STA442 Assignment 1

Hongbo Du
1003568089
hongbo.du@mail.utoronto.ca

September 25, 2019

Contents

Question 1

Statistical Report

Introduction

This report contains all of the primary statistical results obtained from the data set of fruit fly in the library `faraway` via using R programming language. The problem of this research is to determine whether the activity of the fruit fly would affect the lifetime, given the acknowledge that the thorax length of each fruit fly was measured as this affect lifetime. The hypothesis in this research is that the fruit fly activity would affect the life time, that is the higher activity would result in the shorter lifetime for fruit fly.

Methodology

According to the histogram (in Figure 1.1) of the density versus the longevity of the fruit fly, the histogram is nearly a Gamma distributed graph, and the gamma distribution line adequately fits the data since the mode is relatively close to the mean vertical line (the black dashes line). Since the data was distributed in Gamma distribution, the model is Gamma Generalized Linear Model (Gamma GLM) as a function of the thorax length and activity. Now we shall run the diagnostic of model adequacy. The Figure 1.2 shows that the Gamma GLM fits the data. In order to compare the coefficients of estimates on the natural scale, we shall take the exponential of coefficients of the estimates. Mathematically,

$$f(Y_i, \phi, \nu) = \frac{(x/\phi)x^{\nu-1}\exp(-x/\phi)}{\Gamma(\nu)\phi}, \quad \text{for any } Y_i \sim \text{Gamma}(\mu_i/\nu, \nu)$$

$$\log(\mu_i) = X_i\beta \implies \mu_i = \exp(X_i\beta)$$

where β is the estimate.

Result

By fitting the Gamma Generalized Linear Model, we setting the link as log. Therefore we obtain the coefficients of each groups as the Table 1 shown in the following context. First of all, the expontential coefficient of `thorax` is 14.7 with a P-value of 0.00. Since the P-value is 0.00 and is said to be relatively small, which implies that the thorax length of each male was measured affects the lifetime. Secondly, the exponential coefficient of `activityone` is 1.06 with a P-value of 0.30. The p-value can be found in Table 1.1, and the exponential coefficients can be found in Table 1.2. Since the P-value is 0.30 and is said to be relatively big, which implies that the one pregnant female fruit fly does not affect the lifetime. Thirdly, the exponential coefficient of `activitylow` is 0.89 with a P-value of 0.03. Since the P-value is 0.03 and is said to be relatively small, which implies that the one female fruit fly affects the lifetime. Fourthly, the exponential coefficient of `activitymany` is 1.09 with a P-value of 0.13. Since the P-value is 0.13 and is said to be relatively big, which implies that the eight pregnant female fly does not affect the lifetime. Lastly, the exponential coefficient of `activityhigh` is 0.66 with a P-value of 0.00. Since the P-value is 0.00 and is said to be

relatively small, which implies that the eight female fruit fly affects the lifetime. This result can also be showed by the Figure 1.3 and 1.4. Figure 1.3 shows the relation of activity against longevity, which indicates that the high and low groups have the shorter lifetime. Figure 1.4 shows the relation of thorax length against longevity, classified by activity. The trends in this figure shows that the shorter thorax length results in the shorter longevity, as well as the higher activity results the shorter longevity.

Non-technical Summary

In order to find out whether the thorax length and activity of fruit fly would affect the lifetime, there are one hundred and twenty-five fruit flies that were recorded the lifetime, thorax length, and activity, and those fruit flies were divided into five groups by activity. According to the statistical result, the group of high activity comes along with the shorter lifetime, as well as the group of low activity. However, the group of low activity has a longer lifetime than the group of high activity. Similarly, when it comes to the thorax length, the shorter thorax length results in the shorter lifetime for fruit flies as well. Hence the conclusion is that the thorax length and the activity affect the lifetime.

