STA442 Homework 1, Generalized linear models

student number: 1003942326

Yulin WANG

2019-09-25

Question1: Flies

Problem: Do fertile women affect men's lifetime ?

According to a dataset from the Faraway(2005) package. Fruitflies were forced to cohabilitate with either one or many females, either fertile or pregnant (will not mate). The lifetime (in days) of 125 fruitflies were measured controlling for thorax length (which is known to affect lifetime). The mean lifetime(longevity) for each group is listed in the following table:

Table 1: Mean longevity of each fruitfly group

	Longevity (Days)
Isolated	64
With 1 Pregnant Fly	65
With 1 Virgin Fly	57
With 8 Pregnant Flies	65
With 8 Virgin Flies	39

Model and interpretation of coefficients

After nomarlizing thorax, we fit a Gamma generalized linear model with log link function to model the lifetimes as a function of the thorax length and activity. We found that flies cohabilitating with one virgin fly lived 11% less days than flies kept solitary, while flies with 8 virgin flies lived 34%(about a third) less days than flies kept solitary. This can be inferred from the exponentiated parameter estimates given in table 2.

Table 2: Estimated parameters from the Gamma generalized linear model of the fruitflies

	Exp. Estimate	Std. Error	t value	P-Value
Intercept	60.220	0.038	108.333	0.000
Thorax Length	1.226	0.017	11.804	0.000
With 1 Pregnant Fly	1.057	0.053	1.036	0.302
With 1 Virgin Fly	0.890	0.053	-2.184	0.031
With 8 Pregnant Flies	1.085	0.054	1.524	0.130
With 8 Virgin Flies	0.660	0.054	-7.687	0.000

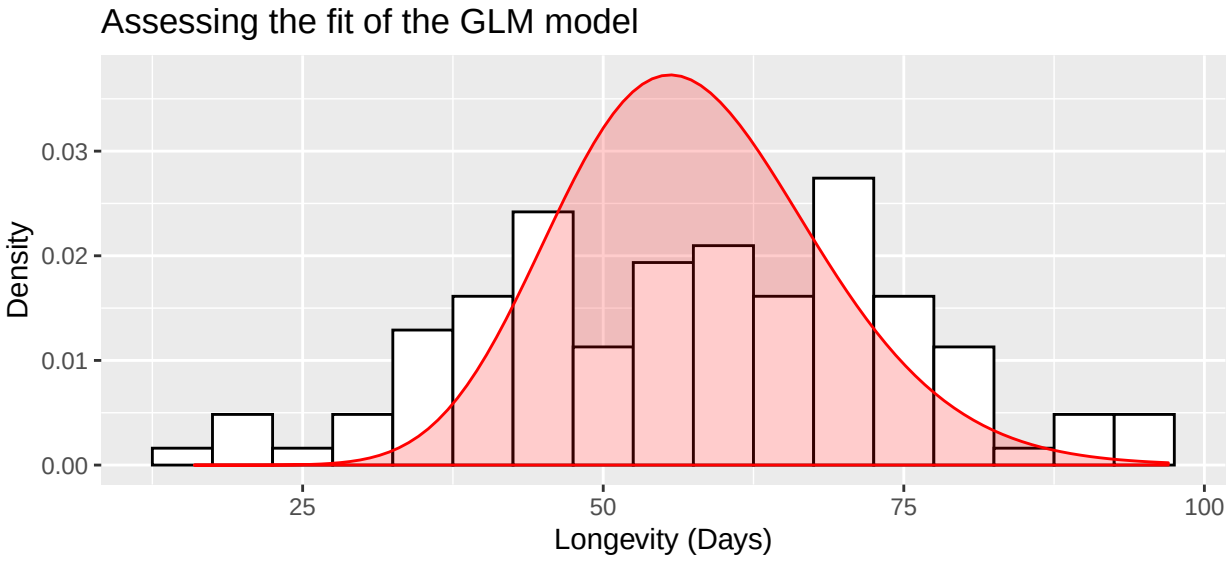

Figure 1: Assessing if the Gamma GLM (shown in red) was a good fit to the friut flies data

Additionally, we present the empirical diatribution along with the model fit to assess how good the fit of the GLM model is to the data.

Non-technical Summary

We divided 125 fruitflies randomly into 5 groups of 25 each by giving them different living conditions(with or without some virgin or pregnant female fruitflies).The thorax length of each male was measured as this was known to affect lifetime. And we also recorded the lifetime of the fruit fly in days. We analyzed the rusults and found that friut flies cohabilitating with virgin flies tended to have shorter lifespan than those living in isolation, and the more virgin flies that lived with, the shorter lifespan the flies had. Therefore, we can conclude that fertile women do affect men's lifetime.

