# STA442_HW1

*Zeyang Zhang*

*2019/9/21*

## The Effect of Sexual Activity to Fruitfly Longevity

### Introduction

We want to investigate the relationship between the sexual activity and the lifetime of fruitflies. From the experiment data,we have 125 fruit flies divided randomly into five groups of 25 each: first group "isolitary" was kept solitary,second group "one" was kept individually with 1 pregnant female each day,third group "low" was kept individually with 1 virgin female each day,fourth group "many" was kept individually with 1 pregnant female each day(there is one missing data in this group), and fifth group "high" was given 8 virgin females per day. Pregnant fruit flies will not mate, so it is an additional control. It is known that to thorax length of each male fruitfly will affect its lifetime, so we measured and added it into our model.

### Analysis methods and coefficients interpretion

The QQ plot (Fig 1, see appendix) indicate longevity is not normally distributed. Also, we note that the longevity should be a positive and continuous number. Thus we should use Gamma rather than Possion GLM. Histagram and density curve (Fig 2, see appendix) below shows the density curve is right-skewness, which is more close to a Gamma distribution. We consider fitting the Gamma GLM with 'log' link function to make coefficients interpretable under 95% Confidence Interval.

Table 1: 95% Confidence Interval coeffient Table for fruitfly sexual activity model

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.887 | 0.194 | 9.726 | 0.000 |
| thorax | 2.688 | 0.228 | 11.804 | 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 |

Table 1 above gives us the details of each coefficient. Except 'Intercept' and 'thorax','activitylow' and 'activityhigh' are signficant under significance level $\alpha = 0.05$. It suffices to show that furitflies whose sexual activity are low or high have different longevity comparing to isolated fuirtflies while other groups are not signficantly different. The final GLM model is:

$$log(\mu) = 1.88722 + 2.68778 \times X_{thorax} - 0.11646 \times \mathbb{I}_{low} - 0.41466 \times \mathbb{I}_{high}$$

Accoring to Table1, we generate Table2 under the 95% confidential interval, which gives us the upper and lower bound of ratio changes of each varibale according to:

$$\mu = \mathrm{e}^{1.88722+2.68778\times X_{thorax}-0.11646\times\mathbb{I}_{low}-0.41466\times\mathbb{I}_{high}}$$

Table 2: How much longevity propotionally changed for fruitflies whose sex activity low and high under 95% Confidence Interval

| | Estimate | Lower | Upper |
|---|---|---|---|
| activitylow | 0.890 | 0.800 | 0.990 |
| activityhigh | 0.661 | 0.593 | 0.736 |

We can see the longevity decreases proportionally if type is low or high. Comparing to isolated group, the longevity of group low decreases 1.0% ~ 20.0% and the longevity of group high decrease 26.4% ~ 40.7%.

## Summary of the Fruitfly Results

In conclusion, we found statistical evidence about association between sexual activity and reduced lifespan of male fruitflies. After reducing the effects of different thorax length, longevity is still very different between isolated male fruitflies and male fruitflies with low and high sexual activities. More sexual activities are related to shorter longevity of male fruitflies, which might indicate significant physiological cost of male sexual activity.

## Summary of the Somkimng Research Results

Based on the data from 2014 American National Youth Tobacco Survey, we investigated two research hypotheses and made the following conclusions: First, comapring with Hispanic-Americans and African-Americans, regular usage of chewing tobacco(including snuff or dip) is much more common for Americans of European. Second, the likelihood of having used a hookah or waterpipe is the same for man and woman if their age and ethnicity are similar.

## Isuues of Tobacco Usage among American School Children

### Introduction

We use the 2014 American National Youth Tobacco Survey to analyze two problems amongst American school children. The first research problem we will investigate is whether regular use of chewing tobacco, snuff or dip is more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans. The second one is whether the likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes when their age, ethnicity, and other demographic characteristics are similar.

### Statistical analysis

In the **FIRST** problem, "chewing_tobacco_snuff_or" is a binary variable. So naturally a binomial GLM should be considered. Also, in order to except the influence of other factor, we include the age and whether living in rural area in our model since white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon.