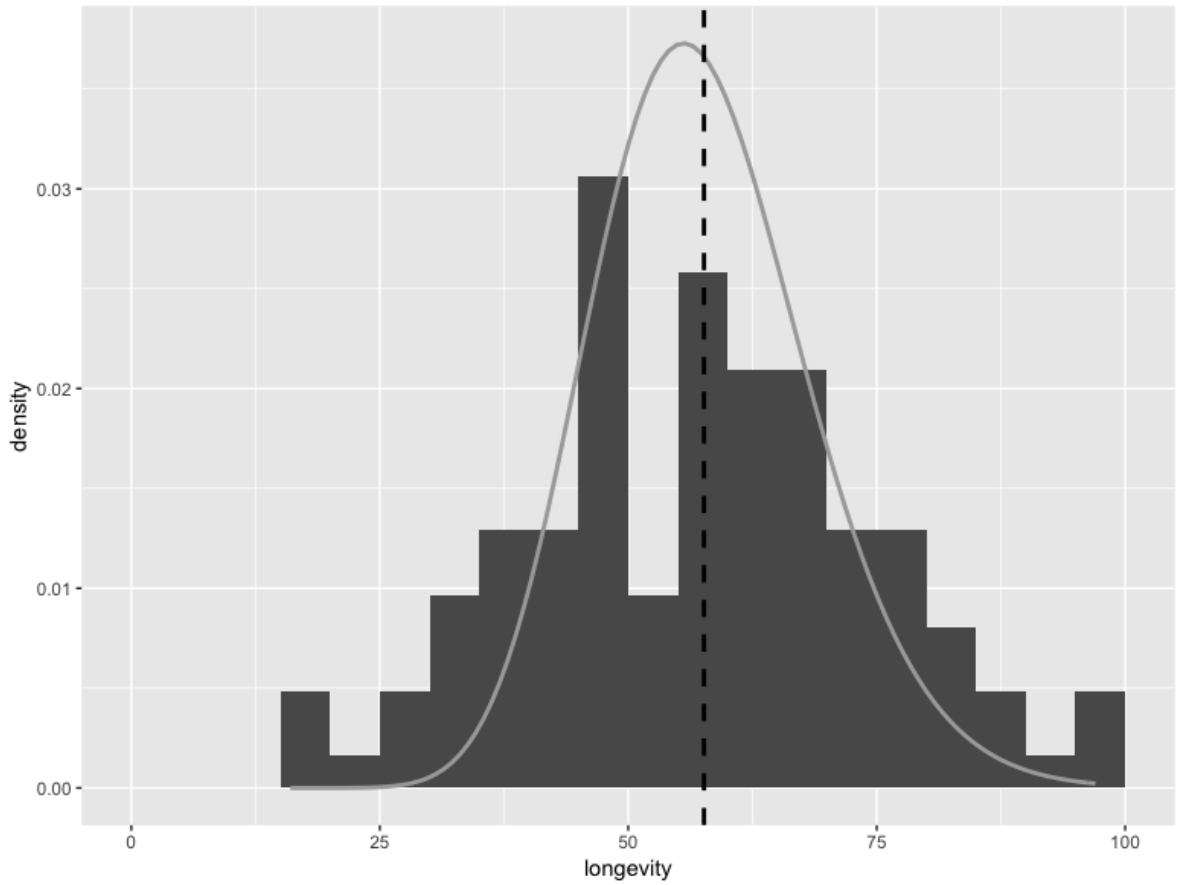
Table and Figures

	Estimate	Std. Error	t-value	P-value
(Intercept)	1.89	0.19	9.73	0.00
thorax	2.67	0.23	11.80	0.00
activityone	0.06	0.05	1.04	0.30
activitylow	−0.12	0.05	−2.18	0.03
activitymany	0.08	0.05	1.52	0.13
activityhigh	−0.41	0.05	−7.69	0.00
shape	28.15	NA	NA	NA

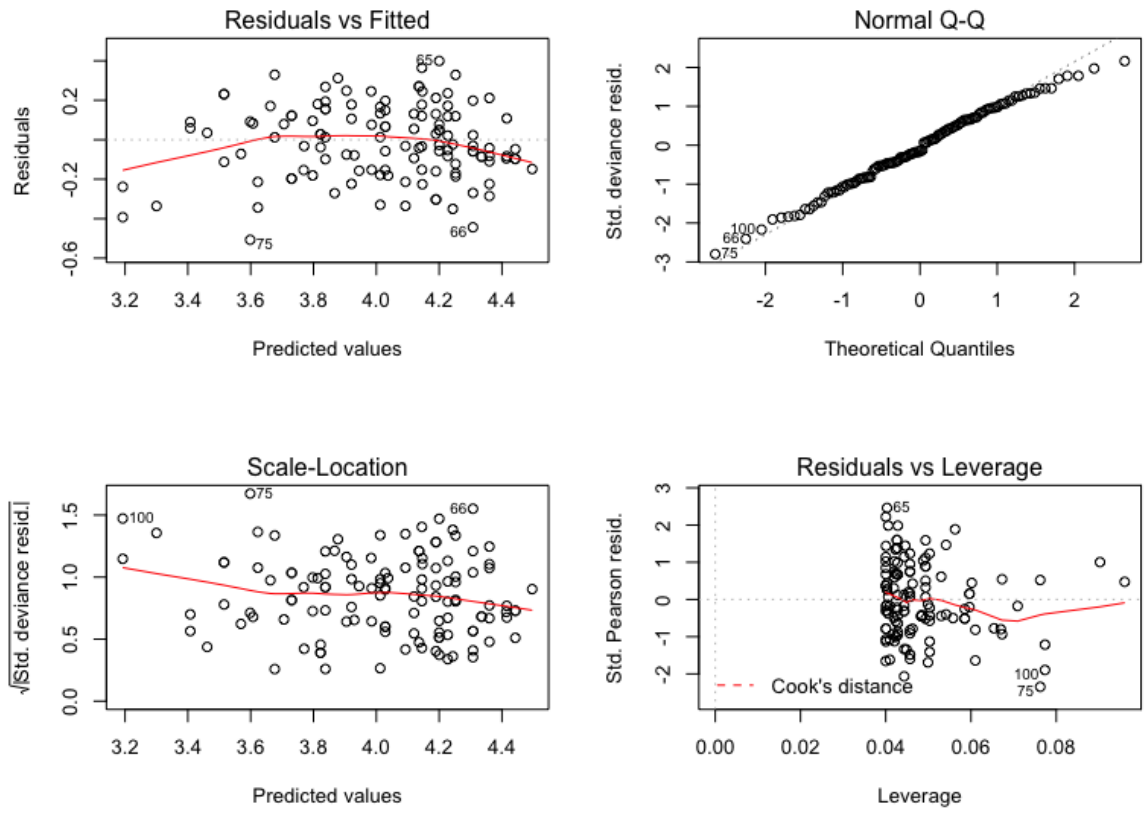
(Table 1.1. The results of coefficients of Gamma GLM.)

intercept	6.60	activitylow	0.89
thorax	14.70	activitymany	1.09
activityone	1.06	activityhigh	0.66

