Homework Assignment 3

Jing Su

Due 8am, 2022-12-10

Total points: 400 points

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, include the code in your solution, along with any plots.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Unit 4

Problem 1. (20 points)

In vertebrates, sweet and savory ("umami") tastes are sensed by receptors termed T1Rs. Most vertebrates have three T1Rs, with T1R2 and T1R3 receptors working together to detect sugars (carbohydrates) and artificial sweeteners, while the T1R1-T1R3 heterodimer mediates umami taste. However, even though birds lack *T1R2* genes, several avian species display high behavioral affinity for nectar or sweet fruit. Receptor expression studies in hummingbirds revealed that the ancestral umami receptor (T1R1-T1R3) has been repurposed to detect sugars.[1]

Researchers investigated whether T1R1-T1R3 function would dictate hummingbird taste behavior. In a series of field tests, hummingbirds were presented simultaneously with two filled containers, one containing test stimuli and a second containing sucrose. The test stimuli included aspartame, erythritol, water, and sucrose. Aspartame is an artificial sweetener that tastes sweet to humans, but is not detected by hummingbird T1R1-T1R3, while erythritol is an artificial sweetener that is known to activate T1R1-T1R3.

Data on how long a hummingbird drank from a particular container for a given trial, measured in seconds, is in the file `hummingbirds.Rdata`. Variable names ending in `1` correspond to the test stimuli, while names ending in `2` correspond to sucrose. For example, in the first field test comparing aspartame and sucrose, a hummingbird drank from the aspartame container for 0.54 seconds and from the sucrose container for 3.21 seconds.

Do the data suggest that T1R1-T1R3 play the described role in hummingbird taste behavior?

To answer this question, analyze the data for each set of trials: aspartame versus sucrose, erythritol versus sucrose, water versus sucrose, and sucrose versus sucrose. Let $\alpha = 0.05$. Write a conclusion summarizing and interpreting the results, referencing numerical results (such as p-values) where appropriate.

Answer:

Here we have taken the null hypothesis as T1R1-T1R3 can detect sugars like surcrose and artificial sweetner like erythritol. Alternative hypothesis is T1R1-T1R3 cannot detect sugars like surcrose and artificial sweetner like erythritol. The p value of aspartame vs sucrose in the test sample is 0.05. So, we reject the null hypothesis. In the study, it was given that T1R1-T1R3 cannot detect aspartame. The p values for erythitol in the test sample is 1. We therefore accept the null hypothesis. This concludes that 1R1-T1R3 play the described role in hummingbird taste behavior.

[1] Baldwin, et al. Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. *Science* 2014; **345**: 929-933.

```
setwd('/users/sreyatummala/downloads/')
load('hummingbirds.Rdata')
str(hummingbirds)
```

```
## 'data.frame':    39 obs. of  8 variables:
##  $ asp.vs.sucr.1 : num  0.54 0.63 0.33 1.3 0.68 0.33 0.43 0.37 0.31 0.28 ...
##  $ asp.vs.sucr.2 : num  3.21 2.48 1.87 2.85 2.42 2.19 4.19 1.94 1.46 1.47 ...
##  $ ery.vs.sucr.1 : num  5.53 1.57 1.1 3.01 1.6 1.58 4.03 1.37 0.58 0.37 ...
##  $ ery.vs.sucr.2 : num  5.13 2.16 2.28 0.72 2.27 1.47 2.25 0.9 0.67 0.27 ...
##  $ wat.vs.sucr.1 : num  0.39 0.33 1.97 0.38 0.47 0.55 0.63 0.31 0.49 0.34 ...
##  $ wat.vs.sucr.2 : num  4.26 3.43 7.07 5.57 4.02 3.42 2.67 0.99 1.7 0.78 ...
##  $ sucr.vs.sucr.1: num  4.93 3.23 1.78 0.17 0.88 ...
##  $ sucr.vs.sucr.2: num  5.73 3.48 1.18 1.67 3.51 1.81 1.28 6.26 3.82 6.37 ...
```

```
# 1. aspartame vs sucrose
# alpha is given as 0.05

# Test sample
#calculating mean of asp Vs sucr1
summary(hummingbirds$asp.vs.sucr.1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.2300  0.3200  0.3700  0.4936  0.5850  1.3000      28
```

```
#calculating 95% confidence interval
t.test(hummingbirds$asp.vs.sucr.1, na.rm = TRUE, conf.level = 0.95)$conf.int
```

```
## [1] 0.2890427 0.6982300
## attr(,"conf.level")
## [1] 0.95
```

```
#calculating t statistic and p value for asp vs sucr test sample
t.test(hummingbirds$asp.vs.sucr.1, mu = 0.6982300, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$asp.vs.sucr.1
## t = -2.2281, df = 10, p-value = 0.05
## alternative hypothesis: true mean is not equal to 0.69823
## 95 percent confidence interval:
##  0.2890427 0.6982300
## sample estimates:
## mean of x
## 0.4936364
```

```
# Sucrose sample
summary(hummingbirds$asp.vs.sucr.2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.390   1.670   2.190   2.315   2.665   4.190      28
```

```
#calculating 95% CI
t.test(hummingbirds$asp.vs.sucr.2, na.rm = TRUE, conf.level = 0.95)$conf.int
```

```
## [1] 1.741615 2.889294
## attr(,"conf.level")
## [1] 0.95
```

```r
#calculating t statistic and p value for asp vs sucr sucrose sample
t.test(hummingbirds$asp.vs.sucr.2, mu = 2.315, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$asp.vs.sucr.2
## t = 0.0017649, df = 10, p-value = 0.9986
## alternative hypothesis: true mean is not equal to 2.315
## 95 percent confidence interval:
##  1.741615 2.889294
## sample estimates:
## mean of x
##  2.315455
```

```r
# 2. erythritol versus sucrose
# test sample
summary(hummingbirds$ery.vs.sucr.1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.250   0.580   1.330   1.508   1.600   5.530      14
```

```r
#calculating t statistic and p value
t.test(hummingbirds$ery.vs.sucr.1, mu = 1.508, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$ery.vs.sucr.1
## t = 0, df = 24, p-value = 1
## alternative hypothesis: true mean is not equal to 1.508
## 95 percent confidence interval:
##  1.002384 2.013616
## sample estimates:
## mean of x
##     1.508
```

```r
#sucrose sample
summary(hummingbirds$ery.vs.sucr.2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.270   0.570   1.470   1.487   2.250   5.130      14
```

```r
#calculating t statistic and p value
t.test(hummingbirds$ery.vs.sucr.2, mu = 1.487, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$ery.vs.sucr.2
## t = -0.00091565, df = 24, p-value = 0.9993
## alternative hypothesis: true mean is not equal to 1.487
## 95 percent confidence interval:
##  1.035993 1.937607
## sample estimates:
## mean of x
##    1.4868
```

```r
# 3. water versus sucrose
# test sample
summary(hummingbirds$wat.vs.sucr.1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.250   0.310   0.350   0.461   0.470   1.970      18
```

```r
#calculating t statistic and p value
t.test(hummingbirds$wat.vs.sucr.1, mu = 0.461, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$wat.vs.sucr.1
## t = -0.00060544, df = 20, p-value = 0.9995
## alternative hypothesis: true mean is not equal to 0.461
## 95 percent confidence interval:
##  0.2968871 0.6250177
## sample estimates:
## mean of x
## 0.4609524
```

```r
#sucrose sample
summary(hummingbirds$wat.vs.sucr.2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.630   1.080   1.810   2.606   4.020   7.070      18
```

```r
#calculating t statistic and p value
t.test(hummingbirds$wat.vs.sucr.2, mu = 2.606, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$wat.vs.sucr.2
## t = -0.00072904, df = 20, p-value = 0.9994
## alternative hypothesis: true mean is not equal to 2.606
## 95 percent confidence interval:
##  1.788218 3.423211
## sample estimates:
## mean of x
##  2.605714
```

```r
# 4. sucrose versus sucrose
# test sample
summary(hummingbirds$sucr.vs.sucr.1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.170   0.720   1.350   1.581   1.855   7.470
```

```r
#calculating t statistic and p value
t.test(hummingbirds$sucr.vs.sucr.1, mu = 1.581, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$sucr.vs.sucr.1
## t = -0.0011404, df = 38, p-value = 0.9991
```

```
## alternative hypothesis: true mean is not equal to 1.581
## 95 percent confidence interval:
##  1.125591 2.035897
## sample estimates:
## mean of x
##  1.580744
```

```
# sucrose sample
summary(hummingbirds$sucr.vs.sucr.2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.070   0.385   1.180   1.608   1.820   6.370
```

```
#calculating t statistic and p value
t.test(hummingbirds$sucr.vs.sucr.2, mu = 1.608 , alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  hummingbirds$sucr.vs.sucr.2
## t = -0.001138, df = 38, p-value = 0.9991
## alternative hypothesis: true mean is not equal to 1.608
## 95 percent confidence interval:
##  1.060324 2.155061
## sample estimates:
## mean of x
##  1.607692
```

```
# correlation testing among the samples
cor.test(hummingbirds$asp.vs.sucr.1,hummingbirds$asp.vs.sucr.2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hummingbirds$asp.vs.sucr.1 and hummingbirds$asp.vs.sucr.2
## t = 1.4259, df = 9, p-value = 0.1876
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2297644  0.8184034
## sample estimates:
##       cor
## 0.4292779
```

```
cor.test(hummingbirds$ery.vs.sucr.1 ,hummingbirds$ery.vs.sucr.2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hummingbirds$ery.vs.sucr.1 and hummingbirds$ery.vs.sucr.2
## t = 3.374, df = 23, p-value = 0.002618
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2333127 0.7907468
## sample estimates:
##       cor
## 0.5753982
```

```
cor.test(hummingbirds$wat.vs.sucr.1,hummingbirds$wat.vs.sucr.2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hummingbirds$wat.vs.sucr.1 and hummingbirds$wat.vs.sucr.2
## t = 3.2227, df = 19, p-value = 0.00448
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2190270 0.8166145
## sample estimates:
##       cor
## 0.5945029
```

```
cor.test(hummingbirds$sucr.vs.sucr.1,hummingbirds$sucr.vs.sucr.2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  hummingbirds$sucr.vs.sucr.1 and hummingbirds$sucr.vs.sucr.2
## t = 0.99802, df = 37, p-value = 0.3248
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1618778  0.4542219
## sample estimates:
##       cor
## 0.1619088
```