Table 3: 95% Confidence Interval coeffient Table for chewing tobacco and race model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.794 | 0.333 | -26.433 | 0.000 |
| RuralUrbanRural | 0.931 | 0.086 | 10.858 | 0.000 |
| Raceblack | -1.559 | 0.169 | -9.246 | 0.000 |
| Racehispanic | -0.721 | 0.101 | -7.101 | 0.000 |
| Raceasian | -1.590 | 0.340 | -4.671 | 0.000 |
| Racenative | 0.168 | 0.274 | 0.613 | 0.540 |
| Racepacific | 1.158 | 0.337 | 3.440 | 0.001 |
| Age | 0.349 | 0.020 | 17.033 | 0.000 |

Accroding to Table 3, under significance level $\alpha = 0.05$, the model can be writtin as

$$log(\frac{\pi}{1-\pi}) = -8.794 + 0.931 \times \mathbb{I}_{Rural} - 1.559 \times \mathbb{I}_{Black} - 0.721 \times \mathbb{I}_{Hispanic} - 1.590 \times \mathbb{I}_{Asian} + 1.158 \times \mathbb{I}_{Pacific} + 0.349 \times \mathbb{X}_{Age}$$

In the **SECOND** problem, "ever_tobacco_hookah_or_wa" is a binary variable. Similarily, a binomial GLM should be considered. Also, we include the age and race as additional influential factor in our model, so the influence of factor sex can be estimated more accurately.

We hope to fit a model like:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_{Sex} \times \mathbb{I}_{Sex} + \beta_{Black} \times \mathbb{I}_{Black} + \beta_{Hispanic} \times \mathbb{I}_{Hispanic} + \beta_{Asian} \times \mathbb{I}_{Asian} + \beta_{Pacific} \times \mathbb{I}_{Pacific} + \beta_{Age} \times \mathbb{X}_{Age}$$

Table 4: 95% Confidence Interval coeffient Table for hookah and sex model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.145 | 0.185 | -44.055 | 0.000 |
| SexF | 0.043 | 0.043 | 1.011 | 0.312 |
| Raceblack | -0.583 | 0.070 | -8.317 | 0.000 |
| Racehispanic | 0.415 | 0.048 | 8.708 | 0.000 |
| Raceasian | -0.510 | 0.117 | -4.367 | 0.000 |
| Racenative | 0.143 | 0.190 | 0.749 | 0.454 |
| Racepacific | 0.997 | 0.268 | 3.724 | 0.000 |
| Age | 0.414 | 0.011 | 36.078 | 0.000 |

However, according to Table 4, under significance level $\alpha = 0.05$, $\beta_{Sex}$is not significantly different from 0, so it indicates the likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes (when their age, ethnicity, and other demographic characteristics are similar).

#### Results

Table 5: How much propotionally changed in the odd ratio of chewing tobacco, snuff or dip for Hispanic-Americans and African-Americans comapring to white Americans

| | Estimate | Lower | Upper |
|---|---|---|---|
| Raceblack | 0.210 | 0.150 | 0.295 |
| Racehispanic | 0.486 | 0.397 | 0.596 |

- **First**, from Table 5, we can see the odds ratio of chewing tobacco, snuff or dip decreases proportionally if the individual is Hispanic-Americans or African-Americans. Comparing to Americans of European ancestry, the odds ratio of chewing tobacco, snuff or dip decreases 70.5% ~ 85.0% for African-Americans and 40.4% ~ 60.3% for Hispanic-Americans.
- **Second**, as stated in the statistical analysis above, we can not reject the null hypothesis that the likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes.

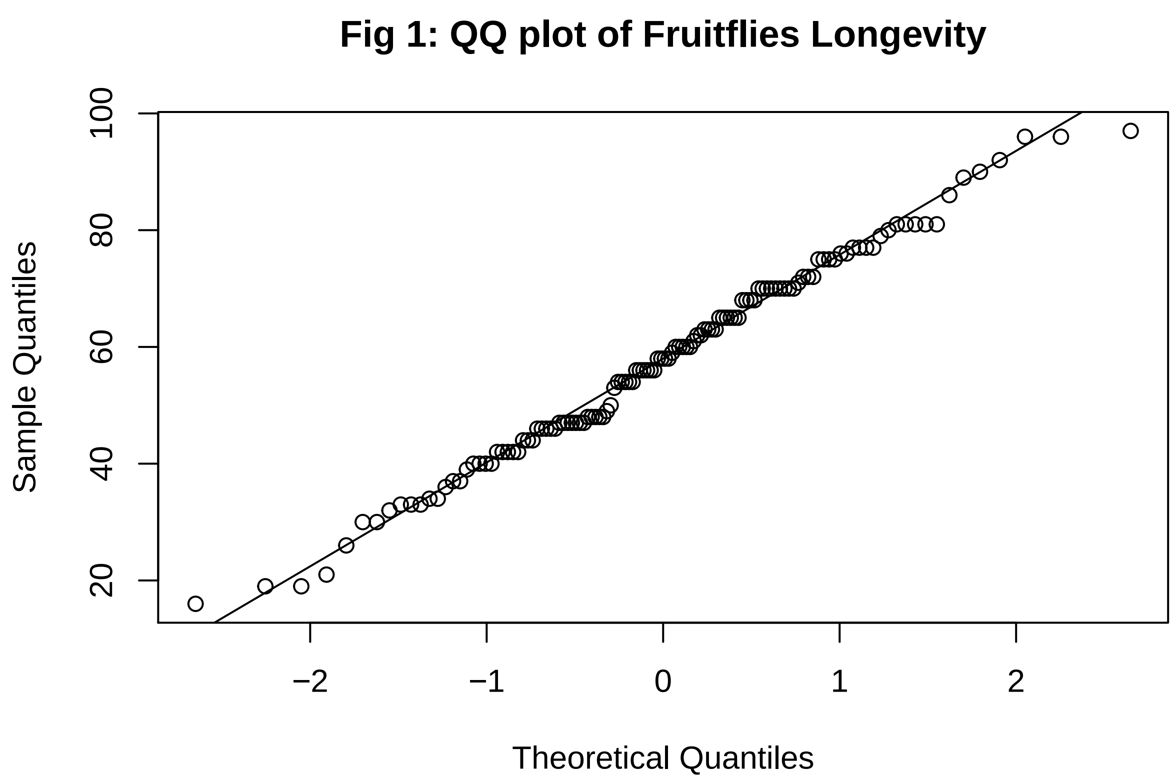
## Appedix

```
# Q1
data('fruitfly', package='faraway')
summary(fruitfly)
```

```
##      thorax          longevity         activity
```

