Assignment2

AmaryahHalo_300417620

31/03/2021

Question 1

a

```r
cont.table <- matrix(data =c(19, 4, 31, 42, 21, 48), nrow = 2, ncol = 3, byrow = TRUE)
cont.table
```

```
##      [,1] [,2] [,3]
## [1,]   19    4   31
## [2,]   42   21   48
```

Yes, it is appropriate to carry out a chi-square test of independence for the data presented in the table above as the collected data meets the conditions for using the test. These conditions are stated below.

1) Ecotourism researcher's generated sample consists of 599 randomly selected tourists hence the subset of 165 respondents is also randomly selected.
2) Both 'level of agreement with the statement' and 'whether or not the respondent did any photography while visiting the national park' variables are categorical.
3) The number of sample observations expected value in each level of the variable is at least 5 or greater.

b

Clearly stated below are the null and alternative hypotheses:

X = Photography action (Photography or No Photography)

Y = Level of agreement in regards to the statement (Disagree, Neutral, Agree)

H0: X and Y are independent

H1: X and Y are not independent

```r
# Expected frequencies for the contingency table
chisq.test(cont.table, correct=FALSE)$expected
```

```
##          [,1]      [,2]     [,3]
## [1,] 19.96364  8.181818 25.85455
## [2,] 41.03636 16.818182 53.14545
```

All expected frequencies are greater than 5 as seen in the table above.

Pearson's Chi-squared Test

```
chisq.test(cont.table, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont.table
## X-squared = 4.7685, df = 2, p-value = 0.09216
```

```
#The test statistic is the X-squared value for both chi-squared tests.
```

Likelihood-ratio Chi-squared Test

```
#install.packages("DescTools")
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.0.5
```

```
GTest(cont.table, correct = "none")
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  cont.table
## G = 5.1486, X-squared df = 2, p-value = 0.07621
```

As seen in results above, the Pearson's chi-squared p-value result is 0.09216 and the Likelihood-ratio chi-squared p-value result is 0.07621. Both p-value's are greater than the alpha value of 0.05, hence, we fail to reject H0 for both tests meaning that there is insufficient evidence to suggest whether a respondent did any photography or not while visiting the national park and the level of agreement with the statement "We have to protect biodiversity for humans in the future, even if it means reducing our standard of living today." are dependent.

Question 2

a

```
cont.table <- matrix(data =c(38, 400, 5, 156), nrow = 2, ncol = 2, byrow = TRUE)
cont.table
```

```
##      [,1] [,2]
## [1,]   38  400
## [2,]    5  156
```

```r
oddsratio<-(38*156)/(5*400)
oddsratio
```

```
## [1] 2.964
```

The odds ratio is 2.964, hence, a respondent is 3 times more likely to engage in photography in the national park and participate in 4WD than not engaging in photography and not 4WD.

b

```r
#confidence interval for theta hat
logoddsratio <- log(oddsratio)
logoddsratio
```

```
## [1] 1.08654
```

```r
#standard error
sum <- ((1/38) + (1/400) + (1/5) + (1/156))
total <- (sqrt(sum))*1.96

upper <- logoddsratio + total
lower <- logoddsratio - total

#log odds ratio to odds ratio
x <- exp(lower)
y <- exp(upper)

c(x, y)
```

```
## [1] 1.145610 7.668659
```

c

```r
#install.packages("exact2x2")
fisher.test(cont.table, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cont.table
## p-value = 0.01107
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.282853      Inf
## sample estimates:
## odds ratio
##   2.959827
```

```r
library(exact2x2)
```

```
## Loading required package: exactci
```

```
## Loading required package: ssanv
```

```r
exact2x2(cont.table, alternative = "greater", midp = TRUE)
```

```
##
##  One-sided Fisher's Exact Test (mid-p version)
##
## data:  cont.table
## p-value = 0.007188
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.37909     Inf
## sample estimates:
## odds ratio
##   2.959827
```

The p-value is 0.007188 hence we reject the null hypothesis. Consequently, we have sufficient evidence to prove our claim that a respondent is more likely to participate in photography and 4WD.

