STA442 Homework1-Generalized Linear Models

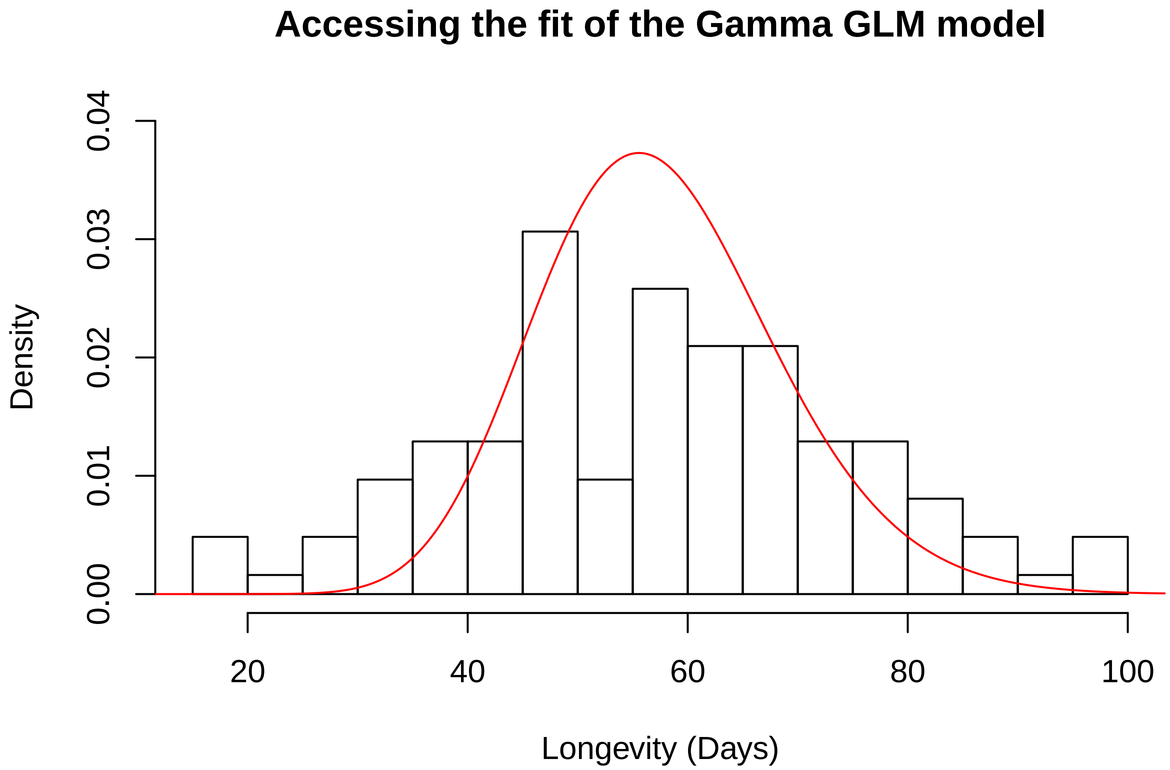
Yingzhu Zhang

24/09/2019

1. Flies

Brief Report

This dataset contains observations on five groups of male fruit flies —- 25 fruit flies in each group —- from an experiment designed to test if sexual activities affect longevity for male fruit flies. We plot the histogram of fruit flies' longevity and the density of Gamma distribution. In the histogram below, the red curve line is the density of a Gamma distribution, which appears to fit the distribution of the dataset. Hence, we can continue with Gamma generalized linear model in the analysis, with longevity being the dependent variables(y), normalized thorax length and sexual activities being the explanatory variables(x).



