STA442 Project1

Aichen Liu

07/10/2020

Affairs

Introduction

This research explores the effect of having children on the chance of having affairs for males and females. It is hypothesized that women with children are less likely to have extramarital sex since nursing children occupy most of the time and making them insular, whereas men are more likely to have affairs because they feel stressed, neglected and constrained. Figure 1 gives a glimpse of the data.

Table 1: Figure 1

	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating	ever	religious	ageC
4	0	male	37	10.00	no	3	18	7	4	FALSE	low	5
5	0	female	27	4.00	no	4	14	6	4	FALSE	med	-5
11	0	female	32	15.00	yes	1	12	1	4	FALSE	anti	0
16	0	male	57	15.00	yes	5	18	6	5	FALSE	high	25
23	0	male	22	0.75	no	2	17	6	3	FALSE	no	-10
29	0	female	32	1.50	no	2	17	5	5	FALSE	no	0

Model

The data contains 601 observations and 12 variables, I have created a new variable called ageC that centred at 32 to make the results easier to interpret. The data contains several key variables: Ever, indicates whether the person ever had extramarital sex; Gender, takes values male and female; Children, indicates whether the person has children; Yearsmarried, a numerical variable indicates the number of years married; Religious, measures the level of devotion to religion. To further analyze the hypothesis, set gender, children, ageC, yearsmarried and religious to be the predictors and ever to be the response, and since response takes the value true and false, a logistics model would best describe the relationship.

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

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

In equation one, Y_i is the number of people who have had extramarital sex N_i is the number of people in subgroup i. μ_i is the estimated probability that the ith individual has had affairs. In the second equation, $X_1, X_2, X_3, X_4$are gender, children, age(adjusted), years married and religious respectively. $logit(p)$ is the log odds of having affairs for each subgroup.

Discussion

Figure 2 contains a statistical analysis of the logistic model. The first column is the estimated value of the odds of having affairs. The second and last columns give the confidence interval for each subgroup. In the

table, levels of devotion to religion are categorized into 5 groups, anti, no religion, low, medium(med), and high. And female and male are each subcategorized into groups with children or no children. In this logistic model, the reference group represents 32 years old males just married, with children, and no devotion to religion.

Table 2: Figure 2

	Odds	lower	upper
Reference Group(baseline)	0.205	0.104	0.404
ageC	0.964	0.930	0.999
yearsmarried	1.111	1.042	1.184
religiousanti	2.025	0.993	4.128
religiouslow	1.317	0.769	2.254
religiousmed	0.483	0.279	0.836
religioushigh	0.515	0.245	1.081
genderfemale:childrenno	0.400	0.186	0.860
gendermale:childrenno	0.777	0.378	1.598
genderfemale:childrenyes	0.768	0.486	1.213

If the confidence interval contains one, then its corresponding subgroup is not significant. Subgroups: genderfemale:childrenyes(female with childern), gendermale:childrenno(male without childern), religioushigh(high level of devotion to religion), and religiouslow(low level of devotion to religion) are not significant, because their confidence intervals contain one. Since genderfemale:childrenyes and gendermale:childrenno are not significantly related to the odds of having affairs, we cannot compare them with the reference group. Thus, there is no enough information to prove the hypothesis. We don't know the effect of having children on the chances that men and women have affairs.

Summary

Recent research explored the effect of having children on the chances that men and women have affairs. Base on a lot of TV sitcoms, some people believe becoming a mother will make women less likely to have affairs, whereas becoming a father will make men become the exact opposite. A logistic model is constructed to address this hypothesis, and the result is unexpected. Research shows that the odds of having affairs for both men and women are not related to whether having children. Therefore, there is no information to support the hypothesis, TV sitcoms cannot reflect reality. However, other factors are actively related to having affairs, such as ages, years married, and religiousness. The research indicates that people who are married for a longer time and older tend to have higher odds of having affairs. More interestingly, the result also shows that people with a medium level of devotion to religion tend to have low odds of having affairs.

