Testing for Causal Associations between Smoking and Sleep Trouble

Jia Yuan Liu 1004793841

22 December 2020

Code and data supporting this analysis is available at https://github.com/ljyljyljyyy/TestingCausality

Abstract

To determine whether smoking is a causal factor for sleep trouble, a logistic regression model with propensity score matching is applied on the 2009-10 NHANES dataset focusing on respondents over 17 years of age. Confounding factors such as depression and demographic measures are controled using propensity score matching in creating a basis non-smoking respondent group and a comparitive smoking respondent group with similar pretreatment factors. After adjusting for differences in baseline risk factors and performing logistic regression, regular smoking behavior is concluded as a statistically significant factor in causing sleep trouble.

Key Words

Causal Inference, Propensity Score, Observational Study, Smoking, Sleep Trouble

Introduction

Frequently having trouble sleeping (insomnia) is a frustrating experience that can seriously impact one's mental and physical health. By identifying the main variables that associate with sleep trouble and analyzing the causal link, individuals can effectively take precaution from developing sleep trouble accordingly. The primary exposures of interest in this study are "Smoke100"- a binary variable indicating whether the participant has smoked more than 100 cigarettes in their life and "Sleep Trouble" - a binary variable indicating whether participants are having trouble sleeping. The primary purpose of this study is to test the hypothesis that the association between smoking and sleep trouble is plausibly a causal association.

The strongest causal modeling framework would be randomly assigning participants into treatment and control groups in order to control the systematic group differences in pre-treatment characteristics that may explain any differences in observation (Harder et. al., 2016). Because experimental design is usually not feasible from both an economic and practical standpoint, an alternative approach is to use propensity score matching to compare the sleep trouble outcomes of smoking participants to those of non-smoking participants whose pre-treatment characteristics on many risk factors are similar.

The "NHANES" dataset collected from the studies performed by the National Center for Health Statistics will be used to investigate how propensity score matching could be used to make inference on the causal link between smoking and sleep trouble. In the Methodology section, the data and the model that was used to perform the propensity score analysis will be introduced. Model diagnosis and model summary results are provided in the Results section. Inferences, conlusions along with limitations of this analysis are presented in Conclusion section.

Methodology

Data

The used "NHANES" dataset was collected from the studies performed by the National Center for Health Statistics, aiming to monitor trends in the prevalence, awareness, treatment and control of selected diseases and its risk factors (Curtin et. al., 2013). The target population of this survey is all resident civilian noninstitutionalized population of United States (Curtin et. al., 2013). A four-stage sample design was used; the frame for each stage included 2007-2010 counties, area segments, dwelling units and household (Curtin et. al., 2013). A total of 6,525 people were sampled and 5,000 people were examined at the end where non-response observations were dicarded (Curtin et. al., 2013). This analysis only focuses on the data of participants of 17 years or older and data collected in the year 2009-10 from the NHANES dataset.

The main advantage of the NHANES dataset is that it is an extensive dataset that encompasses as many as 78 variables that range from demographic measures to health statistics to lifestyle choices. The vast amount of variables is useful when using them for constructing models since diverse factors can be accounted for in the prediction. However, due to the breadth of the dataset, there is limited information on specific topics of interest. For example, the target variable of interest in this study is sleep trouble, but only two variables "Sleep Trouble" and "SleepHrsNight" are related to sleep, which restricts the specificity of the analysis.

Table 1: Summary Statistics based on Sleep Trouble: Median (IQR); n(%)

Smoke	**Overall**, N = 4,828	**0**, N = 2,560	**1**, N = 2,268
Age	49 (35, 64)	46 (32, 62)	52 (38, 66)
Gender			
female	2,432 (50%)	1,466 (57%)	966 (43%)
male	2,396 (50%)	1,094 (43%)	1,302 (57%)
Poverty	2.06 (1.09, 4.08)	2.31 (1.16, 4.37)	1.78 (1.01, 3.55)
Depression	1,214 (25%)	564 (22%)	650 (29%)
PhysActive	2,240 (46%)	1,291 (50%)	949 (42%)
SleepTrouble	1,247 (26%)	557 (22%)	690 (30%)

