Analysis of the Use of Tobacco Products in Subgroups of American School Children

Stephen Zhen Gou
Department of Statistical Science
University of Toronto
Jan, 2018

Introduction

North America has seen a decrease in smoking activities among all groups in recent years. However, cigarette smoking is still the leading cause of death in the world. It kills 6 million people every year [1]. Thousands of young American children start to engage in smoking every year. The emergence of alternative forms of smoking, including chewing tobacco, snuff, dips, e-cigarette, hookah has created new challenges in preventing youth from smoking. Alternative forms of smoking provide an extra sense of coolness, and smokeless ones (chewing, snuff, dip) even give a false impression of being less harmful, but in some case containing more nicotine than non-smokeless tobacco.

Understanding groups of young chilren who are more active in tobacco use is crucial to create effective preventions. This report specifically examines two hyptoheses: 1) regular use of chewing tobacco, snuff or dip is no more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans, once one accounts for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon, as illustrated by Figure 1. 2) The likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes, provided their age, ethnicity, and other demographic characteristics are similar. In addition, the report analyzes how the use of chewing tobacco change with sex, age and ethnicity.

Methods

The analysis is based on the 2014 National Youth Tobacco Survey data, conducted by the Centers for Disease Control and Prevention. The survey contains the responses from 22,007 middle and high school students, representative of all 50 states.

Hypothesis I

The response variable Y is the regular use of chewing, snuff or dip tobacco. A student who smokes 3 days or more in a month is treated as a positive. The dependent variables are: $X_{isWhite}, X_{isRural}, X_{RuralWhite}$. Figure 1 and 2 show that both students from rural area and white students use chewing tobacco more frequently, but rural areas have more white students.To account for this, whether students are from rural areas and its interaction with being white are included in the model.

Since the response variable takes binary value, it is modelled by a generalized linear model with binomial family and logit link function, which is also known as logistic regression model.

$$Y_i \sim Bernoulli(p_i)$$

$$log(\frac{p_i}{1-p_i}) = X^T\beta = \beta_0 + \beta_1 X_{isWhite} + \beta_2 X_{isRural} + \beta_3 X_{Rural} X_{White}$$

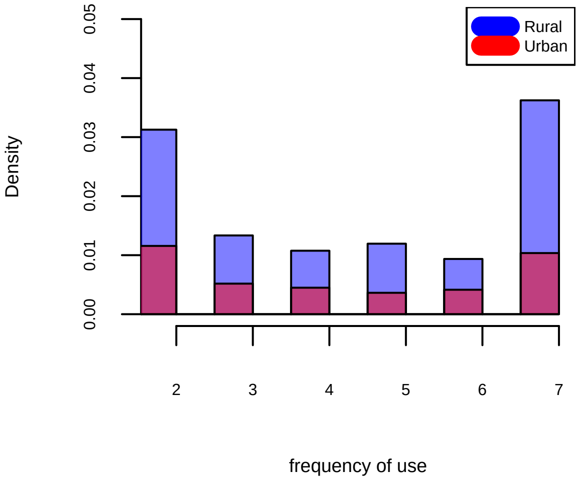

Figure 1: Frequency of using chew tobacco between urban to rural areas in a month.

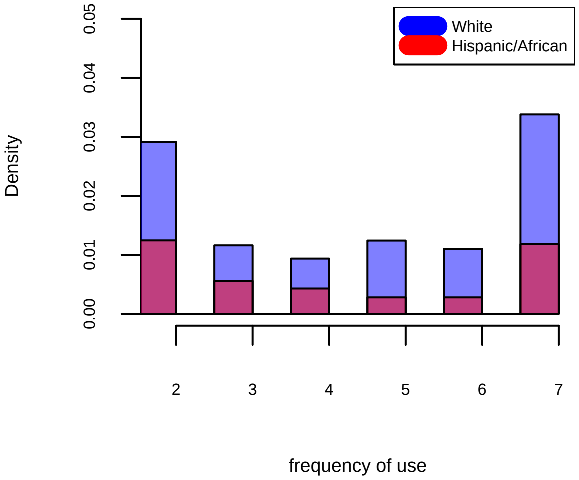

Figure 2: Frequency of using chew tobacco between white to hispanic/African American students in a month.