Smoking

Summary

This research explores the smoking pattern among American youths. Age, sex, ethnicity and demographic characteristics are considered to effect smoking decisions. Before conducting any analysis, there are two hypotheses: White Americans use cigars, cigarillos or little cigars no more common than Hispanic Americans and African Americans do, because cigar smoking is a rural phenomenon and White American more likely live in rural areas. And the second hypothesis is that if age, ethnicity and demographic characteristics are the same, the likelihood of a man using electronic cigarettes is similar to women's. We will use data from the 2019 American National Youth Tobacco Survey to test our hypothesis.

Introduction

The two hypotheses are closely related, both of these two hypotheses involve the same predictors. Demographic characteristics are defined as Rural/Urban and ethnicity is categorized into six groups: White, Black, Hispanic, Asian, native and pacific. The data used contains 152 variables, I have created a new variable called ageC that centred at age 14 to make the results easier to interpret. Keys variables ageC, sex, race, RuralUrban, ever_cigars_cigarillos_or, ever_ecigarette are selected from the original data to fit models. Figure 3 gives a glimpse of the selected data.

Table 3: Figure 3

Age	Sex	Race	RuralUrban	ever_cigars_cigarillos_or	ever_ecigarette	ageC
15	F	NA	Rural	FALSE	FALSE	1
16	M	hispanic	Rural	FALSE	FALSE	2
14	M	hispanic	Rural	FALSE	FALSE	0
14	M	hispanic	Rural	FALSE	FALSE	0
15	M	white	Rural	FALSE	FALSE	1
15	F	hispanic	Rural	FALSE	FALSE	1

Methods

For the above hypotheses, ever_cigars_cigarillos_or and ever_ecigarette are used as our responses. Ever_cigars_cigarillos_or indicates whether the person ever tried smoking cigars, cigarillos, or little cigars. ever_ecigarette indicates whether the person ever used an e-cigarette. AgeC, sex, race, UrbanRural are set to be the predictors. Since both Ever_cigars_cigarillos_or and ever_ecigarette take values True or False, logistic models are better to capture the relationship. Two logistic models, Model_1 and Model_2 are fitted for each hypothesis, as shown below.

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

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Different models have different interpretations of the equations. In Model_1, The first equation gives that Y_i follow Binomial distribution where Y_i is the number of students who tried smoking cigars, cigarillos, or little cigars in group i. N_i is the number of students in group i. μ_i is the estimated probability that the ith individual used cigars, cigarillos, or little cigars. In the second equation, X_1, X_2, X_3, X_4 are the predictors, AgeC, sex, race, UrbanRural respectively.$logit(p)$ represents the log of odds, it is the log of odds of smoking cigars, cigarillos, or little cigars;

In Model_2, Y_i follow Binomial distribution where Y_i is the number of students who tried smoking an electrical cigarette in group i. N_i is the number of students in group i. μ_i is the estimated probability that the ith individual used electrical cigarette. And X_1, X_2, X_3, X_4 are the predictors, AgeC, sex, race, UrbanRural respectively.$logit(p)$ represents the log of odds of trying electrical cigarette. β_is are log odds ratios, that indicate the difference between subgroup and reference group.

Results

For hypothesis one, the result of the statistical analysis is shown in Figure 4. The first column gives the estimated value of the odds of smoking cigar for each subgroup, the second and third columns give the corresponding confidence interval. As indicated in the table, all predictors are significantly related to smoking cigars, except Racehispanic, Racenative, and Racepacific, because their confidence intervals contain one.

In this case, the reference group(baseline) represents 14 years old urban white males. Keeping other conditions unchanged, the odds of rural white males smoking cigars is 1.494 times to those of urban white males. Therefore we know white males who live in rural areas are more likely to smoke cigars. On the other hand, the odds of African-Americans smoking cigars is 1.535 times to those of White-Americans, thus smoking cigars is less common among white-Americans than for African-Americans. Since Racehispanic is not significant, we

can not conclude anything. Hence, for hypothesis one, we have proven that indeed White Americans live in rural areas are more likely to smoke cigars. Smoking cigars, cigarillos or little cigars is less common among White-Americans than for African-Americans, but we don't know whether is the case for Hispanic-Americans.