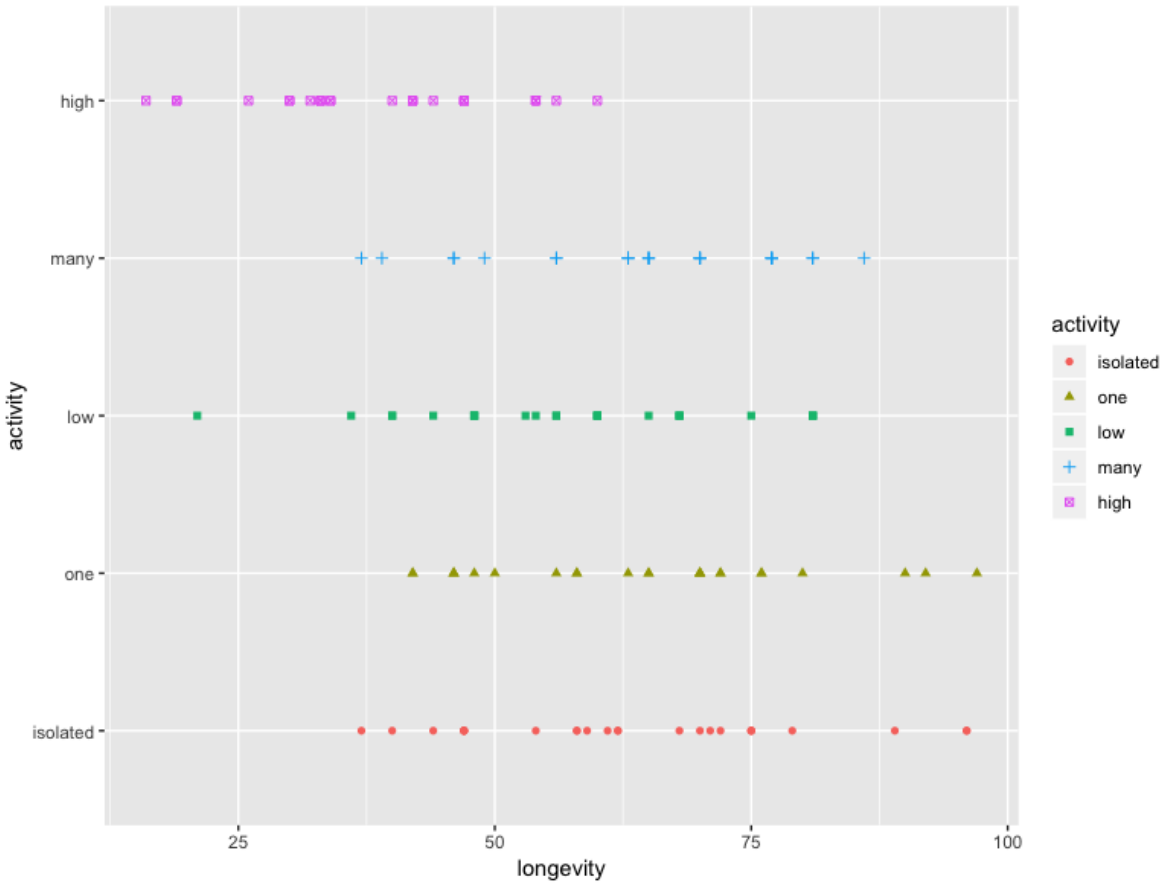
(Table 1.2. The exponential coefficients of GLM for hypothesis 2.)



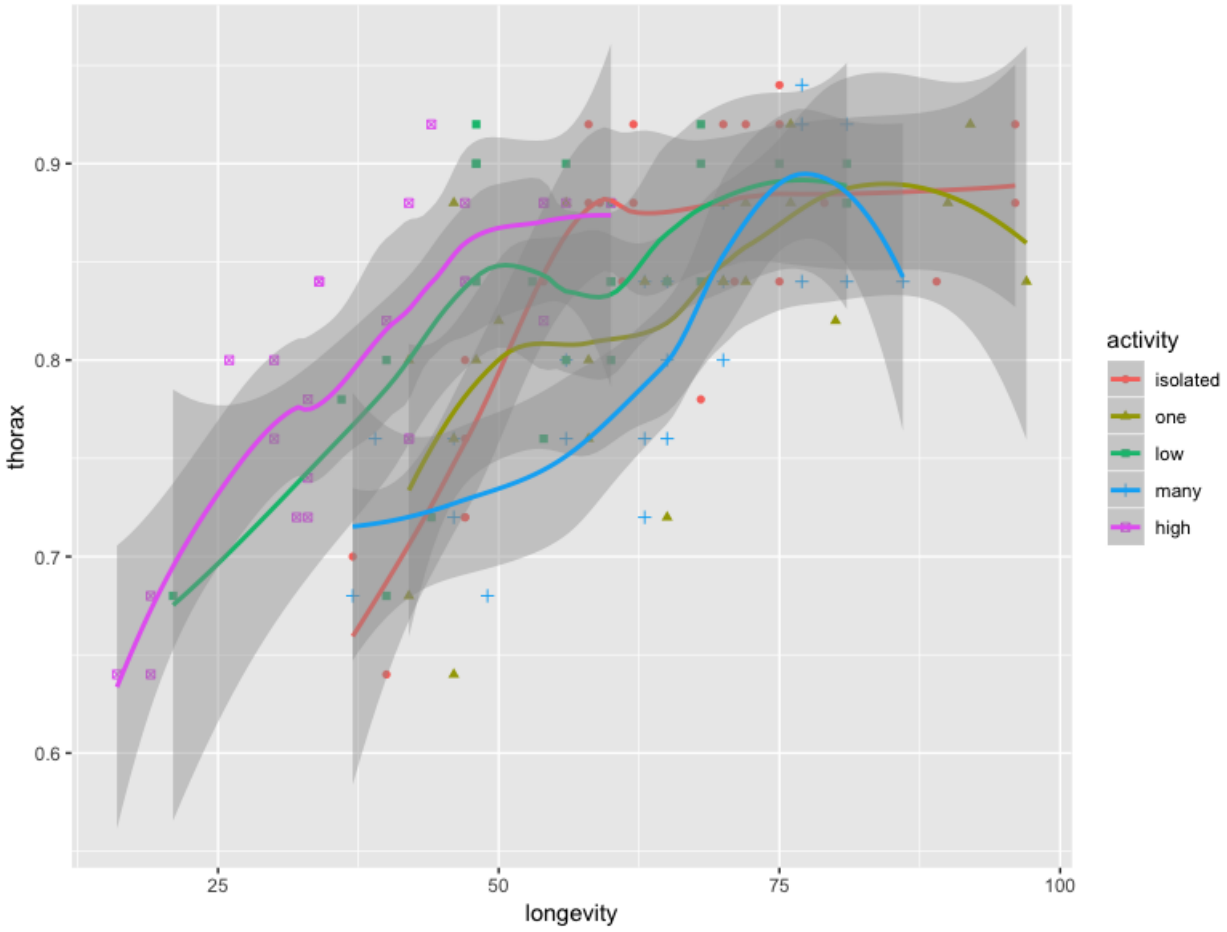
(Figure 1.1. The histogram of longevity versus density.)



