

STA442 Assignment 1

Yingfei Zha

06/10/2020

1 Affairs

Report

To investigate the factors that affect the chance of an individual having affairs, data from a sample of 600 married people in America are analyzed with regard to gender, age, whether or not having children, the number of years married and religiousness of the individual concerned. It is concerned that having children makes a man more likely to have affairs while making a woman less likely to have affairs. A logistic regression model is fitted to study the joint relationship between the chance of having affairs and all variables. It provides the result of odds ratio for each variable.

The statistical results are shown in **Table 1**. Since we are interested in the effect of having children on men and women, let us look at the other variables first. Beginning with the explanatory variable age. As one's age increases by 1 year, the odds significantly decrease by 3.57% ($p = 0.042$). Number of years married also influence the odds of having affairs significantly. As number of years married increases by 1 year, the odds of having affairs increase by 11.1% ($p = 0.001$). Individuals at different religious involvement levels behave differently and two of these levels show significant effect on them. For those who are anti-religious, they have odds of having affairs 2.02 times higher than individuals with no religious ($p = 0.048$). Individuals who involve at a medium level have odds 0.48 times the odds of individuals with no religious ($p = 0.008$). For the last interaction term, it shows that there is a interaction relationship between the effect of having children of men the effect of having children of women. Since the odds ratio is less than 1, the effect of having children of men is smaller than the effect of having children of women. However, since the result is insignificant, there is no difference between men and women under the effect of having children.

In addition, let us look at the effect of gender and having children on the odds of having affairs. Comparing men who have children to women who also have children, the odds of having affairs increases by 30.2% (Odds Ratio = 1.30). Comparing men who have no children to women who have no children, the odds of having affairs increase by 94.3% (Odds Ratio = 1.94). Since the results are insignificant, there is no significant difference between the chance of men and women having affairs. To contrast the effect of having children with not having children on men and women respectively, becoming a father increases the odds of having affairs by 28.7% comparing to men not having children (Odds Ratio = 1.29). Becoming a mother increases the odds of having affairs by 92.2% comparing to women not having children (Odds Ratio = 1.92). However, both genders' contrast results are insignificant, there is no significant difference between the odds of having affairs with or without children.

Table 1: Estimated Parameters from Logistic Regression Model of Affairs Data

	Estimate	Std. Error	Z-value	Pr(> z)	Odds Ratio	Lower (CI)	Upper (CI)
(Intercept)	-2.521	0.398	-6.327	0.000	0.080	0.037	0.175
gendermale	0.664	0.436	1.523	0.128	1.943	0.826	4.571
childrenyes	0.653	0.388	1.683	0.092	1.922	0.898	4.113

	Estimate	Std. Error	Z-value	Pr(> z)	Odds Ratio	Lower (CI)	Upper (CI)
ageNew	-0.036	0.018	-2.030	0.042	0.964	0.931	0.999
yearsmarried	0.105	0.032	3.276	0.001	1.111	1.043	1.183
religiousanti	0.706	0.356	1.981	0.048	2.025	1.008	4.070
religiouslow	0.275	0.269	1.024	0.306	1.317	0.778	2.230
religiousmed	-0.728	0.274	-2.654	0.008	0.483	0.282	0.827
religioushigh	-0.664	0.371	-1.790	0.074	0.515	0.249	1.065
gendermale:childrenyes	-0.401	0.488	-0.820	0.412	0.670	0.257	1.745

Summary

The study showed that becoming parents will not make a woman less likely or a man more likely to have an affair. It is generally believed that becoming a mother will reduce the chance for a woman to have extramarital sex; however, data showed that women chance of having affairs increases by 92.2% after having children and there is not enough evidence saying that becoming a mother makes the chance significantly different. The situation is similar to men. Data showed that the chance of having affairs is increased by 28.7% after a man becoming a father, but there is no significant difference between the chances before or after having children. However, it is not so surprising that the more years the couples marry, the more likelihood to have affairs.

2 Smoke

Summary

In this study, we investigate the smoking behaviors among youth using a sample from 2019 American National Youth Tobacco Survey. As the result of the first analysis with regard to the smoking of cigars, cigarillos or little cigars, European Americans are 40.1% and 30.4% less likely to smoke than African Americans in rural and urban areas respectively; whereas they are 21.1% more common than Hispanic Americans in rural areas. In addition to this, Pacific Americans are most likely and Asian Americans are least likely to smoke cigars among all races. Besides, older men who live in rural area are more likely to smoke cigars.

