A study of smoking behaviors among adolescents based on American National Youth Tobacco Survey

Lan Cheng 1004767102

Dec 21, 2020

Abstract

This study investigated the problem of smoking behaviors among adolescents based on the American National Youth Tobacco Survey data. The study mainly applied logistic models to investigate whether smoking behaviors are different between rural and urban, male and female as well as white and black people. It was found that there is strong evidence shows smoking behaviors among young students are different affected by age, gender and race as well as locations whether in rural or urban. For example, it was found white male youths aged 20 years old smoking with much higher probability compared with black youth with similar other covariates in both rural and urban. With the findings of this study, it suggests for government to make better dicisions on developling a more comprehensive prevention and control programs for America. youth in smoking cigars.

Keywords: cigarettes smoking; American National Youth Tobacco Survey; logistic model;

1 Introduction

Smoking is considered to be very harmful to human health all around the world. There are lots of diseases associated with smoking badly. For examples, fat, diabetes are all related with smoking in a high frequency such as one package per day or even more. Lots of former studies have proved the facts. Doll R. (1996) already showed smoking related to lots of cancers. Murray J and Lopez AD (1997) also have shown the mortality related to smoking cigars are very high.

And it is known smoking cigars is a common behavior among adults, things would be even worse if smoking cigars happens on youth which would affect the youth's heath seriously leading to a series of bad outcomes such as pooring performances on schools, madding moods or even dropping from schools. Thus, it is very necessary to study the behaviors of smoking among the youth to figure out the situations of smoking among them for better decisions to protect them. Also, some former studies have shown that it is well known that rural children smoke more often than urban ones, and children smoking behaviors are different fro different ages and ethnicities. In addition, states and regions may differ significantly in young children's smoking behavior which means smoking behavior in young children may vary from school to school. Under this backgroud, this study aims to investigate the adolescent's smoking behavior based on the American National Youth Tobacco Survey data. In particular, we are interested in the questions that whether smoking behaviors are different between rural and urban, male and female as well as white and black people. To do the tasks,

the study would apply logistic model with random effects, the study is organized as below: first give an introduction, then introduce data sources and models used in the study followed by showing all of the results, at last, we draw conclusions and make discussions of findings, weakness and future work.

The link of report is: https://github.com/CL200306/Smoking-Behavior/pulls.

2 Data

The data source of this study comes from the 2014 and 2019 American National Youth Tobacco Survey (NYTS) data sets. And it is restricted to the youngers with ages no less than 10 years old in this study as we think youngers too young should not and can be considered as not evered smoke at all. The American National Youth Tobacco Survey includes lots of information of the use of cigars and other similar things such as tobacco amongst American school children. As this study, we are only interested in investigating the smoking of cigars, so other things are not studied.

The goal of the survey is to help researchers and public health managers to explore the data so that they can compare the use of cigars or tobacco products across states. And according to the official document, the NYTS data was designed to provide national data for the design, implementation and evaluation of comprehensive tobacco prevention and control programs for youth in America.

The survey is a national wide survey which is a a census survey and it is aimed to ensure that the findings based on the survey could be generalized to the entire population in America. The population is entire population of American school children. The sampling frame is entire population of possible participants of completing the questionares in the survey.

However, there are might be biasness in the survey, as for the youth of American school children. The sampling frame might not close to the entire target population as if the enrollment lists are not current for some schools, the survey can not draw samples from those schools leading to a non-representive sample obtained. So this survey used other methods like selecting a sample exist, including using electric and gas company records to replace the missing enrollment lists in schools. Thus, given the possible biased samples, this study's results might be changed due to different samples, however, this might only affect the magnitude of effects of covariates which should not affect the direction of effects of the covariates as the samples obtained are already large enough. Also, the survey is a face-to-face survey not a telephone survey, so the non-response bias would be smaller.

At last, there is a common challenge in determining the sampling frame and size for the national survey, as there are lots of information needed to be collected. This survey uses sampling method cluster sampling by sampling from communities and regions (states) to draw samples and ensure that each sample should be a representive sample of the major clusters.

3 Models

The basic logistic model is first applied to the data, for the questions interested, the logistic model is as follows:

$$log(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}) = X\beta + \epsilon$$

The response is whether ever smoking cigars, so the $P(Y_i = 1)$ is a binary outcome for the ith individual and X are covariates: sex, rurality, ethnicity, age.

