STA442 Assignment1

Xinyue Jiang

2019/9/18

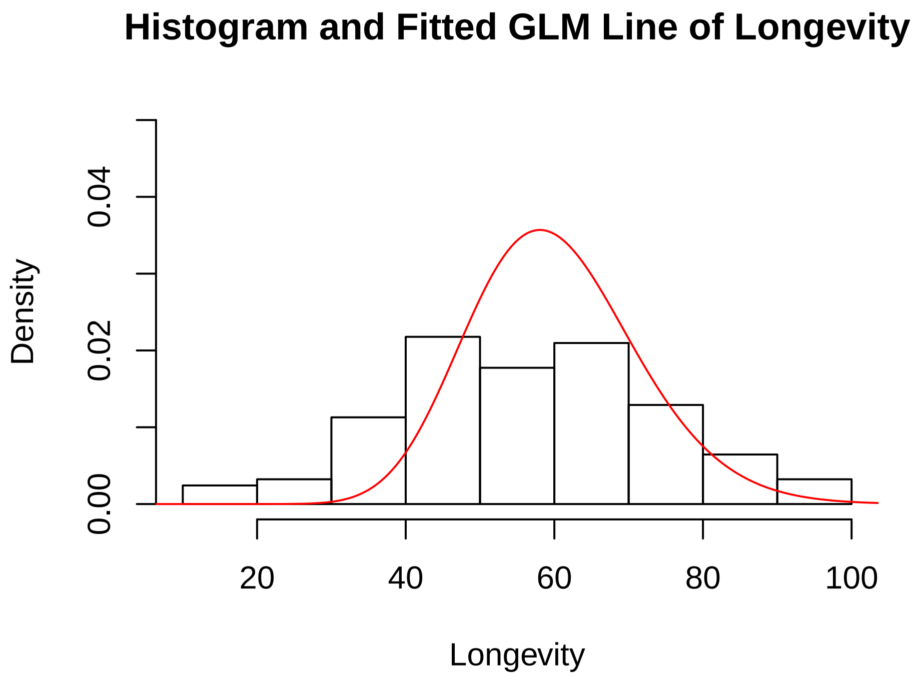
Question1

Problem

We analyzed the data "fruitfly" from the library *faraway*, which collected the information about the lifetimes and thorax lengthes of 125 fruit flies. The 125 fruit flies were randomly divided into five groups of 25 each. One group was kept solitary(Isolated), while another was kept individually with a virgin female each day(ActivityLow). Another group was given eight virgin females per day(ActivityHigh). As an additional control the fourth and fifth groups were kept with one(ActivityOne) or eight (ActivityMany)pregnant females per day (pregnant fruit flies will not mate). We wanted to explore how the throx length and activity influenced the longevity of the fruit fly.

Model

The response variable - longevity, is always greater than 0, so we tried Gamma generalized linear model first. We checked the density function of the Gamma distribution (y ~ Gamma(scale=2.14, shape=28.15)) and the density histogram of longevity, then we found the Gamma distribution captured the bell shape of the histogram. Therefore, Gamma glm was adequate to model the lifetime as a function of the thorax length and activity. We used the following model for the analysis:
$Log(longevity) = \beta_0 + \beta_1 x_{Thorax} + \beta_2 x_{ActivityOne} + \beta_3 x_{ActivityLow} + \beta_4 x_{ActivityMany} + \beta_5 x_{High}$



Result

The summary of the Gamma regression model shows that throax length ($p = 0.00$) and activity ($p_{LowActivity} = 0.03 \quad and \quad p_{HighActivity} = 0.00$) are significant preditors, which means they have effect on expected lifetime.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.098	0.038	108.333	0.000
thorax_norm	0.204	0.017	11.804	0.000
fruitfly$activityone	0.055	0.053	1.036	0.302
fruitfly$activitylow	-0.116	0.053	-2.184	0.031
fruitfly$activitymany	0.082	0.054	1.524	0.130
fruitfly$activityhigh	-0.415	0.054	-7.687	0.000

After the estimated coefficients were adjusted by an exponential fuction, we found fruit flies with low activity lived around 11% shorter than those lived in isolation, and fruit flies with high activity lived 34% shorter than fruit flies lived in isolation.

	x
(Intercept)	60.202
thorax_norm	1.227
fruitfly$activityone	1.057
fruitfly$activitylow	0.890
fruitfly$activitymany	1.086
fruitfly$activityhigh	0.661