The result of the second analysis with regard to the use of electronic cigarettes shows that it has no statistically difference among male and female. On the other hand, older individuals living in rural are more likely to use e-cigarette. Hispanic, African and Asian Americans are less likely to use e-cigarette comparing to European Americans.

Introduction

In this study, data from a nationally representative sample of over 18000 American school children (2019 American National Youth Tobacco Survey) are analyzed with regard to the use of cigars, e-cigarette, and demographic characteristics. First, we investigate whether Hispanic-Americans and African-Americans have higher odds of smoking cigars, cigarillos, or little cigars than European Americans. Second, we are interested in if the odds of using e-cigarette are the same for males and females when keeping other demographic characteristics similar. For both of these analysis, individual's age and living areas (rural or urban) are taken into account.

Methods

To investigate the above hypotheses, the logistics regression model is chosen as a conventional method to model data with dichotomous response variables. We have two logistics regression models fitted for each of the aforementioned hypothesis. The relationship between the response variable and explanatory variables is given by:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

where π indicates the probability when the response variable is 1 (i.e. the response is “yes”) and $\frac{\pi}{1-\pi}$ represents the odds of the event of interest. The explanatory variables are represented by x_i .

The first model predicts the odds that individuals have ever tried smoking cigars, cigarillos, or little cigars. The predictors are individual's sex, age, race and living areas (rural or urban). We also add an interaction term of race and living areas since it is concerned that cigar smoking is more common in rural areas where European Americans are more likely to live in. Ultimately, we contrast Hispanic-Americans, African-Americans with European Americans with regard to the odds of smoking cigars. The second model predicts the odds that individuals have ever used an e-cigarette. Since we are interested in the behaviors of different sexes, the predictors are chosen to be individual's sex, age, race and living areas (rural or urban). The odds of using e-cigarette are contrasted between different sexes while keeping other variables the same. As a rule of thumb, an alpha of 0.05 is used as the cutoff for significance.

Results

1. The statistical results to analyze the first hypothesis are shown in **Table 2**. Sex, age, living areas and some races all show significant influence on the odds of smoking cigars, cigarillos, or little cigars. Beginning with the explanatory variable sex, the odds that women has ever smoked cigars is 0.685 times

the odds that men has ever smoked cigars while keeping other variables the same. As individual's age increases by 1 year, the odds significantly increase by 45.4%. Individuals living in rural areas increase the odds of smoking cigars by 60.2% comparing to those living in urban areas. Individuals of different races have different odds of smoking cigars comparing to European Americans. African Americans, Asian Americans and Pacific Americans shows significant results. The odds of smoking cigars of African Americans are 1.44 times the odds of smoking cigars of European Americans. The odds of Asian Americans are 0.328 times the odds of European Americans. The odds of Pacific Americans are 2.07 times the odds of smoking cigars of European Americans. In the interaction terms, only one has a significant coefficient. It suggests that there is an interaction relationship between the effect of living in rural areas and the effect of being Hispanic Americans.

To answer the first hypothesis, we compare male European Americans who live in rural areas and at age 10 to male African Americans who live in rural areas and at age 10. The result is significant and suggests that the odds of European Americans smoking cigars decrease by 40.1% comparing to the odds of African American smoking cigars at rural areas (Odds Ratio = 0.599). The result is similar at urban areas. The odds of European Americans smoking cigars decrease by 30.4% comparing to the odds of African American smoking cigars at urban areas (Odds Ratio = 0.696). We also compare male European Americans who live in rural areas and at age 10 to male Hispanic Americans who live in rural areas and at age 10. The result is significant as well and the odds of European Americans smoking cigars increase by 21.1% comparing to the odds of Hispanic American smoking cigars at rural areas (Odds Ratio = 1.21). The odds decrease by 7.76% at urban areas, but there is no enough evidence saying there is a significant difference between these two racial groups when living in urban areas.