Problem 2. (20 points)

A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks about family history of cancer. So far, people who sign up complete an average of 4 surveys, with standard deviation 2.2. The research group wants to try a new interface that they think may encourage new enrollees to complete more surveys. They plan to randomize each enrollee to either the old or new interface.

a) How many new enrollees do they need for each group (old or new interface) to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%? Let $\alpha = 0.05$.

5.8 ~ 6 new enrollees are needed for each group. But when the size is determine as 0.5, we have the enrollees as 64.

b) Explain the effect of increasing α on the power of the test. What is one disadvantage to increasing α, from a decision-making standpoint?

When alpha value is increased, it becomes easier to reject the null hypothesis. This reduces the probability of type-2 error. Alpha and power have inverse proportionality. But increasing alpha can result in chances of rejecting the null hypothesis when it is true. This can result in type-1 error which might have effects on data exploration and interpretation.

```
power.t.test(n = NULL, delta = 4 , sd = 2.2, sig.level = 0.05, power = 0.80)
```

```
##
##      Two-sample t test power calculation
##
##               n = 5.883306
##           delta = 4
##              sd = 2.2
```

```
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
power.t.test(d=0.5,sig.level=0.05,power=0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 0.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Unit 5

Problem 3.

Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Suppose a study was conducted to investigate the relationship between coffee consumption and exercise. Participants were randomly recruited from the undergraduate and graduate student populations of universities in the Boston/Cambridge area. Participants were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data, the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The file `coffee_exercise.Rdata` contains simulated MET data for the study participants, based on the amount of coffee consumed. The consumption groups are labeled `A - E`.

- `A`: 1 cup or less of caffeinated coffee consumed per week
- `B`: 2 to 6 cups of caffeinated coffee consumed per week
- `C`: 1 cup of caffeinated coffee consumed per day
- `D`: 2 to 3 cups of caffeinated coffee consumed per day
- `E`: 4 or more cups of caffeinated coffee consumed per day

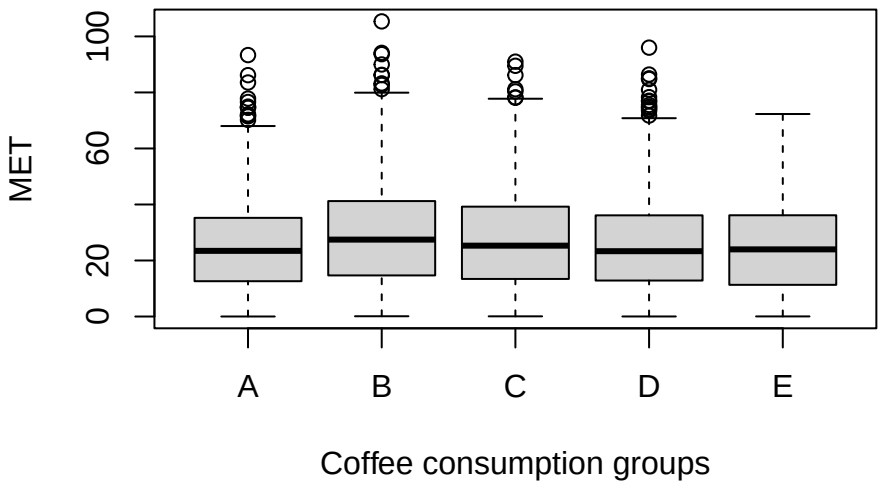
a) Create a plot that shows the association between MET score and coffee consumption. Describe what you see.

When we plot Coffee consumption groups against MET, we can see that group B has maximum association with MET, and has highest correlation. Group E has the least association wit MET and lowest correlation. The median and range for all the groups is almost similar. Group B, C have a similar median at almost 28 and groups D,E have it at almost 26. The distribution shows that all the groups are positively skewed. Group E doesn't have any outliers. Groups A,B,C,D have outliers after an MET value of 60. The MET values are between 10 to 40 for all the groups.

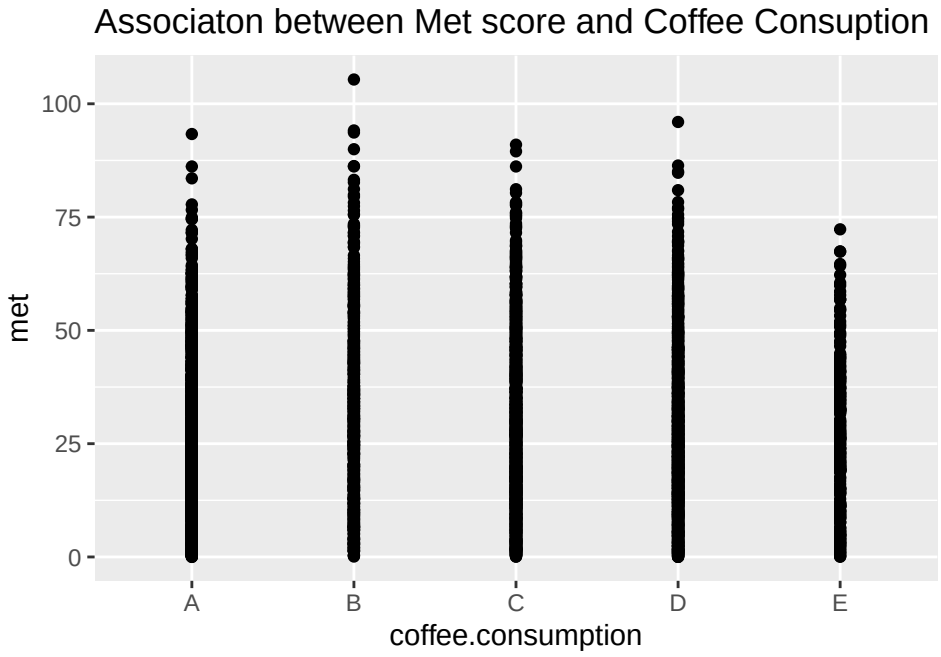
```
setwd('/users/sreyatummala/downloads/')
load('coffee_exercise.Rdata')
str(coffee.exercise)
```

```
## 'data.frame':    5072 obs. of  2 variables:
##  $ met               : num  51.2 25.4 22.4 14.6 37.5 ...
##  $ coffee.consumption: Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

