Homework 1, Generalized linear models

Huifeng Wu

1 Affairs

Introduction

In 1969, the American magazines *Redbook and Psychology Today* collected infidelity data, namely the number of extramarital sex over a certain period, from their 600 married readers. It is of great research interest to investigate whether a specific group has a higher chance of having affairs, because it seems quite prevailing that new fathers are more likely to have affairs than new mothers from our television culture, especially sitcoms. This conjecture is justified that men may lack of belonging once becoming dads, while women may feel overly sensitive. In this report, not only will I evaluate this hypothesis statistically, but also leverage a logistic regression to explore significant factors that lead to the event of love affairs.

Method and Analysis

The Affairs dataset was originally measured as the frequency of extramarital sex. However, this variable was transformed to a binary response for simplicity, with levels of "affairs" and "no affairs". A generalized linear model of logistic regression is simple yet effective to fit the data. The data was collected from 600 different married readers, which satisfies the independence assumption among observations. Continuous variables are all centered with their medians.

For model inclusion, I consider covariates such as age, the number of years married, religiousness faithfulness, gender, children, and the two-way interaction between gender and children, as the first three may be confounding. Model selection will be applied to handle the confounding property between them in the future reports if necessary. It is of great interest to scrutinize how being parents affects the chance of having affairs with different genders, so the interaction term was contained specifically. Low multicollinearity was assumed among the predictor in a general sense. Importantly, the logistic regression models a linear relationship between logit outcome and its predictors as,

$$\begin{aligned} log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) &= X_i\hat{\beta} \\ &= -1.769 + 0.664 * \text{gendermale} + 0.653 * \text{childrenyes} - 0.036 * \text{age} \\ &\quad + 0.105 * \text{yearsmarried} + 0.706 * \text{religiousanti} + 0.275 * \text{religiouslow} \\ &\quad - 0.728 * \text{religiousmed} - 0.664 * \text{religioushigh} - 0.401 * \text{gendermale} * \text{childrenyes} \end{aligned}$$

where $\hat{p}_i$ is the predicted probability of having affairs for an individual, X_i denotes one's features observed, and $\hat{\beta}$ indicates the parameter estimates for the set of covariates involved.

To further interpret, exponential operations were applied to obtain the estimated effects in odds ratios for each predictor, as shown in Table 1. The intercept in terms of the odds ratio of having affairs is 0.171, representing the reference group solely, a 32-year-old married woman who has married for 7 years, has no kids and no religiousness concerned. It is noticeable that the odds ratios are multiplied by 2.026, 1.317, 0.483 or 0.515 if one merely becomes anti-religious, lowly, mediumly, or highly committed from no religious belief, respectively. While fixing other variables, the odds ratio is decreased by a factor of 0.965 for a one-year increase in age, and that is increased multiplicatively by 1.111 for a one-year increase in length of marriage. The model also includes an interaction effect for the group of fathers, and the odds of which is (0.67+1.943+1.921) = 4.534 to the baseline group, if considering the same other features.

In addition, we can evaluate the statistical significance of each predictor by checking whether their respective 95% confidence intervals in Table 1 contain 0 or not. It is evident that all covariates are significant except *religiouslow*, *gendermale*, and the interaction, *gendermale:childrenyes*.

For the research hypothesis, we estimated the odds for women who have children are 1.922 times as those without children, whereas that relation is 1.287 between men with children and men without children. Nevertheless, only the first comparison presents a significant difference (p-value=0.093). Therefore, we conclude no significant difference on the chance of having affairs once males become fathers (p-value=0.484).

Table 1: Estimated effects on odds ratio and 95% CI

	Estimate	2.5 %	97.5 %
(Intercept)	0.171	0.085	0.341
age	0.965	0.931	0.999
yearsmarried	1.111	1.043	1.183
religiousanti	2.026	1.006	4.076
religiouslow	1.317	0.777	2.233
religiousmed	0.483	0.282	0.828
religioushigh	0.515	0.248	1.067
gendermale	1.943	0.825	4.579
childrenyes	1.921	0.897	4.119
gendermale:childrenyes	0.670	0.257	1.748

Summary of the Results

A statistical model, namely logistic regression, was used to interpret the famous Affair dataset, by which researchers found age, number of years married, particular levels of religious commitments are significant variables that contribute to the events of affairs. With such methodology, it is essential to model and predict the probability of having affairs for individuals and that also helps us answer the original research posed earlier - Are human more subject to having affairs than prior to being parents? By comparison, our result provides evidence that there exist no significant difference between males having children and males not having children, yet a significant difference between females having children and females not having children. We estimated the odds for women who have children are 1.922 times as those who do not have one. Therefore, women may have affairs much easily once they become mothers in fact.

