STA442 Assignment 1

Tuoyue Huang Student number:1003906712

Tuesday 24 September 2019

"Report of Fruitfly Dataset"

Summary

We found something really interesting from the fruitfly dataset of Faraway. The fact that male fruitflys kill themselves if they stay with virgin females and have sex activities! Average lifespan of a male fruitfly is about 60 days if they are along or stay with pregnant females. However, their lifetime decreased by 11% if living with one virgin fruitfly per day,and that percentage increase to 34% if thay live with eight virgin females. We also found that one unit increase in thorax length could make thier survial days 23% higher.

Introduction

The problem we are interested is how sexual activity affects the lifetime of fruitflies. The dataset used is from Faraway. There were 125 male fruitflies in the experiment, and were seperated into five groups as showed in Figure 1. In each observation, one male fruitfly stayed with certain number of virgin or pregnent females each day until he was dead. The thorax length of fruitflies were also measured since it's known to affect the lifetime.

Methods

From Figure 1, we can see that the median lifetimes are indeed lower than the others if fruitflies stay with the virgins. All the measured lifetimes are non-negative, so we fit a Gamma Generalized Linear Model with a log-link to take a further look. In the model, the response variable is lifetime of fruitfly, while the explanatory variables are thorax(normalized) and sexual activivy of fruitflies. This is saying our model assumptions are the folowing:

$$Y_i \sim Gamma(\mu_i/v, v), log(\mu_i) = X_i\beta, E(Y_i) = \mu_i$$

Additionally, we plot Figure 2 to check whehter Gamma regression is appropriate to use here. We can see the red line, whcich is the Gamma distribution, follows the trend of the data pretty well, and thus we could get usedul result from the model.

Results

Figure 1: Lifetime of male fruitfly in diffent groups

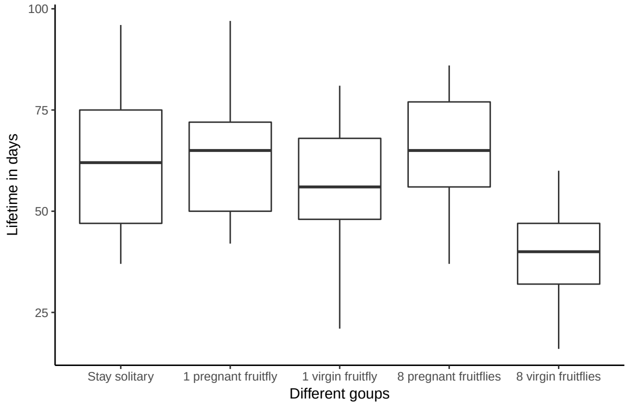


Figure 2: Life Time of Male Fruitflies in Days

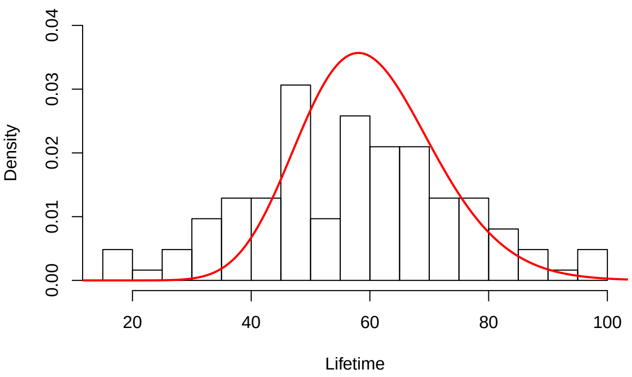


Table 1 gives the exponentiated result of the estimated coefficients from the model due to the log-link. We expect a fruitfly survives about 60 days if he lives alone, and his lifetime decreses by 11% and 34% if he lives with one and eight virgin female correspondingly. The p-value for the three rows are almost all 0, whcih means we accept the alternative hypothesis that the sexual activity will affect the lifetime. Howevere, the p-value for the two pregnent froup are all above 0.05, so we conclude that fruitfly with no sexual activity should live the same period as they live alone. What else, one unit increase in thorax length will increase days of living by 23%.

