hypothesis test

Summary

Two research hypotheses tested in this report are: (1) Regular use of chewing tobacco, snuff or dip is more common amongst Americans of European ancestry, than for Hispanic-Americans and African-Americans, accounting for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon. (2) The likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of different genders, if their age, ethnicity, and other demographic characteristics are similar. **Result**: Our models gave results that agreed with both hypotheses. Both hypotheses were tested through binomial regression model with logit link function. The first hypothesis test gave the result that comparing to 15-years-old male white Americans living at rural area, the Hispanic-Americans and African-Americans decreased the odds by 50.988% and 78.904% respectively to use chewing tobacco, snuff or dip. The second hypothesis test gave the result that a 15-years-old urban white female has a 4.3% increase in odds compare to male with same age, ethnicity, and other demographic characteristics, which was not significantly different at 0.05 significance level.

1. Introduction

Smoking is a major health concern that will lead to various preventable diseases. In this report we used the data from 2014 American National Youth Tobacco Survey to look in to how the use of tobacco varies among American Youth with different characteristics. There are two questions this report focuses: (1) Will regular use of chewing tobacco, snuff or dip is more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans, accounting for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon? (2) Does the likelihood of having used a hookah or waterpipe on at least one occasion vary with gender, if two individuals' age, ethnicity, and other demographic characteristics are similar?

2. Method

2.1 Dataset

The data is from 2014 American National Youth Tobacco Survey. There are 22007 observations with 162 variables. We focused on observations with age over ten since the data for the nine-year-old were suspiciously many as seen in table 1.

Table 1: Age and Cigarette Times

	Once	Twice	None
9	26	6	8
10	3	7	2
11	47	1260	20
12	248	2776	58
13	424	3120	73

After excluding the nine-year-old and removing missing values for variable of interest, there were 21967 observations with 162 variates.

For the first hypothesis, Figure 1 gave some indication of white rural male is having a higher proportion to adapt a regular use of chewing tobacco, snuff or dip.

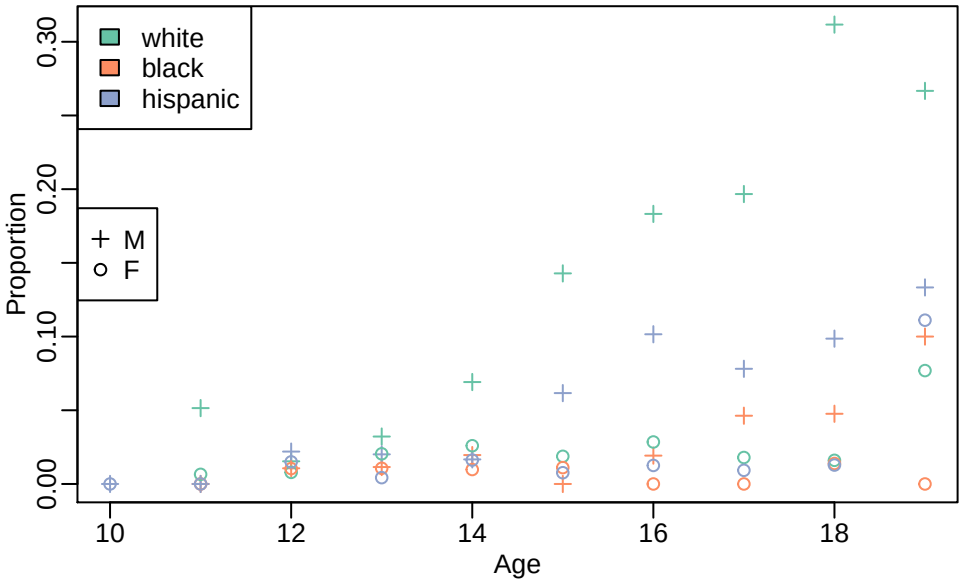

Figure 1: Proportion for regular use of chewing tobacco, snuff or dip for Rural Americans

For the second hypothesis, Figure 2 gave a rough idea that for white urban Americans below eighteen years old, there is not a lot of difference between the likelihood of having used a hookah or waterpipe on at least one occasion. However for those over eighteen years old, some more investigation is needed. Thus for both tests, when building the model we took age, gender, ethnicity, and other demographic characteristics into consideration.