Table 4: Figure 4

	Odds	0.5 %	99.5 %
reference group(baseline)	0.099	0.087	0.113
RuralUrbanRural	1.494	1.329	1.681
Raceblack	1.535	1.300	1.808
Racehispanic	0.945	0.824	1.083
Raceasian	0.283	0.177	0.430
Racenative	1.327	0.759	2.198
Racepacific	1.540	0.708	3.041
SexF	0.683	0.607	0.769
ageC	1.453	1.410	1.499

For hypothesis two, we want to estimate whether sex is a significant factor affecting the smoking decision of E-cigarette. As the result shown in Figure 5, the variables, which confidence interval contains one are Racehispanic, Racenative, Racepacific and SexF. Since confidence interval of sex contains one, therefore, sex is not related to smoking E-cigarette. Thus, we don't have enough information to prove hypothesis two. we don't know whether the likelihood of having used an electronic cigarette for women or man's are the same.

Table 5: Figure 5

	Odds	0.5 %	99.5 %
reference group(baseline)	0.444	0.406	0.484
RuralUrbanRural	1.140	1.046	1.242
Raceblack	0.598	0.520	0.685
Racehispanic	0.915	0.830	1.008
Raceasian	0.364	0.287	0.460
Racenative	1.064	0.712	1.569
Racepacific	1.268	0.720	2.191
SexF	0.942	0.865	1.025
ageC	1.400	1.370	1.431

Appendix

```
library(tidyverse)
library(dplyr)
library(AER)

#Affairs
data('Affairs', package='AER')
#clean data
Affairs$ever = Affairs$affair > 0
Affairs$religious = factor(Affairs$religiousness,levels = c(2,1,3,4,5),
                           labels = c('no','anti','low','med','high'))
#center variable age
Affairs$ageC =Affairs$age - 32
```

```r
#fit model
glm_affairs = glm(ever ~ gender:children + ageC + yearsmarried + religious,
data=Affairs, family='binomial')
Table = as.data.frame(summary(glm_affairs)$coef)
likci = confint(glm_affairs, level = 0.95)
Table$lower = Table$Estimate - 2*Table$`Std. Error`
Table$upper = Table$Estimate + 2*Table$`Std. Error`
OddsRatio = exp(Table[,c('Estimate','lower','upper')])
rownames(OddsRatio)[1]="Reference Group(baseline)"
colnames(OddsRatio)[1]="Odds"

#smoking
#get data
smokeFile = 'smokeDownload.RData'
if(!file.exists(smokeFile)){
    download.file(
            "http://pbrown.ca/teaching/appliedstats/data/smoke.RData",
            smokeFile)
}
(load(smokeFile))

#clean data
#get rid of 9 year olds because their data is suspicious
smokeSub = smoke[which(smoke$Age >= 10), ]
smokeAgg <- smokeSub %>%
  select(Age, Sex, Race, RuralUrban,ever_cigars_cigarillos_or,ever_ecigarette)
smokeAgg$ageC =(smokeAgg$Age - 14)

#fit model
H1 = glm(ever_cigars_cigarillos_or ~ RuralUrban+Race+Sex + ageC,
data=smokeAgg, family='binomial')
#round(summary(H1)$coef,digits = 3)
logOddsMat = cbind(est=H1$coef, confint(H1, level=0.99))

oddsMat = exp(logOddsMat)
rownames(oddsMat)[1] = "reference group(baseline)"
colnames(oddsMat)[1] = "Odds"


H2 = glm(ever_ecigarette ~ RuralUrban+Race+Sex + ageC,
data=smokeAgg, family='binomial')
#round(summary(H2)$coef,digits = 3)
logOddsMat2 = cbind(est=H2$coef, confint(H2, level=0.99))
oddsMat2 = exp(logOddsMat2)
rownames(oddsMat2)[1] = "reference group(baseline)"
colnames(oddsMat2)[1] = "Odds"
```