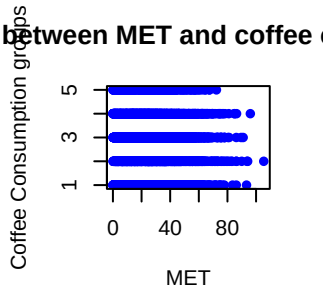
```r
#Boxplot
boxplot(coffee.exercise$met~coffee.exercise$coffee.consumption, xlab = 'Coffee consumption groups', ylab = 'M
```




```r
par(mfrow = c(2, 3))
#creating the ggplot
library(ggplot2)
attach(coffee.exercise)
ggplot(data=coffee.exercise,mapping=aes(coffee.consumption,met))+geom_point() +ggtitle("Associaton between Me
```




```r
#creating scatter plot
plot(met,coffee.consumption, col = 'blue', pch = 16, main = 'Relation between MET and coffee consumption', xl
```



b) Conduct an analysis to determine whether the average physical activity level varies among the different levels of coffee consumption.

i. Assess whether the assumptions for the analysis method are reasonably satisfied.

```
summary(aov(coffee.exercise$met ~ coffee.exercise$coffee.consumption))
```

```
##                                      Df  Sum Sq Mean Sq F value Pr(>F)
## coffee.exercise$coffee.consumption    4   10309  2577.3    8.74  5e-07 ***
## Residuals                          5067 1494241   294.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#analyzing coffee consumption groups with met
coffeeA<-coffee.exercise[coffee.exercise$coffee.consumption=="A",]
summary(coffeeA$met)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01   12.61   23.44   25.41   35.23   93.33
```

```
coffeeB<-coffee.exercise[coffee.exercise$coffee.consumption=="B",]
summary(coffeeB$met)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.10   14.67   27.46   29.72   41.21  105.36
```

```
coffeeC<-coffee.exercise[coffee.exercise$coffee.consumption=="C",]
summary(coffeeC$met)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.07   13.41   25.30   27.28   39.23   90.98
```

```
coffeeD<-coffee.exercise[coffee.exercise$coffee.consumption=="D",]
summary(coffeeD$met)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.02   12.85   23.33   25.72   36.12   95.99
```

```
coffeeE<-coffee.exercise[coffee.exercise$coffee.consumption=="E",]
summary(coffeeE$met)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.04   11.32   23.97   25.26   36.15   72.30
```

```
# pairwise comparison using two-sample t test
pairwise.t.test(coffee.exercise$met, coffee.exercise$coffee.consumption, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  coffee.exercise$met and coffee.exercise$coffee.consumption
##
##   A       B       C      D
## B 2.1e-07 -       -      -
## C 0.0036  0.0019  -      -
## D 0.6575  1.4e-06 0.0148 -
## E 0.9007  0.0006  0.0886 0.7055
##
## P value adjustment method: none
```

```
# Bonferroni correction
pairwise.t.test(coffee.exercise$met, coffee.exercise$coffee.consumption, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  coffee.exercise$met and coffee.exercise$coffee.consumption
##
##   A       B       C     D
## B 2.1e-06 -       -     -
## C 0.036   0.019   -     -
## D 1.000   1.4e-05 0.148 -
## E 1.000   0.006   0.886 1.000
##
## P value adjustment method: bonferroni
```

ii. Summarize the conclusions and comment on the generalizability of the study results.

Answer:

Null hypothesis- There is variation among average physical activity level among different coffee consumption groups Alternative hypothesis- there is no variation in the average physical activity among different levels of coffee consumption Here the p value is less than 0.05. So we reject the null hypothesis.

We have also done the summaries of individual groups of coffee consumption groups with met level and then compared the means. The means for groups A,B, C, D, E are 25.41, 29.72, 27.28, 25.72, 25.26 respectively. The differences in the values suggest that coffee consumption among various groups and average physical activity level are different.

Problem 4. (100 points)

Problem Set 1 introduced data from a study assessing whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children with access to the supply. The

file `water.Rdata` contains data from a study examining 7,257 children in 21 cities from the Flanders region in Belgium.

The fluoride content of the public water supply in each city, measured in parts per million (ppm), is saved as the variable `fluoride`; the number of dental caries per 100 children examined is saved as the variable caries. The number of dental caries is calculated by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth at the time of the study.

a) Create a plot that shows the relationship between fluoride content and caries experience. Add the least squares regression line to the scatterplot.

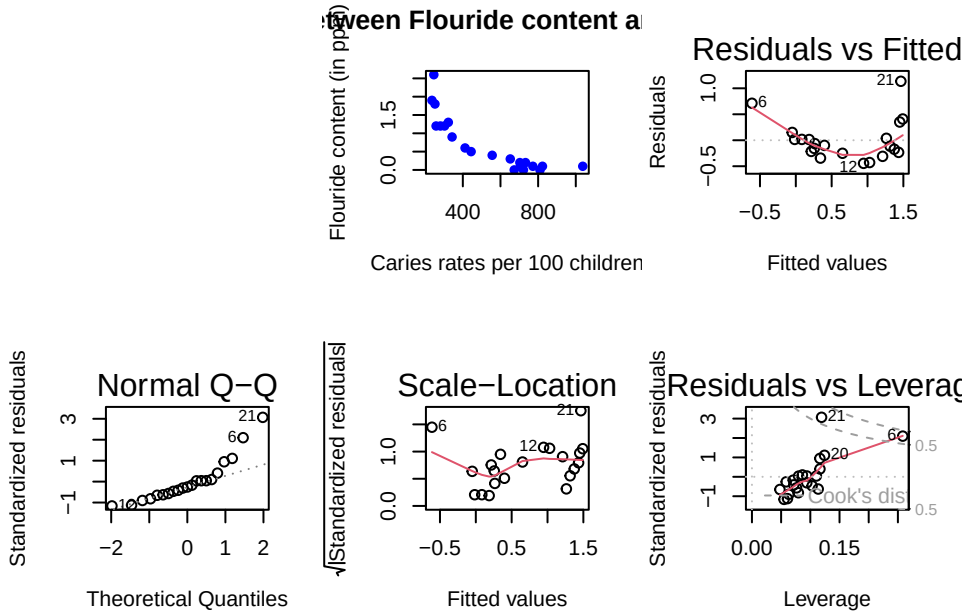
```
par(mfrow = c(2, 3))
load('water.Rdata')
str(water)
```

```
## 'data.frame':    21 obs. of  2 variables:
##  $ fluoride: num  0 0 0 0.1 0.1 ...
##  $ caries  : num  810 673 722 706 823 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "31 Dec 2013 12:03"
##  - attr(*, "formats")= chr [1:2] "%9.2f" "%9.2f"
##  - attr(*, "types")= int [1:2] 255 255
##  - attr(*, "val.labels")= chr [1:2] "" ""
##  - attr(*, "var.labels")= chr [1:2] "fluoride content (ppm)" "caries per 100 children"
##  - attr(*, "version")= int 12
```

```
#scatter plot for flouride content and caries experience
plot.new()
plot(water$fluoride ~ water$caries, col = 'blue', pch = 16, main = 'Relation between Flouride content and Ca

# calculating the linear regression
water_fit <- lm(fluoride ~ caries, data = water)
water_fit
```

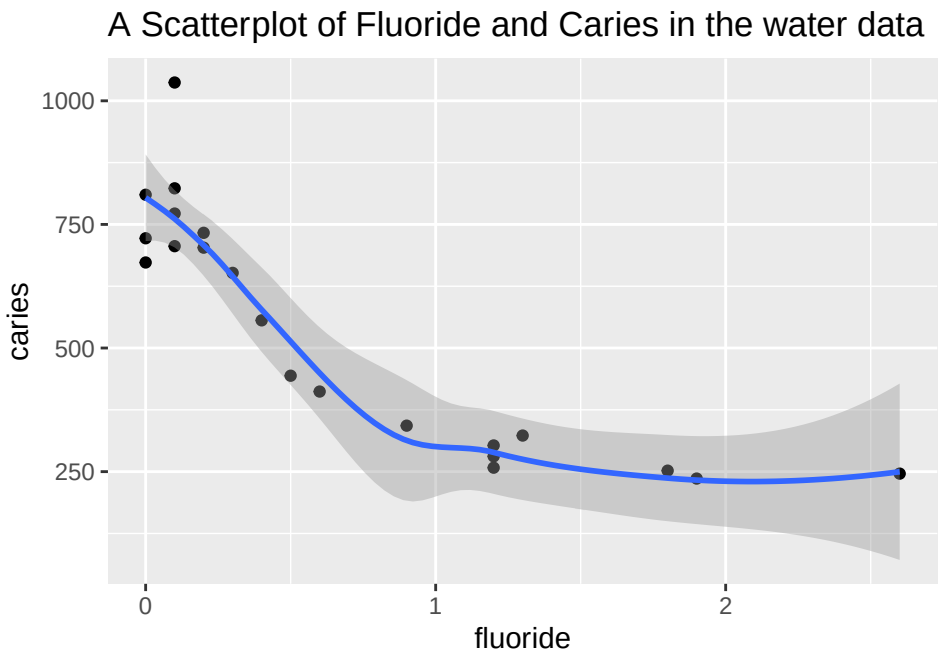
```
##
## Call:
## lm(formula = fluoride ~ caries, data = water)
##
## Coefficients:
## (Intercept)       caries
##    2.110977    -0.002626
```

```
#adding the least squares regression line to scatter plot
plot(water_fit)
```




```
#ggplot
ggplot(data=water,mapping=aes(fluoride,caries))+geom_point()+geom_smooth()+ggtitle("A Scatterplot of Fluoride
```

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



b) Based on the plot from part a), comment on whether the model assumptions of linearity and constant variability seem reasonable for these data.