Table 2: Estimated Parameters from the First Model to Analyze the Usage of Cigars, Cigarillos, or Little Cigars

	Estimate	Std. Error	Z-value	Pr(> z)	Odds Ratio	Lower (CI)	Upper (CI)
(Intercept)	-7.592	0.195	-38.972	0.000	0.001	0.000	0.001
SexF	-0.379	0.046	-8.258	0.000	0.685	0.626	0.749
Age	0.374	0.012	31.317	0.000	1.454	1.420	1.488
Raceblack	0.362	0.094	3.858	0.000	1.436	1.195	1.726
Racehispanic	0.081	0.076	1.060	0.289	1.084	0.934	1.259
Raceasian	-1.116	0.195	-5.717	0.000	0.328	0.224	0.480
Racenative	0.578	0.326	1.772	0.076	1.783	0.941	3.378
Racepacific	0.727	0.351	2.074	0.038	2.070	1.041	4.117
RuralUrbanRural	0.471	0.064	7.346	0.000	1.602	1.413	1.817
Raceblack:RuralUrbanRural	0.151	0.129	1.172	0.241	1.163	0.904	1.496
Racehispanic:RuralUrbanRural	-0.272	0.107	-2.551	0.011	0.762	0.618	0.939
Raceasian:RuralUrbanRural	-0.503	0.417	-1.205	0.228	0.605	0.267	1.370
Racenative:RuralUrbanRural	-0.467	0.418	-1.117	0.264	0.627	0.276	1.423
Racepacific:RuralUrbanRural	-0.701	0.583	-1.203	0.229	0.496	0.158	1.555

2. The statistical results to analyze the second hypothesis are shown in **Table 3**. Age, living areas and some races show significant influence on the odds of using e-cigarette while sex does not show significant effects. The odds that women has ever used e-cigarette is 0.942 times the odds that men has ever used e-cigarette while keeping other variables the same. However, the result is insignificant so it indicates that there is no significant difference between the odds of using e-cigarette of two sexes.

As for the other influential factors, the odds significantly increase by 40.0% as individual's age increases by 1 year. Individuals living in rural areas increase the odds of using e-cigarette by 14.0% comparing to those living in urban areas. Among the five racial groups that compare to European Americans, African Americans, Hispanic Americans and Asian Americans shows significant difference in the odds of using e-cigarette. The odds of using e-cigarette of African Americans are 0.598 times the odds of European Americans. The odds of Hispanic Americans are 0.915 times the odds of European Americans. The

odds of Asian Americans are 0.364 times the odds of using e-cigarette of European Americans.

Table 3: Estimated Parameters from the Second Model to Analyze the Usage of E-cigarette

	Estimate	Std. Error	Z-value	Pr(> z)	Odds Ratio	Lower (CI)	Upper (CI)
(Intercept)	-5.526	0.130	-42.364	0.000	0.004	0.003	0.005
SexF	-0.060	0.033	-1.823	0.068	0.942	0.883	1.005
Age	0.337	0.008	39.901	0.000	1.400	1.377	1.424
RuralUrbanRural	0.131	0.033	3.930	0.000	1.140	1.068	1.217
Raceblack	-0.515	0.054	-9.614	0.000	0.598	0.538	0.664
Racehispanic	-0.089	0.038	-2.369	0.018	0.915	0.849	0.985
Raceasian	-1.009	0.092	-11.023	0.000	0.364	0.305	0.436
Racenative	0.062	0.153	0.406	0.685	1.064	0.788	1.436
Racepacific	0.237	0.215	1.102	0.270	1.268	0.832	1.932

