition Examination Study. R package version 2.1.0. https://CRAN.R-project.org/package=NHANES (https://CRAN.R-project.org/package=NHANES)

9. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)
10. Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/ (http://www.jstatsoft.org/v33/i01/).
11. Gergely Daróczi and Roman Tsegelskyi (2018). pander: An R ‘Pandoc’ Writer. R package version 0.6.3. https://CRAN.R-project.org/package=pander (https://CRAN.R-project.org/package=pander)

12.Frank E Harrell Jr (2020). rms: Regression Modeling Strategies. R package version 6.0-0. https://CRAN.R-project.org/package=rms (https://CRAN.R-project.org/package=rms)

13. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
14. David Robinson and Alex Hayes (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.6. https://CRAN.R-project.org/package=broom (https://CRAN.R-project.org/package=broom)

8. Link to the Github repository

https://github.com/JackieYin323/304final.git (https://github.com/JackieYin323/304final.git)