Summary Statistics from table 1 is grouped by "Smoke" - the treatment variable of this study. There are 2,515 non-smoking respondents and 2,219 smoking respondents. Smoking respondents have a significantly higher average years of age (52) compared to non-smoking respondents (46). In addition, smoking respondents have a significantly higher percentage of male respondents (58%) compared to non-smoking respondents (43%). Also, smoking respondents have a significantly lower income to poverty ratio (1.80) compared to non-smoking respondents (2.31), with a lower poverty ratio representing lower income. A significantly higher percentage of smoking respondents experience depression (29%) than non-smoking respondents (22%). A higher percentage of non-smoking respondents are physically active (51%) than smoking respondents (42%). The outcome variable of interest is "SleepTrouble", which is a binary variable that records whether participants have trouble sleeping. There exists a significantly higher percentage of smoking respondents who have sleep trouble (31%) compared to non-smoking respondents (22%). Propensity score matching is used in the Model section to test whether the association between smoking and sleep trouble is causal.

Model

The purpose of this study is to analyze whether the treatment variable - smoking has a causal effect on the outcome variable - sleep trouble. To isolate the effects of smoking, any group differences on characteristics other than smoking need to be controlled to avoid confounding effects. In addition to demographic measurements including age, gender, race, education and income levels, one important confounding factor is depression. Those who are depressed are more likely to smoke and are thus more likely to experience sleep trouble.

Propensity score matching is used to create matches between a non-smoking baseline group respondent whose pretreatment characteristics are extremely close to a smoking comparison group respondent. A logistic

regression was constructed to predict whether a person received treatment / smoking as a function of a variety of predictor variables that may explain it.

$$log\ (\frac{p_{smo\ i}}{1-p_{smo\ i}}) = \beta_0 + \beta_1 X_{age\ i} + \beta_2 X_{gen\ i} + \beta_3 X_{race\ i} + \beta_4 X_{pov\ i} + \beta_5 X_{edu\ i} + \beta_6 X_{phy\ i} + \beta_7 X_{dep\ i}$$

$$Such\ that\ i = \{1, \ldots, N\}$$

Each respondent is denoted by i. The estimated propensity score represents each individual's probability of smoking based on the values of pretreatment controls. Next, forecast is used to create matches between each smoking respondent and a non-smoking respondent with the closest propensity score. The dataset is then reduced to only keep those matched.

A second logistic regression model was performed to fit the relationship between whether smoking is a significant factor in causing sleep trouble. The dataset used to fit the model now only consists of the matched non-smoking basis group and smoking comparison group. The model was chosen because the outcome variable of this study - sleep trouble is a binary variable. The null hypothesis is that all coefficients equal to zero, indicating that there exists no relationship between various predictors and sleep trouble. Each respondent is denoted by i. The model takes "Age" and "Poverty" as numerical variables instead of grouping respondents into different age or ratio groups. The decision stems from the reason that the majority of the data cluster around the mean, and keeping numerical data ensures precision when making statistical inferences. "Gender", "Depression", "PhyActive" and "Smoke" are categorical variables since the values are binary and there is no intrinsic order to the categories. They are included in the regression as dummy variables where the categorical variables are represented in a numerial form. The estimated regression represents each individual's probability of experiencing sleep trouble based on the values of age, gender, poverty, depression, physical activity and smoking.

$$log\ (\frac{p_{slp\ i}}{1-p_{slp\ i}}) = \beta_0 + \beta_1 X_{age\ i} + \beta_2 X_{gen\ i} + \beta_3 X_{pov\ i} + \beta_4 X_{dep\ i} + \beta_5 X_{phy\ i} + \beta_6 X_{smoke\ i}$$

$$Such\ that\ i = \{1, \ldots, N\}$$

Various model assumptions were checked to ensure that the estimates are unbiased and consistent. "Sleep-Trouble" is a binary variable which fulfills the logistic model assumption of having a binary outcome variable. Since the survey employs stratified random sampling where simple random sampling without replacement in each strata was used, it fulfills the regression assumption of having independent observations and pairwise independent errors. There are 4,438 observations in the matched propensity score dataset used for logistic model construction, which satisfies a large sample size. More model diagnostics such as linearity and multi-collinearity checks are performed to ensure that regression assumptions are fulfilled. Diagnostic results are included in the "Results" Section.

