STA442 Assignment 1

Manyi Luo - 1003799419

2020/10/6

1. Affairs

Statistical Report - One page writing

According to affairs data sampled from 600 married readers of the American magazines Redbook and Psychology Today in 1969, researchers investigate the effect of becoming a parent on the chances that men and women have extramarital affairs. The study hypothesizes that becoming a mother will make a woman less likely to have extramarital affairs, while men are more likely to have affairs once they become fathers.

Logistic regression is the most appropriate model to make regression in testing out multiple factors that may influence the chance of extramarital affair, since having affair is considered as a binary output in this case, either having affair or not. Specifically, the response variable is represented by the logarithmic of affair odds, whereas odds represent the chance of having extramarital affair over the chance of not having extramarital affair. Therefore, higher odds represent extramarital affair is more likely to happen.

Explanatory variables include: 1. the confounding variable between individual's gender and status of having children (both gender and children are binary, whereas male for gender and yes for children represent; 2. individual's age (a continuous variable that is centered to make more sense for the interpretation of regression results, starting from 17.5 years old and more); 3. the number of years married (a continuous variable); 4. religiousness of the individuals (a categorical variable, whereas no religious is considered as the reference level).

From obtaining regression output and getting the odds ratio, it's possible to interpret multiple factors' effect on having affairs through coefficients. Here, only statistically significant factors are considered, including centered age, years married, "med" as well as "anti" for religious, which can be seen from p-value that is smaller than 0.05 and confidence interval for odds ratio that does not include 1. Also, it's important to note that the odds ratio uses 1 as a standard of comparison.

For a continuous variable, if the odds ratio is greater than 1, there exist greater odds when increasing by one unit; and vice versa for odds ratio smaller than 1 to have fewer odds. Considering centered age while keeping all other variables fixed, it's odds ratio is 0.964, which is smaller than 1. This means that, when age increases by 1 year, the odds of affairs will increase by 0.964 times, indicating there is less chance of having affair for greater age. Similarly, years married has an odds ratio of 1.111. So when years married increase by 1 year, the odds of affairs will increase by 1.111 times, meaning that there is a greater chance of having affair for more years of marriage.

For categorical variables like religious, the corresponding odds ratio is compared toward the reference level, which is no religious that does not appear on the table of odds ratio estimates. For med religious that is significant, it's odds ratio is 0.483, which is smaller than 1. This means that, comparing people who are considered as med religious to people with no religious, the odd of having affairs is 0.483 times relative to that of people with no religious, meaning that there is less chance for them to have affairs compared to reference.

It's important to note that the confounding variables between an individual's gender and status of having children considered in our hypothesis are statistically insignificant, specifically for males and females having children. Therefore, we cannot conclude that becoming a mother will make a woman less likely to have extramarital affairs, while men are more likely to have affairs once they become fathers.

Odds ratios and confidences intervals of affairs model

	est	2.5	97.5
Baseline	0.262	0.096	0.711
gendermale	1.943	0.826	4.571
childrenyes	1.922	0.898	4.113
age	0.964	0.931	0.999
yearsmarried	1.111	1.043	1.183
religiousanti	2.025	1.008	4.070
religiouslow	1.317	0.778	2.230
religiousmed	0.483	0.282	0.827
religioushigh	0.515	0.249	1.065
gendermale:childrenyes	0.670	0.257	1.745

Research news: Will becoming a mother make a woman less likely to have extramarital affairs, while a men is more likely to have affairs once he become father?

Based on the frequency data about extramarital sex collected from 600 married readers of the American magazines Redbook and Psychology Today in 1969, researchers test out whether becoming a mother makes a woman less likely to have extramarital affairs, whereas a man is more likely to have affairs once he becomes a father. The result is astonishing: they failed to conclude the relationship between the chance of having affairs and an individual's gender as well as the status of having children due to insignificant data. However, they still find that other factors may contribute to the chance of having affairs, which are: individual's age, years married, and religious. For individuals' age and years married, they obtain results that show less chance of having affair for a greater age and a greater chance of having affair for more years of marriage. Also, they can conclude that specifically for medium religious people, as there is less chance for them to have affair for greater age compared to people with no religious belief. These results obtained about multiple factors that can influence the chance of affairs are intuitive and understandable, which can be seen in real-life events and TV dramas as well. Therefore, we cannot say that, after having children, different gender has a different tendency to cheat, but rather other factors embedded behind to determine the chance of having affairs.