Reseach News

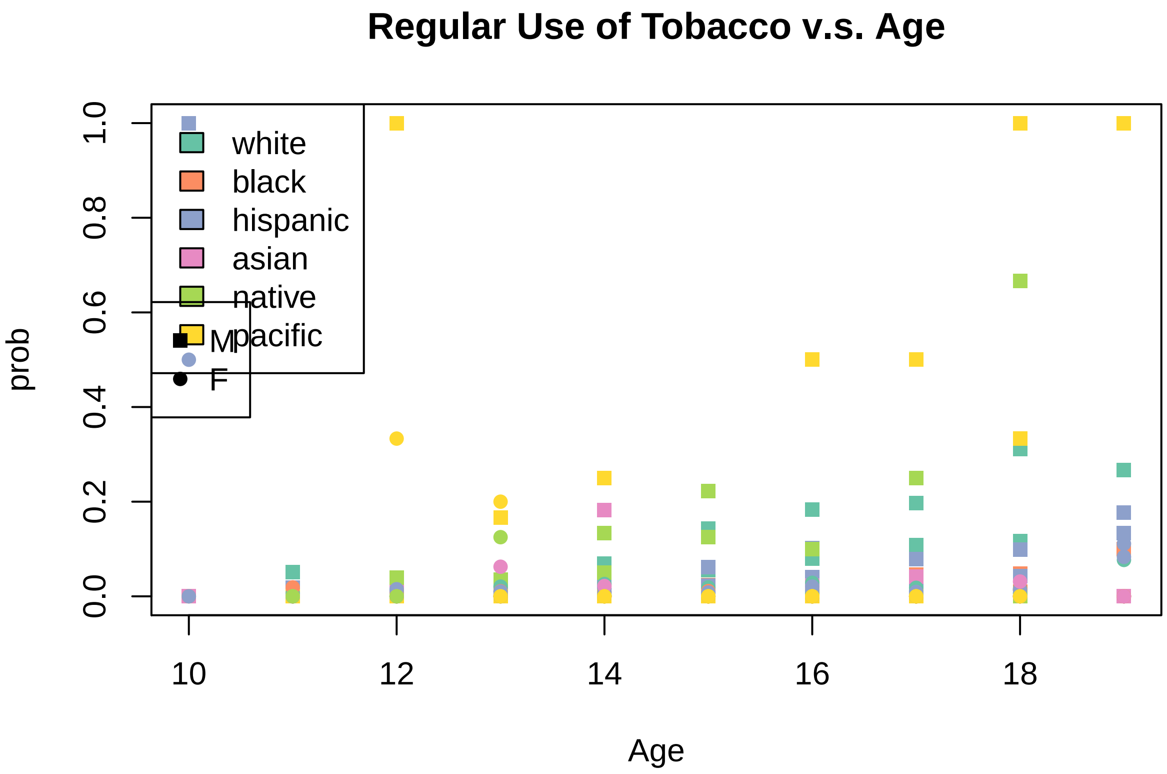
Mating Descrease the Longevity of Males ?

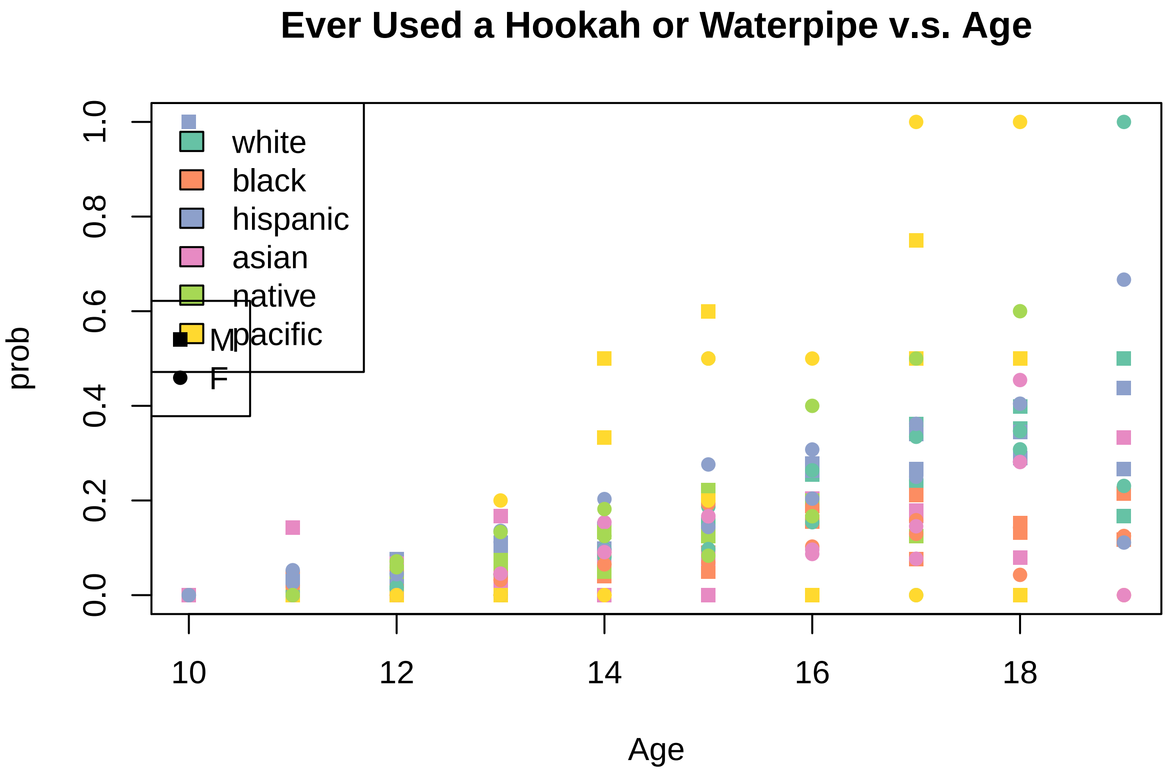
According to the data of 125 fruit flies provided by *faraway*, male fruit flies have shorter lifetimes when they cohabitate with fertile females, giving the thorax lengthes controlled. In the experiment, 125 fruit flies were divided to 5 groups:without any female, with 1 or 8 virgin females, with 1 or 8 pregnant females(pregnant females will not mate). After fitting a Gamma Regression Linear Model, researchers found male fruit flies cohabitated with 1 fertile female lived 11% shorter than those kept solitary; and male fruit flies cohabitated with 8 fertile females had even shorter lifetimes – 34% shorter than those kept solitary.