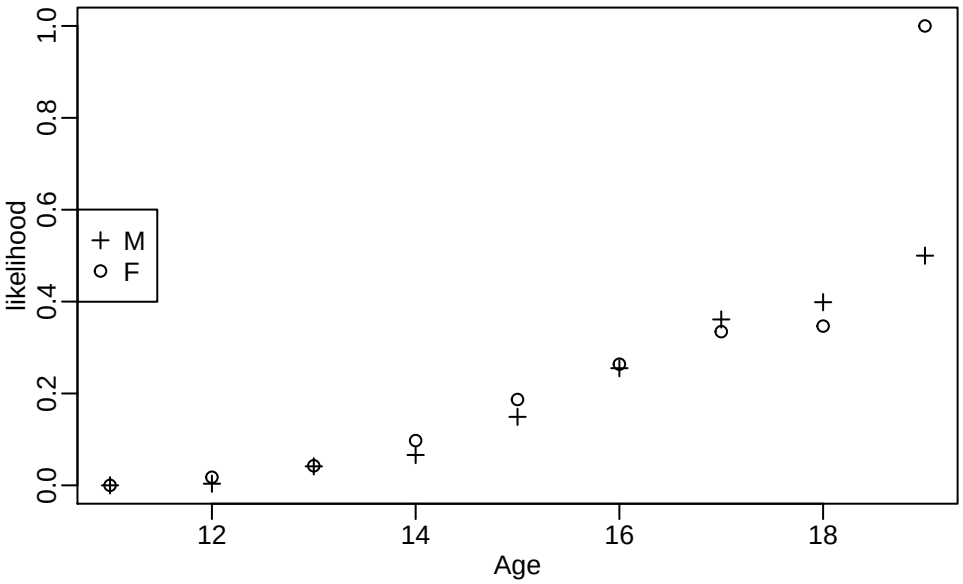

Figure 2: Likelihood of having used a hookah or waterpipe on at least one occasion for White urban Americans

2.2 Models

It was natural to consider a binomial regression for the both hypotheses because both had a two-factor response (0 or 1). The main effects are ethnicity, demographic characteristics (rural or urban) age and sex. Our model for both is the binomial regression with logit link function:

$$Y_i \sim Binomial(\mu_i)$$

$$\log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

Where Y_i is the response of interest, and X_i are the predictor variables(Race, Age, Sex and demographic characteristics). μ_i indicates the probability of the response.

3. Results

3.1 Testing Hypothesis one

The null hypothesis is that $\beta_0 = \beta_1 = \beta_2$, where β_i are the coefficients for white, black and hipanic Americans living in rural area in the model respectively. The baseline is set to be white male Americans living in rural area at age 15. As shown in Table 2, the coefficient is significantly different from the coefficient of Hispanic-Americans and African-Americans living in rural area. Thus we rejected the null hypothesis.

Table 2: Parameter estimates for the binomial regression

	Estimate	Std. Error	p-value	lower	upper
Intercept	-2.081	0.057	-2.195	-1.967	<0.05
age	0.337	0.021	0.295	0.378	<0.05
female	-1.788	0.109	-2.005	-1.571	<0.05
black	-1.556	0.172	-1.899	-1.213	<0.05
hispanic	-0.713	0.104	-0.920	-0.506	<0.05
Urban	-0.951	0.087	-1.126	-0.776	<0.05

Based on Table 3, the baseline probability for rural white American to regularly use of chewing tobacco, snuff or dip is 11.1%.

The Hispanic-Americans and African-Americans living in rural area decreased the odds by 50.988% and 78.904% respectively to use chewing tobacco, snuff or dip, if with same age and gender compared with baseline.

Table 3: MLE's of baseline odds and odds ratios, with 95 conidence intervals.

	Estimate	lower	upper
baseline odds	0.125	0.111	0.140
age	1.400	1.343	1.459
female	0.167	0.135	0.208
black	0.211	0.150	0.297
hispanic	0.490	0.398	0.603
urban	0.386	0.324	0.460

The age and gender are cofounders. One age older will lead to a 40.011% increase in odds and females decrease the odds by 83.277%

Thus, we concluded that regular use of chewing tobacco, snuff or dip is more common amongst Americans of European ancestry, than for Hispanic-Americans and African-Americans, accounting for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon.

3.2 Testing Hypothesis two

The null hypothesis is that $\beta_0 = \beta_1$, where β_i are the coefficients for male and female in the model respectively. The baseline is set to be white male at age of fifteen, living in urban area. As shown in Table 4, the coefficient is not significantly different between male and female. Thus we did not reject the null hypothesis, which means that there is no significant difference between the two genders in case of the likelihood of having used a hookah or waterpipe on at least one occasion.

