Affairs

Nan Jiang

2010.10.09

This data set collects the frequency of extramarital sex from 600 married readers of American magazines Redbook and Psychology Today in 1969. Our hypothesis is that for women, having child will decrease the extramarital sex frequency, for man with a child is the opposite condition. The data is based on the audience of the TV sitcoms.

We bulit a logistic model to conduct the results. With Binomial regression, we can estimate the probability of the i-th person having extramarital sex.

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

The explanatory variables are gender given having children or not, age, years since married, religious, rating of marriage and education level.I change the age baseline from 0 to 32 since half of the samples are over32. And I change the baseline of ratings to average, we can easily see the impact of having a good rating or bad rating of marriage on the frequency of extramarital sex. The baseline is the a 32 year old man with no education and having average rating of marriage and with no children and religious. With the above table, we can see the lager age, medium religious, high religious and higher rating to the marriage, they all have a negative impact on the log odds. For the gender given having children or not, Although they have the negative impact on the estimate, but since their p value are all over 0.05. The other variables have a positive impact on the probability of having affairs.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.790	0.873	-2.051	0.040
agea	-0.042	0.018	-2.267	0.023
yearsmarried	0.092	0.033	2.802	0.005
religiousanti	0.954	0.367	2.601	0.009
religiouslow	0.361	0.279	1.297	0.195
religiousmed	-0.567	0.284	-1.995	0.046
religioushigh	-0.461	0.381	-1.210	0.226
ratings	-0.485	0.093	-5.219	0.000
education	0.035	0.046	0.761	0.447
genderfemale:childrenno	-0.773	0.400	-1.934	0.053
gendermale:childrenno	-0.253	0.368	-0.686	0.492
genderfemale:childrenyes	-0.264	0.255	-1.036	0.300

We collected the data from the The research of interest is having children is more likely to effect the frequency of women and men having affairs. For woman, the effect is negative, for man ,it is positive. We set having children will have negative effect on the affairs to women but positive to men to be the null hypothesis. And the alternative hypothesis is having children does not effect having affairs to women and men.

With the generalized linear model, we can conclude the fact that coefficient shows being a women do have small negative impact on having affairs. Being a male means higher probability having affairs. Having a children do have a larger negative impact on the probability of having affairs. But since the p-value is significant, we have enough evidence to reject the null hypothesis. Such that given having children or not does not have an effect to women and men on the probability of having affairs. With the same way, we can see the evidence to reject that education affects the probability of having affairs is also significant which the p-value is larger than 0.05. The probability of having affairs is affected by the following variables in this investigation, age, years since married, religious and rating to the marriage.

Smoking

nan Jiang

2020/10/10

#Summary
From the data set collect in American schools, we are interested to find out the variables that affect the probability of have ever smoked or not. The first question we are interested with is that is Americans lived in rural more likely to smoke than the urban area. We made a generalized linear model to find out the pattern inside the data. The below table shows the estimated coefficients for both model we made. This is the effect of log odds.

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

$$log(\frac{\mu_i}{1-\mu_i}) = X_i^T \beta$$

. The two model have the same variables, they are race, area sex and age. But for the first model, μ_i means the probability of ever smoked cigarette or not. For the second model, μ_i means the probability of ever smoked e- cigarette or not.
Smoking of cigars, cigars, cigarillos or little cigars not as common as before among Americans of European ancestry. But it is still a major concern among young age teenagers in America., We are interested in finding out the cigar smoking is a rural phenomenon or not and is the white Americans more likely to live in rural areas. The second question we are interest is that if age affects their probability on the have used electronic cigarettes or not. We used the data from the 2019 American National Youth Tobacco Survey towards the American school children which includes different race, sex, area and age.
Since the data is the ever smoked cigar or not, we used a binomial regression rather than the others to predict the probability that we are curious. The first table gave us the estimated coefficients of ever smoked cigar or not. And the second table gave us the estimated coefficients of ever smoked e-cigarette or not.