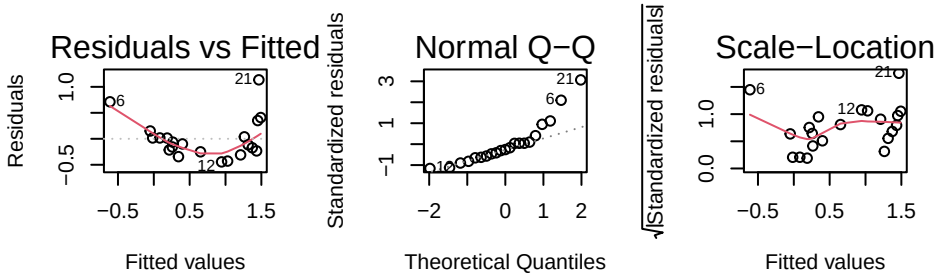
Answer:

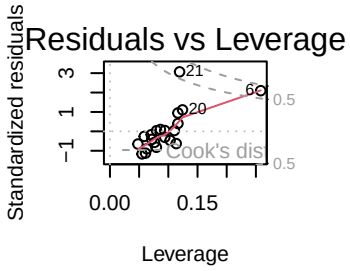
There is non linear relationship between caries and fluoride levels based on the observations from the scatter plot. As the regression line is not linear/ straight, assumptions of linearity doesn't seem reasonable. Change in caries doesn't change the proportional change to fluoride. So there assumptions of constant

variable are also not reasonable.

c) Use a residual plot to assess the model assumptions of linearity and constant variability. Comment on whether the residual plot reveals any information that was not evident from the plot from part b).

```
par(mfrow = c(2, 3))
#residual plot can be shown with the following function
plot(water_fit)


# additionally, we have
#QQ plot
#Scale-Location
#Residuals vs. Leverage
```





Suppose the file `water_new.Rdata` contains data from a more recent study conducted across 175 cities in Belgium (the data are simulated). Repeat the analyses from parts a) - c) with the new data.

d) Create a plot that shows the relationship between fluoride content and caries experience in the new data. Add the least squares regression line to the scatterplot.

```
par(mfrow = c(2, 3))
load('water_new.Rdata')
str(water.new)
```

```
## 'data.frame':    154 obs. of  2 variables:
##  $ fluoride: num  1.92 1.78 0.76 1.55 0.13 0.11 0.02 0.26 1.52 1.92 ...
##  $ caries  : num  79.2 55 322.2 176.3 746.2 ...
```

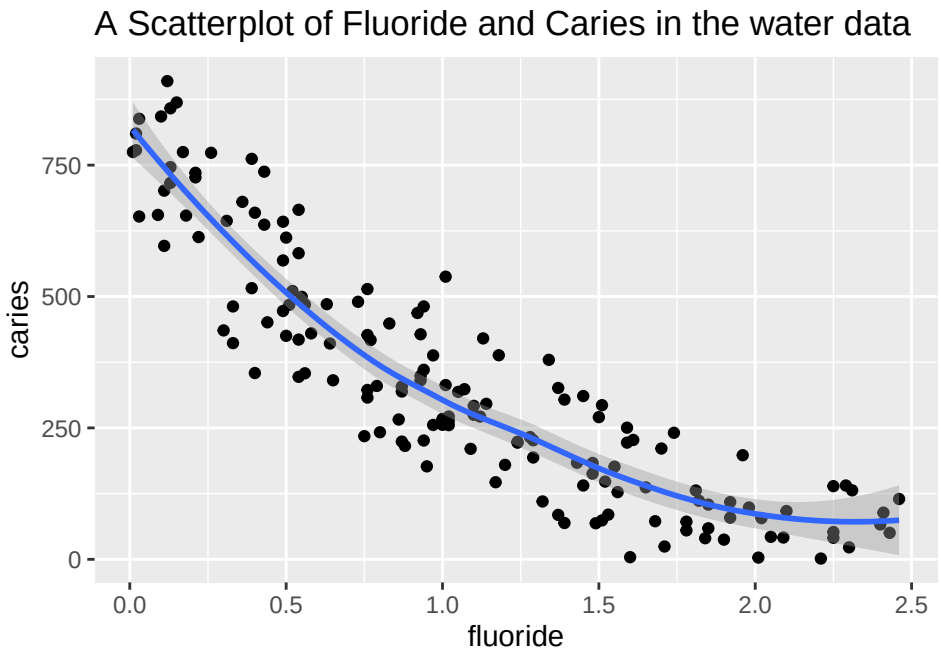
```
plot(water.new$fluoride ~ water.new$caries, col = 'blue', pch = 16, main = 'Relation between Flouride content

# calculating the linear regression
water_fit_new <- lm(fluoride ~ caries, data = water.new)
water_fit_new
```

```
##
```

```
## Call:
## lm(formula = fluoride ~ caries, data = water.new)
##
## Coefficients:
## (Intercept)       caries
##    1.921033    -0.002493
```

```
#adding the least squares regression line to scatter plot
plot(water_fit_new)

#ggplot
ggplot(data=water.new,mapping=aes(fluoride,caries))+geom_point()+geom_smooth()+ggtitle("A Scatterplot of Fluo
```

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



e) Based on the plot from part d), comment on whether the model assumptions of linearity and constant variability seem reasonable for these data.

Answer: There is non linear relationship between caries and fluoride levels based on the observations from the scatter plot. As the regression line is not linear/ straight, assumptions of linearity doesn't seem reasonable. Change in caries doesn't change the proportional change to fluoride. So there assumptions of constant variable are reasonable.

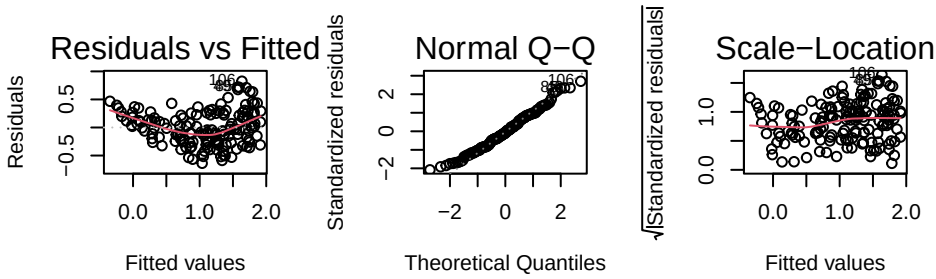
f) Use a residual plot to assess the model assumptions of linearity and constant variability. Comment on whether the residual plot reveals any information that was not evident from the plot from part e).

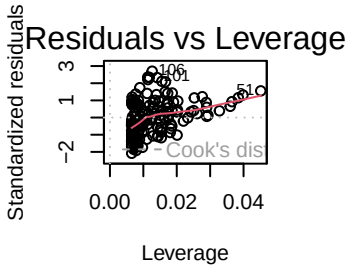
Answer:

The variables are distributed equally for the residuals and fitted values. The spread is rougly equal. This justifies constant variability. This plot shows relation of fitted values to the residuals which is not seen in the above problem e.

```
par(mfrow = c(2, 3))
#residual plot can be shown using the below function
plot(water_fit_new)
```

```
# additionally, we have
#QQ plot
#Scale-Location
#Residuals vs. Leverage
```





Unit 7

Problem 5. (200 points)

In Units 6 and 7, you have become familiar with the Prevention of REnal and Vascular END-stage Disease (PREVEND) study, which took place between 2003 and 2006 in the Netherlands. Clinical and demographic information for 500 individuals are stored as `prevend.samp` in the `oibiostat` package.

The PREVEND data were mainly used throughout the Unit 7 lectures to demonstrate one application of multiple regression: estimating the association between a response variable and primary predictor of interest while adjusting for confounders. Unit 7, Lab 3 discusses a model for the association of RFFT score with statin use that adjusts for age, educational level, and presence of cardiovascular disease. This question uses the PREVEND data in the context of explanatory model building.