The thorax length of each male was measured as this was known to affect lifetime. Hence, we first normalize the 'thorax' variable in order to investigate the correlation between life expectancy and sexual activities only. After we fit the Gamma regression model, we found that fruit flies that cohabitate with virgin flies (whether it's 1 or 8 flies) have p-values (p=0.03 & p=0.00) less than the significant level 0.05, providing strong evidence that sexual activities of fruit flies can affect their longevities. The p-values are given in the summary Table 1 below:

Table 1: Estimated parameters of fruitfly on glm model

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	4.10	0.04	108.33	0.00

	Estimate	Std. Error	t value	Pr(>|t|)
Normalized Thorax Length	0.20	0.02	11.80	0.00
With 1 Pregnant Fly	0.06	0.05	1.04	0.30
With 1 Virgin Fly	-0.12	0.05	-2.18	0.03
With 8 Pregnant Flies	0.08	0.05	1.52	0.13
With 8 Virgin Flies	-0.41	0.05	-7.69	0.00

Furthermore, the Table 2 below shows the exponentiated parameter estimates, we found that flies cohabitating with one virgin fly lived 11% shorter than flies in isolation, and that flies cohabitating with 8 virgin flies can only live 66% of lifespan of flies in isolation.

Table 2: Expoentiated estimated parameters of fruitfly on glm model

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	60.20	1.04	1.117457e+47	1.00
Normalized Thorax Length	1.23	1.02	1.338457e+05	1.00
With 1 Pregnant Fly	1.06	1.05	2.820000e+00	1.35
With 1 Virgin Fly	0.89	1.05	1.100000e-01	1.03
With 8 Pregnant Flies	1.09	1.06	4.590000e+00	1.14
With 8 Virgin Flies	0.66	1.06	0.000000e+00	1.00

Moreover, as shown in Table 3 below, the longevity means of fruit flies that cohabitating with 1 virgin flies live 7 days shorter than flies in isolation, and flies cohabitating with 8 virgin flies live 25 days shorter compare to those in isolation. There is an inverse association between the frequency of sexual activities of fruit flies and their lifespan. The more frequently fruit flies mating, the shorter their lifespans are.

Table 3: Longevity means of each fruit fly group

Sexual activity	Longevity (Days)
isolated	64
one	65
low	57
many	65
high	39

Research News

"Can Having Sex Be Dangerous to Male Fruit Flies?"

The answer might be YES, according to a dataset contains 125 fruit flies' longevity, thorax length, and sexual activities. Male fruit flies were forced to cohabitate with either one or many female flies, either virgin or pregnant (unwilling to mate). The observation recorded each male fruit flies' thorax length, sexual activities and longevities. An isolated male fruit fly can live up to 65 days. Flies that are low sexually active have a 30% (18 days) longer lifespan than those who mating more frequently. A highly sexually active fruit fly can only 39 days! That's 26 days shorter than fly that has never mated in his life! Unfortunately for male fruit flies, having sex can be dangerous, even life-threatening to them. To live a longer life, men fruit flies probably need to sacrifice reproduction for their own good.

2. Smoking

Summary

We analyzed the results of the 2014 American National Youth Tobacco Survey to look for indicators that correlated with increases in the odds of chewing tobacco regularly or trying a hookah. We found that white people are more likely to chew tobacco, and black and Hispanic people, who are 20% and half less likely to chew tobacco regularly. Not surprisingly, older male living in rural area have the most odds of chewing tobacco frequently. For hookah and waterpipe use, unexpectedly, the difference between male and female are not significantly different. Similar to tobacco, older people are more likely to try hookah, as well as Hispanic people. The black people, on the other hand, have less chance of trying hookah. This is a reverse trend as we saw in chewing tobacco.

Introduction

We analyzed a dataset from 2014 American National Youth Tobacco Survey using R studio. The data was collected from American school children. We wanted to discuss the relationships between demographics and smoking habits. In this analysis, we focus on the trend of chewing tobacco, sniff or dip among different races and living area. We also investigated whether the likelihood of trying a hookah or waterpipe on at least one occasion is the same for different sexes, provided their age, ethnicity, and other demographic characteristics are similar. For both of the analysis, we controlled age and area the individual lives (urban or rural) to see the effect of races only. Finally, we use logistic regression models to further investigate the questions discussed above.