Table 1: Gamma GLM Model Summary table of fruitflies

	Survival days	Std.Error	t value	p value	lower	upper
Stay solitary	60.20	0.04	108.33	0.00	55.82	64.93
Thorax length	1.23	0.02	11.80	0.00	1.18	1.27
One pregnant fruitfly	1.06	0.05	1.04	0.30	0.95	1.18
One virgin fruitfly	0.89	0.05	-2.18	0.03	0.80	0.99
Eight pregnant fruitflies	1.09	0.05	1.52	0.13	0.97	1.21
Eight virgin fruitflies	0.66	0.05	-7.69	0.00	0.59	0.74

“Report of American National Youth Tobacco Survey”

Summary

In this report, we analyzed the data from 2014 American National Tobacco Survey. The first result we found is that white American children have the highest odds of chewing tobacco, sniff or dip if we control age, sex and living area. The second is hispanic with nearly half the odds of white, the odds of black is the lowest which is about 80% less compare with the white. In addition, elder children living in rural area tend to have higher odds, leading to the conclusion of chewing tobacco becoming a rural phenomenon. The second result is that odds of ever using hookah or waterpipe is almost the same for males and females fixing other variables.

Introduction

As smoking is a major health concern and is popular among youth, so we analyzed the 2014 American National Youth Tobacco Survey(NYTS2014), which is about use of cigars, hookahs, and chewing tobacco amongst American school children, through the R version of the dataset smoke.RData. It was collected by FDA and CDC from Schools in the U.S by using multi-stage cluster sampling. The first hypothesis we are investigating is whether the probability of chewing tobacco, sniff or dip is identical among white, hispanic and black American children and living in rural is claimed to be the reason of consuming tobacco. The other one is to look at the likelihood of consuming hookah or water pipe for the males and females given the same age, ethnicity and living place.

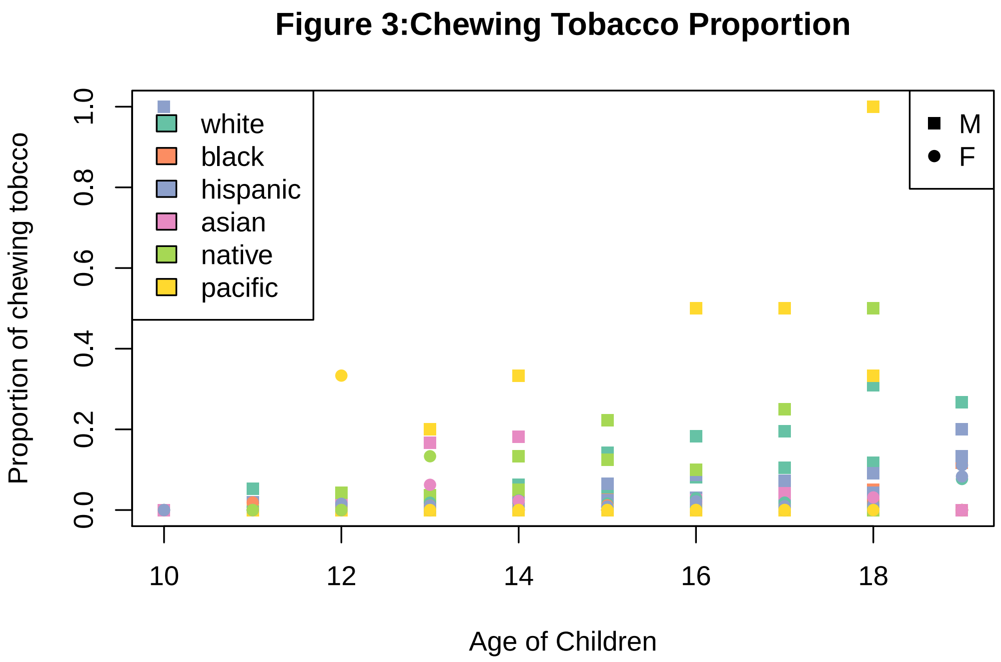
Methods

Our interest variables are both binary, since we are studying yes or no question. Therefore, the model we used here is logistic regression.