Also, as the data is across states and schools, we also considering random effects and build the logistic model with random effects as follows:

$$Y_{ijk} \sim Bernoulli(\lambda_{ijk})$$

$$logit(\lambda_{ijk}) = \mu + X_{ijk}\beta + U_i + Z_{ij}$$

- Y_{ijk} is whether smoking for a young people.
- λ_{ijk} is probability of smoking for a young people.
- X_{ijk} are covariates: sex, rurality, ethnicity, age.
- U_i and Z_{ij} are random effects for states and schools.

First, we build a logistic model as it is very simple and easy to help us understand relations among different characteristics with the response whether smoking or not. We choose logistic model in the first stage, as the response is a binary outcome, and logistic model has a very good interpretion of model results compared with probit model and other models from binomial families.

Second, we build a logistic model with random effects as we notice that the data is collected across states and schools, so there must be random effects of states and schools, we need to add them based on the basic logistic model. And for the questions of interest, we interpret the model results mainly on the fixed effects estimations of the logistic model with random effects.

4 Results

Table 1 shows the logistic model estimates of the possibility of cigar smoking. As it can be seen from the table, since the p values of the three factors are all less than 0.05, the three factors are all significant at the

Table 1: logistic model of cigar smoking

	Estimate	Std. Error	z value	Pr(>|z|)	OR	2.5 %	97.5 %
(Intercept)	-6.80	0.17	-39.16	0.00	0.00	0.00	0.00
RuralUrbanRural	0.64	0.04	14.91	0.00	1.90	1.75	2.07
Raceblack	-0.39	0.07	-5.53	0.00	0.68	0.59	0.78
Racehispanic	-0.08	0.05	-1.74	0.08	0.92	0.84	1.01
Raceasian	-0.99	0.14	-7.02	0.00	0.37	0.28	0.49
Racenative	0.61	0.17	3.57	0.00	1.83	1.30	2.54
Racepacific	-0.18	0.31	-0.59	0.55	0.83	0.44	1.47
Age	0.32	0.01	29.70	0.00	1.38	1.35	1.41