Method

For our analysis, the questions focus on White, Black, and Hispanic people. However, we still included the rest of the races since there is no significant difference upon the exclusion of the other races in the dataset. After checking the data, we found that the minimum of participants' age is 9 years old, which looks suspicious and implausible, so we remove the 9-year-old dataset from the model. Since we are investigating the odds of chewing tobacco and ever using a hookah, it is clear that we should use logistic regression model. Mathematically, the GLM model is described as:

$$log(odds) = \beta_0 + \beta_1 X_{age} + \beta_2 I_{female} + \beta_3 I_{black} + \beta_4 I_{hispanic} + \beta_5 I_{asian} + \beta_6 I_{native} + \beta_7 I_{pacific} + \beta_8 I_{rural}$$

where:

- odds = the odds of regularly chewing tobacco OR the odds of ever using a hookah, depending on context
- β_i = estimated parameters of each demographic group, where β_0 is the intercept representing white population
- X_{age} = the age of participants in the study
- I = indicators of each demographic group

Specifically, we used hypothesis test to see whether race is a significant predictor of chewing tobacco:

$$H_0 : \beta_0 = \beta_3 = \beta_4 = 0$$

$$H_a : at\ least\ one\ of\ \beta_0, \beta_3, \beta_4\ is\ not\ 0$$

The hypothesis test of whether gender is a significant predictor of using a hookah:

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

Furthermore, to get the likelihood difference among groups, we exponentiated the estimated parameters of our models to check the odds ratio increase/decrease between groups (lower and upper bound are included in the tables at the end of appendix). Then, we check the p-values of each predictor to see whether it is a significant predictor of the event. Finally, we compare the value of exponentiated estimated parameters (β_i) from each model to determine which predictor has the most significant effect on predicting chewing tobacco or using hookah respectively.

Results

Table 4: Estimated parameters of regular use of Chewing tobacco on glm model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-8.08	0.34	-24.00	0.0
Age	0.34	0.02	16.20	0.0
Female	-1.79	0.11	-16.48	0.0
Black	-1.56	0.17	-9.06	0.0
Hispanic	-0.71	0.10	-6.88	0.0
Asian	-1.55	0.34	-4.52	0.0
Native	0.11	0.28	0.39	0.7
Pacific	1.01	0.36	2.81	0.0
Rural	0.95	0.09	10.88	0.0

Table 4 above shows the p-values of each predictors, for the three races (white, black, and Hispanic) we need to discuss, all of them are significant predictors of chewing tobacco. However, in order to see the odds ratio increase/decrease between groups, we need to check the exponentiated estimated parameters, which are shown in Table 5 below:

Table 5: Modeling odds of regular use of Chewing tobacco

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.00	1.40	0.00	1.00
Age	1.40	1.02	10900445.05	1.00
Female	0.17	1.11	0.00	1.00
Black	0.21	1.19	0.00	1.00
Hispanic	0.49	1.11	0.00	1.00
Asian	0.21	1.41	0.01	1.00
Native	1.11	1.32	1.47	2.01
Pacific	2.75	1.43	16.56	1.01
Rural	2.59	1.09	52876.79	1.00

As we can see in Table 5, older people are 40% more likely to chew tobacco regularly then younger people, and the odd for female chewing tobacco is only 17% likely as male. For the black people and Hispanic people, the are 21% and 49% as likely to chew tobacco as whites respectively. Additionally, people who lived in rural area are 2.59 times likely to chew tobacco, sniff or dip regularly compare to the people who lived in urban area.

Table 6: Estimated parameters of regular use of having used a hookah or waterpipe on glm model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-8.00	0.19	-43.11	0.00
Age	0.42	0.01	36.27	0.00
Female	0.04	0.04	0.98	0.33
Black	-0.63	0.07	-9.01	0.00
Hispanic	0.35	0.05	7.14	0.00
Asian	-0.63	0.12	-5.36	0.00
Native	0.16	0.19	0.84	0.40
Pacific	0.96	0.27	3.57	0.00
Rural	-0.39	0.04	-8.77	0.00

After analyzing the estimated parameters of regular use of hookah or waterpipe in Table 6 above, we found a similar result as chewing tobacco, that race is a significant predictor. However, the pvalue for female is 0.33, which is larger than 0.05. This provided weak evidence that the use of hookah is different between genders, we cannot reject the null, hence concluded that gender does not have a significant effect on hookah or waterpipe usage.