Appendix

```
## Question 1 ##
library(knitr)

# Load 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'))

# median of age is 32
Affairs$age = Affairs$age - median(Affairs$age)
# median of year married is 7
Affairs$yearsmarried = Affairs$yearsmarried - median(Affairs$yearsmarried)

# Model fitting
model <- glm(ever~age+yearsmarried+religious+gender*children,
             data=Affairs,
             family='binomial')

summary(model)
round(model$coefficients, 3)

# Breif summary of model stats
Estimate <- round(summary(model)$coefficients,3)
Estimate <- round(exp(Estimate), 3) # exponential
Estimate <- as.data.frame(Estimate)$Estimate # Estimates

table1 <- confint.lm(model) # Confidence interval
table1 <- round(exp(table1), 3) # exponential
table1 <- cbind(table1, Estimate)
table1 <- table1[,c(3,1,2)] # reorder

# Table 1
kable(table1,
      booktabs = TRUE,
      caption = "Estimated effects on odds ratio and 95% CI")

# Table 2
table2 <- contrast::contrast(model,
  list(gender='female', children='yes', age=37, religious='no', yearsmarried=15.0),
  list(gender='female', children='no', age=37, religious='no', yearsmarried=15.0))
exp(table2$Contrast)

# Table 3
table3 <- contrast::contrast(model,
  list(gender='male', children='yes', age=37, religious='no', yearsmarried=15.0),
  list(gender='male', children='no', age=37, religious='no', yearsmarried=15.0))
exp(table3$Contrast)
```

2 Smoking

Introduction

Smoking is one of the most common causes of lung diseases and cardiovascular diseases worldwide, and may result in deaths which could have been preventable. Nowadays, smoking seems a growing trend among the youth. Using logistic regression, this report conducts a statistical analysis to the data according to the 2019 American National Youth Tobacco Survey. With taking account into similar demographic characteristics, we examined the odds of smoking on white Americans versus Hispanic-Americans and white Americans versus African-American. In addition, the effect of gender on smoking electronic cigarettes was also investigated.

Method

The incidence of smoking cigarettes/e-cigarettes, say Y, was measured in a binary scale, so the data are considered to follow a binomial distribution. Observations missing the response was excluded in the further analysis. Moreover, I filtered out subjects who are younger than 10 years old, due to the suspicious nature of their data. For better interpretability, here the continuous variables are centered by subtracting the median in every observation.

Owing to large sample provided in the survey, logistic regressions were proposed to model the two binary responses separately by assuming linearity between logit outcome and a set of predictors, *RuralUrban, Race, Sex* and *Age.* This approach requires each observation to be independent; the assumption holds as the 19018 records all come from different students. No interactions are considered in both models, and the underlying covariates assumed to exhibit low multicolinearity.

$$Y_i \sim Binomial(n, p)$$

$$log\left(\frac{\widehat{p}_i}{1-\widehat{p}_i}\right) = X_i\widehat{\beta}$$

where $\widehat{p}_i$ is the predicted probability of smoking/e-smoking, X_i denotes the given demographic features of a student, and $\widehat{\beta}$ indicates the coefficients estimates along the used covariates. The research hypotheses focus on the effect on smoking between difference races while excluding their living areas, and the effect on e-smoking between different genders. Without interaction effects, it is convenient to compare the desired groups directly via p-values of the corresponding coefficient estimates, race indicators in the former model and age in the latter model.

Results

Two models are fitted with the same covariates, The signs of coefficient estimates are consistent, and the size of estimates are similar, which implies a positive correlation between the two responses. Exponential transformation is appiled to explain the effects on odds ratio. The intercept terms both indicate the odds ratios of smoking/e-smoking for the baseline group, a 14-year-old white male student from urban, as 0.099 and 0.444, respectively.

$$\begin{aligned} log(\frac{\widehat{p}_{cigs}}{1-\widehat{p}_{cigs}}) &= -2.310 + 0.401 * \text{RuralUrbanRural} + 0.429 * \text{Raceblack} - 0.057 * \text{Racehispanic} \\ &\quad - 1.262 * \text{Raceasian} + 0.283 * \text{Racenative} + 0.432 * \text{Racepacific} - 0.381 * \text{SexF} + 0.374 * \text{Age} \\ log(\frac{\widehat{p}_{e-cigs}}{1-\widehat{p}_{e-cigs}}) &= -0.813 + 0.131 * \text{RuralUrbanRural} - 0.515 * \text{Raceblack} - 0.089 * \text{Racehispanic} \\ &\quad - 1.009 * \text{Raceasian} + 0.062 * \text{Racenative} + 0.237 * \text{Racepacific} - 0.060 * \text{SexF} + 0.337 * \text{Age} \end{aligned}$$

Next, interpretation is made on the premise of keeping other variables constant.