Table 4: Parameter estimates for the binomial regression

	Estimate	Std. Error	lower	upper	p-value
intercept	-1.724	0.044	-1.811	-1.636	<0.05
age	0.419	0.012	0.396	0.442	<0.05
female	0.042	0.043	-0.044	0.128	>0.05
black	-0.635	0.070	-0.776	-0.494	<0.05
hispanic	0.346	0.048	0.249	0.442	<0.05
asian	-0.631	0.118	-0.866	-0.396	<0.05
native	0.160	0.190	-0.221	0.540	>0.05
pacific	0.964	0.270	0.423	1.504	<0.05
Rural	-0.388	0.044	-0.477	-0.300	<0.05

Based on Table 5, the probability for 15 years old white male to have used a hookah or waterpipe on at least one occasion is 0.151. A 15-years-old urban white female has a 4.3% increase in odds compare to male with same age, ethnicity, and other demographic characteristics, which was not significantly different at 0.05 significance level.

Table 5: MLE's of baseline odds and odds ratios, with 95 conidence intervals.

	Estimate	lower	upper
baseline odds	0.178	0.163	0.195
age	1.520	1.485	1.555
female	1.043	0.957	1.136
black	0.530	0.460	0.610
hispanic	1.413	1.282	1.557
asian	0.532	0.421	0.673
native	1.173	0.802	1.717
pacific	2.621	1.527	4.500
Rural	0.678	0.621	0.741

Thus, we concluded that the likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of different genders, if their age, ethnicity, and other demographic characteristics are similar.