Table 7: Modeling odds of having used a hookah or waterpipe

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.00	1.20	0.000000e+00	1.00
Age	1.52	1.01	5.623341e+15	1.00
Female	1.04	1.04	2.670000e+00	1.39
Black	0.53	1.07	0.000000e+00	1.00
Hispanic	1.41	1.05	1.258300e+03	1.00
Asian	0.53	1.12	0.000000e+00	1.00
Native	1.17	1.21	2.310000e+00	1.49
Pacific	2.62	1.31	3.539000e+01	1.00
Rural	0.68	1.05	0.000000e+00	1.00

Similarly, in Table 7, we found that older people are over half more likely to chew tobacco regularly then younger people, black people are half less likely to try hookah, and Hispanic people are 41% more likely than white people to try hookah. One opposite trend we found in this analysis is that people who live in rural area actually are 68% as likely to try hookah as people in the urban area. This phenomenon maybe caused by the diversity and multicultural in the cities.

Finally, to recall, although female is about 4% more likely to try hookah than male, the p-value is not statistically significant. Hence, we can conclude that the likelihood of having used a hookah or waterpipe on at least one occasion is the same for the different sexes, controlling age, ethnicity, and other demographic characteristics.

Appendix

```
# Load the data
data('fruitfly', package='faraway')
# Normalize thorax
thorax_norm = ((fruitfly$thorax - mean(fruitfly$thorax))/sqrt(var(fruitfly$thorax)))
# Fit the dataset to gamma glm
longFit = glm(fruitfly$longevity ~ thorax_norm + fruitfly$activity,
              family=Gamma(link=log), data=fruitfly)
names(longFit$coefficients) <- c('Intercept', 'Normalized Thorax Length',
                                 'With 1 Pregnant Fly',
                                 'With 1 Virgin Fly', 'With 8 Pregnant Flies',
                                 'With 8 Virgin Flies')
knitr::kable(summary(longFit)$coef, digits = 2,
             cap = 'Estimated parameters of fruitfly on glm model')
knitr::kable(exp(summary(longFit)$coef), digits = 2,
             cap = 'Expoentiated estimated parameters of fruitfly on glm model')
# Get the longeivity means of each groups
mean_long = aggregate(fruitfly$longevity, by=list(fruitfly$activity), mean)
knitr::kable(mean_long, digits = 0, col.names=c('Sexual activity',
                                                'Longevity (Days)'),
             cap = 'Longevity means of each fruit fly group')
# Plot histogram to check the gamma model
shape = 1/summary(longFit)$dispersion
scale = mean(fruitfly$longevity)/shape
hist(fruitfly$longevity, prob = TRUE, ylim = c(0, 0.04),
     breaks=15,
     xlab='Longevity (Days)',
     main='Accessing the fit of the Gamma GLM model')
x = seq(0, 120, len=1000)
lines(x, dgamma(x, shape = shape, scale = scale), col = 'red')

# Load the smoking data
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext='.RData'))
download.file(smokeUrl, smokeFile)
(load(smokeFile))
# Hypothesis 1: Regular use of chewing tobacco, snuff or dip
smokeFormats[smokeFormats$colName == 'chewing_tobacco_snuff_or', ]
smoke$chew = factor(smoke$chewing_tobacco_snuff_or,
                         levels=c('TRUE', 'FALSE'), labels=c('yes','no'))
# Nine year olds look suspicious, get rid of missings and age 9
smokeSub1 = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                    !is.na(smoke$chew) &