Suppose that you have accepted a request to do some consulting work for a friend. Your task is to develop a prediction model for RFFT score based on the following possible predictor variables and the data in `prevend.samp`.

Variable	Description
`Age`	age in years
`Gender`	gender, coded `0` for males and `1` for females
`Education`	highest level of education
`DM`	diabetes status, coded `0` for absent and `1` for present
`Statin`	statin use, coded `0` for non-users and `1` for users
`Smoking`	smoking, coded `0` for non-smokers and `1` for smokers
`BMI`	body mass index, in kg/m^2
`FRS`	Framingham risk score, measure of risk for cardiovascular event with 10 years

The variable `Education` is coded `0` for primary school, `1` for lower secondary education, `2` for higher sec-

ondary school, and 3 for university. A higher FRS indicates higher risk of a cardiovascular event.

Your friend has requested that your final model have no more than two predictor variables. Additionally, your friend would like you to predict the mean RFFT score for a female individual of age 55 with a university education, no diabetes, no statin use, who is not a smoker, has BMI of 24, and FRS of 5. Use only the information necessary to make a prediction from your model.

In your solution, briefly explain the work done at each step of developing the final model and evaluate the final model's strengths and weaknesses.

Please consider the following sections for your solution:

```
# loading the data
library(oibiostat)
data("prevend.samp")
str(prevend.samp)
```

```
## 'data.frame':    500 obs. of  31 variables:
##  $ Casenr       : int  2266 3235 1068 3422 3570 1932 3134 3573 1103 868 ...
##  $ Age          : int  55 65 46 68 70 53 64 70 46 44 ...
##  $ Gender       : int  1 1 0 1 0 0 0 0 1 0 ...
##  $ Ethnicity    : int  3 0 2 0 0 0 2 0 0 0 ...
##  $ Education    : int  2 1 3 2 2 0 1 3 2 3 ...
##  $ RFFT         : int  62 79 89 70 35 14 31 47 88 91 ...
##  $ VAT          : int  -1 11 6 5 10 7 8 5 11 11 ...
##  $ CVD          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DM           : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Smoking      : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Hypertension : int  0 1 0 0 0 0 0 1 1 0 ...
##  $ BMI          : num  39.7 29 23.2 22.3 32.4 ...
##  $ SBP          : num  122 108 120 114 114 ...
##  $ DBP          : num  63.5 66.5 75 67 72.5 75 77 76.5 99 70 ...
##  $ MAP          : num  86 82.5 92.5 85 89.5 ...
##  $ eGFR         : num  83.3 76.5 76.4 61.2 88.1 ...
##  $ Albuminuria.1: int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Albuminuria.2: int  0 0 2 1 0 0 1 0 1 0 ...
##  $ Chol         : num  3.86 5.64 6.83 7.11 5.04 3.05 4.9 5.5 3.92 5.75 ...
##  $ HDL          : num  1.54 1.53 1.04 1.85 1.4 0.79 1.23 1.57 1.39 1.18 ...
##  $ Statin       : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Solubility   : int  2 1 2 2 2 2 2 2 2 2 ...
##  $ Days         : int  -1 1672 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ Years        : num  -1 4.58 -1 -1 -1 ...
##  $ DDD          : num  0 1373 0 0 0 ...
##  $ FRS          : int  8 11 -1 9 12 8 12 15 11 8 ...
##  $ PS           : num  0.3743 0.2559 0.1285 0.0942 0.1934 ...
##  $ PSquint      : int  5 4 3 2 4 3 3 4 3 2 ...
##  $ GRS          : int  1 1 0 0 1 1 0 1 0 0 ...
##  $ Match_1      : int  816 727 -1 838 -1 15 200 -1 -1 -1 ...
##  $ Match_2      : int  113 242 -1 -1 276 121 -1 -1 -1 -1 ...
```

Data Exploration

Initial data exploration reveled that RFFT is the residual with 8 predictors. To determine the significance of correlation of the predictors to the RFFT, we have done correlation and identified, Age, Education, BMI and FRS as potential predictors.We have visualized the data to find out the skewness od distribution and determine ouliers. BMI has been skewed.

```
#numerical summeries
summary(prevend.samp$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36.00   46.00   54.00   54.82   64.00   81.00
```

```
summary(prevend.samp$Gender)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.47    1.00    1.00
```

```
summary(prevend.samp$Education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.798   3.000   3.000
```

```
summary(prevend.samp$RFFT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    11.0    46.0    67.0    68.4    88.0   136.0
```

```
summary(prevend.samp$DM)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.066   0.000   1.000
```

```
summary(prevend.samp$Smoking)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.00    0.00    0.00    0.22    0.00    1.00
```

```
summary(prevend.samp$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.11   23.87   26.11   26.90   29.00   60.95
```

```
summary(prevend.samp$Statin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.23    0.00    1.00
```

```
summary(prevend.samp$FRS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.000   5.000  10.000   9.946  15.000  29.000
```

```
cor(prevend.samp$RFFT,prevend.samp$Age)
```

```
## [1] -0.5338617
```

```
cor(prevend.samp$RFFT,prevend.samp$Gender)
```

```
## [1] 0.02877565
```

```
cor(prevend.samp$RFFT,prevend.samp$Education)
```

```
## [1] 0.553066
```

```
cor(prevend.samp$RFFT,prevend.samp$DM)
```

```
## [1] -0.1534254
```

```
cor(prevend.samp$RFFT,prevend.samp$Smoking)
```

```
## [1] -0.1061164
```

```
cor(prevend.samp$RFFT,prevend.samp$BMI)
```

```
## [1] -0.1869802
```

```
cor(prevend.samp$RFFT,prevend.samp$Statin)
```

```
## [1] -0.1545881
```

```
cor(prevend.samp$RFFT,prevend.samp$FRS)
```

```
## [1] -0.437095
```

```
![](Assignment-3_files/figure-latex/unnamed-chunk-12-1.pdf)<!-- --> ![](Assignment-3_files/figure-latex/unnam
```

Initial Model Fitting In the initial model we have included all the 8 predictors. But later fitted with only the potential predictors.

```
# based on the initial model, we have decided to take Age, Education, BMI, FRS
# here we transform BMI variable to log BMI as it shows positive skewness.
prevend.samp$BMI <- log(prevend.samp$BMI)

#including all the variables to the model
initial_model <- lm(RFFT ~ Age+Education+Gender+DM+Smoking+Statin+BMI+FRS, data = prevend.samp)
initial_model
```

```
##
## Call:
## lm(formula = RFFT ~ Age + Education + Gender + DM + Smoking +
##     Statin + BMI + FRS, data = prevend.samp)
##
## Coefficients:
## (Intercept)          Age    Education       Gender           DM      Smoking
##    150.3234      -1.1043      10.6405       0.9722      -3.0905      -7.3711
##      Statin          BMI          FRS
##      4.1616     -13.4830       0.4133
```

```
# fitting the model
model_fit <- lm(RFFT ~ Age+Education+BMI+FRS, data = prevend.samp)
model_fit
```

```
##
## Call:
## lm(formula = RFFT ~ Age + Education + BMI + FRS, data = prevend.samp)
##
## Coefficients:
## (Intercept)          Age    Education          BMI          FRS
##   126.52854     -0.91312     10.99288     -8.71889      0.07603
```

```
summary(model_fit)
```

```
##
## Call:
## lm(formula = RFFT ~ Age + Education + BMI + FRS, data = prevend.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.014 -15.133  -1.169  13.901  60.827
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126.52854   21.89144   5.780 1.32e-08 ***
## Age          -0.91312    0.12256  -7.450 4.18e-13 ***
## Education    10.99288    1.03086  10.664  < 2e-16 ***
## BMI          -8.71889    6.26574  -1.392    0.165
## FRS           0.07603    0.21762   0.349    0.727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.76 on 495 degrees of freedom
## Multiple R-squared:  0.4304, Adjusted R-squared:  0.4258
## F-statistic: 93.53 on 4 and 495 DF,  p-value: < 2.2e-16
```

