Homework 1

Xinqi Shen

Question 1

Does Sexual Activity Affect the Lifetime of Male Fruit Flies?

Intruduction

The dataset is coming from the Faraday package. Fruit flies were randomly divided into 5 different groups: live alone, live with either one or eight virgin flies and live with either one or eight pregnant flies. The pregnant fruit flies will not mate with male flies compared with virgin flies. Thus, we want to analysis the effect of the sexual activity and thorax length on the life span of fruit flies.

Method

We fit a Gamma Regression model to control the effect of throax size in different groups. Given the following model:

$$Y_i \sim Gamma(\mu_i/\nu, \nu)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{thorax} + \beta_2 I_{1Pregnant} + \beta_3 I_{1Virgin} + \beta_4 I_{8Pregnant} + \beta_5 I_{8Virgin}$$

The Gamma GLM was a roughly good fit to the fly data, we may choose better distribution in the future.

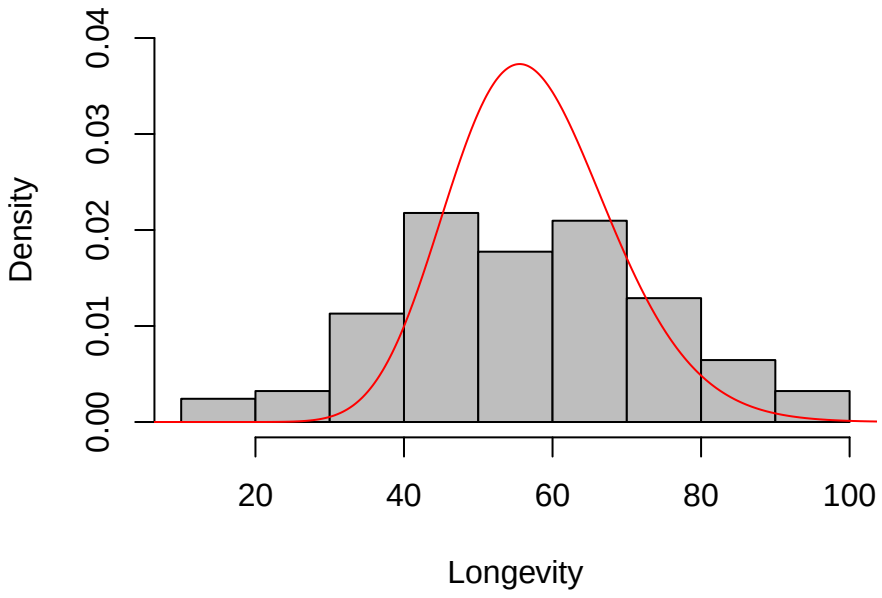

Figure 1: The density of the fitted Gamma distribution

Results

The boxplot shows that fruit flies that live alone or live with pregnant flies have larger mean lifespan.

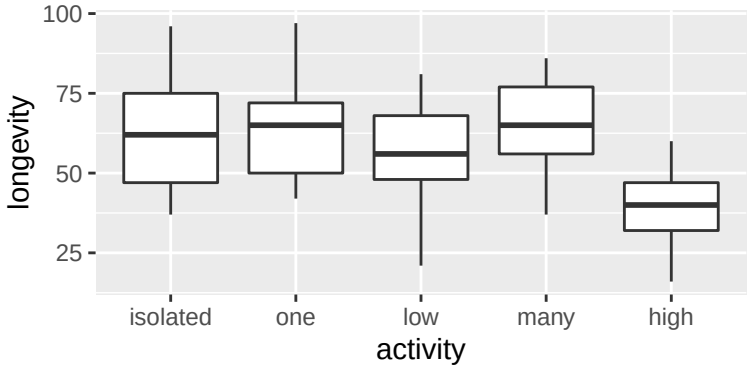

Figure 2: The longevity of each group

We can discuss the p values and exponential estimated values based on the output from Gamma GLM. Firstly, there is no significant difference between flies that live with pregnant flies (no matter how many, 1 or 8) and live alone, if we pick the significant level 5%. Noted that both p values are greater than 0.05. In other words, living with pregnant flies have the similar lifespan with flies in isolation. However, living with virgin flies do significantly affect the lifespan of flies, given their p value less than 0.05. Secondly, in terms of the exponential estimated value of parameters, the estimated lifespan will decrease around 11% for flies live with 1 virgin fly and drcrease around 34% for flies live with 8 virgin flies compared with isolated flies. Thus, it can imply that living with more virgin flies will lead to shorter lifespan. In other words, the mating frequency of fruit flies will have extremely negative effect on their lifespan.Finally, the longer throax length also leads to longer lifespan based on the output.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	60.20	0.04	108.33	0.00
Thorax Length	1.23	0.02	11.80	0.00
Live with 1 Pregnant Fly	1.06	0.05	1.04	0.30
Live with 1 Virgin Fly	0.89	0.05	-2.18	0.03
Live with 8 Pregnant Fly	1.09	0.05	1.52	0.13
Live with 8 Virgin Fly	0.66	0.05	-7.69	0.00