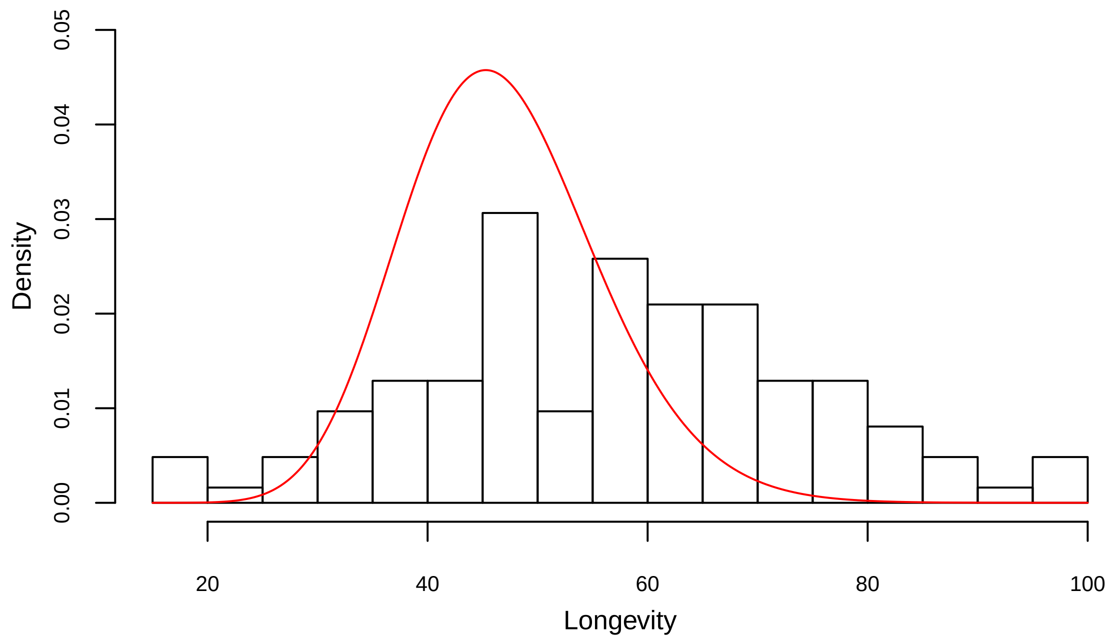
```
##  Min.   :0.6400   Min.   :16.00   isolated:25
##  1st Qu.:0.7600   1st Qu.:46.00   one     :25
##  Median :0.8400   Median :58.00   low     :25
##  Mean   :0.8224   Mean   :57.62   many    :24
##  3rd Qu.:0.8800   3rd Qu.:70.00   high    :25
##  Max.   :0.9400   Max.   :97.00
```

```r
qqnorm(fruitfly$longevity, main='Fig 1: QQ plot of Fruitflies Longevity')
qqline(fruitfly$longevity)
```




```r
fruitfly$tho = fruitfly$thorax - 0.73
m3 = glm(longevity ~ tho + activity, family = Gamma(link = 'log'), data = fruitfly)
shape = 1/summary(m3)$dispersion
scale = exp(m3$coef["(Intercept)"]) / shape
hist(fruitfly$longevity, prob = T, main = " ", xlab = " "
, ylab = " ", breaks = 15, ylim = c(0, 0.05), cex.axis = 0.7)
title(main = "Figure 2: Histogram of Fruitflies Longevity", cex.main = 0.8, line = 0.5)
title(xlab="Longevity", ylab="Density", line=2, cex.lab=0.8)
xSeq = seq(15,100,length = 1000)
lines(xSeq, dgamma(xSeq, shape = shape, scale = scale), col = "red")
```