Question 2: Smoking

Summary

We analyzed the results of the 2014 American National Youth Tobacco Survey amongst American school children to look for indicators that correlated with inceases in the odds of chewing tobacco regularly or trying a hookah. After accounting for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon, we found that the odds of regular use of chewing tobacco, snuff or dip is different amongst different races and White people were the most likely to chew tobacco followed by Hispanics and Black people. How about the odds of people using a hookah or waterpipe at least once? There is no statistically significant difference between men and women given both of them have similar other characteristics.

Introduction

We analyzed the 2014 American National Youth Tobacco Survey by using an R version of the dataset available at pbrown.ca. We wanted to analyze the relationship between smoking habits and age, sex, living area and ethnic group. This survey defined chewing tobacco regularly to be at least once in the past 30 days. We explored whether the odds of regular use of chewing tobacco, snuff or dip is different amongst different races and whether the odds of people using a Hookah or waterpipe at least once was affected by gender.

Methods

After centering Age(so intercept is age 15), we fit a Binomial(Logistic) generalized linear model with logit link function to model (the regular use of chewing tobacco) or (people using a Hookah or waterpipe at least once) as a function of the age, sex, race and rural/urban as follows:

$$\ln(odds) = \beta_0 + \beta_1 x_{Age} + \beta_2 I_{Female} + \beta_3 I_{Black} + \beta_4 I_{Hisp} + \beta_5 I_{Asian} + \beta_6 I_{Native} + \beta_7 I_{Pacific} + \beta_8 I_{Rural}$$

where β_i denotes as coefficient, I denotes as indicator and the odds could be either the odds of regular use of chewing tobacco or the odds of people using a Hookah or waterpipe. For the former, we use the null hypothesis $H_0 \colon \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ to test whether race is a significant predictor of regular use of chewing tobacco. When it comes to the latter, we use the null hypothesis $H_0 \colon \beta_2 = 0$ to test whether gender is a significant predictor of using a hookah.

Results

Table 3: ANOVA summary table for modelling odds of regular use of chewing tobacco

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.665	197.894	2	199	0.000
ageC	1	0.051	5.373	2	199	0.005
Sex	1	0.072	7.770	2	199	0.001
Race	5	0.692	21.171	10	400	0.000
RuralUrban	1	0.048	4.977	2	199	0.008
Residuals	200	NA	NA	NA	NA	NA

From the ANOVA table from our first model, we found that race was a significant predictor of regular use of chewing tobacco, even after controlling for age, sex and whether in rural or urban. It's also obvious that the use of chewing tobacco were significantly different between ages, the sexes and between whether living in rural or urban.

Table 4: Estimated odds of regular use of chewing tobacco

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.048	0.083	-36.483	0.000
Age	1.400	0.021	16.204	0.000
Female	0.167	0.109	-16.481	0.000
Black	0.211	0.172	-9.064	0.000
Hispanic	0.490	0.104	-6.884	0.000
Asian	0.213	0.342	-4.519	0.000
Native	1.113	0.278	0.385	0.700
Pacific	2.751	0.361	2.807	0.005
Rural	2.588	0.087	10.876	0.000

After looking at the exponentiated coefficients of our first model(table 4: Estimated odds of regular use of chewing tobacco), we can see that Black people and Hispanics are about 21 percent and half as likely to chew tobacco as White people respectively, with all the other aforementioned covariates held fixed. Additionally, we can see that women are only about 17 percent as likely to chew tobacco as men, and Americans regularly tended to smoke about 40 more percent tabacco every year than the last year. And as we already accounted the fact that chewing tabacco is a rural phenomenon, without any surprise, we see that people living in rural areas are over 2.5 times more likely to chew tobacco than those in urban.

Table 5: Estimated odds of ever using hookah or waterpipe

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.178	0.044	-39.226	0.000
Age	1.520	0.012	36.266	0.000
Female	1.043	0.043	0.980	0.327
Black	0.530	0.070	-9.005	0.000
Hispanic	1.413	0.048	7.138	0.000
Asian	0.532	0.118	-5.362	0.000
Native	1.173	0.190	0.838	0.402
Pacific	2.621	0.270	3.566	0.000
Rural	0.678	0.044	-8.769	0.000

After looking at the exponentiated coefficients of our second model(table 5: Estimated odds of ever using hookah or waterpipe), we see that the odds of using a hookah are about 4 percent higher for women then men, but this difference is not statistically significant (since p-value = 0.327) so we cannot conclude that women and men are no more likely to use a hookah provided their age, ethnicity, and other demographic characteristics are similar.

Appendix

Include all code used in this paper.