Question 3

a

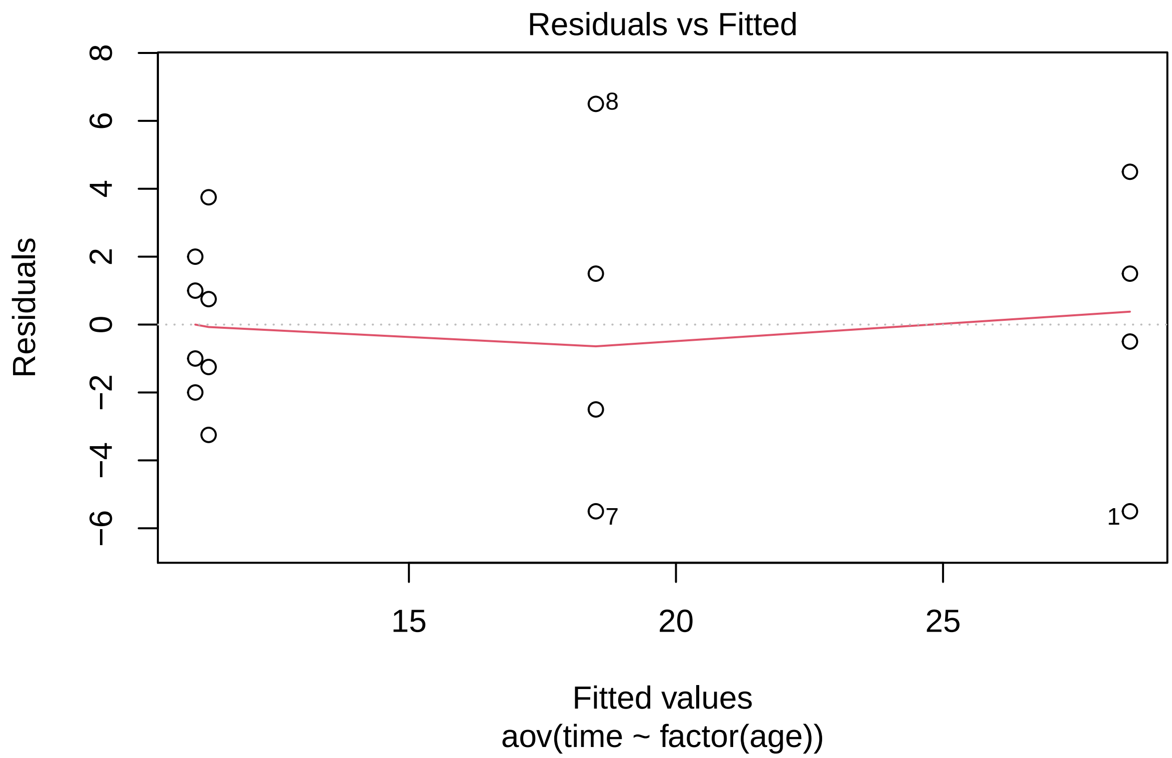
```r
time <- c(23, 30, 28, 33, 20, 16, 13, 25, 10, 8, 15, 12, 10, 12, 9, 13)

age <- rep(c(3, 5, 7, 9), c(4, 4, 4, 4))
```