Hyptohesis II

The response variable Y is whether the student has ever used hookah or waterpipe at least once. The dependent variables included in the model are: $X_{isFemale}, X_{Age}, X_{isRural}, X_{RuralFemale}$. Age is centered. Whehter students living in rural areas and its interaction with sex are included to account for the effect of rural areas on smoking behaviour. Since the response variable takes binary value, it is modelled by a logistic regression model.

$$Y_i \sim Bernoulli(p_i)$$

$$log(\frac{p_i}{1-p_i}) = X^T\beta = \beta_0 + \beta_1 X_{isFemale} + \beta_2 X_{Age} + \beta_3 X_{isRural} + \beta_4 X_{Rural} X_{Female}$$

Effects of age, sex and ethnicity on the use of chewing tobacco

The response variable Y is the number of days the student uses chewing, snuff, and dip tobacco in a month. Sex, age (centered) and ethnicity group are the dependent variables.

Since the response variable has non-negative integer values, it is modelled by a generalized linear model with Poisson family and log link function as following:

$$Y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = X^T\beta = \beta_0 + \beta_1 X_{isFemale} + \beta_2 X_{Age} + \beta_3 X_{isBlack} + \beta_4 X_{isHispanic} + \beta_5 X_{isAsian} + \beta_6 X_{isNative} + \beta_7 X_{isPacific}$$

Results

Hypothesis I

Table 1: Model coefficients estimates for testing hypothesis I.

	Estimate	Std. Error	z value	Pr(>|z|)	exp(Estimate)
(Intercept)	-4.525	0.130	-34.864	0.000	0.011
is_whiteTRUE	0.544	0.172	3.172	0.002	1.723
is_ruralTRUE	0.526	0.179	2.937	0.003	1.692
is_whiteTRUE:is_ruralTRUE	0.632	0.219	2.879	0.004	1.880

The estimates for the model is shown in Table 1. The intercept is estimated at -4.525 with a significant p-value. It means that the base case (students Hispanic/African American students living in urban areas), has a 0.011 odds of being a regular user of chewing tobacco. In urban areas, white students has 0.544 higher log odds ratio with significant p-value (equivalent to 1.723 odds ratio), comparing to Hispanic/African American. In rural areas, the contrast between white and Hispanic/African American is represented by $\beta_{isWhite} + \beta_{isWhite:isRural} = 1.176$ (standard error 0.137) higher log odds ratio(equivalent to 3.24 odds ratio) than Hispanic/African American students. This result shows that white students are more likely to be regular users of chewing (snuff, dip) tobacco than Hispanic/African American students even when living area is accounted for. The hypothesis is thus rejected.

Hypothesis II

Table 2: Model coefficients estimates for testing hypothesis II.

	Estimate	Std. Error	z value	Pr(>|z|)	exp(Estimate)
(Intercept)	-7.6075	0.1729	-43.9924	0.0000	0.0005
SexF	0.1467	0.0538	2.7238	0.0065	1.1580
Age	0.3870	0.0108	35.7363	0.0000	1.4726
RuralUrbanRural	-0.2547	0.0582	-4.3743	0.0000	0.7752
SexF:RuralUrbanRural	-0.2502	0.0841	-2.9766	0.0029	0.7787

The estimates for the model is shown in Table 2. The intercept is estimated at -7.608 with a significant p-value, meaning that the base case (average aged female students living in urban areas), has a 0.0005 odds of having ever tried hookah or waterpipe. Age has a siginificant p-value as well, showing that every extra year of age gives a 0.387 higher log odds ratio (1.4726 higher odds ratio). In urban areas, female students has 0.1467 higher log odds ratio (equivalent to 1.158 odds ratio) with significant p-value , comparing to male students. In rural areas, the contrast between female and male is represented by $\beta_{isFemale} + \beta_{isFemale:isRural} = -0.104 \pm 0.13$ difference in log odds ratio (0.791 to 1.026 odds ratio with 95% confidence) than male students. It shows that sex has a sigfinicant effect on the odds of having ever tried hookah or waterpipe in rural areas, but not for urban areas.

Effects of age, sex and ethnicity on the use of chewing tobacco

Table 3: Model coefficients estimates for testing effects of age, sex and ethnicity.

