STA442 Homework 1

Kevin Dang - 1003205079

Due 24 September 2019

Fruit Flies

Fruit flies have relatively short lifespans, which makes them a good candidate for experiments that involve collecting data on their lifetimes and other factors that may affect their lifetimes, including sexual activity and thorax length. Lifespan is measured in days, and we are not told what units thorax length is measured in however it is likely to be very small units such as millimeters. Sexual activity is divided into five categories, with 25 male fruit flies in each group: fly kept alone (`isolated`), one pregnant female (`activityone`), eighty pregnant females (`activitymany`), one virgin female (`activitylow`), and eight virgin females (`activityhigh`). The first three groups are the control groups which involve no mating while the latter two categories with virgin female fruit flies involve mating.

The model that is used to answer our questions regarding fruit fly lifetimes is a Gamma generalized linear model. In Gamma regression, the data follow a Gamma distribution with range parameter $\phi = \mu_i/\nu$ and shape parameter ν, i.e. $Y_i \sim Gamma(\mu_i/\nu)$ with $E(Y_i) = \mu_i$. We are using a log-link so we have a model of:

$$\log(\mu_i) = X_i\beta = \beta_{intercept} + \beta_{I(activity)}X_1 + \beta_{thorax}X_2$$

Table 1 below contains the output for the Gamma regression model, and Table 2 contains the 95% confidence intervals for the parameters. The intercept represents flies that were kept solitary, and this will be our baseline. Starting with β_1 (`activityone`), we have a $100(\exp(0.055) - 1) = 5.7\%$ chance of increase in lifespan if the fruit fly is kept with one pregnant female. If the fruit fly is kept with one virgin female (`activitylow`), then it represents a lifespan decrease of -11.0% ($100(\exp(-0.116) - 1) = -11.0\%$). Next is `activitymany`, also known as flies kept with eight pregnant females, which has an increase of 8.5% (see previous calculations). Finally, we have a decrease in longevity by 34.0% for `activityhigh`. The p-values for `activityone` and `activitymany` are quite large compared to the standard 0.05 significance level so these results are not significant. This makes sense because those two groups are also control groups which involve no mating. Thorax length appears to have the greatest effect, because it increases lifespan by a whopping 1370%. In Figure 1 we have a histogram with a Gamma density line as a visual aid to help us check the model fit. The red Gamma line follows the histogram shape quite well so the Gamma generalized linear model is a good fit to the fruit fly data.

Table 1: Gamma GLM, lifetimes as a function of the thorax length and activity

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.145	0.038	109.918	0.000
activityone	0.055	0.053	1.036	0.302
activitylow	-0.116	0.053	-2.184	0.031
activitymany	0.082	0.054	1.524	0.130
activityhigh	-0.415	0.054	-7.687	0.000
length	2.688	0.228	11.804	0.000