ANOVA Test

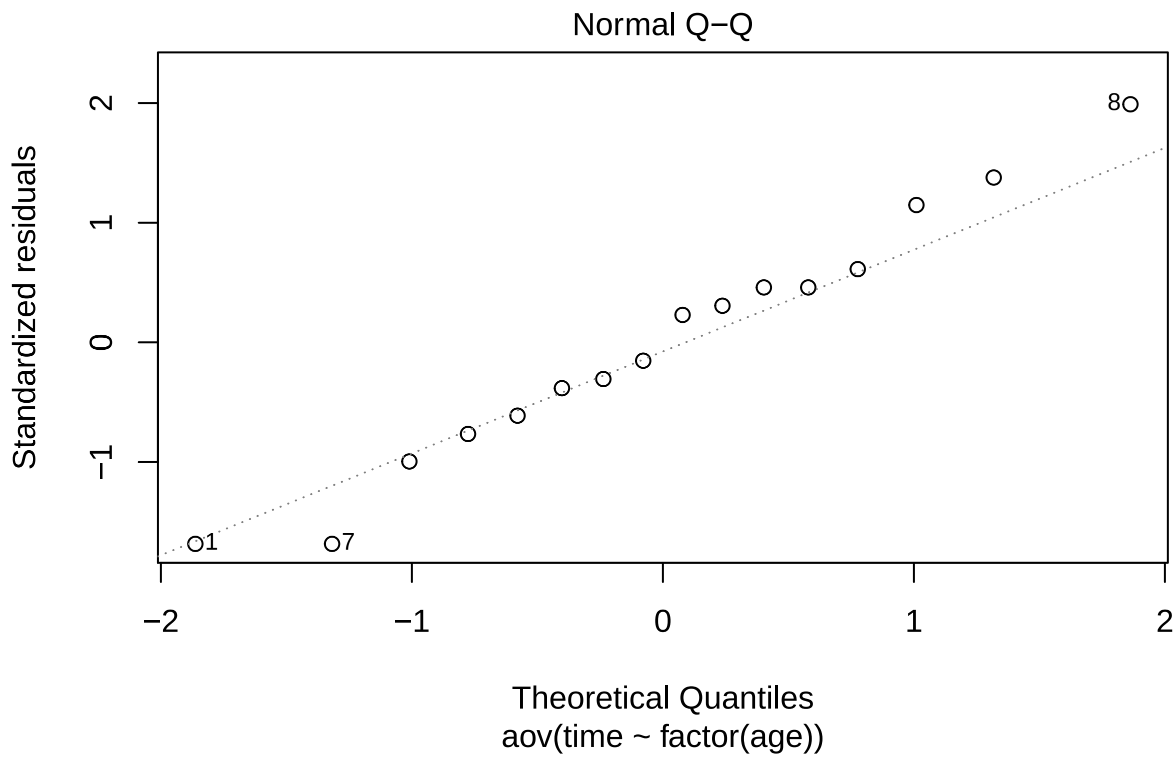
```r
test_ANOVA <- aov(time~factor(age))
summary(test_ANOVA)
```