(Figure 1.2. The diagnostic plot.)



(Figure 1.3. The scatter plot of longevity versus the activity.)



(Figure 1.4. The scatter plot with smooth lines of longevity versus the thorax length.)

Question 2

Non-technical Summary

The 2014 American National Youth Tobacco Survey contains the use of cigars, hookahs, and chewing tobacco amongst American school children. Using R programming language to analyze this data set would give us some information about the use of tobacco amongst American school children, such as the use of chewing tobacco and the use of hookahs. We found that chewing tobacco is more popular amongst white people than amongst African-American and Hispanic-American, and this phenomenon is found popular in rural area. Another conclusion is that there is very small difference between male and female in the use of hookah or waterpipe, given the similar age, ethnicity, and other demographic characteristics backgrounds.

Statistical Report

Introduction

This report contains the statistical results obtained from the data set of The 2014 American National Youth Tobacco Survey via using R programming language. The data contains the variables of race, age, sex, and varies uses of tobaccos. The goal of this report is to investigate two hypothesises. The first hypothesis is that the regular use of chewing tobacco, snuff or dip is no more common amongst white Americans than for Hispanic-Americans and African-Americans, once one accounts for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon. The second hypothesis is that given the age, ethnicity, and other demographic characteristics are similar, the likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes.

Methodology

In order to analyze the data, we fit two Logistic Generalized Linear Models to model the used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days as a function of race, age and demographic characteristics, and ever used a hookah or waterpipe as a function of sex, race, age and demographic characteristics. Mathematically, the first Logistic Generalized Linear Models can be written as

$$\log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) = \beta_0+\beta_1 I_{\texttt{Rural}_i}+\beta_2 I_{\texttt{Black}_i}+\beta_3 I_{\texttt{Hispanic}_i}+\beta_4 I_{\texttt{Asian}_i}+\beta_5 I_{\texttt{Native}_i}+\beta_6 I_{\texttt{Pacific}_i}, +\beta_7 x_{\texttt{Age}_i}$$

and the second Logistic Generalized Linear Models can be written as

$$\log\left(\frac{\hat{\pi}_2}{1-\hat{\pi}_2}\right) = \beta_0+\beta_1 I_{\texttt{Rural}_i}+\beta_2 I_{\texttt{Black}_i}+\beta_3 I_{\texttt{Hispanic}_i}+\beta_4 I_{\texttt{Asian}_i}+\beta_5 I_{\texttt{Native}_i}+\beta_6 I_{\texttt{Pacific}_i}, +\beta_7 x_{\texttt{Age}_i}+\beta_8 I_{\texttt{SexF}_i}$$

where $\hat{\pi}_1$ is the estimated probability that the i-th person chewing tobacco and $\hat{\pi}_2$ is the estimated probability that the i-th person ever used hookah or waterpipe. The $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$, and β_7 in both models are the coefficients of whether the i-th person lives in rural, is black, Hispanic, Asian, native, pacific, and the age, respectively. β_8 is whether the i-th person is female. According to the definition of logistic regression, we shall take the exponential of the coefficient of estimates, that is,

$$\log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) = \sum_{p=1}^{P} X_{ip}\beta_p \implies \left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) = \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}}$$

where β_p is the coefficient of estimates.

Result

According to the four tables in the following context, we can get the conclusions for this research. Let us begin with the test of hypothesis 1. According to the exponential coefficient of estimate of `RuralUrbanRural`, 2.54 (in Table 2), and the p-value of 0.00 (in Table 1), this

indicates that people living in rural area still have a high rate of use of chewing tobacco, snuff or dip. The variables `Raceblack` and `Racehispanic` have exponential coefficient of estimates 0.21 and 0.49 (in Table 2), respectively. Both the p-value of those two variables are 0.00 (in Table 1). Therefore, we can conclude that the popularity of the use of chewing tobacco, snuff or dip is not common as amongst the white Americans. In terms of age, age seems to have zero effect on the result of this hypothesis, however, older population might used to live in rural area. Hence considering the age is essential. The exponential coefficient of `age` is 1.42 (in Table 2) and the p-value is 0.00 (in Table 1), which indicates that older people use the chewing tobacco more often. This result also satisfies the fact that once white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon. Therefore, we may not reject the null hypothesis. For hypothesis 2, the exponential coefficient of the estimate of `SexF` is 1.04 (in Table 4) with a p-value of $0.33 > 0.05$ (in Table 3). We shall reject the null hypothesis, which implies that there is no difference in likelihood between male and female in the use of tobacco by smoking hookah and waterpipe, given the similar age, ethnicity, and other demographic characteristics backgrounds.