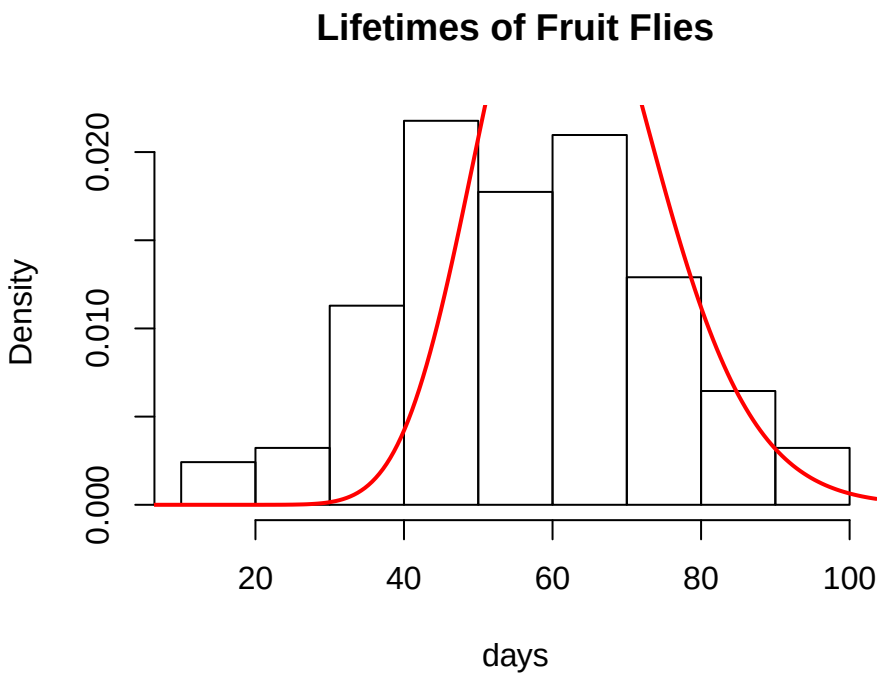

Figure 1: Histogram of Fruit Fly lifespan with density of a Gamma distribution

Table 2: 95% Confidence Intervals for Fly Gamma GLM

	2.5 %	97.5 %
(Intercept)	4.072	4.220
activityone	-0.049	0.160
activitylow	-0.221	-0.012
activitymany	-0.024	0.189
activityhigh	-0.520	-0.309
length	2.231	3.141

Fruit Flies Summary

A scientific laboratory conducted a study on how the level of sexual activity and thorax length affects the lifetimes of male fruit flies. Thorax length was measured for each fly and it had quite a large effect, increasing the probability of a longer than average lifespan by 1370%. Sexual activity also had an impact on the lifespans. The male fruit flies who were kept with one pregnant female or eight pregnant females are more likely to slightly live longer than the average isolated fruit fly, by 5.7% and 8.5% respectively, however these values may be due to random chance as the model showed that there was no significant difference. These groups are not able to mate so they are less sexually active. However the groups which were kept with one virgin female or eight virgin females have on average a decrease in lifetime by 11% and 34% respectively. These two groups are more sexually active, particularly the second group. In conclusion, fruit flies with lower sexual activity levels tend to live longer compared to the groups who have very high sexual activity levels.

Appendix A

Flies Code

```
library(tidyverse)
data('fruitfly', package='faraway')

fruitfly$length <- fruitfly$thorax - median(fruitfly$thorax)
flyglm <- glm(longevity ~ activity+length, family=Gamma(link='log'), data=fruitfly)
knitr::kable(summary(flyglm)$coef, digits=3,
             caption="Gamma GLM, lifetimes as a function of the thorax length and activity") # Table 1

knitr::kable(confint(flyglm), digits=3,
             caption="95% Confidence Intervals for Fly Gamma GLM ") # Table 2

# Code below borrowed from cars.r
shape = 1/summary(flyglm)$dispersion
hist(fruitfly$longevity, prob=TRUE,
     xlab='days', main = "Lifetimes of Fruit flies")
xSeq = seq(par('usr')[1], par('usr')[2], len=200)
lines(xSeq,
      dgamma(xSeq, shape=shape,
             scale = exp(flyglm$coef['(Intercept)'])/shape),
      col='red', lwd=2
)
```

Smoking

Introduction

The 2014 American National Youth Tobacco Survey is a rich data source that will allow us to answer various research hypotheses. The first problem that we want to investigate is whether the regular use of chewing tobacco, snuff or dip among three specific ethnic groups in America is different. The ethnic groups are European-American, Hispanic-American, and African-American. In this analysis, we must account for whether they live in an urban or rural area. The second hypothesis to investigate involves determining how likely it is for two youths of the opposite sex with similar characteristics to have used a hookah or waterpipe at least once. A sample of the data that will be used in the analysis is in Table 3 below.

Table 3: Smoking Data

Age	Sex	RuralUrban	Race	chewing_tobacco_snuff_or	ever_tobacco_hookah_or_wa
13	M	Urban	hispanic	FALSE	FALSE
12	F	Urban	hispanic	FALSE	FALSE
14	M	Urban	native	FALSE	FALSE
13	M	Urban	hispanic	FALSE	TRUE
14	M	Urban	native	FALSE	FALSE