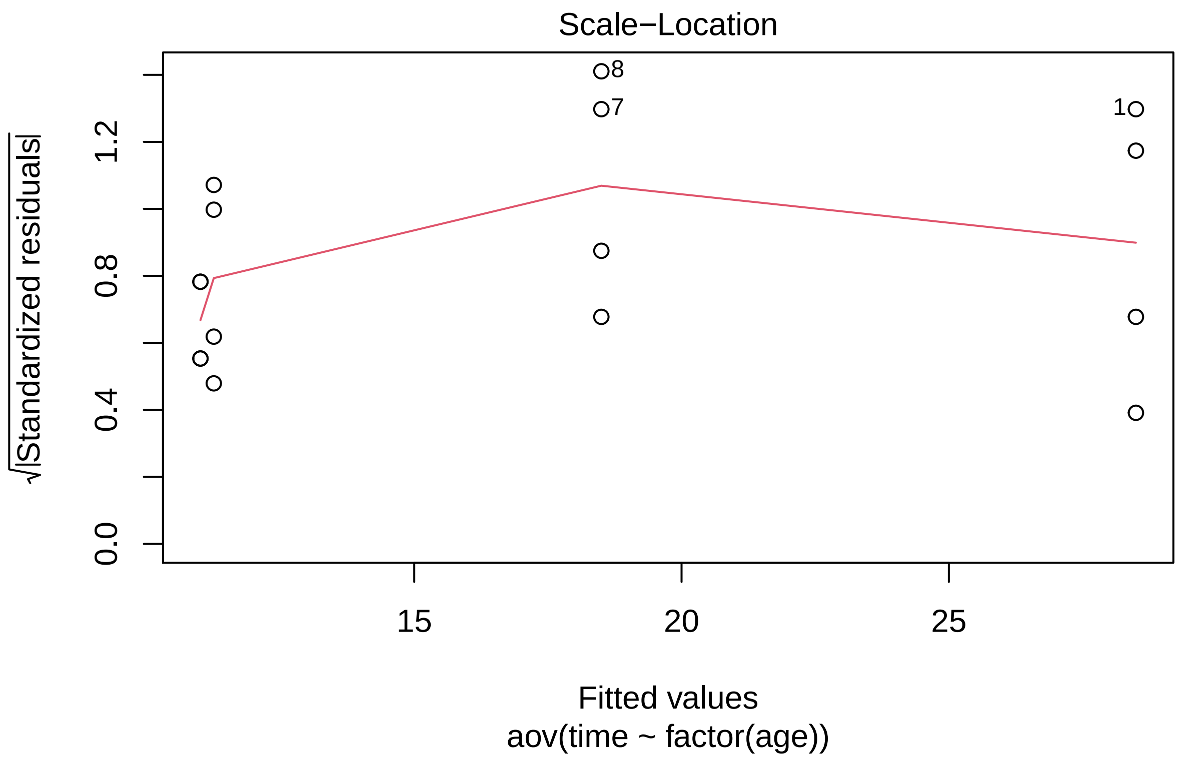
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## factor(age)  3  812.7  270.90   19.04 7.41e-05 ***
## Residuals   12  170.8   14.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