For model 1 (top), the odds of smoking for African, Hispanic, Asian, native and Pacific American to white American, are 1.536, 0.945, 0.283, 1.327, and 1.527, correspondingly. Also, the odds ratio is increased by a factor of 1.454 for a one-year increase in age. The odds ratio is increased multiplicatively by 1.49 from rural to urban, and that is decreased multiplicatively by 0.683 from male to female.

For model 2 (bottom), similar interpretation can be used to explain the change of odds of e-smoking for different race groups. It is worth noting that the odds ratio is increased multiplicatively by 1.401 for a unit change in Age, and that is decreased multiplicatively by 0.942 from male to female, with fixing other variables.

With the significance level of 5%, Table 2 and 3 are sufficient to evaluate the research hypothesis addressed earlier: estimated coefficients show statistical significance between white and African American (p-value=0), but no significance between white and Hispanic Americans (p-value=0.286). No significant effects found on the chance of e-smoking in terms of gender (p-value=0.068). The models share similar significant covariates including living areas, age, African-American and Asian-American identity, and model 2 has marked Hispanic-American identity significant additionally.

Table 2: Summary of Model 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.310	0.050	-46.446	0.000
RuralUrbanRural	0.401	0.046	8.796	0.000
Raceblack	0.429	0.064	6.702	0.000
Racehispanic	-0.057	0.053	-1.066	0.286
Raceasian	-1.262	0.172	-7.355	0.000
Racenative	0.283	0.205	1.377	0.168
Racepacific	0.432	0.280	1.541	0.123
SexF	-0.381	0.046	-8.318	0.000
Age	0.374	0.012	31.324	0.000

Table 3: Summary of Model 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.813	0.034	-23.961	0.000
RuralUrbanRural	0.131	0.033	3.930	0.000
Raceblack	-0.515	0.054	-9.614	0.000
Racehispanic	-0.089	0.038	-2.369	0.018
Raceasian	-1.009	0.092	-11.023	0.000
Racenative	0.062	0.153	0.406	0.685
Racepacific	0.237	0.215	1.102	0.270
SexF	-0.060	0.033	-1.823	0.068
Age	0.337	0.008	39.901	0.000

Summary

Based on the 2019 American National Youth Tobacco Survey data, the report takes advantage of logistic regression modelling to learn which specific populations by contrast are more probabilistically subject to smoking cigarettes/e-cigarettes. Particular races, age and living areas are shown to be significant predictors as indicated by our model. Also, we find African-Americans are more prone to smoking than white Americans. Nevertheless, the underlying data does not show solid evidence to prove the difference between Hispanic-Americans and white Americans, nor the difference in e-smoking probability in different genders. From the perspective of healthcare, it is still crucial to track down the e-smoking rate in youth of different genders, as our test still indicates weak significance (p-value=0.068). Thus, we can establish education on tobacco to youth in school and make an early intervention if needed.

Appendix

```r
## Question 2 ##
rm(list = ls())

dataDir = "/data"
smokeFile = file.path(dataDir,"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), ]

# Data cleaning
smokeSub$Age = smokeSub$Age - median(smokeSub$Age) # median(smokeSub$Age) is 14

model1 <- glm(ever_cigars_cigarillos_or~RuralUrban+Race+Sex+Age,
             family=binomial,
             data=smokeSub,
             na.action=na.omit)

model2 <- glm(ever_ecigarette~RuralUrban+Race+Sex+Age,
             family=binomial,
             data=smokeSub,
             na.action=na.omit)

summary(model1)
summary(model2)

# Table 1
summary1 <- round(summary(model1)$coefficients,3)
summary1 <- as.data.frame(summary1)
kable(summary1,
      booktabs = TRUE,
      caption = "Summary of Model 1")

# Table 2
summary2 <- round(summary(model2)$coefficients,3)
summary2 <- as.data.frame(summary2)
kable(summary2,
      booktabs = TRUE,
      caption = "Summary of Model 2")
```