                    !is.na(smoke$chewing_tobacco_snuff_or) &
                    !is.na(smoke$Sex), ]
smokeAgg1a = reshape2::dcast(smokeSub1,
                             Age + Sex + Race + RuralUrban ~ chew, length)
smokeAgg1b = na.omit(smokeAgg1a)
smokeAgg1b$y = cbind(smokeAgg1b$yes, smokeAgg1b$no)
# Fit the model
smokeFit1 = glm (y ~ Age + Sex + Race + RuralUrban,
                 family = binomial(link='logit'), data = smokeAgg1b)
```

```r
names(smokeFit1$coefficients) <- c('Intercept', 'Age', 'Female',
                                   'Black', 'Hispanic', 'Asian',
                                   'Native', 'Pacific', 'Rural')
knitr::kable(summary(smokeFit1)$coef, digit=2,
             cap='Estimated parameters of regular use of
             Chewing tobacco on glm model')
knitr::kable(exp(summary(smokeFit1)$coef), digit=2,
             cap='Modeling odds of regular use of Chewing tobacco')
# Hypothesis 2: The likelihood of having used a hookah or waterpipe
smokeFormats[smokeFormats$colName == 'ever_tobacco_hookah_or_wa', ]
smoke$ever = factor(smoke$ever_tobacco_hookah_or_wa,
                    levels=c('TRUE', 'FALSE'), labels=c('yes','no'))
# Nine year olds look suspicious, get rid of missings and age 9
smokeSub2 = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                    !is.na(smoke$ever) &
                    !is.na(smoke$ever_tobacco_hookah_or_wa) &
                    !is.na(smoke$Sex), ]
smokeAgg2a = reshape2::dcast(smokeSub2,
                             Age + Sex + Race + RuralUrban ~ ever, length)
smokeAgg2b = na.omit(smokeAgg2a)
smokeAgg2b$y = cbind(smokeAgg2b$yes, smokeAgg2b$no)
# Fit the model
smokeFit2 = glm (y ~ Age + Sex + Race + RuralUrban,
                 family = binomial(link='logit'), data = smokeAgg2b)
names(smokeFit2$coefficients) <- c('Intercept', 'Age', 'Female',
                                   'Black', 'Hispanic', 'Asian',
                                   'Native', 'Pacific', 'Rural')
knitr::kable(summary(smokeFit2)$coef, digit=2,
             cap='Estimated parameters of regular use of
             having used a hookah or waterpipe on glm model')
knitr::kable(exp(summary(smokeFit2)$coef), digit=2,
             cap='Modeling odds of having used a hookah or waterpipe')
# Odds ratio of regular use of Chewing tobacco
smokeTable1 = as.data.frame(summary(smokeFit1)$coef)
smokeTable1$lower = smokeTable1$Estimate - 2*smokeTable1$'Std. Error'
smokeTable1$upper = smokeTable1$Estimate + 2*smokeTable1$'Std. Error'
smokeOddsRatio1 = exp(smokeTable1[,c('Estimate','lower','upper')])
rownames(smokeOddsRatio1)[1] = 'baseline prob'
smokeOddsRatio1[1,] = smokeOddsRatio1[1,]/(1+smokeOddsRatio1[,1])
knitr::kable(smokeOddsRatio1, digits=2,
             cap="Odds ratio of regular use of Chewing tobacco")
# Odds ratio of regular use of having used a hookah or waterpipe
smokeTable2 = as.data.frame(summary(smokeFit2)$coef)
smokeTable2$lower = smokeTable2$Estimate - 2*smokeTable2$'Std. Error'
smokeTable2$upper = smokeTable2$Estimate + 2*smokeTable2$'Std. Error'
smokeOddsRatio2 = exp(smokeTable2[,c('Estimate','lower','upper')])
rownames(smokeOddsRatio2)[1] = 'baseline prob'
smokeOddsRatio2[1,] = smokeOddsRatio2[1,]/(1+smokeOddsRatio2[,1])
knitr::kable(smokeOddsRatio2, digits=2,
             cap="Odds ratio of regular use of having used a hookah or waterpipe")
```