Table 1: Estimated parameters from Gamma Regression model

Summary

From the analysis of the result, Fruit flies will live shorter if they live with more virgin flies. In other words, the result implies the lifespan of the male fruit flies will remarkablely decrease with the increase of mating behaviours. However, It does not mean the male flies will reduce their lifespan when living with female flies(pregnant flies).Actually, living with one or many female pregnant flies do not affect their lifespan. Besides, in terms of thorax legnth, the longer thorax length will cause longer lifespan.

Appendix

```r
data('fruitfly', package='faraway')
thorax_mean = mean(fruitfly$thorax)
thorax_var = var(fruitfly$thorax)
thorax_new = (fruitfly$thorax - thorax_mean) / sqrt(thorax_var)
flyfit = glm(fruitfly$longevity ~ thorax_new + fruitfly$activity,
             family = Gamma(link = 'log'),data = fruitfly)
hist(fruitfly$longevity,prob=TRUE,ylim=c(0,0.04),col="grey",
     border="black",main="",xlab = "Longevity")
shape1 = 1/summary(flyfit)$dispersion
scale1 = mean(fruitfly$longevity)/shape1
xSeq = seq(1,120,len = 1000)
lines(xSeq, dgamma(xSeq,shape = shape1,scale = scale1),col = "red")
library(ggplot2)
ggplot(data=fruitfly,aes(x=activity, y=longevity)) +
  geom_boxplot()
new_data = summary(flyfit)$coef
new_data[,1] = exp(new_data[,1])
rownames(new_data) = c("Intercept", "Thorax Length","Live with 1 Pregnant Fly",
                       "Live with 1 Virgin Fly","Live with 8 Pregnant Fly",
                       "Live with 8 Virgin Fly")
knitr::kable(new_data, digits=2, format='latex')
```

Question 2

Whether gender and ethnicity have an impact on smoking habits?

Summary

Based on the analysis of the result of the dataset from the 2014 American National Youth Tobacco Survey, we found some intereting results. White people are more likely chewing tobacco than hispanic and black people, with the odds are 1/2 and 1/5 compared with white people. Additionally, people living in rural areas are more likely chewing tobacco. In terms of the odds of having used a hookah or waterpipe on at least one occasion, we explored that sex has no effect. Also, Pacific people are more likely to use hookah, but Black and Asian people are less likely. Finally, older people have higher odds of using hookah, which is not surprising.

Introduction

The dataset is coming from the 2014 American National Youth Tobacco Survey. We want to analyze the problem related to the use of cigars, hookahs, and chewing tobacco amongst American school children. In this report, we will mainly focus on the relationship between the habit of smoking and race which we controlled for living place(rural/urban area) and age. Also, we will investigate whether sex affect the use of hookah or waterpipe on at least one occasion, given their age, ethnicity, and other demographic characteristics.

Methods

We will fit a logistic regression model to analysis these relationships, since the response variable are both binary — "chewing_tobacco_snuff_or" and "ever_tobacco_hookah_or_wa". Thus, we will study the odds of regularly chewing tobacco, snuff or dip, as well as the odds of ever smoked tobacco out of a hookah or waterpipe to explore their relations. Given the following full model:

$$\log Odds = \beta_0 + \beta_1 x_{Age} + \beta_2 I_{Female} + \beta_3 I_{Black} + \beta_4 I_{Hispanic} + \beta_5 I_{Asian} + \beta_6 I_{Native} + \beta_7 I_{Pacific} + \beta_8 I_{Rural}$$

Based on our full model, we can test the significant level for each estimator to determine the effect of different predictors. Also, we can test different null hypothesis compared with reduced model to analyze the result. In order to test whether the use of chewing tobacco,snuff or dip is different between white people and hispanic/black people, the null hypothesis is: $H_0 : \beta_3 = 0$ or $\beta_4 = 0$. Additionally, in order to test whether gender affect using hookah or waterpipe, the null hypothesis is: $H_0 : \beta_2 = 0$. Finally, we will calculate the odds among different groups to get more intuitive result.