Model Comparison

Model which contains Age, Education, BMI performs better than other models with R squared value 42.7%

```
# excluding variable age from model
model_no_age <- lm(RFFT ~ Education+BMI+FRS, data = prevend.samp)

summary(model_no_age)$adj.r.squared
```

```
## [1] 0.3627493
```

```
#excluding education
model_no_education <- lm(RFFT ~ Age+BMI+FRS, data = prevend.samp)

summary(model_no_education)$adj.r.squared
```

```
## [1] 0.2953658
```

```
# excluding BMI
model_no_BMI <- lm(RFFT ~ Age+Education+FRS, data = prevend.samp)
summary(model_no_BMI)$adj.r.squared
```

```
## [1] 0.4247593
```

```
# excluding FRS
model_no_FRS <- lm(RFFT ~ Age+Education+BMI, data = prevend.samp)
summary(model_no_FRS)$adj.r.squared
```

```
## [1] 0.4268594
```

```
# exclude BMI and FRS
model_Age_Edu <- lm(RFFT ~ Age+Education, data = prevend.samp)
summary(model_Age_Edu)$adj.r.squared
```

```
## [1] 0.4259148
```

Here we have visualized the viables age and education to detrmine the distribution of values in them.

```r
#load color package
library(RColorBrewer)

boxplot(RFFT ~ Age, data = prevend.samp,
    main = "Age correspondance to RFFT",
    col = brewer.pal(5, "Blues"))
```

```
![](Assignment-3_files/figure-latex/unnamed-chunk-15-1.pdf)<!-- -->

```r
boxplot(RFFT ~ Education, data = prevend.samp,
    main = "Education correspondance to RFFT",
    col = brewer.pal(5, "Reds"))
```

```
![](Assignment-3_files/figure-latex/unnamed-chunk-15-2.pdf)<!-- -->



```r
#create the age.binary variable
prevend.samp$age.binary = prevend.samp$Age

#redefine the factor levels of grazing.binary
levels(prevend.samp$age.binary) = list(Below_sixty = c(36, 39, 42, 45, 48, 51,54,57),Sixty_and_above = c(60,63,6

# fitting the model with age.binary variable
model_age.binary <- lm(RFFT ~ age.binary+Education+BMI, data = prevend.samp)
summary(model_age.binary)$adj.r.squared
```

```
## [1] 0.4268594
```

```r
#create education.binary variable
prevend.samp$education.binary = prevend.samp$Education
levels(prevend.samp$education.binary) = list(lesser_education = c(0,1), higher_education = c(2,3))

#fitting the model with education.binary
model_edu.binary <- lm(RFFT ~ age.binary+education.binary+BMI, data = prevend.samp)
summary(model_edu.binary)$adj.r.squared
```

```
## [1] 0.4268594
```

There is no improvement in the model performance by creating binaries. We therefore use the interaction terms. The model performance has improved with 43% prediction of RFFT.

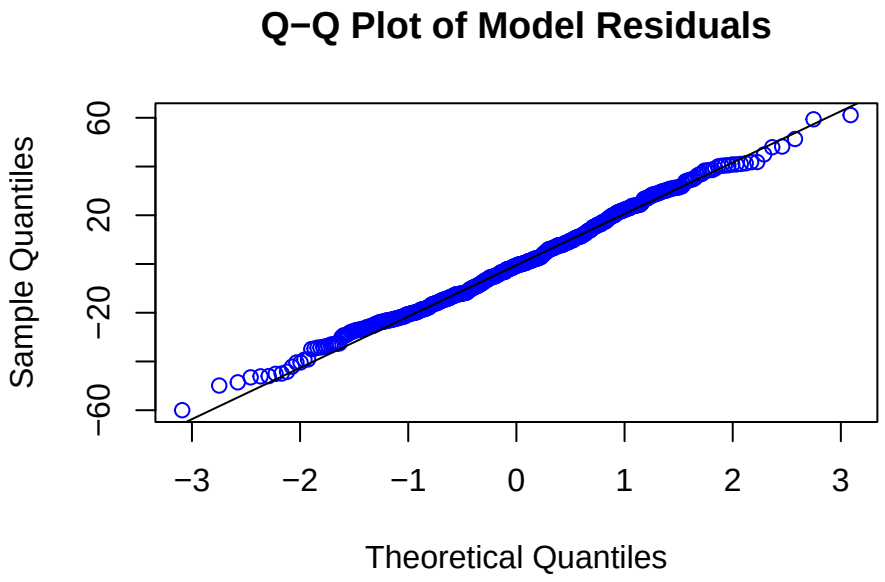
```r
# fit the model with interaction terms
model_interaction <- lm(RFFT ~ Age*Education+BMI, data = prevend.samp)
summary(model_interaction)$adj.r.squared
```

```
## [1] 0.4349873
```

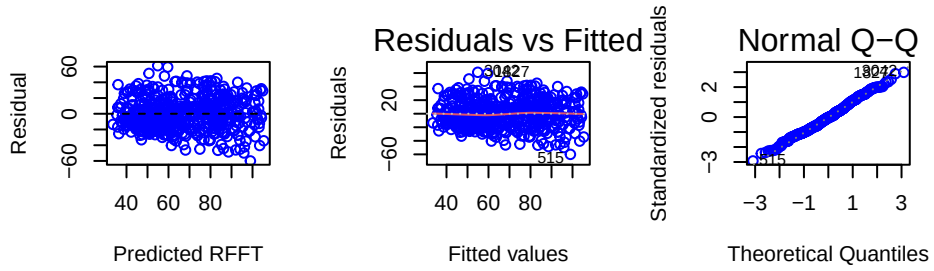
Model Assessment

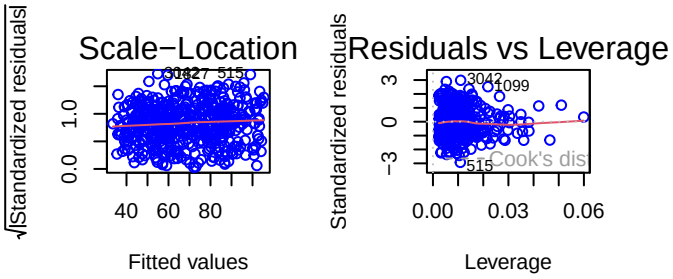
The final model shows normal distribution as there is no much spread except at the tails.

```r
#create a residual plot for the model
final_model = model_interaction

#create a Q-Q plot for residual
qqnorm(resid(final_model),
       pch = 21, col = 'blue',
       main = "Q-Q Plot of Model Residuals")

# add a straight diagonal line to the plaot
qqline(resid(final_model))
```




```r
#distribution of residuals against variables in the model
par(mfrow = c(2, 3))
#plot residuals vs fitted
plot(resid(final_model) ~ fitted(final_model),
     pch = 21, col = 'blue',
     xlab = "Predicted RFFT", ylab = "Residual")
abline(h = 0, lty = 2)

#plotting the fonal model
plot(final_model,pch = 21,  col = 'blue')
```





Conclusions

Initial data exploration reveled that RFFT is the residual with 8 predictors. To determine the significance of correlation of the predictors to the RFFT, we have done correlation and identified, Age, Education, BMI and FRS as potential predictors.We have visualized the data to find out the skewness od distribution and determine ouliers. BMI has been skewed. In the initial model we have included all the 8 predictors. But later fitted with only the potential predictors. Model which contains Age, Education, BMI performs better than other models with R squared value 42.7% Here we have visualized the viables age and education to detrmine the distribution of values in them. The final model shows normal distribution as there is no much spread except at the tails.

The predicted outcome of RFFT is 99.489. The model predicts only 43% of RFFT .

```
#predicting based on given criteria
predict(initial_model, newdata = data.frame(Age=55,Gender=1,Education=3,DM=0,Smoking=0,Statin=0,BMI=24,FRS=5)
```

```
##         1
## -199.0467
```

```
#final model summary
summary(final_model)
```

```
##
## Call:
## lm(formula = RFFT ~ Age * Education + BMI, data = prevend.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.993 -14.696  -0.651  13.717  61.111
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   94.51699   23.06634   4.098 4.88e-05 ***
## Age           -0.47020    0.16851  -2.790  0.00547 **
## Education     24.48995    4.85413   5.045 6.37e-07 ***
```

```
## BMI           -6.66941    6.08461  -1.096  0.27356
## Age:Education -0.23896    0.08378  -2.852  0.00452 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.59 on 495 degrees of freedom
## Multiple R-squared:  0.4395, Adjusted R-squared:  0.435
## F-statistic: 97.04 on 4 and 495 DF,  p-value: < 2.2e-16
```

R codes and visualization

Unit 8

Problem 6. (200 points)

Biological ornamentation refers to features that are primarily decorative, such as the elaborate tail feathers of a peacock. The evolution of ornamentation in males has been extensively researched; there are many studies exploring how male ornamentation functions as a signal of phenotypic and/or genetic quality to potential mates. In contrast, there are few studies investigating female ornamentation.[2]

Some biologists have hypothesized that there is strong natural selection against overly conspicuous female ornaments. Bright or colorful plumage in females might be expected to increase the incidence of predation on nests for species in which females incubate eggs. Female ornamentation might also undergo positive selection, functioning in sexual signaling like male ornamentation, and indicating desirable qualities such as high immune function.