Results

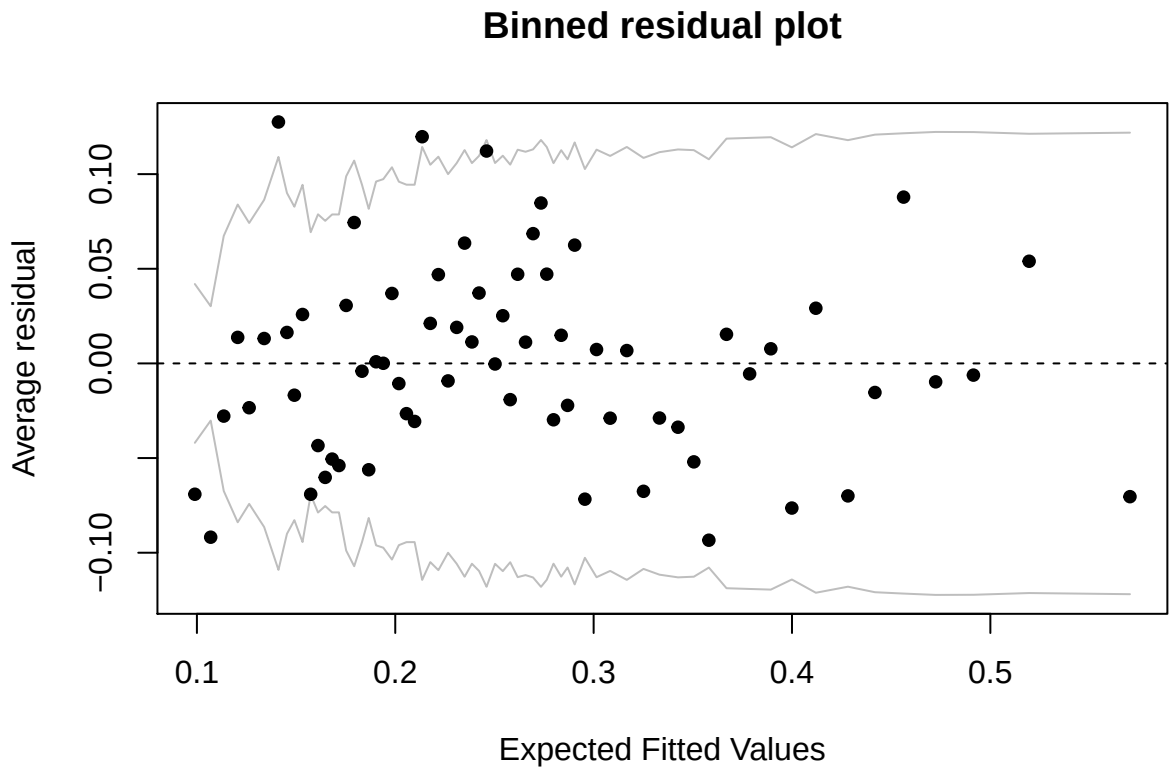
Table 2: Statistical Summary of Logistic Model for Sleep Trouble

Estimated_Parameters	Estimated_Coefficient	p_value
Age	0.0217	0.000
Gendermale	-0.6579	0.000
Poverty	0.0322	0.194
DepressionYes	0.9783	0.000
PhysActiveYes	0.0228	0.774
SmokeYes	0.6279	0.000

With higher values representing increased log odds of experiencing sleep trouble, logistic model summary from table 2 results that Smoking has a coefficient of 0.6279 and a p-value < 0.001, indicating causal relationship between smoking and sleep trouble. Model summary also shows that age, gender and depression are significant factors for developing sleep trouble. The coefficient is 0.0217 for age, -0.6579 for male, and 0.9783 for depression. They all have a p-value < 0.001. However, "Poverty" and "PhysActive" have p-values > 0.05. The dataset is based off the propensity score matching of a basis group of non-smoking respondents and a comparative group of smoking respondents with close pretreatment characteristics. Based on the basis and treatment groups, a logistic regression model was constructed, which accounted for "Age", "Gender", "Poverty", "Depression", "PhysActive", and "Smoke" variables.