Question2

Summary

We analyzed the 2014 American National Youth Tobacco Survey to investigate what elements are correlated to the odds of regular use of chewing tobacco and trying a hookah or waterpipe. The dataset is available from the Survey's webpage. We found race was an significant factor that influenced the regular use of chewing tobacco. Pasific people liked chewing tobacco most, followed by white, hispanic and black people. Pasific people were 1.75 times more likely to regularly chew tobacco than white people; hispanic and black peole were only a half and one fifth respectively as likely to regularly chew tobacco compared to white people. Also, we found older males living in rural area were more likely to chew tobacco regularly. Different from regular use of chewing tobacco, gender did not affact the odds of trying hookah or waterpipe. However, race was still an significant factor. Pacific and hispanic people were more likely to try hookah or waterpipe, 2.6 and 1.4 times respectively as likely as white people;whereas, black people were only a half as likely to try hookah or waterpipe as white people. In addition, the rate of trying a hookah or waterpipe increased as age increased./





Introduction

We analyzed the 2014 American National Youth Tobacco Survey(the dataset is available from the Survey's webpage) to investigate the following two hypoyheses:
1.Regular use of chewing tobacco(1 or more days in the past 30 days) is no more common amongest American of European ancestry than for Hispanic-American and Aferican_American.
2.The likelihood of trying hookah or waterpipe (ever smoked tobacco out of a hookah or waterpipe)is same for males and females.
In addition, we considered the effets of age and whether or not the respondants live in rural or urban area for both hypotheses

Methods

We used logistic regression modle to study the two hypotheses because Regular Use of Tobacco follows binomial distribution (y~ Binomial(N,μ)), and Have Used a Hookah or Waterpipe follows binomial distribution (y~ Binomial(N,μ)). We let Odds1 be the Odds of regular use of Tobacco, Odds2 be the Odds of having used a hookah or waterpipe. Based on the analysis, we used the following two modles to investigate the two hypotheses respectively.
ln Odds1 $= \beta_0 + \beta_1 x_{age} + \beta_2 x_{Female} + \beta_3 x_{Balck} + \beta_4 x_{Hisp} + \beta_5 x_{Asian} + \beta_6 x_{Native} + \beta_7 x_{pacif} + \beta_8 x_{Rural}$
We tested whether or not race (especially white, balck and hispanic),is a significant predictor of regular use of Tobacco : $H_0 : \beta_3 = \beta_4 = 0$
ln Odds2 $= \beta_0 + \beta_1 x_{age} + \beta_2 x_{Female} + \beta_3 x_{Balck} + \beta_4 x_{Hisp} + \beta_5 x_{Asian} + \beta_6 x_{Native} + \beta_7 x_{pacif} + \beta_8 x_{Rural}$
We tested whether or not gender is a significant predictor of having used a hookah or waterpipe: $H_0 : \beta_2 = 0$

