Jiahao Cheng Assignment 1 October 9, 2020

1. Affairs Report: Does having children affect married people to have extramartial sex?

Introduction

We used data collected from 600 married readers of the American magazines *Redbook* and *Psychology Today* in 1969 to analyze how having children would affect married people to have affairs. This survey focused on the frequency married readers had extramarital sex. We wanted to figure out whether married people have affairs after they had children, and other factors that might influence people to have affairs.

Method

In order to do this analysis, we built a model using the Affairs dataset which changed the frequency of affairs to a yes or no question. Thus, a logistic regression model could be used to explore our investigation, which

$$Y_i \sim Binomial(N_i, \mu_i)$$
$$log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

- Y_i indicates whether the i^{th} person has affairs;
- μ_i is the probability of the i^{th} person that has affairs;
- X_i represents a collection of personal information for the i^{th} person including the interaction between the sex and the number of children, the sex, the age, the number of children, the number of years married, and the religious;
- β are coefficients for all covariates X_i, which revels the influences of each factor.

For the data used for the model, we redefined the religious to 5 levels: no religious, anti religious, low religious, mid religious, and high religious. Also, we centered the age and the number of married years to 32 and 4 respectively.

Result

Figure 1 shows the exponential coefficients of our model with its confidence intervals, which is convenient to do comparison by computing the odd ratio. We explored that fathers are more like to have affairs which is about 11 ($\frac{0.242+1.922+0.67}{0.242}$) times more than married men who do not have children, and mothers follow similar trend, which is about 16.5 ($\frac{0.124+1.922}{0.124}$) times more than married women who do not have children regardless of other factors.

The huge differences are caused by the high weight for the parameter children. After we took into account other factors, we found that the gender, age, the number of years married, and some types of religious are statistically significant because of low p-value (< 0.05). In addition, the confidence intervals of the children, low religious, high religious, and the interaction of gender and children include 0, which gives that we can ignore those factors. Since it seems like men are often older than his wife in the last century, we can only certain that mothers are more likely to have affairs after they have children. However, it is certainly true that females are about half ($\frac{0.124}{0.242}$) as likely to have affairs as males. According to Figure 1, we can see that the age and the married years play a significant role in our investigation, which reveals that as years go on, people are more likely to have affairs. Moreover, for their religious status, we can see that anti religious people have more probabilities to have affairs than people who are mid religious with the ratio of 4.2 ($\frac{2.025}{0.483}$).

	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %
genderfemale	0.124	0.354	-5.890	0.000	-2.821	-1.426
gendermale	0.242	0.345	-4.113	0.000	-2.126	-0.767
childrenyes	1.922	0.388	1.683	0.092	-0.087	1.445
ageC	0.964	0.018	-2.030	0.042	-0.072	-0.002
yearsmarriedC	1.111	0.032	3.276	0.001	0.043	0.169
religiousanti	2.025	0.356	1.981	0.048	0.001	1.403
religiouslow	1.317	0.269	1.024	0.306	-0.252	0.804
religiousmed	0.483	0.274	-2.654	0.008	-1.271	-0.193
religioushigh	0.515	0.371	-1.790	0.074	-1.421	0.042
gendermale:childrenyes	0.670	0.488	-0.820	0.412	-1.369	0.555

Figure 1: The table of odds of the affairs model with confidence intervals.

Summary

We analyzed the frequency of affairs from 600 married readers of the American magazines *Redbook* and *Psychology Today* in 1969. We explored that people are more likely to have affairs as they become older. Especially for the married men, they are about doubled probabilities to have extramarital sex than women. In addition, religious people are less likely to have affairs than anti-religious people. Moreover, the number of children is not significant in this investigation once we included the age and the number of married years as the factors.

Appendix

We used R to built the models, which can be accessed in Appendix 1.

2. Smoking Report: The investigation of the distribution of smoking people among American Youth

Summary

We analyzed the smoking situation amongst American students by two logistic regression models using the data from 2019 American National Youth Tobacco Survey. We explored that African-Americans smoke cigars, cigarillos, or little cigars the most, followed by whites, and finally Hispanic-Americans from the comparison of these three ethnic groups. In addition, there is high certainly that white Americans are more likely to smoke cigars in rural areas. Moreover, in terms of the group of using a electronic cigarettes, females and males are mostly equally distributed if they have the same age, race, and other demographic characteristics.

Introduction

We used the data collected from the 2019 American National Youth Tobacco Survey to analyze the smoking situation among American youth who are older than 9 in terms of their ethnicity, region, sex, and age. This survey focused on the use of cigars, hookahs, and chewing tobacco among American school children. We wanted to figure out which ethnic group from the native American, Hispanic-Americans, and African-Americans, is more likely to smoke real cigarettes like cigars, cigarillos, or little cigars. Especially for the white Americans, whether the region, living in the rural or the urban, influences the behaviours of smoking interested our research. In addition, we compared the differences of chewing the electronic cigarettes between male and female.