Methods

The method used to tackle the first hypothesis is a logistic regression model. Looking at Tables 4 and 5 below, we can see that the potential predictors for our model are binary, which makes sense to use logistic regression in this scenario. The model is as follows: $Y_i \sim Bin(N_i, \mu_i)$ with $\log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$. Y_i is the number of people who have used chewing tobacco, snuff or dip on 1 or more days in the past 30 days and N_i is the number of observations. The probability of a person chewing tobacco, snuff or dip given covariates X_i (race, urban/rural location) is μ_i. The parameter of interest is β_{Race} and the confounders are $\beta_{RuralUrban}$ and $\beta_{intercept}$. Notice that in Table 4 the proportion of European-Americans who have tried chewing tobacco is largest, but in Table 5 the percentage of European-Americans who live in rural areas is also the highest. Chewing tobacco is more common in rural areas so we must account for that in the model. The model will be fit via a two way interaction:

$$\log(\frac{\mu_i}{1-\mu_i}) = \beta_{intercept} + \beta_{Race}X_1 + \beta_{RuralUrban}X_2 + \beta_{Race}\beta_{RuralUrban}X_3$$

For the second research hypothesis we also fit a logistic regression model. This time, Y_i represents the number of people who ever smoked tobacco out of a hookah or waterpipe. So the probability of a person who has smoked tobacco out of a hookah or waterpipe given X_i (sex, age, race) is μ_i. The parameter of interest is β_{Sex}, while the confounders are $\beta_{Age}, \beta_{Race}, \beta_{RuralUrban}, \beta_{intercept}$. The model is as follows:

$$\log(\frac{\mu_i}{1-\mu_i}) = \beta_{intercept} + \beta_{Sex}X_1 + \beta_{Age}X_2 + \beta_{Race}X_3 + \beta_{RuralUrban}X_4$$

Table 4: Chewing tobacco, snuff or dip, by Race

	FALSE	TRUE
white	9290	524
black	3312	39
hispanic	5796	130
asian	953	9
native	317	15
pacific	71	11

Table 5: Urban vs Rural, by Race

	Urban	Rural
white	4440	5440
black	1972	1452
hispanic	3713	2320
asian	758	213
native	146	188
pacific	44	41

Results

The logistic regression model for chewing tobacco is shown in Table 6 below along with the 95% confidence intervals in Table 7. The intercept represents European-Americans who live in an urban areas which is our reference. The odds ratio of an African-American in an urban area using chewing tobacco, snuff or dip regularly is exp(-1.006)=0.366, which means that the their odds are 100(exp(-1.006)-1)=-63.4% lower compared to the odds for urban European-Americans. For urban Hispanic-Americans the odds decrease is -30.7% (see previous calculation). If a European-American lives in a rural area then the percentage odds of using chewing tobacco, snuff or dip regularly increases by 220.9%. For a rural African-American the percentage odds changes by -50.6% and for rural Hispanic-Americans the percentage decrease is -48.1%. The most noticeable difference here is the huge increase in odds by more than three times for European-Americans in rural areas compared to urban areas. This is due to the fact chewing tobacco is a rural phenomenon. There was also an increase for the African-Americans, however the case is different for Hispanic-Americans who saw a decrease in likelihood. Perhaps Hispanic-Americans who live in rural areas happen to be more watchful over their children compared to those in the busy cities because of the fact that tobacco is more common there, but we need more data to confirm this. From these numbers, we can see that every group with the exception of rural European-Americans is less likely to have used chewing tobacco relative to the urban European-Americans. While accounting for urban and rural areas, we can see that regular use of chewing tobacco is more common amongst European-Americans than for Hispanic-Americans and African-Americans.

Table 6: Logistic Regression Model, Regular use of Chewing Tobacco

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.657	0.096	-38.042	0.000
Raceblack	-1.006	0.256	-3.936	0.000
Racehispanic	-0.367	0.159	-2.314	0.021
Raceasian	-1.573	0.510	-3.083	0.002
Racenative	0.318	0.465	0.683	0.495
Racepacific	1.092	0.607	1.800	0.072
RuralUrbanRural	1.166	0.109	10.701	0.000
Raceblack:RuralUrbanRural	-0.706	0.341	-2.070	0.038
Racehispanic:RuralUrbanRural	-0.656	0.208	-3.149	0.002
Raceasian:RuralUrbanRural	0.346	0.684	0.506	0.613
Racenative:RuralUrbanRural	-0.694	0.570	-1.218	0.223
Racepacific:RuralUrbanRural	0.013	0.726	0.018	0.986