Appendix

```
#####Set Up#####
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

#####Affairs Data#####
data('Affairs', package='AER')
Affairs$ever = Affairs$affair > 0
Affairs$religious = factor(Affairs$religiousness,
levels = c(2,1,3,4,5), labels = c('no','anti','low','med','high'))
Affairs$ageNew = Affairs$age - mean(Affairs$age) #Center age
attach(Affairs)
fit = glm(ever ~ gender*children + ageNew + yearsmarried + religious, data=Affairs, family="binomial")
summary(fit)

#####Result Table for Affairs#####
theCiMat = Pmisc::ciMat(0.95)
parTable = summary(fit)$coef[,rownames(theCiMat)]%*%theCiMat
rownames(parTable)[1]="Baseline"
x = cbind(summary(fit)$coef,exp(parTable))
knitr::kable(x, digits=3, col.names = c('Estimate', 'Std. Error', 'Z-value', 'Pr(>|z|)',
'Odds Ratio', 'Lower (CI)', 'Upper (CI)'),
caption="Estimated Parameters from Logistic Regression Model of Affairs Data ")

#####Get Age Range#####
max(ageNew)
min(ageNew)

#####Contrast#####
contrast::contrast(fit,
list(gender='male', children='yes', ageNew=1, religious='low', yearsmarried=10),
list(gender='female', children='yes', ageNew=1, religious='low', yearsmarried=10))
exp(0.263872) #1.301962

#same as gendermale coef
contrast::contrast(fit,
list(gender='male', children='no', ageNew=1, religious='low', yearsmarried=10),
list(gender='female', children='no', ageNew=1, religious='low', yearsmarried=10))

contrast::contrast(fit,
list(gender='male', children='yes', ageNew=1, religious='low', yearsmarried=10),
list(gender='male', children='no', ageNew=1, religious='low', yearsmarried=10))
exp(0.2525844) #1.287348

#same as childrenyes coef
contrast::contrast(fit,
list(gender='female', children='yes', ageNew=1, religious='low', yearsmarried=10),
list(gender='female', children='no', ageNew=1, religious='low', yearsmarried=10))

#####Smoke Data#####
dataDir = "~/Downloads"
smokeFile = file.path(dataDir, "smokeDownload.RData")
if (!file.exists(smokeFile)) {
```

```

download.file("http://pbrown.ca/teaching/appliedstats/data/smoke.RData", smokeFile) }
(load(smokeFile))
# get rid of 9 year olds because their data is suspicious
smokeSub = smoke[which(smoke$Age >= 10), ]
attach(smokeSub)

#####Smoke Table for Hypothesis 1#####
mod_1 = glm(ever_cigars_cigarillos_or ~ Sex + Age + Race * RuralUrban , family=binomial, data=smokeSub)
theCiMat = Pmisc::ciMat(0.95)
parTable = summary(mod_1)$coef[,rownames(theCiMat)]%*%theCiMat
rownames(parTable)[1]="Baseline"
y = cbind(summary(mod_1)$coef,exp(parTable))
knitr::kable(y, digits=3, col.names = c('Estimate', 'Std. Error', 'Z-value', 'Pr(>|z|)',
                                         'Odds Ratio', 'Lower (CI)', 'Upper (CI)'),
             caption="Estimated Parameters from the First Model to Analyze the Usage of Cigars, Cigaril.

#####Contrast#####
contrast::contrast(mod_1,
  list(RuralUrban='Rural', Race='white', Age=10, Sex='M'),
  list(RuralUrban='Rural', Race='black', Age=10, Sex='M'))
exp(-0.5125319) #0.5989771

contrast::contrast(mod_1,
  list(RuralUrban='Urban', Race='white', Age=10, Sex='M'),
  list(RuralUrban='Urban', Race='black', Age=10, Sex='M'))
exp(-0.3618487) #0.6963877

contrast::contrast(mod_1,
  list(RuralUrban='Rural', Race='white', Age=10, Sex='M'),
  list(RuralUrban='Rural', Race='hispanic', Age=10, Sex='M'))
exp(0.1916315) #1.211224

contrast::contrast(mod_1,
  list(RuralUrban='Urban', Race='white', Age=10, Sex='M'),
  list(RuralUrban='Urban', Race='hispanic', Age=10, Sex='M'))
exp(-0.08075049) #0.9224238

#####Smoke Table for Hypothesis 2#####
mod_2 = glm(ever_ecigarette ~ Sex + Age + RuralUrban + Race, family=binomial, data=smokeSub)
theCiMat = Pmisc::ciMat(0.95)
parTable = summary(mod_2)$coef[,rownames(theCiMat)]%*%theCiMat
rownames(parTable)[1]="Baseline"
y = cbind(summary(mod_2)$coef,exp(parTable))
knitr::kable(y, digits=3, col.names = c('Estimate', 'Std. Error', 'Z-value', 'Pr(>|z|)',
                                         'Odds Ratio', 'Lower (CI)', 'Upper (CI)'),
             caption="Estimated Parameters from the Second Model to Analyze the Usage of E-cigarette ")

#####Contrast#####
contrast::contrast(mod_2,
  list(RuralUrban='Rural', Race='white', Age=10, Sex='M'),
  list(RuralUrban='Rural', Race='white', Age=10, Sex='F'))
exp(0.06010364) #1.061947

```