Appendix

```
## ----import, include=FALSE-----------------------------------------------
library(dplyr)
library(kableExtra)
knitr::opts_chunk$set(echo = TRUE)
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext='.RData'))
download.file(smokeUrl, smokeFile)
(load(smokeFile))
dim(smoke)
smokeSub = smoke[smoke$Age >= 10, ]
# exlcude NA
smokeAgg = smokeSub[complete.cases("chewing_tobacco_snuff_or", "Race","sex","age","RuralUrban"),]
# set Rural as baseline
smokeAgg$RuralUrban <- factor(smokeAgg$RuralUrban, levels=c("Rural","Urban"))
# centerize age at 15
smokeAgg$ageC = smokeAgg$Age - 15
smokeFit = glm(chewing_tobacco_snuff_or ~ ageC + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg)

# coefficient table
parTable =  data.frame(summary(smokeFit)$coef[c(1,2,3,4,5,9),c(1,2,4)])
parTable$lower = parTable[,1] - 2*parTable[,2]
parTable$upper = parTable[,1] + 2*parTable[,2]
parTable[,c(1,2,4,5)] = round(parTable[,c(1,2,4,5)],3)
parTable$p_val = ifelse(parTable[,3]<0.05,"<0.05",">0.05")
parTable = parTable[,c(1,2,4,5,6)]
rownames(parTable) = c("Intercept","age","female","black","hispanic","Urban")
colnames(parTable) = c( "Estimate", "Std. Error", "p-value","lower","upper")

# OddsRatio table
smokeTable1 = as.data.frame(summary(smokeFit)$coef)
smokeTable1$lower = smokeTable1$Estimate - 2*smokeTable1$'Std. Error'
smokeTable1$upper = smokeTable1$Estimate + 2*smokeTable1$'Std. Error'
smokeOddsRatio1 = exp(smokeTable1[,c('Estimate','lower','upper')])
# focus only on interested variables
smokeOddsRatio1 = smokeOddsRatio1[c(1,2,3,4,5,9),]
rownames(smokeOddsRatio1) = c('baseline odds', 'age', 'female','black','hispanic','urban')
#smokeOddsRatio1[1,] = smokeOddsRatio1[1,]/(1+smokeOddsRatio1[,1]) # probability


## ----exclude9, echo=F----------------------------------------------------
table = table(smoke$Age, smoke$Tried_cigarette_smkg_even,exclude =NULL)[1:5,]
colnames(table) = c("Once","Twice","None")
knitr::kable(table, caption = "Age and Cigarette Times", booktabs = TRUE,longtable = T)


## ----hypoplot,message=F, echo=F, fig.height=3, fig.width=5, fig.cap='Proportion for regular use of ch
smokeplotAgg = reshape2::dcast(smokeSub,
                               Age + Sex + Race + RuralUrban ~ chewing_tobacco_snuff_or,length)
smokeplotAgg = smokeplotAgg[complete.cases("chewing_tobacco_snuff_or", "Race","Sex","Age","RuralUrban")
smokeplotAgg = smokeplotAgg[which(smokeplotAgg$Race == 'white'|smokeplotAgg$Race == 'black'|smokeplotAg
```

```r
smokeplotAgg$Race = factor(smokeplotAgg$Race)
smokeplotAgg$total = smokeplotAgg$"TRUE" + smokeplotAgg$"FALSE"
smokeplotAgg$prop = smokeplotAgg$"TRUE" / smokeplotAgg$total
smokeplotAgg = smokeplotAgg[which(smokeplotAgg$RuralUrban == "Rural" & smokeplotAgg$prop<1),]
#'
#'
#+ smokeExplPlot
Spch = c('M' = 3, 'F'=1)
Scol = RColorBrewer::brewer.pal(nlevels(smokeplotAgg$Race), 'Set2')
names(Scol) = levels(smokeplotAgg$Race)
par(mar=c(2.5,2.5,0.1,0.1),
mgp=c(1.5, 0.5, 0), cex=0.8)
plot(smokeplotAgg$Age, smokeplotAgg$prop, pch = Spch[as.character(smokeplotAgg$Sex)],
     col = Scol[as.character(smokeplotAgg$Race)],
     xlab = "Age",
     ylab = "Proportion")
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))
```

```r
## ----hypoplot2,message=F, echo=F, fig.height=3, fig.width=5, fig.cap="Likelihood of having used a hoo
smokeplotAgg = reshape2::dcast(smokeSub,
                               Age + Sex + Race + RuralUrban ~ ever_tobacco_hookah_or_wa,length)
smokeplotAgg = smokeplotAgg[complete.cases("ever_tobacco_hookah_or_wa", "Race","Sex","Age","RuralUrban"
smokeplotAgg$total = smokeplotAgg$"TRUE" + smokeplotAgg$"FALSE"
smokeplotAgg$prop = smokeplotAgg$"TRUE" / smokeplotAgg$total
smokeplotAgg = smokeplotAgg[which(smokeplotAgg$RuralUrban == "Urban" & smokeplotAgg$Race=="white"),]
#'
#'
#+ smokeExplPlot
Spch = c('M' = 3, 'F'=1)
par(mar=c(2.5,2.5,0.1,0.1),
mgp=c(1.5, 0.5, 0), cex=0.8)
plot(smokeplotAgg$Age, smokeplotAgg$prop, pch = Spch[as.character(smokeplotAgg$Sex)],
     xlab = "Age",
     ylab = "likelihood")
legend('left', pch=Spch, legend=names(Spch))
```

```r
## ----Q2cof1,results='markup',echo=FALSE----------------------------------
kable(parTable, caption = "Parameter estimates for the binomial regression", booktabs = TRUE,longtable =


## ----Q2oddsratio1, results='markup',echo=FALSE---------------------------
knitr::kable(smokeOddsRatio1, digits=3, caption = " MLE's of baseline odds and odds ratios, with 95 con


## ----echo=F--------------------------------------------------------------
smokeAgg2 = smokeSub[complete.cases("ever_tobacco_hookah_or_wa", "Race","RuralUrban", "Age","Sex"),]
smokeAgg2$ageC = smokeAgg2$Age - 15

smokeFit2 = glm(ever_tobacco_hookah_or_wa ~ ageC + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg2)


# coefficient table
```

```r
parTable2 =  data.frame(summary(smokeFit2)$coef)
parTable2$lower = parTable2[,1] - 2*parTable2[,2]
parTable2$upper = parTable2[,1] + 2*parTable2[,2]
parTable2[,c(1,2,5,6)] = round(parTable2[,c(1,2,5,6)],3)
parTable2$p_val = ifelse(parTable2[,4]<0.05,"<0.05",">0.05")
parTable2 = parTable2[,c(1,2,5,6,7)]
rownames(parTable2) = c("intercept","age","female","black","hispanic","asian",
                        "native","pacific","Rural")
colnames(parTable2) = c( "Estimate", "Std. Error","lower","upper","p-value")


smokeTable2 = as.data.frame(summary(smokeFit2)$coef)
smokeTable2$lower = smokeTable2$Estimate - 2*smokeTable2$'Std. Error'
smokeTable2$upper = smokeTable2$Estimate + 2*smokeTable2$'Std. Error'
smokeOddsRatio2 = exp(smokeTable2[,c('Estimate','lower','upper')])
rownames(smokeOddsRatio2) = c('baseline odds',"age","female","black","hispanic","asian","native","pacifi
#smokeOddsRatio2[1,]/(1+smokeOddsRatio2[,1]) # probability


## ----Q2cof2,results='markup',echo=FALSE-----------------------------------
kable(parTable2, caption = "Parameter estimates for the binomial regression",
      booktabs = TRUE,longtable = T)


## ----Q2oddsratio2, results='markup',echo=FALSE----------------------------
knitr::kable(smokeOddsRatio2, digits=3, caption = " MLE's of baseline odds and odds ratios, with 95 con
```