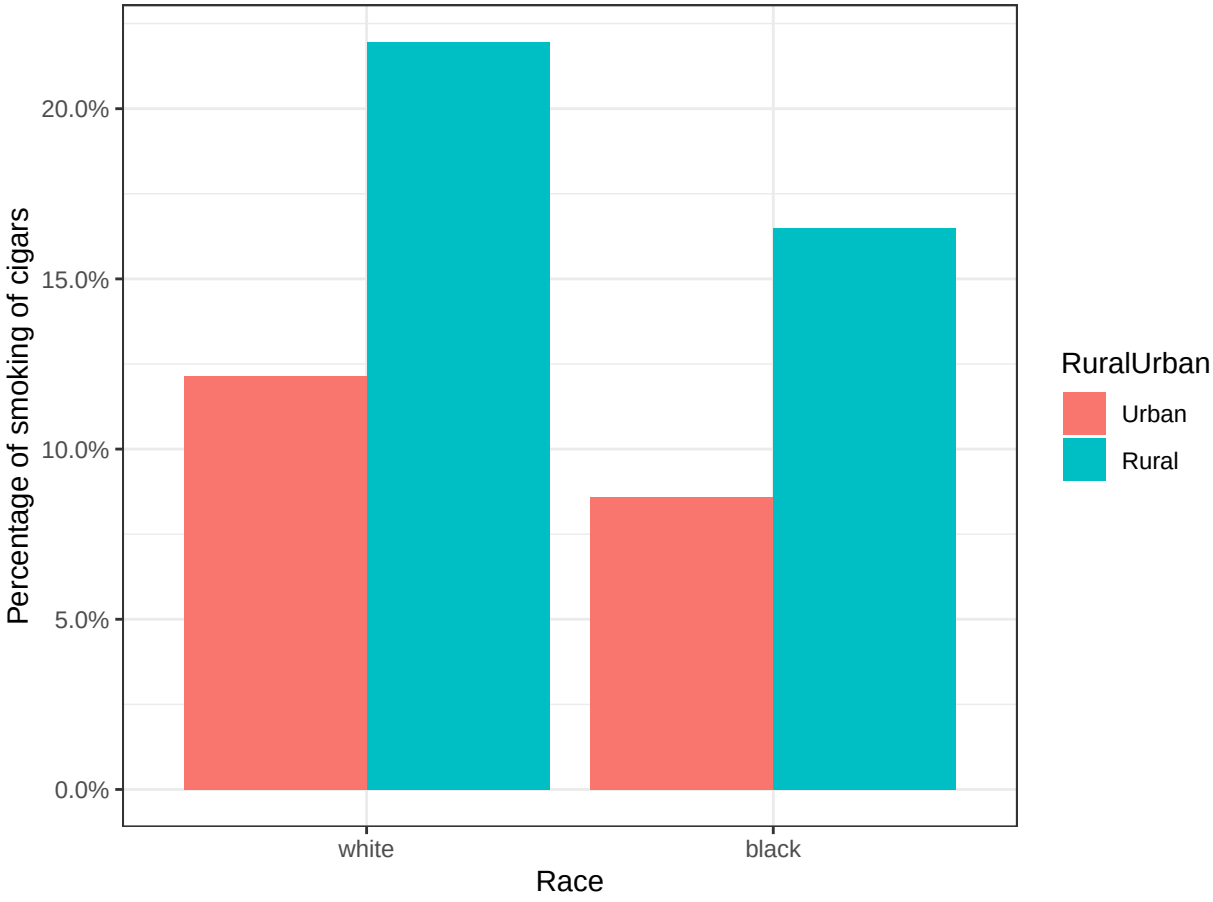

Figure 1: Cigar smoking among youth by race, rural

5% level. Figure 1 shows consistent results that black people smoked more cigars than white people whether in rural or urban areas.