The data in the file `rubythroats.Rdata` are from a study of 83 female rubythroats, a bird species in which both males and females exhibit a brightly colored red patch on the throat and breast (referred to as a "bib"). In rubythroats, females incubate the eggs, while males provide food to females to facilitate uninterrupted incubation.

- `survival`: records whether the bird survived to return to the nesting site the subsequent year, `yes` if the female was observed and `no` if the female was not observed
- `weight`: weight of the bird, measured in grams
- `wing.length`: wing length of the bird, measured in millimeters
- `tarsus.length`: tarsus (i.e., leg) length of the bird, measured in millimeters
- `first.clutch.size`: number of eggs in the first clutch laid during the first year that the bird was observed
- `nestling.fate`: whether the nestlings from the first clutch survived to fledging (`Fledged`) or were lost to predation (`Predated`)
- `second.clutch`: whether the bird laid a second clutch during the first year that the bird was observed, recorded as `Yes` for laying a second clutch and `No` for otherwise
- `carotenoid.chroma`: a measure of the abundance of red carotenoid pigment in feathers, as measured from a sample of four feathers taken from the center of the bird's bib. Larger numbers indicate higher levels of pigment in the feathers and a more saturated red color.
- `bib.area`: the total area of the bird's bib, measured in millimeters squared
- `total.brightness`: a measure of bib brightness, calculated from spectrometer analyses. Larger numbers indicate a brighter red color.

You will be conducting an analysis of the results in order to investigate how bib attributes and other phenotypic characteristics of female birds are associated with measures of fitness.

[2] Freeman-Gallant, et al., J Evol. Biol. (2014) 27: 982-991 doi:10.1111/jeb.12369.

```
#load data
setwd('/users/sreyatummala/downloads/')
load('rubythroats.Rdata')
str(rubythroats)
```

```
## 'data.frame':    85 obs. of  10 variables:
##  $ survival         : Factor w/ 2 levels "no","yes": 2 1 2 1 2 1 1 1 2 2 ...
##  $ weight           : num  10.1 9.6 11.6 12.8 10.7 11.1 11.5 10.9 11.4 11.7 ...
##  $ wing.length      : num  53 53 52 50.5 48.5 49 51 52 54 52 ...
##  $ tarsus.length    : num  20.2 19.8 19.3 19.6 18.6 18.7 18.9 19.3 19 19.1 ...
##  $ first.clutch.size: int  5 NA 4 5 5 4 3 3 4 4 ...
##  $ nestling.fate    : Factor w/ 2 levels "Predated","Fledged": 2 NA 2 2 2 1 2 2 2 1 ...
##  $ second.clutch    : Factor w/ 2 levels "No","Yes": 2 NA 2 2 2 1 2 1 2 1 ...
##  $ carotenoid.chroma: num  0.992 0.901 0.921 1.043 0.976 ...
##  $ bib.area         : num  529 389 413 508 363 ...
##  $ total.brightness : num  17.3 22.7 19.9 11 13.2 ...
```

a) Fit a model to predict nestling fate from female bib characteristics (carotenoid chroma, bib area, total brightness) and female body characteristics (weight, wing length, tarsus length). Identify the slope coefficients significant at $\alpha = 0.10$, and provide an interpretation of these coefficients in the context of the data.

Answer:

The slope coefficients of predictors Total Brightness (p value = 0.00273) and Wing length (p value = 0.02627) are significant at alpha 0.10. When there is increase in total brightness, the estimated log odds of nestling fate is reduced by 0.130880 when other variables are constant. Similarly when the wing length is increased, the estimated log odds of nestling fate is increased by 0.521821. This provides an interpretation that total brightness and wing length are the most important and necessary predictors for the residual value.

```
rubythroats$nestling.fate <- as.factor(rubythroats$nestling.fate)
log_reg_fit <- glm(nestling.fate ~ carotenoid.chroma+bib.area+total.brightness+weight+wing.length+tarsus.leng
log_reg_fit
```

```
##
## Call:  glm(formula = nestling.fate ~ carotenoid.chroma + bib.area +
##     total.brightness + weight + wing.length + tarsus.length,
##     family = binomial(link = "logit"), data = rubythroats)
##
## Coefficients:
##       (Intercept)  carotenoid.chroma           bib.area   total.brightness
##         -24.057818           -4.774799           -0.001272           -0.130880
##            weight        wing.length      tarsus.length
##          -0.358164            0.521821            0.476708
##
## Degrees of Freedom: 70 Total (i.e. Null);  64 Residual
##   (14 observations deleted due to missingness)
## Null Deviance:       97.74
## Residual Deviance: 79.65     AIC: 93.65
```

```
summary(log_reg_fit)
```

```
##
## Call:
## glm(formula = nestling.fate ~ carotenoid.chroma + bib.area +
##     total.brightness + weight + wing.length + tarsus.length,
```

```
##     family = binomial(link = "logit"), data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8550  -0.8988  -0.4402   1.0220   1.7890
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -24.057818  13.046850  -1.844  0.06519 .
## carotenoid.chroma  -4.774799   3.302978  -1.446  0.14829
## bib.area           -0.001272   0.002833  -0.449  0.65341
## total.brightness   -0.130880   0.043669  -2.997  0.00273 **
## weight             -0.358164   0.419582  -0.854  0.39332
## wing.length         0.521821   0.234826   2.222  0.02627 *
## tarsus.length       0.476708   0.484340   0.984  0.32500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.736  on 70  degrees of freedom
## Residual deviance: 79.655  on 64  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 93.655
##
## Number of Fisher Scoring iterations: 4
```

b) Investigate the factors associated with whether a female lays a second clutch during the first year that she was observed.

i. Is there evidence of a significant association between nestling fate and whether a female lays a second clutch? If so, report the direction of association.

Answer: As the p value is 5.499e-05 and less than 0.05, we reject the null hypothesis. There is association. The The slope of the interaction term indicates the association between nestling fate and second clutch to be positive as the value is 3.4045 and p value is less than 0.05. The proportion of birds predated when they did not lay second clutch is 0.6825397 and the proportion of birds fledged when they do not lay second clutch is 0.3174603.

```
table_ns <- table(rubythroats$nestling.fate,rubythroats$second.clutch)
table_ns
```

```
##
##            No Yes
##   Predated 43   1
##   Fledged  20  14
```

```
# null hypothesis- no association between nestling fate and second clutch
# alternative hypothesis - there is association between nestling fate and second clutch
chisq.test(rubythroats$nestling.fate,rubythroats$second.clutch)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  rubythroats$nestling.fate and rubythroats$second.clutch
## X-squared = 16.268, df = 1, p-value = 5.499e-05
```

```r
#INTERACTION TESTING
nestling_sencondeggs <- glm(nestling.fate~second.clutch,family = binomial(link='logit'),data = rubythroats)

nestling_sencondeggs
```

```
##
## Call:  glm(formula = nestling.fate ~ second.clutch, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Coefficients:
##      (Intercept)  second.clutchYes
##          -0.7655            3.4045
##
## Degrees of Freedom: 77 Total (i.e. Null);  76 Residual
##   (7 observations deleted due to missingness)
## Null Deviance:       106.8
## Residual Deviance: 86.09     AIC: 90.09
```

```r
summary(nestling_sencondeggs)
```

```
##
## Call:
## glm(formula = nestling.fate ~ second.clutch, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.327  -0.874  -0.874   1.229   1.515
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.7655     0.2707  -2.828  0.00468 **
## second.clutchYes   3.4045     1.0699   3.182  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106.85  on 77  degrees of freedom
## Residual deviance:  86.09  on 76  degrees of freedom
##   (7 observations deleted due to missingness)
## AIC: 90.09
##
## Number of Fisher Scoring iterations: 5
```

ii. Fit a model to predict whether a female lays a second clutch from nestling fate and bib characteristics. Identify the two predictors that are most statistically significantly associated with the response variable.