Figure 2: Histogram of Fruitflies Longevity




```r
m1 <- glm(longevity ~ thorax + activity, family=Gamma(link = 'log'), data=fruitfly)
knitr::kable(summary(m1)$coef, digits=3)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.887 | 0.194 | 9.726 | 0.000 |
| thorax | 2.688 | 0.228 | 11.804 | 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 |

```r
FruitflyTable = as.data.frame(summary(m1)$coef)
FruitflyTable$Lower = FruitflyTable$Estimate - 2*FruitflyTable$'Std. Error'
FruitflyTable$Upper = FruitflyTable$Estimate + 2*FruitflyTable$'Std. Error'
Ratio = exp(FruitflyTable[,c('Estimate','Lower','Upper')])
knitr::kable(Ratio, digits=3)
```

| | Estimate | Lower | Upper |
|---|---|---|---|
| (Intercept) | 6.601 | 4.478 | 9.731 |
| thorax | 14.699 | 9.322 | 23.177 |
| activityone | 1.057 | 0.950 | 1.176 |
| activitylow | 0.890 | 0.800 | 0.990 |
| activitymany | 1.086 | 0.975 | 1.210 |
| activityhigh | 0.661 | 0.593 | 0.736 |

```
# Q2
load('/Users/65421/R/STA442/smoke.RData')
smokeSub = smoke[smoke$Age > 9 & !is.na(smoke$Race) &
                  !is.na(smoke$chewing_tobacco_snuff_or)&
                    !is.na(smoke$RuralUrban), ]
sm1<- glm(chewing_tobacco_snuff_or ~ RuralUrban + Race + Age, family=binomial, data=smokeSub)
knitr::kable(summary(sm1)$coef, digits=3, align = 'l')
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.794 | 0.333 | -26.433 | 0.000 |
| RuralUrbanRural | 0.931 | 0.086 | 10.858 | 0.000 |
| Raceblack | -1.559 | 0.169 | -9.246 | 0.000 |
| Racehispanic | -0.721 | 0.101 | -7.101 | 0.000 |
| Raceasian | -1.590 | 0.340 | -4.671 | 0.000 |
| Racenative | 0.168 | 0.274 | 0.613 | 0.540 |
| Racepacific | 1.158 | 0.337 | 3.440 | 0.001 |
| Age | 0.349 | 0.020 | 17.033 | 0.000 |

```
smokeSub = smoke[smoke$Age > 9 & !is.na(smoke$Race) &
                   !is.na(smoke$ever_tobacco_hookah_or_wa)&
                   !is.na(smoke$Sex), ]
sm2<- glm(ever_tobacco_hookah_or_wa ~ Sex + Race + Age, family=binomial, data=smokeSub)
knitr::kable(summary(sm2)$coef, digits=3, align = 'l')
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.145 | 0.185 | -44.055 | 0.000 |
| SexF | 0.043 | 0.043 | 1.011 | 0.312 |
| Raceblack | -0.583 | 0.070 | -8.317 | 0.000 |
| Racehispanic | 0.415 | 0.048 | 8.708 | 0.000 |
| Raceasian | -0.510 | 0.117 | -4.367 | 0.000 |
| Racenative | 0.143 | 0.190 | 0.749 | 0.454 |
| Racepacific | 0.997 | 0.268 | 3.724 | 0.000 |
| Age | 0.414 | 0.011 | 36.078 | 0.000 |

```
SomkeTable = as.data.frame(summary(sm1)$coef)
SomkeTable$Lower = SomkeTable$Estimate - 2*SomkeTable$'Std. Error'
SomkeTable$Upper = SomkeTable$Estimate + 2*SomkeTable$'Std. Error'
OddRatio = exp(SomkeTable[,c('Estimate','Lower','Upper')])
knitr::kable(OddRatio, digits=3,align = 'l')
```

| | Estimate | Lower | Upper |
|---|---|---|---|
| (Intercept) | 0.000 | 0.000 | 0.000 |
| RuralUrbanRural | 2.537 | 2.137 | 3.011 |
| Raceblack | 0.210 | 0.150 | 0.295 |
| Racehispanic | 0.486 | 0.397 | 0.596 |
| Raceasian | 0.204 | 0.103 | 0.403 |
| Racenative | 1.183 | 0.684 | 2.047 |
| Racepacific | 3.185 | 1.624 | 6.244 |
| Age | 1.417 | 1.361 | 1.477 |