Result

We checked the table for regular use of tobacco first. We found that the p-values of estimated coefficients of Afarican-Americans and Hispanic-Americans were all 0.00, which meant the likelyhood of regular use of tobacco is not the same among white, black, and hispanic teenagers in America.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.032	0.083	-36.483	0.000
ageC	0.337	0.021	16.204	0.000
SexF	-1.788	0.109	-16.481	0.000
Raceblack	-1.556	0.172	-9.064	0.000
Racehispanic	-0.713	0.104	-6.884	0.000
Raceasian	-1.546	0.342	-4.519	0.000
Racenative	0.107	0.278	0.385	0.700
Racepacific	1.012	0.361	2.807	0.005
RuralUrbanRural	0.951	0.087	10.876	0.000

We adjusted the estimated coefficients by an exponetial function to study how exactly predictors influenced the odds of regular use of chewing tobacco. We found that regular use of tobacco was more common among white people than among black and hispanic people, who regularly chewed tobacco at only 21% and 49% of the rate for white people respectively.

	x
(Intercept)	0.048
ageC	1.400
SexF	0.167
Raceblack	0.211
Racehispanic	0.490
Raceasian	0.213
Racenative	1.113
Racepacific	2.751
RuralUrbanRural	2.588

We then looked at the table for having used a hookah or waterpipe and its adjusted version(adjusted by exponential), found women were 4% more likely to try a hookah or waterpipe than men; however, the p-value of the difference was not significant (p-value = 0.33).We could not conclude that the likelyhood of trying a hookah or waterpipe is the same for different genders, giving their age, race, and other demographic characteristics are controlled.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.724	0.044	-39.226	0.000
ageC	0.419	0.012	36.266	0.000
SexF	0.042	0.043	0.980	0.327
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

	x
(Intercept)	0.178
ageC	1.520
SexF	1.043
Raceblack	0.530
Racehispanic	1.413
Raceasian	0.532
Racenative	1.173
Racepacific	2.621
RuralUrbanRural	0.678

Appendix

Question1

```r
install.packages("faraway")
data('fruitfly', package='faraway')
head(fruitfly)

#normalize thorax
summary(fruitfly)
thorax_mean<- mean(fruitfly$thorax)
thorax_var<- var(fruitfly$thorax)
thorax_norm= (fruitfly$thorax-thorax_mean)/sqrt(thorax_var)

#fit glm model
fruitflyglm<-glm(fruitfly$longevity ~ thorax_norm + fruitfly$activity, family=Gamma(link= "log"))
glm(fruitfly$longevity ~ thorax_norm + fruitfly$activity, family=Gamma(link= "log"))

#find the shape and scale of the glm model
shape = 1/summary(fruitflyglm)$dispersion
shape
scale = exp(fruitflyglm$coef["(Intercept)"])/shape
scale

#draw the histogram and glm pdf
scale = mean(fruitfly$longevity)/shape
hist(fruitfly$longevity,prob=TRUE,ylim = c(0,0.05), main = "Histogram and Fitted GLM Line of Longevity"
xSeq = seq(0,120,len=1000)
lines(xSeq,dgamma(xSeq,shape=shape,scale=scale),col="red")

#summary the glm model
knitr::kable(summary(fruitflyglm)$coef, digit =3)
knitr::kable((exp(summary(fruitflyglm)$coef)[1:6,c("Estimate")]), digit =3)
```