	Estimate	Std. Error	z value	Pr(>|z|)
Racewhite	-2.310	0.050	-46.446	0
Raceblack	-1.881	0.065	-28.823	0
Racehispanic	-2.367	0.055	-42.738	0
Raceasian	-3.572	0.171	-20.837	0
Racenative	-2.027	0.206	-9.862	0
Racepacific	-1.878	0.280	-6.709	0
RuralUrbanRural	0.401	0.046	8.796	0
SexF	-0.381	0.046	-8.318	0
agere	0.374	0.012	31.324	0

```
Waiting for profiling to be done...
```

	est	2.5 %	97.5 %
Baseline prob	0.090	0.083	0.099
Raceblack	0.152	0.134	0.173

	est	2.5 %	97.5 %
Racehispanic	0.094	0.084	0.104
Raceasian	0.028	0.020	0.039
Racenative	0.132	0.087	0.194
Racepacific	0.153	0.086	0.258
RuralUrbanRural	1.494	1.366	1.634
SexF	0.683	0.624	0.747
agere	1.453	1.420	1.488

	Estimate	Std. Error	z value	Pr(>|z|)
Racewhite	-0.813	0.034	-23.961	0.000
Raceblack	-1.327	0.054	-24.625	0.000
Racehispanic	-0.902	0.038	-23.901	0.000
Raceasian	-1.822	0.091	-20.075	0.000
Racenative	-0.751	0.153	-4.904	0.000
Racepacific	-0.576	0.215	-2.680	0.007
RuralUrbanRural	0.131	0.033	3.930	0.000
SexF	-0.060	0.033	-1.823	0.068
agere	0.337	0.008	39.901	0.000

```
Waiting for profiling to be done...
```

	est	2.5 %	97.5 %
Baseline prob	0.307	0.293	0.322
Raceblack	0.265	0.238	0.295
Racehispanic	0.406	0.377	0.437
Raceasian	0.162	0.135	0.193
Racenative	0.472	0.348	0.635
Racepacific	0.562	0.367	0.853
RuralUrbanRural	1.140	1.068	1.217
SexF	0.942	0.883	1.005
agere	1.400	1.377	1.424

We can see that for people lived in rural tends to have a 1.4918 higher probability than the Americans lived in urban. Such that the hypothesis one is true since we do not have enough evidence to reject the hypothesis. For the second question, we use the second model and the estimated coefficients. Since the p-value is below 0.05, we can see there will be a 0.33 effect on the log odds with every age increased. Such that is hypothesis for the second question is true.

#Report

The data set we collected is from 2019 American National Youth Tobacco Survey with questions about age, chewing tobaccos and chewing e-cigarettes. The sample is taken from the American school students. The research we are interested is the area of smoke habit difference. The rural students tends to smoke more cigarette than the urban students. And the age influence on chewing e-cigarettes is also one problem.

Since it is a m success in n trays problem, we used a binomial regression instead of poison and gamma regression. And since the problem is based on the probability, we used the generalized linear model to solve the problem. I recentered the age to 14 since the medium is 14 for the age.

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

$$log(\frac{\mu_i}{1-\mu_i}) = X_i^T \beta$$

The explanatory variable are sex, age, area and race. Except sex, the other variable all have the p-value below 0.05. The other race all have an negative impact on the probability. And female tends to have negative impact on the probability. The others all have the positive influence.

We make an null hypothesis that the rural area people tends to have a higher probability of smoking than in the urban area. The alternative hypothesis is that the area does not affects the probability. Since the data is not significant, we do not have enough evidence to reject the null hypothesis. Such that the hypothesis one is true.

For the second question Assume all conditions are equal, one age increase will conduct an 0.33 increase on the probability since 0.33 is the coefficients for the age variable. Null hypothesis that the older people to have a higher probability of smoking e-cigerattes than younger people with all other conditions equal. The alternative hypothesis that the age does not affects the probability. And with the p-value below 0.05, we do not have enough evidence to reject the null hypothesis.