Linearity

One of the main assumptions of logistic regression is that the log odds of the outcome variable and the predictors are assumed to have a linear relationship.Binned residual plots are obtained from categorizing the data into bins based on their fitted values, and then plotting the average residual versus the average fitted value for each bin (Gelman and Hill, 2007). Binned residual plot shows that there is almost no average residual outside the ± 2 SE bands and the average residuals are scattered around the horizontal axis, fulfilling the linearity assumption.



Multicollinearity

Another assumption of logistic regression is that little or no multicollinearity among the predictor variables is required, which means that the predictor variables should not be highly correlated with each other. Variance inflation factor is a tool that quantifies the severity of multicollinearity in regression by calculating the increase in variance of an estimated coefficient due to collinearity. Table 3 shows that all predictors of the propensity score regression model has VIF scores close to 1, indicating that the no multicollinearity assumption is fulfilled.

Table 3: VIF of Propensity Score Regression Model

	x
Age	1.087365
Gender	1.040466
Poverty	1.112328
Depression	1.069495
PhysActive	1.091115
Smoke	1.029748

Discussion

Summary

In this study, a logistic regression model with propensity score matching is applied in predicting the causality of smoking on sleep trouble. First, descriptive statistics of 2009-10 NHANES dataset on respondents over 17 years old were summarized based on whether respondents have smoked 100 cigarattes in their life (NHANES, 2019). The comparison demonstrates a significant higher percentage of smokers who experience sleep trouble (31%) compared to non-smokers (22%). In order to investigate whether this association is causal, confounding factors such as depression and physical activity need to be controled. Propensity score matching is used to create a basis non-smoking respondent group and a comparitive smoking respondent group with similar pretreatment factors. Several predictor variables including age, gender, race,poverty, education, physical activity and depression are used to predict the probability of smoking based on a logistic regression performed using R (R Core Team, 2019b). After creating matches between smoker and non-smoker who have closest propensity scores using the arm package, the dataset is then reduced to only contain the matches of basis and comparison groups (Gelman and Hill, 2007). A logistic regression is then performed on predicting sleep trouble based on the predictor variables age, gender, poverty, depression, physical activity and smoking. Various model diagnosis are performed to ensure that regression assumptions are fulfilled. Results interpretation as well as study limitations and potential improvements are discussed.

Interpreting the Model

Table 4: Statistical Summary of Logistic Model for Sleep Trouble

Estimated_Parameters	Estimated_Coefficient	Std_error	z_value	p_value
Age	0.0217	0.0021	6.648	0.000
Gendermale	-0.6579	-0.0719	-6.874	0.000
Poverty	0.0322	0.0231	1.389	0.194
DepressionYes	0.9783	0.0773	11.021	0.000
PhysActiveYes	0.0228	0.0743	0.315	0.774
SmokeYes	0.6279	0.0718	6.440	0.000

By looking at the coefficients of the predictor variables for predicting sleep trouble from table 4, the estimations closely align with the data distribution discovered in exploratory summary statistics analysis. The fitted logistic model reflects that holding all else constant, the log odds for experiencing sleep trouble is 0.6279 (p-value < 0.001) for smoking, verifying the association between the difference in sleep trouble percentages between smoking and non-smoking groups as seen in table 1 to be causal. In addition, age, gender and depression are also statistically significant factors that associate with sleep trouble. Holding all else constant, the log odds for experiencing sleep trouble increases by 0.0217 (p-value < 0.001) with an increase in one year of age. Confirming the observation of higher percentage of females experiencing sleep trouble than male from table 1, modeling results show that holding all else constant, the log odds of male experiencing

sleep trouble is -0.6579. Depression has the strongest association with sleep trouble as its coefficient is 0.9783 (p-value < 0.001), indicating that holding all else constant, the log odds for experiencing sleep trouble for depressed individuals is as high as 0.9783. However, poverty and physical activity are not statistically significant factors in predicting sleep trouble since their p-values > 0.05 and the null hypothesis of having no relationship between the predictor and the outcome variable has failed to reject. These statistics show that older, depressed smoking females are significantly more likely to experience sleep trouble.