$$Y_i \sim Binomial(N_i, \mu_i), log(\mu_i/(1-\mu_i)) = log(Odds) = X_i\beta, E(Y_i) = N_i\mu_i$$

Odds and probability transformation is monotonic, then a higher odds means a relatively higher probability. We can simply fit two models with our interested response variables, whcih are ‘if chewing tobacoo’ and ‘if ever using hookah or water pipe’. Explanatory variables used here in bothe models are ‘age’,‘sex’,‘rural or urban’ and ‘races’. After that, we compute the estimated coefficients and take the exponential, then we can simply compare the odds we get to see the differences between groups.

Results



From Figure 3,we are able to roughly infer that males tends to get tobacco more than females. Among males, pacific and native American children tends to chewing more tobacco, although they are out of our interst. Followling after, it's white and hispanic, while the points represent black are all located at the vrey bottom.

Table 2: OddsRatio and Confidence Interval of Chewing tobacco

	Odds ratio	Lower bound	Upper bound
Baseline odds	0.00	0.00	0.00
Age	1.41	1.35	1.47
Female	0.17	0.13	0.21
Rural	2.56	2.14	3.07
Black	0.22	0.16	0.32
Hispanic	0.48	0.39	0.59
Asian	0.22	0.11	0.43
Native	1.11	0.62	1.96
Pacific	2.43	1.10	5.37

We use Table 2 to analysis for the first hypothesis. The column Odds ratio is from the exponentiated estimates of summary table of the fitted model. We could see that black has odds ratio equal to 0.22 between groups, which is saying the odds ratio of black decreases by 78% compare to the white. Because white is our baseline group and black is the indicator variable, thus we need to multiply the odds of the white and 0.22 together to get odds for the black. Similarly, for hispanic, they are half less likely to take tobacco respect to white due to odds ratio of 0.48. Further, we can see row rural is 2.56 ,which means if people live in rural then the baseline odds ratio will be multiply by about 2.5 times. This confirms people in rural are more likely to have tobacco and different races also have different probability.In addition, we have their 95% confidence interval in the table.For age, that is a confidence interval for the odds of chewing tobacco incresed between the range of 45% and 47% with a unit increase in age.

Table 3: OddsRatio and Confidence Interval of using hookah or waterpipe

	Odds ratio	Lower bound	Upper bound
Baseline odds	0.00	0.00	0.00
Age	1.52	1.49	1.56
Female	1.04	0.96	1.14
Rural	0.68	0.62	0.74
Black	0.52	0.46	0.61
Hispanic	1.42	1.28	1.56
Asian	0.52	0.41	0.66
Native	1.18	0.80	1.72
Pacific	2.75	1.60	4.72

We focus on row female here since we are studying the difference of sex for the second hypothesis. The odds ratio is 1.04 for female and 1 for male, then there is only 4% difference between them if we control age, race and living place to be the same. This will lead a really small difference of probability if we do transfer.As a result, we believe there is no significant difference between sexes on trying hookah or water pipe.

On the other hand, we can see hispanic Americans are more likely to use hookah or water pipe or people living in rural with older age have higher probability to get hookah or water pipe from the odds ratio column.

Appendix(code used)

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

ggplot(data = fruitfly, aes(x=activity,y=longevity)) +
  geom_boxplot() +
  labs(title= "Figure 1: Lifetime of male fruitfly in diffent groups",
       x="Different goups",y="Lifetime in days") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5,face = "bold")) +
  scale_x_discrete(labels=c("Stay solitary","1 pregnant fruitfly",
                            "1 virgin fruitfly","8 pregnant fruitflies",
                            "8 virgin fruitflies"))

fruitfly_scaled <- fruitfly %>%
  mutate_at(("thorax"),~(.x - mean(.x))/sd(.x))

mod <- glm(longevity ~ thorax + activity,family = Gamma(link = "log"),data = fruitfly_scaled)

shape = 1/summary(mod)$dispersion
hist(fruitfly_scaled$longevity, prob=TRUE,
    xlab='Lifetime', breaks=20, plot = TRUE, ylim = c(0,0.04),
     main = "Figure 2: Life Time of Male Fruitflies in Days")
xSeq = seq(par('usr')[1], par('usr')[2], len=200)
lines(xSeq,
    dgamma(xSeq, shape=shape,
        scale = exp(mod$coef['(Intercept)'])/shape),
    col='red', lwd=2)
coeTable = as.data.frame(summary(mod)$coef)
coeTable$lower = exp(coeTable$Estimate - 2*coeTable$`Std. Error`)
```