Tables

	Estimate	Std. Error	t-value	P-value
(Intercept)	−8.79	0.33	−26.43	0.00
RuralUrbanRural	0.93	0.09	10.86	0.00
Raceblack	−1.56	0.17	−9.25	0.00
Racehispanic	−0.72	0.10	−7.10	0.00
Raceasian	−1.59	0.34	−4.67	0.00
Racenative	0.17	0.27	0.61	0.54
Racepacific	1.16	0.34	3.44	0.00
Age	0.35	0.02	17.03	0.00
shape	1.00	NA	NA	NA

(Table 1. The results of coefficients of GLM for hypothesis 1)

intercept	0.00	Racenative	1.18
RuralUrbanRural	2.54	Racepacific	3.18
Raceblack	0.21	Age	1.42
Racehispanic	0.49	shape	1.00
Raceasian	0.20		

(Table 2. The exponential coefficients of estimates of GLM for hypothesis 1)

	Estimate	Std. Error	t-value	P-value
(Intercept)	−8.00	0.19	−43.11	0.00
RuralUrbanRural	−0.39	0.04	−8.77	0.00
Raceblack	−0.63	0.07	−9.01	0.00
Racehispanic	0.35	0.05	7.14	0.00
Raceasian	−0.63	0.12	−5.36	0.00
Racenative	0.16	0.19	0.84	0.40
Racepacific	0.96	0.27	3.57	0.00
Age	0.42	0.01	36.27	0.00
SexF	0.04	0.04	0.98	0.33
shape	1.00	NA	NA	NA

(Table 3. The results of coefficients of estimates of GLM for hypothesis 2)

intercept	0.00	Racenative	1.17
RuralUrbanRural	0.58	Racepacific	2.62
Raceblack	0.53	Age	1.52
Racehispanic	1.41	SexF	1.04
Raceasian	0.53	shape	1.00

(Table 4. The exponential coefficients of GLM for hypothesis 2)

Appendix

The appendix contains all the R code and its output.

STA442 Assignment 1 R Code

```
library(ggplot2)
```

Question 1

```
# Load data
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
```

```
# Construct a Gamma GLM to model lifetime as a function of thorax length and activity.
glmGamma <- glm(longevity ~ thorax + activity, family=Gamma(link = 'log'),
                data=fruitfly)

summary(glmGamma)
```

```
##
## Call:
## glm(formula = longevity ~ thorax + activity, family = Gamma(link = "log"),
##     data = fruitfly)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.50718  -0.15216  -0.02833   0.12434   0.39938
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.88722    0.19405   9.726  < 2e-16 ***
## thorax        2.68778    0.22769  11.804  < 2e-16 ***
## activityone   0.05527    0.05337   1.036   0.3024
## activitylow  -0.11646    0.05332  -2.184   0.0309 *
## activitymany  0.08250    0.05413   1.524   0.1302
## activityhigh -0.41466    0.05394  -7.687 4.93e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.0355297)
##
##     Null deviance: 13.2803  on 123  degrees of freedom
## Residual deviance:  4.3151  on 118  degrees of freedom
## AIC: 942.29
##
## Number of Fisher Scoring iterations: 4
```