Table 7: 95% Confidence Intervals for Tobacco GLM

	2.5 %	97.5 %
(Intercept)	-3.851	-3.474
Raceblack	-1.539	-0.532
Racehispanic	-0.682	-0.060
Raceasian	-2.757	-0.702
Racenative	-0.734	1.128
Racepacific	-0.338	2.126
RuralUrbanRural	0.956	1.383
Raceblack:RuralUrbanRural	-1.374	-0.029
Racehispanic:RuralUrbanRural	-1.065	-0.248
Raceasian:RuralUrbanRural	-1.009	1.765
Racenative:RuralUrbanRural	-1.774	0.508
Racepacific:RuralUrbanRural	-1.333	1.605

The logistic regression model for hookah or waterpipe is shown in Table 8 below along with the 95% confidence intervals in Table 9. The parameter of interest corresponds is `SexF`, which says that holding all other demographic characteristics fixed, there is a 100(exp(0.042)-1)=4.3% increase in the odds of using a hookah or waterpipe if the person is female. The p-value of 0.327 is quite large and the increase in odds is rather small so there is no significant difference between two very similar individuals of the opposite sex in the likelihood of using a hookah or waterpipe.

Table 8: Logistic Regression Model, Use of hookah or waterpipe

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.003	0.186	-43.111	0.000
SexF	0.042	0.043	0.980	0.327
Age	0.419	0.012	36.266	0.000
Raceblack	-0.635	0.070	-9.005	0.000
Racehispanic	0.346	0.048	7.138	0.000
Raceasian	-0.631	0.118	-5.362	0.000
Racenative	0.160	0.190	0.838	0.402
Racepacific	0.964	0.270	3.566	0.000
RuralUrbanRural	-0.388	0.044	-8.769	0.000

Table 9: 95% Confidence Intervals for Hookah GLM

	2.5 %	97.5 %
(Intercept)	-8.369	-7.642
SexF	-0.042	0.126
Age	0.396	0.441
Raceblack	-0.774	-0.498
Racehispanic	0.251	0.440
Raceasian	-0.867	-0.405
Racenative	-0.230	0.519
Racepacific	0.416	1.480
RuralUrbanRural	-0.475	-0.302

Smoke Summary

From this study, we conclude that regular use of chewing tobacco is more common amongst European-Americans than for Hispanic-Americans and African-Americans. An interesting takeaway from the results is that European-Americans in rural areas are more than three times as likely to use chewing tobacco than European-Americans in urban areas. The study also found that given two individuals of the opposite sex with very similar demographic characteristics such as age, race and location, the likelihood of having used a hookah or waterpipe at least once is the same.

Appendix B

Smoking Code

```
# Load in Data
dataDir = "../data"
smokeFile = file.path(dataDir, "smokeDownload.RData")
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/appliedstats/data/smoke.RData",
smokeFile)
}
load(smokeFile)

knitr::kable(smoke[1:5,c('Age','Sex','Grade','RuralUrban','Race','chewing_tobacco_snuff_or',
                         'ever_tobacco_hookah_or_wa')], caption="Smoking Data") # Table 3
smokeSub = smoke[smoke$Age >= 10, ] # 9 yr olds are suspicious, remove

knitr::kable(table(smokeSub$Race, smokeSub$chewing_tobacco_snuff_or),
             caption="Chewing tobacco, snuff or dip, by Race") # Table 4
knitr::kable(table(smokeSub$Race, smokeSub$RuralUrban),
             caption="Urban vs Rural, by Race") # Table 5

# Binomial GLM Regular use of Chewing Tobacco.
smokeglm1 <- glm(chewing_tobacco_snuff_or ~ Race*RuralUrban,data=smokeSub,
                 family=binomial(link="logit"))
knitr::kable(summary(tobaccoglm)$coef, digits=3,
             caption="Logistic Regression Model, Regular use of Chewing Tobacco") # Table 6
knitr::kable(confint(tobaccoglm), digits=3,
             caption="95% Confidence Intervals for Tobacco GLM ") # Table 7

# Binomial GLM Likelihood of hookah or waterpipe
smokeglm2 <- glm(ever_tobacco_hookah_or_wa ~ Sex+Age+Race+RuralUrban,data=smokeSub,
                 family=binomial(link="logit"))
knitr::kable(summary(hookahglm)$coef, digits=3,
             caption="Logistic Regression Model, Likelihood of hookah or waterpipe") # Table 8
knitr::kable(confint(hookahglm), digits=3,
             caption="95% Confidence Intervals for Hookah GLM") # Table 9
```