2. Smoking

```
## [1] "smoke"        "smokeFormats"
```

Consulting Report - More detailed codes provided in Appendix

Introduction

To carry about multiple research questions about smoking behaviors among children with different origins and sex, the 2019 American National Youth Tobacco Survey is used to conduct the following analysis, as it includes data of the use of cigars, hookahs, and chewing tobacco amongst American school children. There are two main research hypotheses: The first hypothesis is that given white Americans more likely to live in rural areas and cigar smoking is a rural phenomenon, the chance of smoking of cigars, cigarillos or little cigars is no more common

amongst Americans of European ancestry than for Hispanic-Americans and African-Americans. The second hypothesis is that the chance of used electronic cigarettes on at least one occasion is the same among people of different sex when controlling their age, ethnicity, and other demographic characteristics.

Statistical Method

According to the research hypothesis, two variables are used, 1. “ever_cigars_cigarillos_or”, representing the binary output from the question: “Have you ever tried smoking cigars, cigarillos, or little cigars, such as Swisher Sweets, Black and Mild, Garcia y Vega, Cheyenne, White Owl, or Dutch Masters, even one or two puffs?”; 2. “ever_ecigarette”, representing the binary output from the question: “Have you ever used an e-cigarette, even once or twice?” Therefore, logistic regression is the most appropriate model to test out multiple factors that may influence dichotomous outputs (Yes = 1, No = 0).

In logistic model, the response variable (Y, with $P(Y = 1) = p$) is represented by the logarithmic of smoking odds, either ever tried smoking cigars, cigarillos, little cigars or ever used an e-cigarette. Specifically, odds represents the chance of ever tried smoking over the chance of never tried smoking ($logit[Y = 1] = log(\frac{P(Y=1)}{P(Y=0)}) = log(\frac{p}{1-p})$).

Explanatory variables (X_1 to X_4) include: X_1. RuralUrban: a catagorical variable representing whether the school the respondent attended was rural or urban (binary output, Rural = 1); X_2. variable representing Race: also catagorical, including Black (binary, Black = 1), Hispanic (binary, Hispanic = 1), Asian (binary, Asian = 1), Native (binary, Native = 1), Pacific (binary, Pacific = 1), whereas White is considered as the reference level; X_3. Sex: a catagorical variable representing male and female (binary output, Female = 1); X_4. Age: a categorical variable that has been converted to years. Note that the counfounding variable between RuralUrban and Race may appear in the first hypothesis to facilitate our regression, since the hypothesis considers the joint affect between Race and RuralUrban, specifically under Rural.

Therefore, the model for the first hypothesis can be expressed as: $logit[Y = 1] = \alpha + \beta_1 X_1 * X_2 + \beta_2 X_3 + \beta_3 X_4$, whereas the model for the second hypothesis can be expressed as: $logit[Y = 1] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

Results - First Research Question

For the first hypothesis, regression output and odds ratio can be obtained from the table below to see how region (RuralUrban), Race, Sex, and Age may influence the ever smoking odds of children. Here, all four independent variables are statistically significant related to the ever smoking odds, as their p-values are smaller than 0.05 and confidence intervals for odds ratio do not include 1, specifically RuralUrbanRural, Raceblack, Raceasian, Racepacific, SexF, Age.

Using RuralUrban as an example of significant categorical variables, the corresponding odds ratio for Rural is compared toward the reference level, Urban. Its odds ratio is 1.602, which is greater than 1, meaning that there exist greater odds when increasing by one unit. While keeping all other variables fixed, this means that comparing children from Rural region to children from Urban region, the odds of ever smoking is 1.602 times more relative to that of children from Urban region, meaning that there is more chance for them to ever tried smoking cigars, cigarillos, or little cigars. Similarly, for Race, Black children odds of ever smoking is 1.436 times more and Pacific children odds of ever smoking is 2.070 times more relative to that of white children; Asian children odds of ever smoking is 0.328 times less relative to that of white children. For sex, female children’s odds of ever smoking is 0.685 times less relative to that of male children.

For continuous variable specifically Age, it’s odds ratio is 1.454, which is greater than 1. This means that, while keeping all other variables fixed, when age increases by 1 year, the odds of ever smoking will increase by 1.454 times, indicating there is more chance of smoking for greater age.