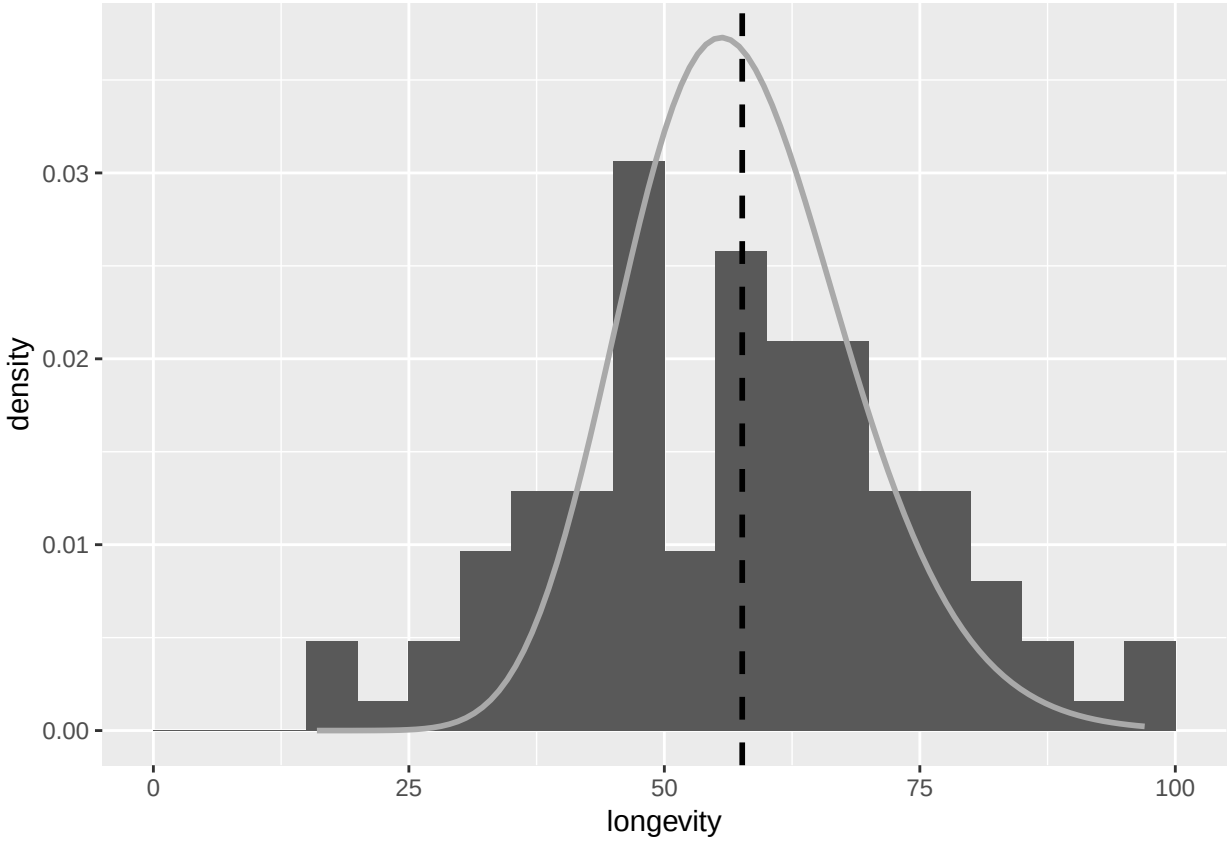
```
# Make a table of the coefficients
knitr::kable(rbind(summary(glmGamma)$coef,
                   shape = c(1 / summary(glmGamma)$dispersion, NA, NA, NA)),
             digits=2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.89	0.19	9.73	0.00
thorax	2.69	0.23	11.80	0.00
activityone	0.06	0.05	1.04	0.30
activitylow	-0.12	0.05	-2.18	0.03
activitymany	0.08	0.05	1.52	0.13
activityhigh	-0.41	0.05	-7.69	0.00
shape	28.15	NA	NA	NA

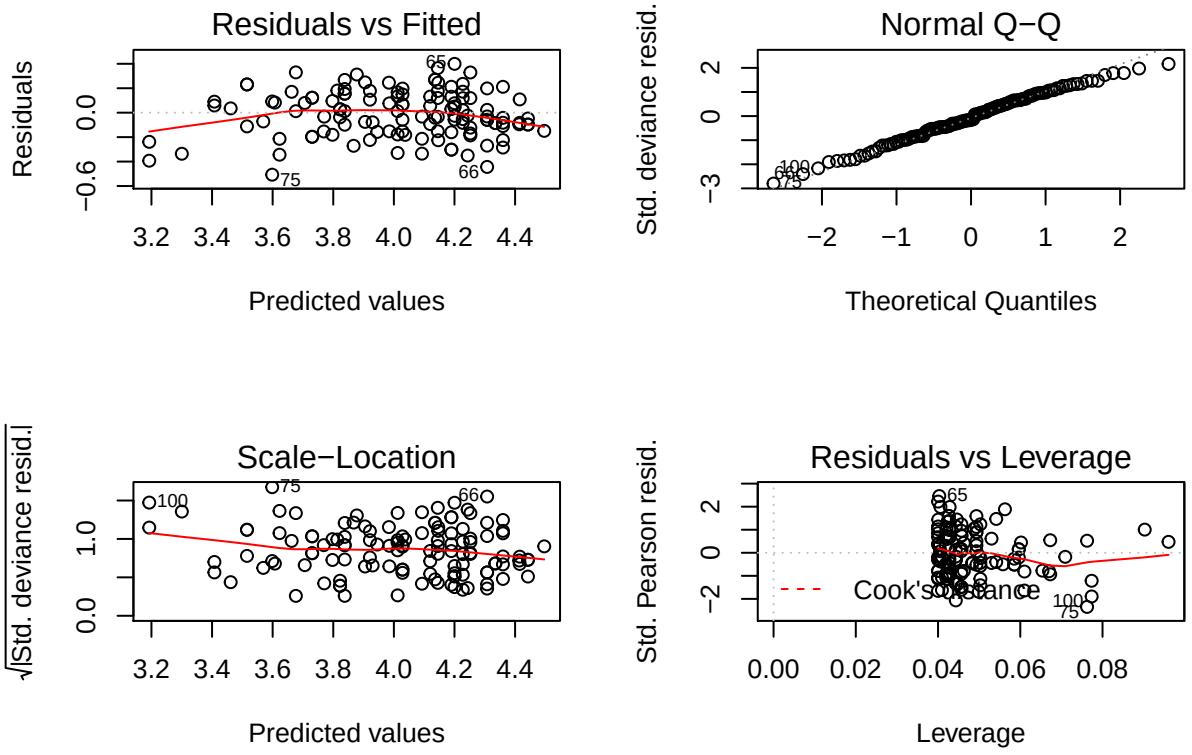
```
glmGamma_coef <- summary(glmGamma)$coef
exp(glmGamma_coef[,1])
```