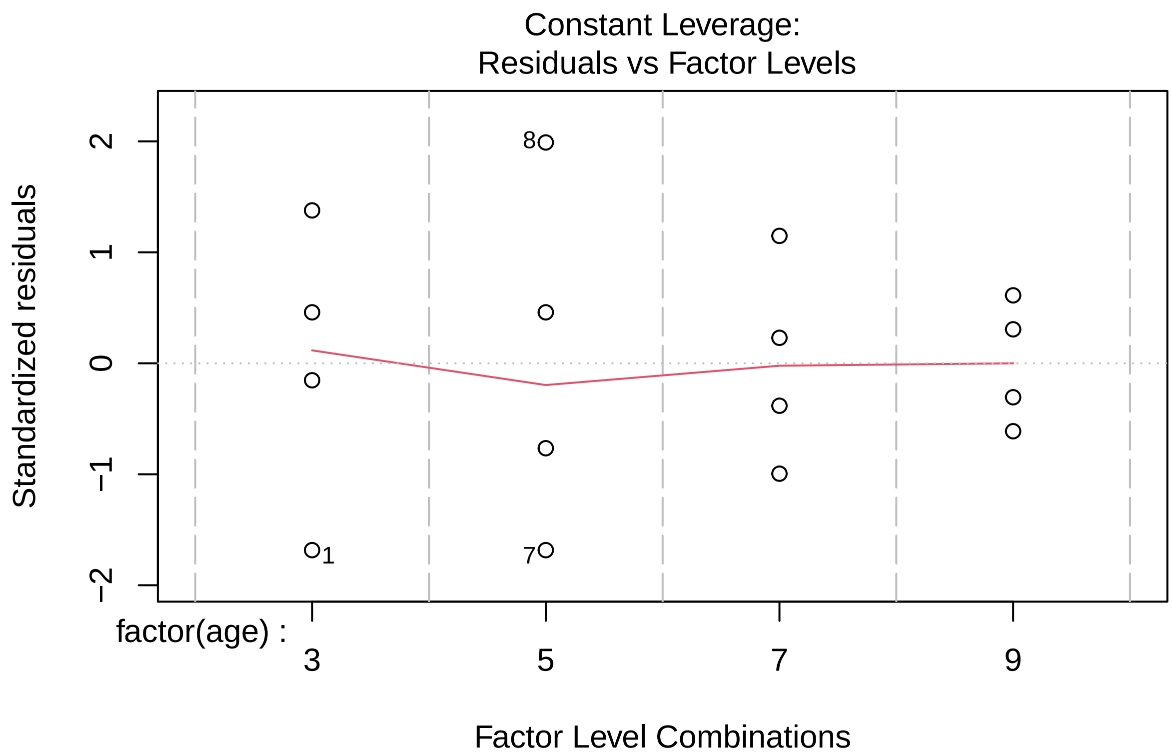
```r
plot(test_ANOVA)
```



Residuals vs Fitted
Residuals
8
6
4
2
0
−2
−4
−6
8
7
1
15
20
25
Fitted values
aov(time ~ factor(age))


Normal Q−Q
Standardized residuals
2
1
0
−1
8
1
7
−2
−1
0
1
2
Theoretical Quantiles
aov(time ~ factor(age))


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
8
7
1
15
20
25
Fitted values
aov(time ~ factor(age))




```r
#install package car
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:DescTools':
##
##     Recode
```

```r
# Checking for assumptions using Levene's test
leveneTest(time~factor(age))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.2514 0.3346
##       12
```

```r
# Checking for assumptions using Shapiro's  test
shapiro.test(time)
```

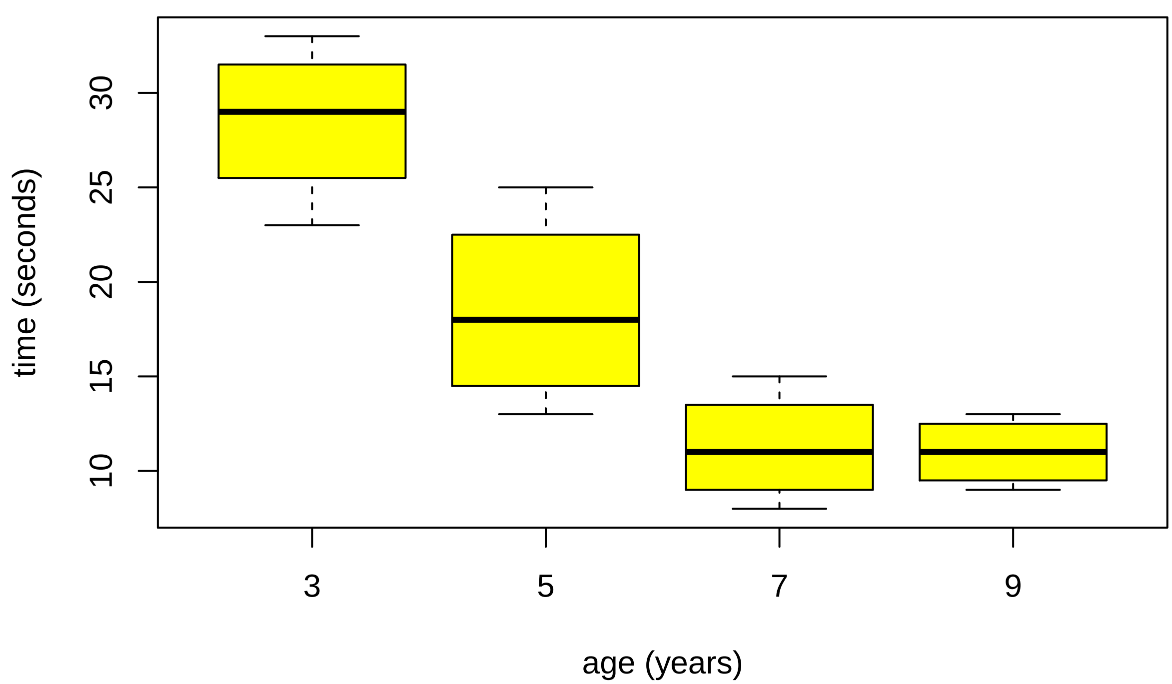
```
##
##  Shapiro-Wilk normality test
##
## data:  time
## W = 0.89143, p-value = 0.05868
```