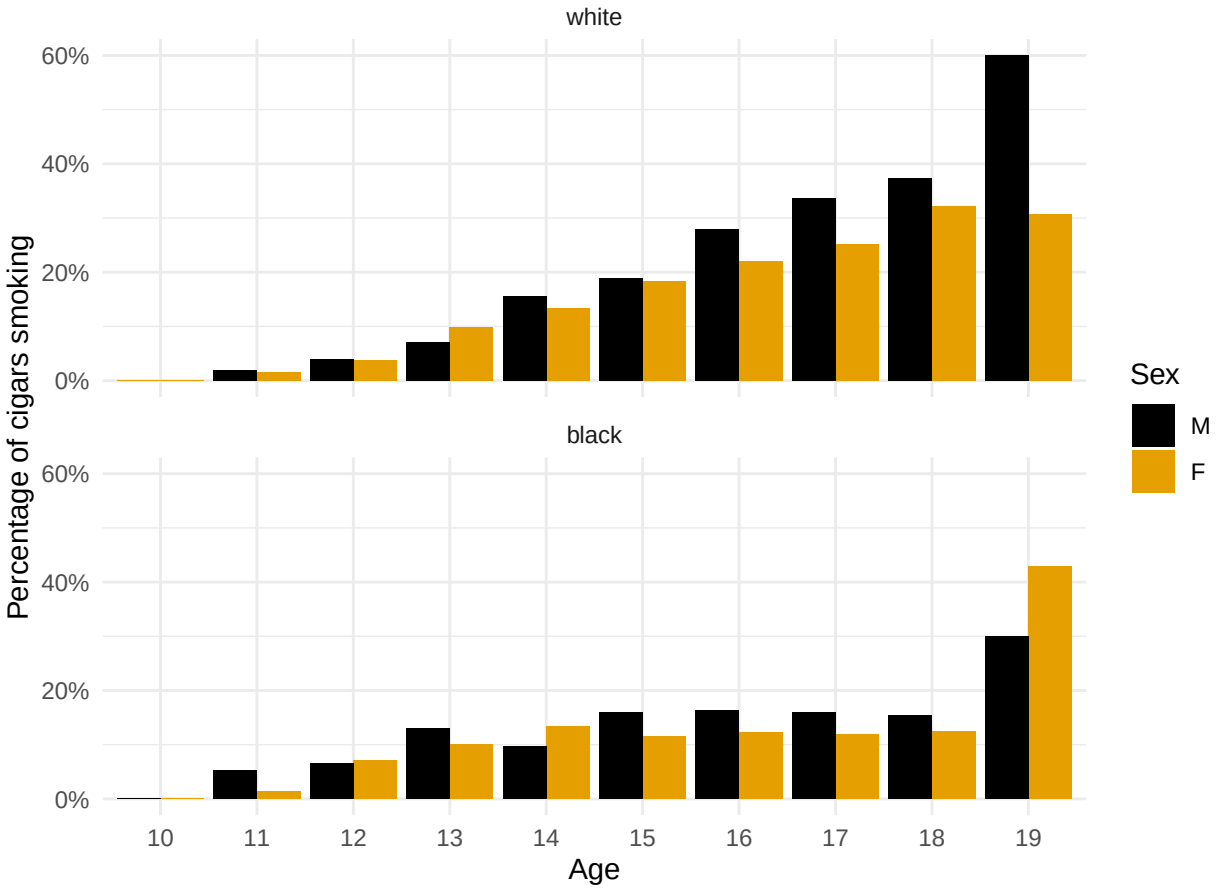

Figure 2: Cigar smoking among youth by race, gender and Age

Figure 2 shows the consistent results, the percentage of smoking. There was no significant difference in smoking behavior between men and women when their age and race were similar. This means that the possibility of using cigarettes is the same for both genders and they are similar in age and race.

Figure 3 shows the effect of age on smoking for the races - white, Black americans grouped by gender and race. There are some clear patterns that there are some differences among the behaviors of age 20 years old across the 4 groups.