Question1

```
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     fruitfly
```

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

#change the factor levels as below
levels(fruitfly$activity) <- c("Solitary", "1 Preg Fly", "1 Vig Fly", "8 Preg Flies", "8 Vig Flies")

aggdata = aggregate(longevity~activity, fruitfly, mean )
aggdata = round(aggdata[,2])
aggdata = as.matrix(aggdata,5)
colnames(aggdata)<-c("Longevity (Days)")
rownames(aggdata)<- c("Isolated","With 1 Pregnant Fly","With 1 Virgin Fly","With 8 Pregnant Flies","Wit
#create table1
knitr::kable(aggdata, cap="Mean longevity of each fruitfly group")
```

Table 6: Mean longevity of each fruitfly group

	Longevity (Days)
Isolated	64
With 1 Pregnant Fly	65
With 1 Virgin Fly	57
With 8 Pregnant Flies	65
With 8 Virgin Flies	39

```
#normalize thorax
c = mean(fruitfly$thorax)
d = var(fruitfly$thorax)
new_thorax = ((fruitfly$thorax)-c)/sqrt(d)
fitmod=glm(longevity~new_thorax + activity,family=Gamma(link='log'), data=fruitfly)
coeffdata = round(summary(fitmod)$coef,3)
coeffdata[,1] = round(exp(coeffdata[,1]),3) #to use a more natural Scale
colnames(coeffdata) <- c("Exp. Estimate", "Std. Error", "t value", "P-Value")
rownames(coeffdata) <- c("Intercept" ,"Thorax Length", "With 1 Pregnant Fly","With 1 Virgin Fly","With 8
"With 8 Virgin Flies")
#create table2
knitr::kable(coeffdata, cap="Estimated parameters from the Gamma  generalized linear model of the fruit
```

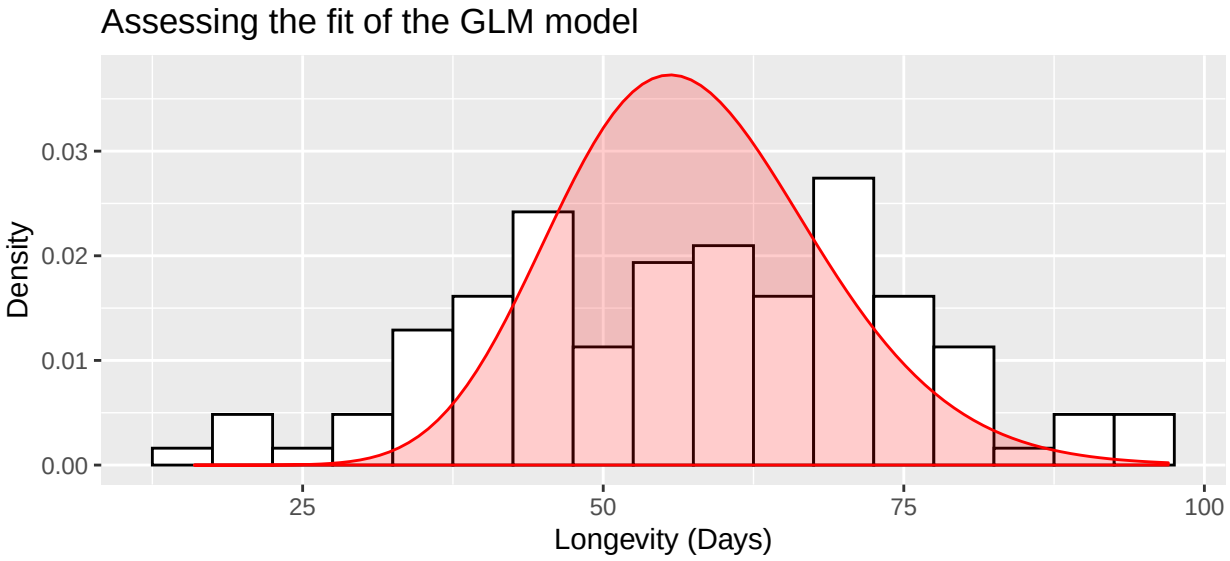

Figure 2: Assessing if the Gamma GLM (shown in red) was a good fit to the friut flies data

Table 7: Estimated parameters from the Gamma generalized linear model of the fruitflies

	Exp. Estimate	Std. Error	t value	P-Value
Intercept	60.220	0.038	108.333	0.000
Thorax Length	1.226	0.017	11.804	0.000
With 1 Pregnant Fly	1.057	0.053	1.036	0.302
With 1 Virgin Fly	0.890	0.053	-2.184	0.031
With 8 Pregnant Flies	1.085	0.054	1.524	0.130
With 8 Virgin Flies	0.660	0.054	-7.687	0.000

```
library(ggplot2)
shape = 1/summary(fitmod)$dispersion
scale = mean(fruitfly$longevity)/shape
ggplot(fruitfly, aes(x=longevity)) +
  geom_bar(binwidth = 5, colour="black", fill="white",aes(y=..density..)) +
  stat_function(fun=dgamma, args = list(shape = shape, scale = scale), colour="red",fill="red", geom="r
  labs(title="Assessing the fit of the GLM model") +
                labs(x="Longevity (Days)", y="Density")
```