Kruskal-Wallis Test

```
kruskal.test(time~factor(age))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  time by factor(age)
## Kruskal-Wallis chi-squared = 11.638, df = 3, p-value = 0.008733
```

```
boxplot(time~factor(age),
        xlab = "age (years)",
        ylab = "time (seconds)",
        col = "yellow")
```



b

Model Equation

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Null and alternative Hypotheses

H0 : All group means are equal

H1 : At least one of the group means is not equal

Assumptions about the data, and comments about whether diagnostic graphs support those assumptions

There are three assumptions that must be met for an ANOVA which include equal variance, normality and sample independence.

Levene's test had a p-value of 0.3346 which is greater than the 0.05 significance level hence we can assume that equal variance is met.

Shapiro's test had a p-value of 0.05868 which is slightly greater than the 0.05 significance level hence we can assume that normality is met.

When we look at the Residuals vs Fitted graph we can see that equal variance is met due to the distribution of the dots being equally distributed. The Normal Q-Q graph shows the distribution of dots is linear hence we can say that normality is met. Both of these diagnostic graphs support the assumptions made by the tests above.

Kruskal-Wallis Test The Kruskal-Wallis p-value is 0.008733 which is less than the 0.05 significance level hence we reject the null hypothesis as we have evidence that at least one of the group means is not equal.

From the box plot graph above we can see that the first two box plots of age 3 and 5 have different group means in comparison to the last two box plots of age 7 and 9 where the group means are fairly similar. The first box plot is statistically significantly higher with a mean near 30 seconds while the two box plots on the right have the lowest mean of around 10 seconds. This box plot graph supports the ANOVA results.

In conclusion, we can see that children who are younger will take longer to complete the manipulative task which requires hand to eye coordination than children who are older