```r
coeTable$upper = exp(coeTable$Estimate + 2*coeTable$`Std. Error`)
coeTable[,1] = exp(coeTable[,1])

rownames(coeTable) = c("Stay solitary","Thorax length",
                       "One pregnant fruitfly","One virgin fruitfly",
                       "Eight pregnant fruitflies","Eight virgin fruitflies")
colnames(coeTable) = c("Survival days","Std.Error","t value","p value","lower","upper")
knitr::kable(coeTable, digits=2, caption = "Gamma GLM Model Summary table of fruitflies")

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

subsmoke <- smoke[c('Age','Sex','RuralUrban','Race','chewing_tobacco_snuff_or',
                    'ever_tobacco_hookah_or_wa')]
subsmoke = subsmoke[subsmoke$Age >= 10,]
smokeSub = subsmoke[!is.na(subsmoke$Race) & !is.na(subsmoke$Age) & !is.na(subsmoke$Sex) &
                      !is.na(subsmoke$RuralUrban) &
                      !is.na(subsmoke$ever_tobacco_hookah_or_wa) &
                      !is.na(subsmoke$chewing_tobacco_snuff_or),]

#if reshape the data, it will give the same result
#smokeAgg = reshape2::dcast(smokeSub,Age + Sex + Race + RuralUrban
#           ~chewing_tobacco_snuff_or,length)
#colnames(smokeAgg) = c("Age","Sex","Race","RuralUrban","Yes","No")
#smokeAgg$y = cbind(smokeAgg$No, smokeAgg$Yes)

fitmod1 <- glm(smokeSub$chewing_tobacco_snuff_or ~ smokeSub$Age + smokeSub$Sex +
                  smokeSub$RuralUrban + smokeSub$Race,data = smokeSub,
                family = binomial(link='logit'))
fitmod2 <-glm(smokeSub$ever_tobacco_hookah_or_wa ~ smokeSub$Age + smokeSub$Sex +
                  smokeSub$RuralUrban + smokeSub$Race,data = smokeSub,
                family = binomial(link='logit'))

smokeTable1 = as.data.frame(summary(fitmod1)$coef)
smokeTable2 = as.data.frame(summary(fitmod2)$coef)
smokeTable1$lower = smokeTable1$Estimate - 2*smokeTable1$`Std. Error`
smokeTable1$upper = smokeTable1$Estimate + 2*smokeTable1$`Std. Error`
smokeTable2$lower = smokeTable2$Estimate - 2*smokeTable2$`Std. Error`
smokeTable2$upper = smokeTable2$Estimate + 2*smokeTable2$`Std. Error`

smokeOddsRatio1 = exp(smokeTable1[,c('Estimate','lower','upper')])
smokeOddsRatio2 = exp(smokeTable2[,c('Estimate','lower','upper')])

rownames(smokeOddsRatio1) = c("Baseline odds","Age","Female","Rural","Black",
                              "Hispanic","Asian","Native","Pacific")
colnames(smokeOddsRatio1) = c("Odds ratio","Lower bound",
                              "Upper bound")
rownames(smokeOddsRatio2) = c("Baseline odds","Age","Female","Rural","Black",
                              "Hispanic","Asian","Native","Pacific")
colnames(smokeOddsRatio2) = c("Odds ratio","Lower bound",
                              "Upper bound")
```

```
knitr::kable(smokeOddsRatio1, digits=2, caption = "OddsRatio
             and Confidence Interval of Chewing tobacco")
knitr::kable(smokeOddsRatio2, digits=2, caption = "OddsRatio
             and Confidence Interval of using hookah or waterpipe")
```