```
##  (Intercept)       thorax  activityone  activitylow activitymany
##    6.6010112   14.6990246    1.0568300    0.8900653    1.0859945
## activityhigh
##    0.6605655
```

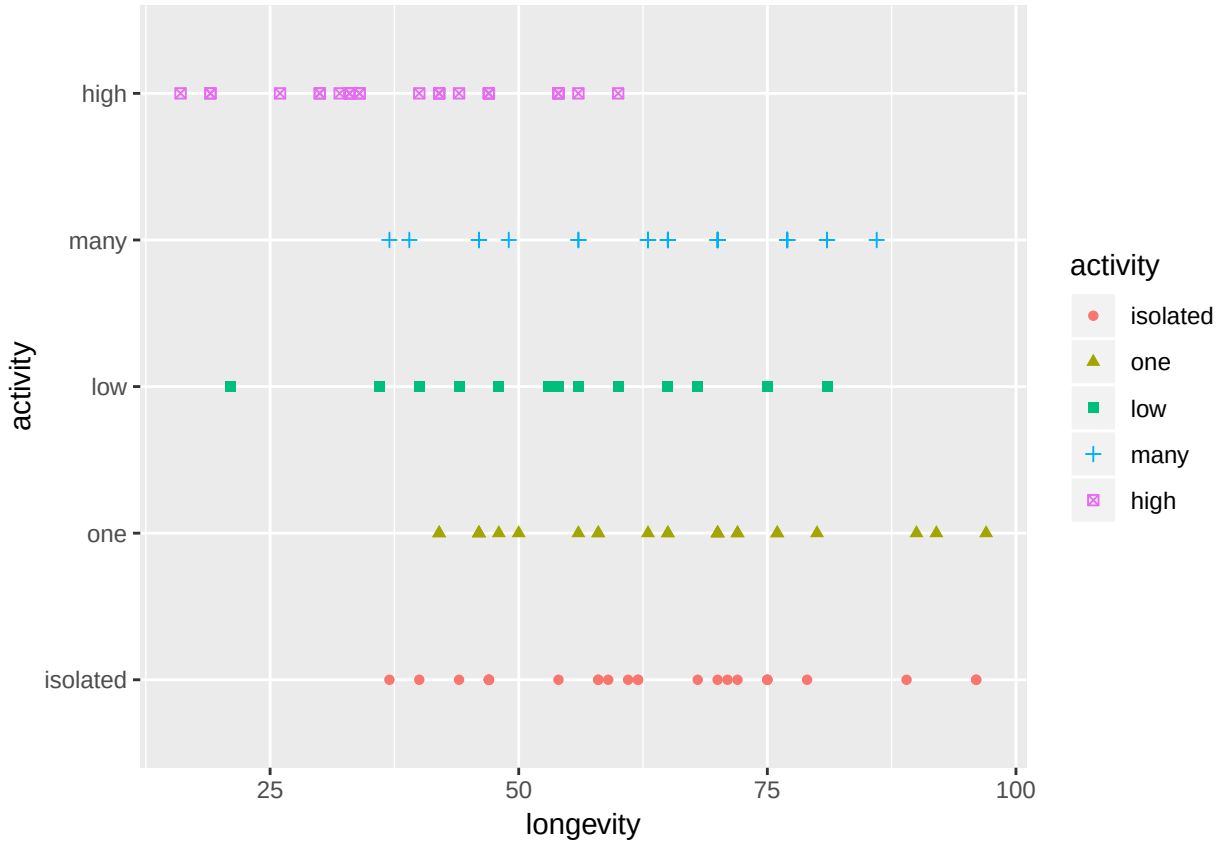
```
#Set the shape and the scale
shape <- 1/summary(glmGamma)$dispersion
# By taking the hint, consider centering and rescaling variables
scale <- mean(fruitfly$longevity)/shape

#Histogram of longevity
ggplot(glmGamma, aes(x = longevity)) +
  geom_histogram(aes(y=..density..), breaks = seq(0, 100, by = 5)) +
  stat_function(fun = dgamma, args = list(shape = shape, scale = scale),
                color = 'darkgray', size = 1) +
  geom_vline(aes(xintercept = mean(fruitfly$longevity)),
             linetype="dashed", size = 1)
```