```
#plot the hist and assess the fitting
```

Question2

```
load("/Users/mac/Desktop/2019 Fall/STA442/assignments/smoke.rdata")
#reponse of two models
smoke$Reg_chew_tob = factor(smoke$chewing_tobacco_snuff_or, levels=c('TRUE','FALSE'), labels=c('yes','n
smoke$Ever_hookah = factor(smoke$ever_tobacco_hookah_or_wa, levels=c('TRUE','FALSE'), labels=c('yes','n

#' nine year olds look suspicious
#' get rid of missings(NA) and age 9
#+ smokeSub1 and smokeSub2
smokeSub1 = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                    !is.na(smoke$Reg_chew_tob), ]
smokeSub2 = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                     !is.na(smoke$Ever_hookah), ]

#reshape the data by using reshape2
smokeAgg1 = reshape2::dcast(smokeSub1,
                            Age + Sex + Race + RuralUrban ~ Reg_chew_tob,
                            length)
smokeAgg2 = reshape2::dcast(smokeSub2,
                             Age + Sex + Race + RuralUrban ~ Ever_hookah,
                             length)

# data collection finished
smokeAgg1 = na.omit(smokeAgg1)
smokeAgg2 = na.omit(smokeAgg2)

# center Age so intercept is age 15 instead of 0
smokeAgg1$ageC = smokeAgg1$Age - 15
smokeAgg2$ageC = smokeAgg2$Age - 15

# fit two binomial(logistic) models
smokeAgg1$y = cbind(smokeAgg1$yes, smokeAgg1$no)
smokeAgg2$y = cbind(smokeAgg2$yes, smokeAgg2$no)
smokeFit1 = glm(y ~ ageC + Sex + Race + RuralUrban,
                family=binomial(link='logit'), data=smokeAgg1)
smokeFit2 = glm(y ~ ageC + Sex + Race + RuralUrban,
                 family=binomial(link='logit'), data=smokeAgg2)

#create the one-way anova of the first model
knitr::kable(anova(aov(smokeFit1)), digits=3,
             cap="ANOVA summary table for modelling odds of regular use of chewing tobacco")
```

Table 8: ANOVA summary table for modelling odds of regular use of chewing tobacco

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.665	197.894	2	199	0.000
ageC	1	0.051	5.373	2	199	0.005
Sex	1	0.072	7.770	2	199	0.001
Race	5	0.692	21.171	10	400	0.000
RuralUrban	1	0.048	4.977	2	199	0.008
Residuals	200	NA	NA	NA	NA	NA

```
#convert to odds
smokeTable1 = summary(smokeFit1)$coef
smokeTable1[,1] = round(exp(smokeTable1[,1]),3)
smokeTable2 = summary(smokeFit2)$coef
smokeTable2[,1] = round(exp(smokeTable2[,1]),3)

# make row names nicer
rownames(smokeTable1) <- c("Intercept" ,"Age", "Female","Black","Hispanic","Asian","Native","Pacific",
                           "Rural")
rownames(smokeTable2) <- c("Intercept" ,"Age", "Female","Black","Hispanic","Asian","Native","Pacific",
                             "Rural")
#make column names nicer
colnames(smokeTable1) <- c("Exp. Estimate", "Std. Error", "z value", "P-Value")
colnames(smokeTable2) <- c("Exp. Estimate", "Std. Error", "z value", "P-Value")
#create the table1
knitr::kable(smokeTable1, cap="Estimated odds of regular use of chewing tobacco",digits = 3)
```

Table 9: Estimated odds of regular use of chewing tobacco

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.048	0.083	-36.483	0.000
Age	1.400	0.021	16.204	0.000
Female	0.167	0.109	-16.481	0.000
Black	0.211	0.172	-9.064	0.000
Hispanic	0.490	0.104	-6.884	0.000
Asian	0.213	0.342	-4.519	0.000
Native	1.113	0.278	0.385	0.700
Pacific	2.751	0.361	2.807	0.005
Rural	2.588	0.087	10.876	0.000

```
#create the table2
knitr::kable(smokeTable2, cap="Estimated odds of ever using hookah or waterpipe",digits = 3)
```

Table 10: Estimated odds of ever using hookah or waterpipe

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.178	0.044	-39.226	0.000
Age	1.520	0.012	36.266	0.000
Female	1.043	0.043	0.980	0.327
Black	0.530	0.070	-9.005	0.000
Hispanic	1.413	0.048	7.138	0.000
Asian	0.532	0.118	-5.362	0.000
Native	1.173	0.190	0.838	0.402
Pacific	2.621	0.270	3.566	0.000
Rural	0.678	0.044	-8.769	0.000