Appendix

nan Jiang

2020/10/10

```
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'))
#knitr::kable(summary(glm(ever ~ gender:children + age + yearsmarried + religious,data=Affairs, family=
#knitr::kable(quantile(Affairs$age))
#knitr::kable(quantile(Affairs$yearsmarried))
#knitr::kable(quantile(Affairs$education))

Affairs$ratings = Affairs$rating - 3
agea = Affairs$age - 32
knitr::kable(summary(glm(
ever ~ gender:children + agea + yearsmarried + religious + ratings + education,
data=Affairs, family='binomial'))$coef, digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.790	0.873	-2.051	0.040
agea	-0.042	0.018	-2.267	0.023
yearsmarried	0.092	0.033	2.802	0.005
religiousanti	0.954	0.367	2.601	0.009
religiouslow	0.361	0.279	1.297	0.195
religiousmed	-0.567	0.284	-1.995	0.046
religioushigh	-0.461	0.381	-1.210	0.226
ratings	-0.485	0.093	-5.219	0.000
education	0.035	0.046	0.761	0.447
genderfemale:childrenno	-0.773	0.400	-1.934	0.053
gendermale:childrenno	-0.253	0.368	-0.686	0.492
genderfemale:childrenyes	-0.264	0.255	-1.036	0.300

```
#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))
#quantile(smokeSub$Age,na.rm = TRUE)
load("C:\\Users\\jn405\\Downloads\\smoke.RData")
smokeSub = smoke[which(smoke$Age >= 10), ]
smokeSub$agere = smokeSub$Age - 14
```

```
smokemodel = glm(ever_cigars_cigarillos_or ~ 0+ Race+RuralUrban+Sex + agere ,
family=binomial, data=smokeSub)
knitr::kable(summary(smokemodel)$coef, digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
Racewhite	-2.310	0.050	-46.446	0
Raceblack	-1.881	0.065	-28.823	0
Racehispanic	-2.367	0.055	-42.738	0
Raceasian	-3.572	0.171	-20.837	0
Racenative	-2.027	0.206	-9.862	0
Racepacific	-1.878	0.280	-6.709	0
RuralUrbanRural	0.401	0.046	8.796	0
SexF	-0.381	0.046	-8.318	0
agere	0.374	0.012	31.324	0

```
logOdds = cbind(est = smokemodel$coef, confint(smokemodel,level = 0.95))
```

```
## Waiting for profiling to be done...
```

```
oddsmat = exp(logOdds)
oddsmat[1,] = oddsmat[1,]/(1 + oddsmat[1,])
rownames(oddsmat)[1] = 'Baseline prob'
knitr::kable(oddsmat, digits = 3)
```

	est	2.5 %	97.5 %
Baseline prob	0.090	0.083	0.099
Raceblack	0.152	0.134	0.173
Racehispanic	0.094	0.084	0.104
Raceasian	0.028	0.020	0.039
Racenative	0.132	0.087	0.194
Racepacific	0.153	0.086	0.258
RuralUrbanRural	1.494	1.366	1.634
SexF	0.683	0.624	0.747
agere	1.453	1.420	1.488

```
esmokemodel = glm(ever_ecigarette ~ 0+ Race+RuralUrban+Sex + agere ,
family=binomial, data=smokeSub)
knitr::kable(summary(esmokemodel)$coef, digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
Racewhite	-0.813	0.034	-23.961	0.000
Raceblack	-1.327	0.054	-24.625	0.000
Racehispanic	-0.902	0.038	-23.901	0.000
Raceasian	-1.822	0.091	-20.075	0.000
Racenative	-0.751	0.153	-4.904	0.000
Racepacific	-0.576	0.215	-2.680	0.007
RuralUrbanRural	0.131	0.033	3.930	0.000

	Estimate	Std. Error	z value	Pr(>|z|)
SexF	-0.060	0.033	-1.823	0.068
agere	0.337	0.008	39.901	0.000

```
elogOdds = cbind(est = esmokemodel$coef, confint(esmokemodel,level = 0.95))
```

```
## Waiting for profiling to be done...
```

```
eoddsmat = exp(elogOdds)
eoddsmat[1,] = eoddsmat[1,]/(1 + eoddsmat[1,])
rownames(eoddsmat)[1] = 'Baseline prob'
knitr::kable(eoddsmat, digits = 3)
```

	est	2.5 %	97.5 %
Baseline prob	0.307	0.293	0.322
Raceblack	0.265	0.238	0.295
Racehispanic	0.406	0.377	0.437
Raceasian	0.162	0.135	0.193
Racenative	0.472	0.348	0.635
Racepacific	0.562	0.367	0.853
RuralUrbanRural	1.140	1.068	1.217
SexF	0.942	0.883	1.005
agere	1.400	1.377	1.424