Odds ratios and confidences intervals of smoking model 1

	est	2.5	97.5
Baseline	0.001	0.000	0.001
RuralUrbanRural	1.602	1.413	1.817
Raceblack	1.436	1.195	1.726
Racehispanic	1.084	0.934	1.259
Raceasian	0.328	0.224	0.480
Racenative	1.783	0.941	3.378
Racepacific	2.070	1.041	4.117
SexF	0.685	0.626	0.749
Age	1.454	1.420	1.488
RuralUrbanRural:Raceblack	1.163	0.904	1.496
RuralUrbanRural:Racehispanic	0.762	0.618	0.939
RuralUrbanRural:Raceasian	0.605	0.267	1.370
RuralUrbanRural:Racenative	0.627	0.276	1.423
RuralUrbanRural:Racepacific	0.496	0.158	1.555

When comparing Rural Hispanic children toward Rural White children, it's log odds ratio is -0.192 (odds ratio = 0.826), which is smaller than 1 with a significant p-value smaller than 0.05. Therefore we can conclude on our hypothesis that accounting for other variables such as age and sex for smoking， in rural areas, Hispanic children have 82.6% fewer odds to ever tried smoking cigars, cigarillos, or little cigars than White children, which makes a part conclusion on the first hypothesis.

```
## glm model parameter contrast
##
##    Contrast       S.E.      Lower      Upper     t    df Pr(>|t|)
##  -0.1916315 0.07480418 -0.3382547 -0.0450084 -2.56 18396   0.0104
```

On the other hand, the comparison between Rural Black children toward Rural White children also concludes on the first hypothesis, as log odds ratio is 0.513 (odds ratio = 1.670), which is greater than 1 with a significant p-value smaller than 0.05. Therefore we can conclude on our hypothesis that accounting for other variables such as age and sex for smoking， in rural areas, Black children have 167% more odds to ever tried smoking cigars, cigarillos.

```
## glm model parameter contrast
##
##   Contrast       S.E.     Lower     Upper    t    df Pr(>|t|)
##  0.5125319 0.08800525 0.3400334 0.6850304 5.82 18396        0
```

Results - Second Research Question

For the second hypothesis, regression output and odds ratio can be obtained similarly from the table below to see how region (RuralUrban), Race, Sex, and Age may influence the odds of ever used an e-cigarette, even once or twice. Here, also all four independent variables are statistically significant related to the ever used an e-cigarette odds, as their p-values are smaller than 0.05 and confidence intervals for odds ratio do not include 1, specifically RuralUrbanRural, Raceblack, Racehispanic, Raceasian, Age.

Similarly, for RuralUrban, the corresponding odds ratio for Rural is 1.140, which is greater than 1. This means that comparing children from Rural region to children from Urban region, the odds of ever used an e-cigarette odds is 1.140 times more relative to that of children from Urban region. Also for Race, Black children odds of ever used an e-cigarette is 0.598 times less, Hispanic children odds of ever used an e-cigarette is 0.915 times less, and Asian children odds of ever used an e-cigarette is 0.364 times less relative to that of white children.

For the continuous variable Age, it's odds ratio is 0.337, which is less than 1 as well. This means that, while keeping all other variables fixed, when age increases by 1 year, the odds of ever smoking will decrease by 0.337 times, indicating there is less chance of smoking for greater age.

Note that for the variable Sex that is truly concerned in the second hypothesis, it is not even considered as statistically significant when examining its p-value from logistic regression.

Odds ratios and confidences intervals of smoking model 2

	est	**2.5**	**97.5**
Baseline	0.004	0.003	0.005
RuralUrbanRural	1.140	1.068	1.217
Raceblack	0.598	0.538	0.664
Racehispanic	0.915	0.849	0.985
Raceasian	0.364	0.305	0.436
Racenative	1.064	0.788	1.436
Racepacific	1.268	0.832	1.932
SexF	0.942	0.883	1.005
Age	1.400	1.377	1.424

The statistically insignificant Sex can also be shown when comparing Rural Asian males toward females, as it's p-value is greater than 0.05 in model parameter contrast. Even though it's log odds ratio is -0.060 (odds ratio = 1.062), which is greater than 1, meaning that accounting for other variables such as age and region for smoking, female children have 1.062% more odds to ever used an e-cigarette. But due to insignificance, we failed to reject the second hypothesis that the chance of used electronic cigarettes on at least one occasion is the same among people of different sex, when controlling their age, ethnicity, and other demographic characteristics.

```
## glm model parameter contrast
##
##     Contrast      S.E.      Lower       Upper     t    df Pr(>|t|)
##  -0.06010364 0.0329749 -0.1247375 0.004530222 -1.82 18392   0.0684
```