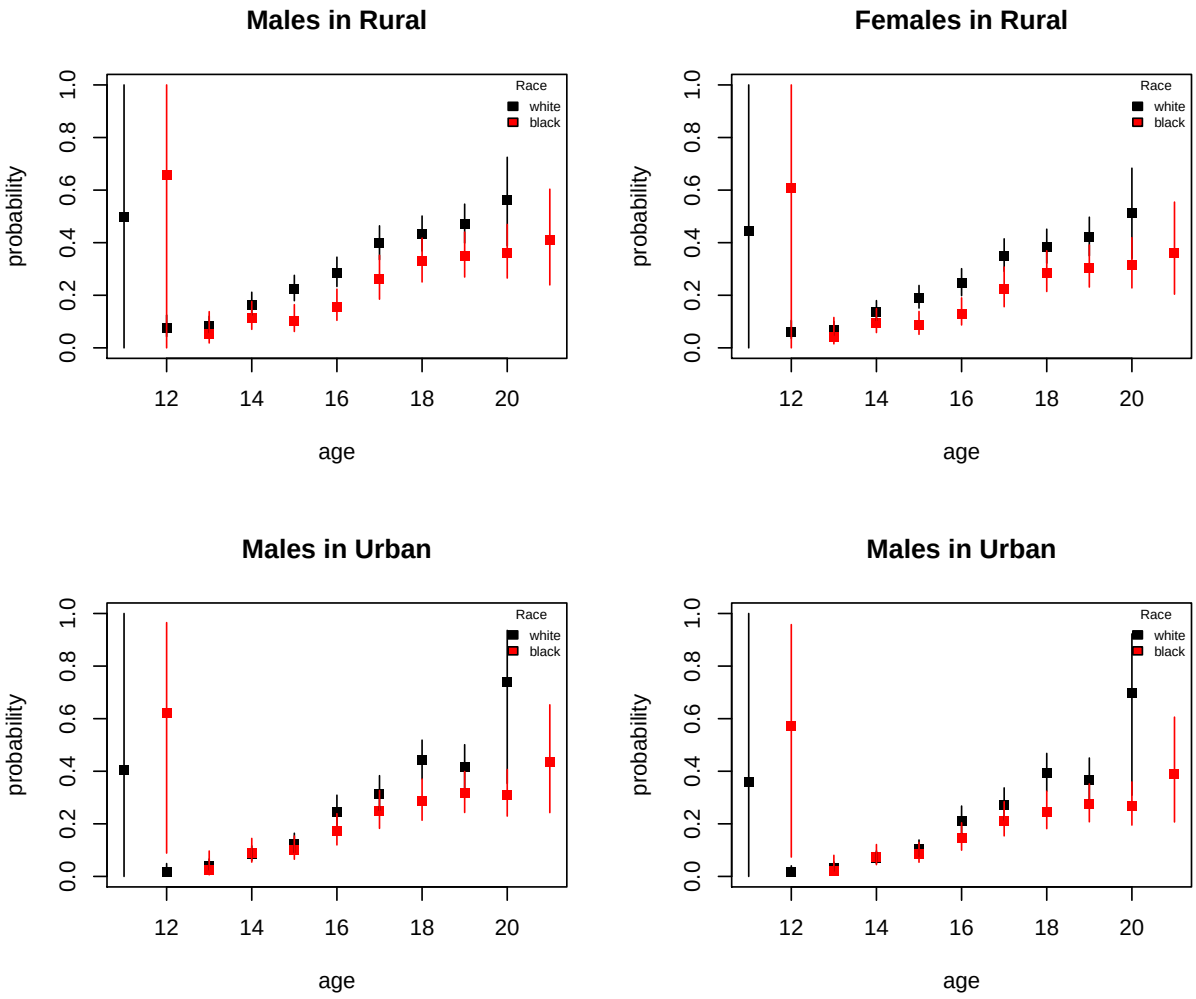

Figure 3: Age's effect on smoking grouped by gender and race

5 Discussion and conclusion

Table 1 shows the logistic model estimates of the possibility of cigar smoking. As it can be seen from the table, since the p values of the three factors are all less than 0.05, the three factors are all significant at the 5% level. Therefore, in addition to the factors of interest, living in rural or urban areas, age is an important confounding factor. It is estimated that the chance of smoking cigars increases by 41% with age.

Also, from the OR estimated, rural people are 2.07 times more likely to smoke cigars than urban people. The probability of black adolescents smoking cigars was 2.07 times higher than that of white adolescents, P value was less than 0.05, the difference was statistically significant. The probability of smoking cigars among Hispanic adolescents was 95% of that of white adolescents. Due to the p value of 0.08, which was greater than 0.05, there was no significant difference between Hispanic and white adolescents in smoking cigars.

Figure 1 shows consistent results that black people smoked more cigars than white people whether in rural or urban areas. Cigar smoking rates in cities are relatively close, so smoking is not just a rural phenomenon. So as the table 1 shows that cigar smoking is no more common among white youth than among white youth Hispanic Americans and African Americans. Although the proportion of cigar smoking in rural areas is higher, the proportion of cigar smoking in rural areas is higher.

Figure 2 shows the consistent results, the percentage of smoking. There was no significant difference in smoking behavior between men and women when their age and race were similar. This means that the possibility of using cigarettes is the same for both genders and they are similar in age and race.

Finally, as shown in Figure 3, the effect of age on smoking is really different for white, black Americans grouped by gender and rural, the results are also different. for example, black urban males are much more likely in smoking at the age of 20 compaed with white urban males and it is the same that black urban females are much more likely in smoking at the age of 20 compaed with white urban females. However, for ages under 12, there are unusual patterns which might be due to unusual data collected, but there are no differences among all of the 4 groups, so we can ignore the differences in this age group. At last, for other groups, we can find there are no obvious differences among probability of smoking for young people in different races, gender and locations.

Besides the findings of the study, there are also some weaknesses. First, the data used across different years, so the conclusions might not be appropriate to draw together which means the inferences on the 2014 data might not be appropriate for the year 2019. Second, the data sets are national wide survey based on face-to-face questionares, the responses might not be appropriate for all responsers as there might be differences due to different cultures of whites and blacks which leading to biasness of responses especially for blacks.

Also, as the sampling method is a clustering sampling method used in the survey, and in each of the clusters, students might be correlated with each other, for examples, students who are friends in same schools might

be affected by behaviors from each other, it indicates if one of these students began to smoke cigars, then the probabilities of their friends in smoking cigars would be much higher than normal classmate relationship. So this would also makes the results biased and not reliable.

Finally, in next steps, we can do lots of things to improve the results obtained in this study. For examples, in future work, we can include clustering errors in estmating the models based on relations of friends among the students. Also, we could add more covariates such as family income, students' performances(such as GPA) which might be also affect the behavior of smoking cigars to reduce the biased introduced by omitted variables.

6 References

1. Doll R. Cancers weakly related to smoking. Br Med Bull. 1996;52:35–49.
2. D. Kourounis, A. Fuchs, and O. Schenk, Towards the next generation of multiperiod optimal power flow solvers, IEEE Transactions on Power Systems, vol. PP, no. 99, pp. 1-10, 2018.
3. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
4. Hadley Wickham, Romain Franois, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
5. Jeffrey B. Arnold (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes
6. Murray J, Lopez AD. Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. Lancet. 1997;349:1498–504.
7. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
8. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2014. National Youth Tobacco Survey (NYTS) data.
9. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2014. National Youth Tobacco Survey (NYTS) data.
10. Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595