#####Question2

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

```r
(load(smokeFile))
dim(smoke)
#'
#'
#+ explore regular chewing tobacco
smoke[1:5,c('Age','Sex','Grade','RuralUrban','Race', 'chewing_tobacco_snuff_or')]
smokeFormats[smokeFormats$colName == 'chewing_tobacco_snuff_or', ]
smoke$chewTobacco = factor(smoke$chewing_tobacco_snuff_or, labels=c('no','yes'))
smoke$chewTobacco
table(smoke$Grade, smoke$Age, exclude=NULL)
table(smoke$Race, smoke$chewTobacco, exclude=NULL)

#' nine year olds look suspicious
#' get rid of missings and age 9
#+ smokeSub1
smokeSub1 = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                   !is.na(smoke$chewTobacco), ]
table(smokeSub1$Race, smokeSub1$chewTobacco, exclude=NULL)
dim(smokeSub1)

#'
#'
#+
smokeAgg1 = reshape2::dcast(smokeSub1,
                            Age + Sex + Race + RuralUrban ~ chewTobacco,
                            length)
dim(smokeAgg1)
smokeAgg1 = na.omit(smokeAgg1)
dim(smokeAgg1)

smokeAgg1[which(smokeAgg1$Race == 'white' &
                 smokeAgg1$Sex == 'M' & smokeAgg1$RuralUrban == 'Urban'),]
smokeAgg1$total = smokeAgg1$no + smokeAgg1$yes
smokeAgg1$prop = smokeAgg1$yes / smokeAgg1$total

#'
#+ chewing tobacco plot
Spch = c('M' = 15, 'F'=16)
Scol = RColorBrewer::brewer.pal(nlevels(smokeAgg1$Race), 'Set2')
names(Scol) = levels(smokeAgg1$Race)
plot(smokeAgg1$Age, smokeAgg1$prop, pch = Spch[as.character(smokeAgg1$Sex)],
     col = Scol[as.character(smokeAgg1$Race)])
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))
#'
#' Which races smoke the least?
#' ... age is a confounder
#' ... as is urban/rural.
#+ coefficient table for chewig tobacco without changing the intercept
smokeAgg1$y = cbind(smokeAgg1$yes, smokeAgg1$no)
smokeFit = glm(y ~ Age + Sex + Race + RuralUrban,
                family=binomial(link='logit'), data=smokeAgg1)
knitr::kable(summary(smokeFit)$coef, digits=3)
```

```r
#' Intercept is age zero
#' center Age so intercept is age 15
#+ smokeFit2
smokeAgg1$ageC = smokeAgg1$Age - 15
smokeFit2 = glm(y ~ ageC + Sex + Race + RuralUrban,
                family=binomial(link='logit'), data=smokeAgg1)
knitr::kable(summary(smokeFit2)$coef, digits=3)

#adjusted by exponetial function
knitr::kable((exp(summary(smokeFit2)$coef)[ ,c("Estimate")]), digit =3)





#'
#'
#+ explore Having used a hookah or waterpipe
smoke[1:5,c('Age','Sex','Grade','RuralUrban','Race', 'ever_tobacco_hookah_or_wa')]
smokeFormats[smokeFormats$colName == 'ever_tobacco_hookah_or_wa', ]
table(smoke$Race, smoke$ever_tobacco_hookah_or_wa, exclude=NULL)
smoke$everTobacco = factor(smoke$ever_tobacco_hookah_or_wa, labels=c('no','yes'))
smoke$everTobacco
table(smoke$Grade, smoke$Age, exclude=NULL)
table(smoke$Race, smoke$everTobacco, exclude=NULL)
#'
#' nine year olds look suspicious
#' get rid of missings and age 9
#+ smokeSub
smokeSub = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                   !is.na(smoke$everTobacco), ]
table(smokeSub$Race, smokeSub$everTobacco, exclude=NULL)
dim(smokeSub)

#'
#'
#+
smokeAgg2 = reshape2::dcast(smokeSub,
                            Age + Sex + Race + RuralUrban ~ everTobacco,
                            length)
dim(smokeAgg2)
smokeAgg2 = na.omit(smokeAgg2)
dim(smokeAgg2)

smokeAgg2[which(smokeAgg2$Race == 'white' &
                  smokeAgg2$Sex == 'M' & smokeAgg2$RuralUrban == 'Urban'),]
smokeAgg2$total = smokeAgg2$no + smokeAgg2$yes
smokeAgg2$prop = smokeAgg2$yes / smokeAgg2$total
#'
#'
#+ Hookah Plot
Spch = c('M' = 15, 'F'=16)
```

```r
Scol = RColorBrewer::brewer.pal(nlevels(smokeAgg2$Race), 'Set2')
names(Scol) = levels(smokeAgg2$Race)
plot(smokeAgg2$Age, smokeAgg2$prop, pch = Spch[as.character(smokeAgg2$Sex)],
     col = Scol[as.character(smokeAgg2$Race)])
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))

#' ... age is a confounder
#' ... as is urban/rural.
#+ Hookah glm without changing intercept
smokeAgg2$y = cbind(smokeAgg2$yes, smokeAgg2$no)
smokeFit = glm(y ~ Age + Sex + Race + RuralUrban,
               family=binomial(link='logit'), data=smokeAgg2)
knitr::kable(summary(smokeFit)$coef, digits=3)

#'
#' Intercept is age zero
#' center Age so intercept is age 15
#+ smokeFit4
smokeAgg2$ageC = smokeAgg2$Age - 15
smokeFit4 = glm(y ~ ageC + Sex + Race + RuralUrban,
                family=binomial(link='logit'), data=smokeAgg2)
knitr::kable(summary(smokeFit4)$coef, digits=3)

#adjust the coefficient table by an exponetial function
knitr::kable((exp(summary(smokeFit4)$coef)[ ,c("Estimate")]), digit =3)
```