```
# Run diagnostic of the model
par(mfrow = c(2, 2))
plot(glmGamma)
```

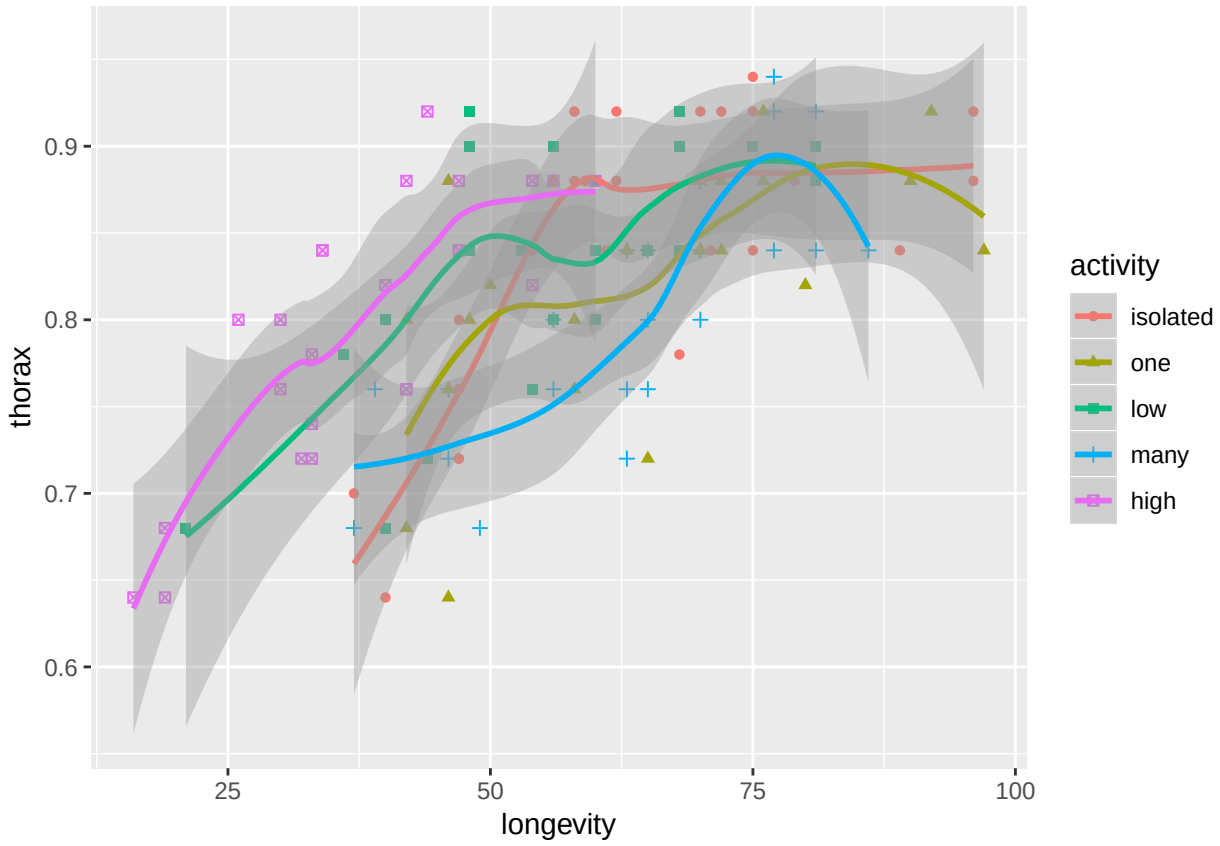


```
# This plot is to see whether the longevity has correlation to activity.
ggplot(data = fruitfly,
       aes(x = longevity, y = activity, color = activity, shape = activity)) +
  geom_point()
```




```
# This plot is to see whether the longevity has correlation to thorax length.
ggplot(data = fruitfly,
       aes(x = longevity, y = thorax, color = activity, shape = activity)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Question 2

```
# Load data
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext = '.RData'))
```

```
## [1] "/var/folders/zk/1674vlt15wgd_w_7ldmdgdpw0000gn/T//RtmptbnZgK/file7d41010ebeb.RData"
```

```
download.file(smokeUrl, smokeFile, made = "wb")
(load(smokeFile))
```

```
## [1] "smoke"        "smokeFormats"
```

```
# Get rid of 9 year olds because their data is suspicious.
smokeSub = smoke[smoke$Age >= 10, ]

# GLM of hypothesis 1
glm1 <- glm(chewing_tobacco_snuff_or ~ RuralUrban + Race + Age,
            family = binomial(link = 'logit'), data = smokeSub)

# Make a table of the coefficients of hypothesis 1
knitr::kable(rbind(summary(glm1)$coef,
                   shape = c(1 / summary(glm1)$dispersion,
                             NA, NA, NA)), digits = 2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.79	0.33	-26.43	0.00

	Estimate	Std. Error	z value	Pr(>|z|)
RuralUrbanRural	0.93	0.09	10.86	0.00
Raceblack	-1.56	0.17	-9.25	0.00
Racehispanic	-0.72	0.10	-7.10	0.00
Raceasian	-1.59	0.34	-4.67	0.00
Racenative	0.17	0.27	0.61	0.54
Racepacific	1.16	0.34	3.44	0.00
Age	0.35	0.02	17.03	0.00
shape	1.00	NA	NA	NA

```
# Taking the exponential of the coefficients of estimates.
glm1_coef <- summary(glm1)$coef
exp(glm1_coef[,1])
```

```
##     (Intercept) RuralUrbanRural       Raceblack    Racehispanic
##    0.0001516506    2.5367216911    0.2104106311    0.4864772004
##       Raceasian      Racenative     Racepacific             Age
##    0.2039272970    1.1830170736    3.1845058749    1.4173936694
```

```
# GLM of hypothesis 2
glm2 <- glm(ever_tobacco_hookah_or_wa ~ RuralUrban + Race + Age + Sex,
            family = binomial(link = 'logit'), data = smokeSub)

# Make a table of the coefficients of hypothesis 2
knitr::kable(rbind(summary(glm2)$coef,
                   shape = c(1/summary(glm2)$dispersion,
                             NA, NA, NA)), digits = 2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.00	0.19	-43.11	0.00
RuralUrbanRural	-0.39	0.04	-8.77	0.00
Raceblack	-0.63	0.07	-9.01	0.00
Racehispanic	0.35	0.05	7.14	0.00
Raceasian	-0.63	0.12	-5.36	0.00
Racenative	0.16	0.19	0.84	0.40
Racepacific	0.96	0.27	3.57	0.00
Age	0.42	0.01	36.27	0.00
SexF	0.04	0.04	0.98	0.33
shape	1.00	NA	NA	NA

```
# Taking the exponential of the coefficients of estimates.
glm2_coef <- summary(glm2)$coef
exp(glm2_coef[,1])
```

```
##     (Intercept) RuralUrbanRural       Raceblack    Racehispanic
##    0.0003343861    0.6781289818    0.5300868553    1.4128802035
##       Raceasian      Racenative     Racepacific             Age
##    0.5321126589    1.1730570246    2.6212523192    1.5199046171
##            SexF
##    1.0429494606
```