Question 4

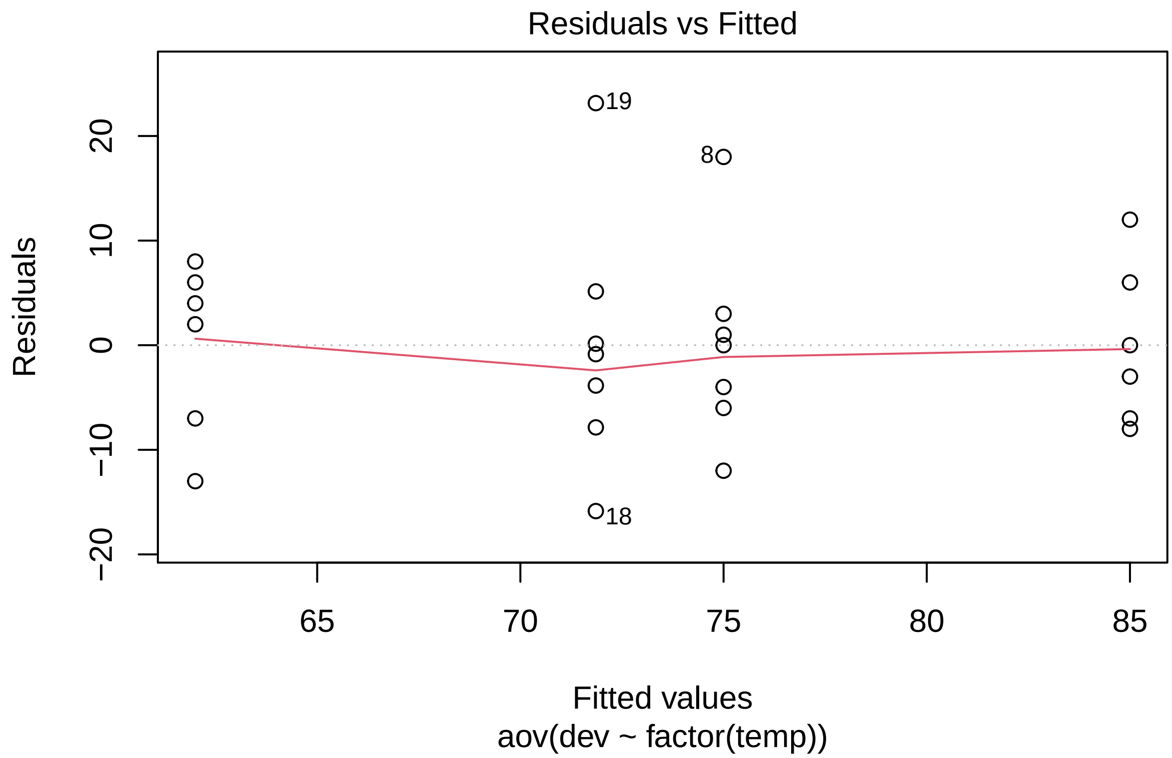
a

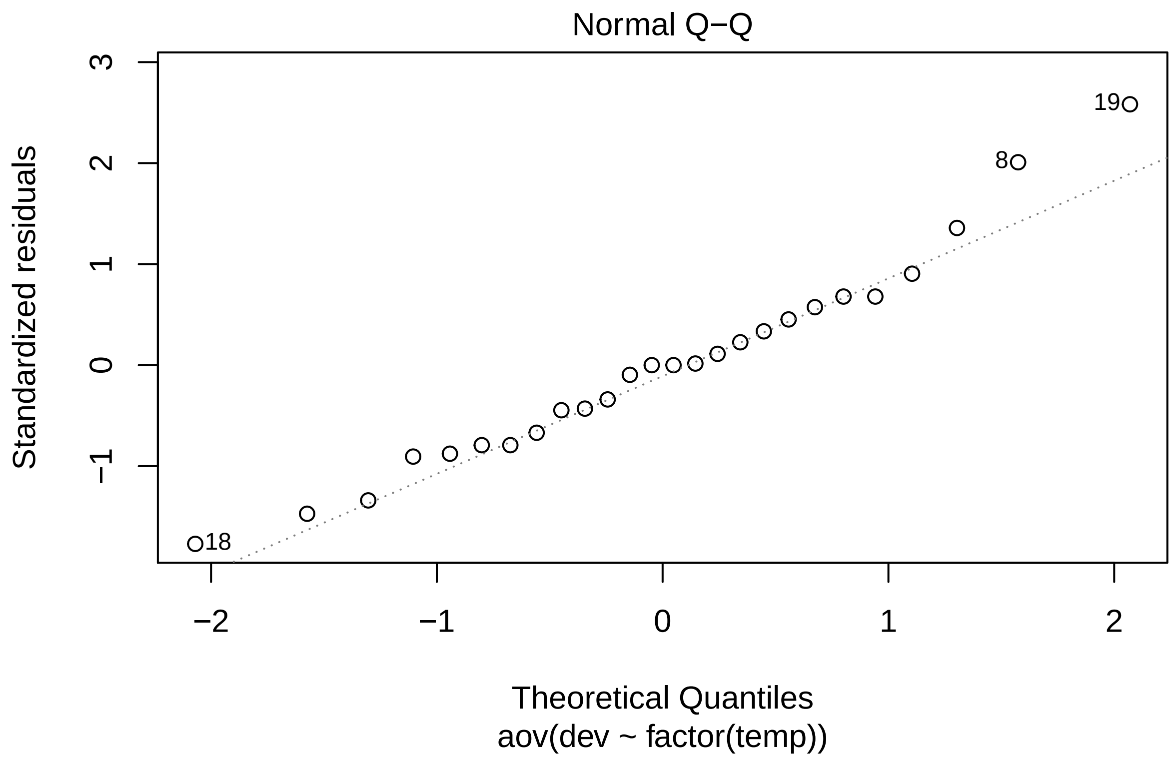
```r
dev <- c(78, 91,97,82,85,77,75,93,78,71,63,76,69,64,72,68,77,56,95,71,55,66,49,64,70,68)

temp <- rep(c(12, 16, 20, 24), c(6, 7, 7, 6))

test_ANOVA <- aov(dev~factor(temp))
summary(test_ANOVA)
```