Conclusion

Based on logistic regression estimation with propensity score matching, regular smoker respondents are 62.8% (p-value < 0.001) more likely to experience sleep trouble than non-smoker respondents who has similar pretreatment characteristics including age, gender, race, education, poverty, physical activity and depression. The analysis is based off the propensity score matching of a basis group of non-smoking respondents and a comparative group of smoking respondents with close pretreatment characteristics. Based on the basis and treatment groups, a logistic regression model was constructed, which accounted for "Age", "Gender", "Poverty", "Depression", "PhysActive", and "Smoke" variables. As a result, it can be conluded that smoking is a statistically significant factor in causing sleep trouble.

Weakness & Next Steps

The dominant weakness of this analysis is that some significant pretreatment characteristics may be inadvertently omitted which may lead to omitted variable bias. Unlike an ideal randomized experiment, some significant predictor that is not measured in the survey may influence the prediction of smoking behavior and thus cause bias under propensity score matching. Although a number of predictor variables are already included in the model in predicting smoking as a treatment, the model can be improved by putting more variables into consideration. For instance, A.D. McNeill suggests that parental smoking behaviour and beliefs about the effects of smoking on health are potential factors that influence one's behavior of smoking. One future step of the study is to perform a follow-up survey on respondents to gain more detailed information on potential factors that may influence their behavior of smoking.

Another limitation of the study is the reliance on respondents' self report on whether they are experiencing sleep trouble. This may involve self report bias due to different individuals' interpretation of sleep trouble. For example, some respondents may characterize falling asleep 15 minutes after going to bed as experiencing sleep trouble, while others may define sleep trouble as being unable to fall asleep for more than 30 minutes after they go to bed. The interpretation of sleep trouble may also differ on the frequency. Some people may report experiencing sleep trouble when they struggle to fall asleep for one or two days, while others may define having sleep trouble if they struggle with falling asleep for more than seven days. To improve the estimation of the analysis, one can redefine the survey problems and clearly categorize sleep trouble based on the severeness and frequency. The severity of sleep trouble can be answered using an ordinal type of variable where respondents identify their time to fall asleep within certain ranges of time, for example, less than ten minutes / ten to thirty minutes / thirty to sixty minutes etc. Sleep trouble can also be specified based on the frequency. For instance, respondents will be asked whether they suffer from short-term (less than seven days) sleep trouble or long-term (more than seven days) sleep trouble. Properly defining the survey questions can help focus the scope of the study and produce more accurate results.

References

"National Health and Nutrition Examination Survey (NHANES)." National Health and Nutrition Examination Survey (NHANES) | HealthData.gov, 18 Nov. 2019, healthdata.gov/dataset/national-health-and-nutrition-examination-survey-nhanes-%E2%80%93-vision-and-eye-health-surveillance.

Curtin L. R., Mohadjer L.K, Dohrmann S. M., Kruszon-Moran D., Mirel L.B. (2013). National Health and Nutrition Examination Survey: Sample Design, 2007–2010. National Center for Health Statistics. https://www.cdc.gov/nchs/data/series/sr_02/sr02_160.pdf

Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-15.

R Core Team (2019b). R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Xie Y (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, https://yihui.org/knitr/.

Harder, V. S., Morral, A. R., & Arkes, J. (2006). Marijuana use and depression among adults: testing for causal associations. Addiction, 101(10), 1463–1472. https://doi.org/10.1111/j.1360-0443.2006.01545.x

McNeill, A. D., Jarvis, M. J., Stapleton, J. A., Russell, M. A., Eiser, J. R., Gammage, P., & Gray, E. M. (1989). Prospective study of factors predicting uptake of smoking in adolescents. Journal of epidemiology and community health, 43(1), 72–78. https://doi.org/10.1136/jech.43.1.72

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge; New York: Cambridge University Press.