Method

In order to do these analysis, we built two models using the SmokeSub dataset which eliminating people who are younger than 10. The first one deals with the problem whether American youth chew cigars, cigarillos, or little cigars. Since the response is a yes or no question, we used a logistic regression model to explore the factors influencing the chance that an individual chews cigars. cigarillos or little cigars, which

$$Y_i \sim Binomial(N_i, \mu_i)$$
$$log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

- Y_i indicates whether the i^{th} student who is greater than 10 years old chews cigars, cigarillos, or little cigars at least once;
- μ_i is the probability of the i^{th} student chewing cigars, cigarillos, or little cigars - X_i represents a collection of personal information for the i^{th} student including the age, the gender, the race, and the living place whether it is rural or urban;
- β are coefficients for all covariates X_i, which revels the influences of each factor.

For the second model, We built it to explore the differences of having used a electronic cigarettes on at least one occasion in terms of the gender using the same covariates as the first one. Similarly, we used a logistic regression model since the response is a yes or no question, which

$$Y_i \sim Binomial(N_i, \mu_i)$$
$$log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

where Y_i indicates whether the i^{th} student who is greater than 10 years old chews electronic cigarettes at least once, μ_i is the probability of the i^{th} student chewing electronic cigarettes, and X_i and β represent the same things as the first model.

Result

In order to to do the comparison between students in different region, we set the urban male students as the baseline of our model. Thus, according to Figure 2, which shows the exponential coefficients of the first logistic model, we can easily do comparison by computing the odds ratio between different groups. We can get that comparing with the native students, black people are more likely to smoke cigars, cigarillos, or little cigars with around 15.6% ($\frac{1.535}{1.327}$) more, while Hispanic Americans are about 28.8% ($\frac{0.945}{1.327}$) less to smoke. Focusing on the white Americans, since the coefficient of the rural is 1.494, we can conclude that students in the rural area are as one and a half times as those in the urban area for smoking.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.001	0.192	-39.216	0.000
RuralUrbanRural	1.494	0.046	8.796	0.000
Raceblack	1.535	0.064	6.702	0.000
Racehispanic	0.945	0.053	-1.066	0.286
Raceasian	0.283	0.172	-7.355	0.000
Racenative	1.327	0.205	1.377	0.168
Racepacific	1.540	0.280	1.541	0.123
SexF	0.683	0.046	-8.318	0.000
Age	1.453	0.012	31.324	0.000

Figure 2: The table of odds of the model about smoking cigars, cigarillos, or little cigars.

	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %
RuralUrbanUrban	0.004	0.130	-42.364	0.000	-5.783	-5.272
RuralUrbanRural	0.005	0.130	-41.538	0.000	-5.651	-5.142
Raceblack	0.598	0.054	-9.614	0.000	-0.620	-0.410
Racehispanic	0.915	0.038	-2.369	0.018	-0.163	-0.015
Raceasian	0.364	0.092	-11.023	0.000	-1.191	-0.832
Racenative	1.064	0.153	0.406	0.685	-0.242	0.358
Racepacific	1.268	0.215	1.102	0.270	-0.191	0.654
SexF	0.942	0.033	-1.823	0.068	-0.125	0.005
Age	1.400	0.008	39.901	0.000	0.320	0.353

Figure 3: The table of odds of the model about smoking electronic cigarettes with confidence intervals.

We use the second model to figure out how likely students have used a electronic cigarettes between their genders with similar age, ethnicity, and living place. According to Figure 3, we can see that the percentage of smoking electronic cigarettes for females are very close to that for males with approximately 6% lower rate. However, it is not statistically significant because the p-value is 0.06 and 0 is included in its confidence interval. So, there is high certainly that females and males have similar

probabilities to smoke electronic cigarettes.

Appendix

We used R to built the models, which can be accessed in Appendix 2.

Appendix

Appendix 1 R code for Question 1

```
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$ageC = Affairs$age - 32
Affairs$yearsmarriedC = Affairs$yearsmarried - 4
Affairs$ratingC = Affairs$rating - 1

glmAffairs <- glm(ever ~ 0 + gender:children + gender + children + ageC + yearsmarriedC
    + religious ,
                  data=Affairs, family='binomial'(link='logit'))
parTableAffairs <- summary(glmAffairs)$coef
parTableAffairs[, 1] <- exp(parTableAffairs[, 1])
confTableAffairs <- na.omit(confint(glmAffairs))
knitr::kable(cbind(parTableAffairs, confTableAffairs), digits=3)
```

Appendix 2 R code for Question 2

```
smokeFile = file.path("smoke.RData")
if (!file.exists(smokeFile)) {
  download.file("http://pbrown.ca/teaching/appliedstats/data/smoke.RData", smokeFile)
}
(load(smokeFile))

smokeSub = smoke[which(smoke$Age >= 10&!is.na(smoke$Race)), ]
glmSmoke1 <- glm(ever_cigars_cigarillos_or ~ RuralUrban + Race + Sex + Age,
                 family=binomial, data=smokeSub)
glmSmoke2 <- glm(ever_ecigarette ~ 0 + RuralUrban + Race + Sex + Age,
                 family=binomial, data=smokeSub)
parTable1 <- summary(glmSmoke1)$coef
parTable1[, 1] <- exp(parTable1[, 1])
knitr::kable(parTable1, digits=3)

parTable2 <- summary(glmSmoke2)$coef
parTable2[, 1] <- exp(parTable2[, 1])
confTable2 <- confint(glmSmoke2)
knitr::kable(parTable2, digit=3)
knitr::kable(cbind(parTable2, confTable2), digit=3)
```