Results

	Estimate	Std. Error	z value	Pr(>|z|)	lower	upper
Intercept	0.048	0.083	-36.483	0.000	-3.198	-2.866
Age	1.400	0.021	16.204	0.000	0.295	0.378
Female	0.167	0.109	-16.481	0.000	-2.005	-1.571
Black	0.211	0.172	-9.064	0.000	-1.899	-1.213
Hispanic	0.490	0.104	-6.884	0.000	-0.920	-0.506
Asian	0.213	0.342	-4.519	0.000	-2.231	-0.862
Native	1.113	0.278	0.385	0.700	-0.448	0.662
Pacific	2.751	0.361	2.807	0.005	0.291	1.733
Rural	2.588	0.087	10.876	0.000	0.776	1.126

Table 2: The odds ratios of using chewing tobacco, snuff or dip

From the output of table 2, we find that age, sex, race and living area all affect the odds of chewing tobacco, snuff or dip, given their p values less that 0.05(ignore the Native one if we only consider race as one factor). More specific for our research hypotheses, the odds of chewing tobbcco among black people(African-Americans) and hispanic people(Hispanic-Americans) only account for about 22% and 53% relative to white people(European-Americans). Additionly, the result also verify the fact that chewing tobacco is a rural phenomenon for white people, since the odds of chewing tobacco are almost 2.6 times than people living in urban areas.

	Estimate	Std. Error	z value	Pr(>|z|)	lower	upper
Intercept	0.178	0.044	-39.226	0.000	-1.811	-1.636
Age	1.520	0.012	36.266	0.000	0.396	0.442
Female	1.043	0.043	0.980	0.327	-0.044	0.128
Black	0.530	0.070	-9.005	0.000	-0.776	-0.494
Hispanic	1.413	0.048	7.138	0.000	0.249	0.442
Asian	0.532	0.118	-5.362	0.000	-0.866	-0.396
Native	1.173	0.190	0.838	0.402	-0.221	0.540
Pacific	2.621	0.270	3.566	0.000	0.423	1.504
Rural	0.678	0.044	-8.769	0.000	-0.477	-0.300

Table 3: The odds ratios of ever using a hookah or waterpipe

From the output of table 3, age, race and living areas are all statistically significant, however, there is no significant difference between male and female by controlling other factors (p value:0.327 $>>$ 0.05). Thus, we do not have evidence to show the difference of having used a hookah or waterpipe in terms of gender. If we investigate age and living areas, older people are more likely to use a hookah or waterpipe and the odds of using hookah in rural areas is only the 68% of the odds of using hookah in urban areas. Finally, if we look at the race, Pacific peole are most likely using hookah or waterpipe(2.6 times the odds than whites) but Black and Asian people are most unlikely using it(0.53, half the odds compared with whites).

Appendix

```
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext='.RData'))
download.file(smokeUrl, smokeFile, mode='wb')
(load(smokeFile))

smokeSub = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                   !is.na(smoke$chewing_tobacco_snuff_or) & !is.na(smoke$Sex), ]
smokeAgg = reshape2::dcast(smokeSub,
    Age + Sex + Race + RuralUrban ~ chewing_tobacco_snuff_or,
    length)
smokeAgg = na.omit(smokeAgg)
smokeAgg$y = cbind(smokeAgg$'TRUE', smokeAgg$'FALSE')
smokeAgg$ageC = smokeAgg$Age - 15
smokeFit2 = glm(y ~ ageC + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg)
smoke_summary = as.data.frame(summary(smokeFit2)$coef)
smoke_summary$lower = smoke_summary$Estimate -
  2*smoke_summary$`Std. Error`
smoke_summary$upper = smoke_summary$Estimate +
  2*smoke_summary$`Std. Error`
smoke_summary[,1] = exp(smoke_summary[,1])
```

```
rownames(smoke_summary) = c("Intercept","Age","Female","Black","Hispanic","Asian",
                            "Native","Pacific","Rural")
knitr::kable(smoke_summary, digits=3,format='latex')

smokeSub1 = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                  !is.na(smoke$ever_tobacco_hookah_or_wa), ]
smokeAgg1 = reshape2::dcast(smokeSub1,
    Age + Sex + Race + RuralUrban ~ ever_tobacco_hookah_or_wa,
    length)
smokeAgg1 = na.omit(smokeAgg1)
smokeAgg1$y = cbind(smokeAgg1$'TRUE', smokeAgg1$'FALSE')
smokeAgg1$ageC = smokeAgg1$Age - 15
smokeFit2_1 = glm(y ~ ageC + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg1)
smoke_summary_1 = as.data.frame(summary(smokeFit2_1)$coef)
smoke_summary_1$lower = smoke_summary_1$Estimate -
  2*smoke_summary_1$`Std. Error`
smoke_summary_1$upper = smoke_summary_1$Estimate +
  2*smoke_summary_1$`Std. Error`
smoke_summary_1[,1] = exp(smoke_summary_1[,1])
rownames(smoke_summary_1) = c("Intercept","Age","Female","Black","Hispanic","Asian",
                              "Native","Pacific","Rural")
knitr::kable(smoke_summary_1, digits=3,format='latex')
```