Answer: The two predictors most statistically significantly associated with laying a second clutch are total brightness ($p = 0.030$) and nestling fate ($p = 0.0015$).

```r
#fitting new model
# predictors- nestlingfate, carotid chroma, bib area, total brightness
# response variable- second clutch
# second cluth as factor
```

```
rubythroats$second.clutch <- as.factor(rubythroats$second.clutch)
new_model_ruby <- glm(second.clutch ~ nestling.fate+carotenoid.chroma+bib.area+total.brightness, family = bi
summary(new_model_ruby)
```

```
##
## Call:
## glm(formula = second.clutch ~ nestling.fate + carotenoid.chroma +
##     bib.area + total.brightness, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.87779  -0.25235  -0.09075  -0.02136   2.10799
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -21.397205   7.593148  -2.818  0.00483 **
## nestling.fateFledged   5.527158   1.740419   3.176  0.00149 **
## carotenoid.chroma     11.585799   6.085778   1.904  0.05694 .
## bib.area               0.007019   0.003801   1.847  0.06480 .
## total.brightness       0.149085   0.068533   2.175  0.02960 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.503  on 75  degrees of freedom
## Residual deviance: 39.076  on 71  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 49.076
##
## Number of Fisher Scoring iterations: 7
```

iii. Fit a new model to predict whether a female lays a second clutch using the two predictors identified in part ii. and their interaction. Interpret the model coefficients in the context of the data.

The coefficients for nestling.fate is 0.00103 and total brightness is 0.16532. The p value of the model is 0.00136 signifying that both the predictor variables influence the model. Nestling fate shows positive correlation while total brightness has no correlation with second clutch. The individual p value of total brightness shows that it is not significantly associated with second clutch.

```
model_secondclutch <- glm(second.clutch ~ nestling.fate+total.brightness, family = binomial(link='logit'),dat
summary(model_secondclutch)
```

```
##
## Call:
## glm(formula = second.clutch ~ nestling.fate + total.brightness,
##     family = binomial(link = "logit"), data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2778  -0.7369  -0.1949  -0.1434   2.4354
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)          -5.45275    1.70247  -3.203  0.00136 **
## nestling.fateFledged  3.80377    1.15864   3.283  0.00103 **
## total.brightness      0.05957    0.04294   1.387  0.16532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.940  on 76  degrees of freedom
## Residual deviance: 53.558  on 74  degrees of freedom
##   (8 observations deleted due to missingness)
## AIC: 59.558
##
## Number of Fisher Scoring iterations: 6
```

c) Investigate the factors associated with whether a female survives to return to the nesting site the subsequent year.

i. Fit a model to predict survival from bib characteristics, female body characteristics, first clutch size, and whether a second clutch was laid. Identify factors that are positively associated with survival for the observed birds.

The factors that are positively associated with survival of these birds include, weight - 0.605506 wing length- 0.842892 first clutch- 2.620596 second clutch size- 1.664275

```
#assigning survival as.factor
rubythroats$survival <- as.factor(rubythroats$survival)

#predicting the model
model_female_survives <- glm(survival ~ carotenoid.chroma+bib.area+total.brightness+weight+wing.length+tarsus
summary(model_female_survives)
```

```
##
## Call:
## glm(formula = survival ~ carotenoid.chroma + bib.area + total.brightness +
##     weight + wing.length + tarsus.length + first.clutch.size +
##     second.clutch, family = binomial(link = "logit"), data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9501  -0.5706  -0.2036   0.4403   2.5038
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -24.322666  22.759433  -1.069   0.2852
## carotenoid.chroma -18.517610   6.804872  -2.721   0.0065 **
## bib.area           -0.002710   0.004754  -0.570   0.5686
## total.brightness   -0.108595   0.063858  -1.701   0.0890 .
## weight              0.605506   0.648570   0.934   0.3505
## wing.length         0.842892   0.430007   1.960   0.0500 *
## tarsus.length      -0.886169   0.837672  -1.058   0.2901
## first.clutch.size   2.620596   1.132561   2.314   0.0207 *
## second.clutchYes    1.664275   1.075925   1.547   0.1219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.438  on 48  degrees of freedom
## Residual deviance: 34.764  on 40  degrees of freedom
##   (36 observations deleted due to missingness)
## AIC: 52.764
##
## Number of Fisher Scoring iterations: 6
```

ii. Fit a new model with only the significant predictors from the previous model; let $\alpha = 0.10$. Comment on whether this model is preferable to the one fit in part i.

Here the p value is 0.05 and the one in part 1 is 0.2852. So, this model predicts better than that in model 1. Also the predictors in this model have significance levels less than alpha contributing to the prediction better.

```
# predictors that fit the given criteria - carotenoid chroma, total brightness, wing.length, first clutch siz

#model fitting
model_significant <- glm(survival ~ carotenoid.chroma+total.brightness+wing.length+first.clutch.size, family
summary(model_significant)
```

```
##
## Call:
## glm(formula = survival ~ carotenoid.chroma + total.brightness +
##     wing.length + first.clutch.size, family = binomial(link = "logit"),
##     data = rubythroats)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0203  -0.7386  -0.4338   0.8047   2.4505
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -20.15330   10.45426  -1.928   0.0539 .
## carotenoid.chroma  -9.49612    3.89478  -2.438   0.0148 *
## total.brightness   -0.07515    0.04582  -1.640   0.1010
## wing.length         0.46054    0.22795   2.020   0.0433 *
## first.clutch.size   1.56415    0.70098   2.231   0.0257 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 72.997  on 55  degrees of freedom
## Residual deviance: 54.389  on 51  degrees of freedom
##   (29 observations deleted due to missingness)
## AIC: 64.389
##
## Number of Fisher Scoring iterations: 5
```

```
#better parsimonious model of the ones fit in parts i. and ii
AIC(model_female_survives)
```

```
## [1] 52.76392
```

```
AIC(model_significant)
```

```
## [1] 64.38912
```

For parts iii. and iv., use the better parsimonious model of the ones fit in parts i. and ii.

iii. Compare the odds of survival for a female who laid 5 eggs in her first clutch to the odds of survival for a female who laid 3 eggs in her first clutch, if the females are physically identical and both laid a second clutch.

Answer: We use the better parsimonious model, i.e model_significant. The female who laid 5 eggs in the first clutch has odds of survival 3.901085. The female who laid 5 eggs in the first clutch has odds of survival 0.1708379.

Female who laid 5 eggs in the first clutch has 22.83501 times survival to female who laid 3 eggs in the first clutch.

```
#first clutch size 5

log.odds.eggs_5 = predict(model_significant, data.frame(carotenoid.chroma=0.991766,total.brightness=17.25410,

exp(log.odds.eggs_5)
```

```
##        1
## 3.901085
```

```
#first clutch size 3

log.odds.eggs_3 = predict(model_significant, data.frame(carotenoid.chroma=0.991766,total.brightness=17.25410,

exp(log.odds.eggs_3)
```

```
##         1
## 0.1708379
```

```
#comparing the odds of survival between 2 females
exp(log.odds.eggs_5)/exp(log.odds.eggs_3)
```

```
##        1
## 22.83501
```

iv. Suppose female A has bib area 350 mm^2, total brightness of 35, carotenoid chroma 0.90, tarsus length of 19.5 mm, wing length 51 mm, weighs 10.8 g, lays 4 eggs in her first clutch, and lays a second clutch. Female B has bib area 300 mm^2, total brightness of 20, carotenoid chroma 0.85, tarsus length of 19.0 mm, wing length 50 mm, weighs 10.9 g, lays 3 eggs in her first clutch, and lays a second clutch. Compare the odds of survival for females A and B.

The odds of survival for Female- A = 0.2047139 The odds of survival for Female-B = 0.1341531 The odds of survival of female A is 1.525972 times greater than female B

```
#female-A odds or survival

p_1 = predict(model_significant, newdata = data.frame(carotenoid.chroma=0.90,total.brightness=35,wing.length=
exp(p_1)
```

```
##         1
## 0.2047139
```

```
#female-B odds of survival
```

```
p_2 = predict(model_significant, newdata = data.frame(carotenoid.chroma=0.85,total.brightness=20,wing.length=
exp(p_2)
```

```
##         1
## 0.1341531
```

```
# Comparing the odds of survival between female A and B

exp(p_1)/exp(p_2)
```

```
##        1
## 1.525972
```

d) Biological fitness refers to how successful an organism is at surviving and reproducing. Based on the results of your analysis, briefly discuss whether female ornamentation seems beneficial for fitness in this bird species. Limit your response to at most ten sentences. You do not need to reference specific numerical results/models from the analysis.

The female ornamentation has shown to have effects on the nestling fate and survival of the female rubythroats.

Nestling fate: The total brightness seems to have negative effects with the nestling fate and wing length shows to have positive correlation with the nestling fate. The female happening to have laid second clutch of eggs has more nestling fate.

Survival: The survival of the female is influenced by laying of first and second clutches, wing length and weight of the female birds. The female which happens to lay more eggs in the first clutch has more chances of survival. Weight and wing length have positive relation with survival of female.