Summary

Based on the data about smoking behaviors among children with different origins and sex the 2019 American National Youth Tobacco Survey, the following analysis is conducted to test out two research questions: 1. Given white Americans more likely to live in rural areas and cigar smoking is a rural phenomenon, whether the chance of smoking of cigars, cigarillos, or little cigars is no more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans; 2. Whether the chance of using electronic cigarettes on at least one occasion is the same among people of different sex when controlling their age, ethnicity, and other demographic characteristics. The result for the first research question indicates the difference in behavior between Hispanic and Black children compared to White children, whereas accounting for other variables such as age and sex for smoking，in rural areas, Hispanic children have 82.6% fewer odds to ever tried smoking cigars, cigarillos, or little cigars than White children; On the other hands, accounting for other variables such as age and sex for smoking，in rural areas, Black children have 167% more odds to ever tried smoking cigars, cigarillos. For the result of the second research question, even though accounting for other variables such as age and region for smoking, female children have 1.062% more odds to ever used an e-cigarette, this result is statistically insignificant, meaning that we failed to conclude that the chance of used electronic cigarettes on at least one occasion is different among people in different sex, when controlling their age, ethnicity, and other demographic characteristics.

```
#1. Affairs
#install.packages('AER')
#install.packages('Pmisc', repo = "http://r-forge.r-project.org")

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'))

#center parameter: make more sense for the interpretation of regression intercept, start
ing from 17.5 years old and more
Affairs$ageC <- Affairs$age - min(Affairs$age)
#Affairs$yearsmarriedC <- Affairs$yearsmarried - min(Affairs$yearsmarried)

#age, years married, some religious significant
options(scipen = 999)
affairs_model <- glm(ever ~ gender*children + age + yearsmarried + religious, data=Affai
rs, family='binomial')
summary(affairs_model)
summary(affairs_model)$coef

#interpretation: considering years married while keeping all other variables fixed, it's
odds ratio is 1.111, which is greater than 1. Therefore, when years married increases by
1 year, the odds of affairs will increase by 1.111 times, meaning that here is greater c
hance of having affair for more years of marriage.

knitr::kable(summary(affairs_model)$coef, digits = 3, caption = "Statistics Summary of A
ffairs Model")

#odds ratio
theCiMat = Pmisc::ciMat(0.95)
parTable = summary(affairs_model)$coef[,rownames(theCiMat)] %*% theCiMat
rownames(parTable)[1] <- "Baseline"
knitr::kable(exp(parTable), digits = 3, caption = "Odds ratios and confidences intervals
of affairs model")

#2. Smoking
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))

smokeSub = smoke[which(smoke$Age >= 10), ]

#First Research Question
smoking_model <- glm(ever_cigars_cigarillos_or ~ RuralUrban * Race + Sex + Age, family=b
inomial, data=smokeSub)
summary(smoking_model)
summary(smoking_model)$coef

#install.packages("kableExtra")
#odds ratio
```

```
theCiMat = Pmisc::ciMat(0.95)
parTable = summary(smoking_model)$coef[,rownames(theCiMat)] %*% theCiMat
rownames(parTable)[1] <- "Baseline"
knitr::kable(exp(parTable), digits = 3, caption = "Odds ratios and confidences intervals
of smoking model 1")

contrast::contrast(smoking_model,
list(RuralUrban = "Rural", Race = "hispanic", Age = 16, Sex = "F"),
list(RuralUrban = "Rural", Race = "white", Age = 16, Sex = "F"))
exp(-0.1916315)

contrast::contrast(smoking_model,
list(RuralUrban = "Rural", Race = "black", Age = 16, Sex = "F"),
list(RuralUrban = "Rural", Race = "white", Age = 16, Sex = "F"))
exp(0.5125319)

#Second Research Question
smoking_model_2 <- glm(ever_ecigarette ~ RuralUrban + Race + Sex + Age, family=binomial,
data=smokeSub)
summary(smoking_model_2)
summary(smoking_model_2)$coef

knitr::kable(summary(smoking_model_2)$coef, digits = 3)

#odds ratio
theCiMat = Pmisc::ciMat(0.95)
parTable = summary(smoking_model_2)$coef[,rownames(theCiMat)] %*% theCiMat
rownames(parTable)[1] <- "Baseline"
knitr::kable(exp(parTable), digits = 3, caption = "Odds ratios and confidences intervals
of smoking model 2")

contrast::contrast(smoking_model_2,
list(RuralUrban = "Rural", Race = "asian", Age = 16, Sex = "F"),
list(RuralUrban = "Rural", Race = "asian", Age = 16, Sex = "M"))
exp(0.06010364)
```