	Estimate	Std. Error	z value	Pr(>|z|)	exp(Estimate)
(Intercept)	0.062	0.014	4.442	0.000	1.064
SexF	-2.227	0.034	-65.031	0.000	0.108
center_age	0.344	0.005	65.752	0.000	1.411
Raceblack	-1.730	0.047	-37.155	0.000	0.177
Racehispanic	-0.876	0.026	-33.559	0.000	0.417
Raceasian	-1.940	0.094	-20.652	0.000	0.144
Racenative	-0.236	0.082	-2.876	0.004	0.790
Racepacific	-0.020	0.117	-0.170	0.865	0.980

Resulting coefficiets estimates are shown in Table 3. The base case is an average age white male student, whose expected days of using chewing tobacco is 1.064 days in a month. A female student has a 0.108 ratio with significant p-value (89.2% less) expected days of use.Every year of extra age causes a 1.411 ratio with significant p-value (41.1% more) expected days of use. All non-white ethnic groups (except Pacific) has a less than 1 ratio with significant p-value of expected days of use of chewing tobacco.

Discussion

This report analyzed and compared the tendency of using of certain tobacco products between subgroups of American school children. The analysis is based on the *2014 National Youth Tobacco Survey* data [2]. The result shows that in urban areas, white students have 72.3% higher odds of engaging in regular use of chewing (including snuff and dip) tobacco comparing to Hispanic/African American students. In rural areas, the contrast is even more extreme, showing that white students have 324% higher odds. Interestingly, in urban areas, female students are more likely (15.8% higher odds) to have tried hookah or waterpipe than males, but no sifinicant evidence for students in rural areas. In general, sex, age and ethnicity have strong effects on the expected frequency of using chewing tobacco. For example, every extra year of age has 41.1% higher exptected frequency, while Asian students have 85.6% less.

Reference

1. Johanne Harvey, Nicholas Chadi; Canadian Paediatric Society, Adolescent Health Committee. Preventing smoking in children and adolescents: Recommendations for practice and policy. The Canadian Paediatric Society, 2016.

2. Office on Smoking and Health. 2014 National Youth Tobacco Survey: Methodology Report. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2015.

3. Brown, P. Methods of Applied Stats 2, Generalized Linear Models [PowerPoint presentation]. Retrieved from http://darjeeling.pbrown.ca/teaching/astwo/slides/glm.pdf, 2018.

Appendix

R code for modelling

```
#load data
load('smoke.RData')
data = smoke
data$regular_chew_use = (data$days_use_chew_tobacco_snu >= 3) * 1

#Testing Hypothesis I
data1 = data[data$Race == 'white' | data$Race == 'hispanic' |data$Race=='black',]
data1$is_white = as.factor(data1$Race == 'white')
data1$is_rural = as.factor(data1$RuralUrban == 'Rural')
model1 = glm(regular_chew_use ~ is_white + is_rural + is_white * is_rural, data = data1, family = binomi

#contrast of white rural vs hispanic rural
A = c(0,1,0,1)
c(est=crossprod(A, model1$coef), stderr=sqrt(crossprod(A, summary(model1)$cov.scaled) %*% A))

#testing hypothesis 2
data2 = data
data2 = data2[!is.na(data2$Sex),]
data2$is_white = as.factor(data2$Race == 'white')
data2$ever_hookah = as.numeric(data2$EVER_TRIED_hookah_waterpi ==1)
data2$ever_hookah[is.na(data2$ever_hookah)] = 0
data2 = data2[!is.na(data2$ever_hookah),]

model2 = glm(ever_hookah ~ Sex + Age + RuralUrban + Sex * RuralUrban, family = binomial(link = 'logit')

A2 = c(0,1,0,0,1)
c(est=crossprod(A2, model2$coef), stderr=sqrt(crossprod(A2, summary(model2)$cov.scaled) %*% A2))

#Model for effects of sex, age, ethnicity on expected frequency of using chewing tobacco.
```

```
data3 = data
data3$days_chew = 0
data3$center_age = data3$Age - 14.5
data3 = data3[!is.na(data3$days_use_chew_tobacco_snu),]
data3[data3$days_use_chew_tobacco_snu == 7,]$days_chew = 30
data3[data3$days_use_chew_tobacco_snu == 6,]$days_chew = 25
data3[data3$days_use_chew_tobacco_snu == 5,]$days_chew = 15
data3[data3$days_use_chew_tobacco_snu == 4,]$days_chew = 7
data3[data3$days_use_chew_tobacco_snu == 3,]$days_chew = 4
data3[data3$days_use_chew_tobacco_snu == 2,]$days_chew = 1
data3[data3$days_use_chew_tobacco_snu == 1,]$days_chew = 0

model3 = glm(days_chew ~ Sex + center_age + Race, family = poisson(link = 'log'),data=data3)
summary(model3)
```