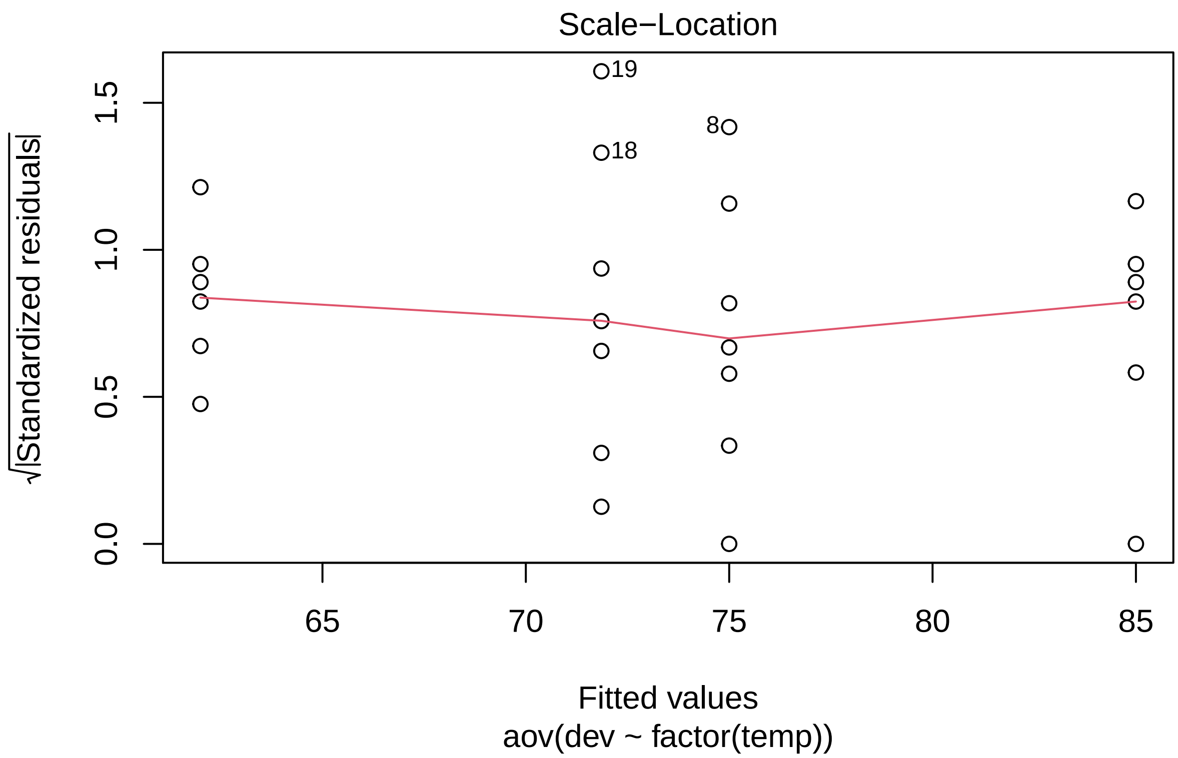
```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## factor(temp)  3   1622   540.5    5.77 0.00455 **
## Residuals    22   2061    93.7
```

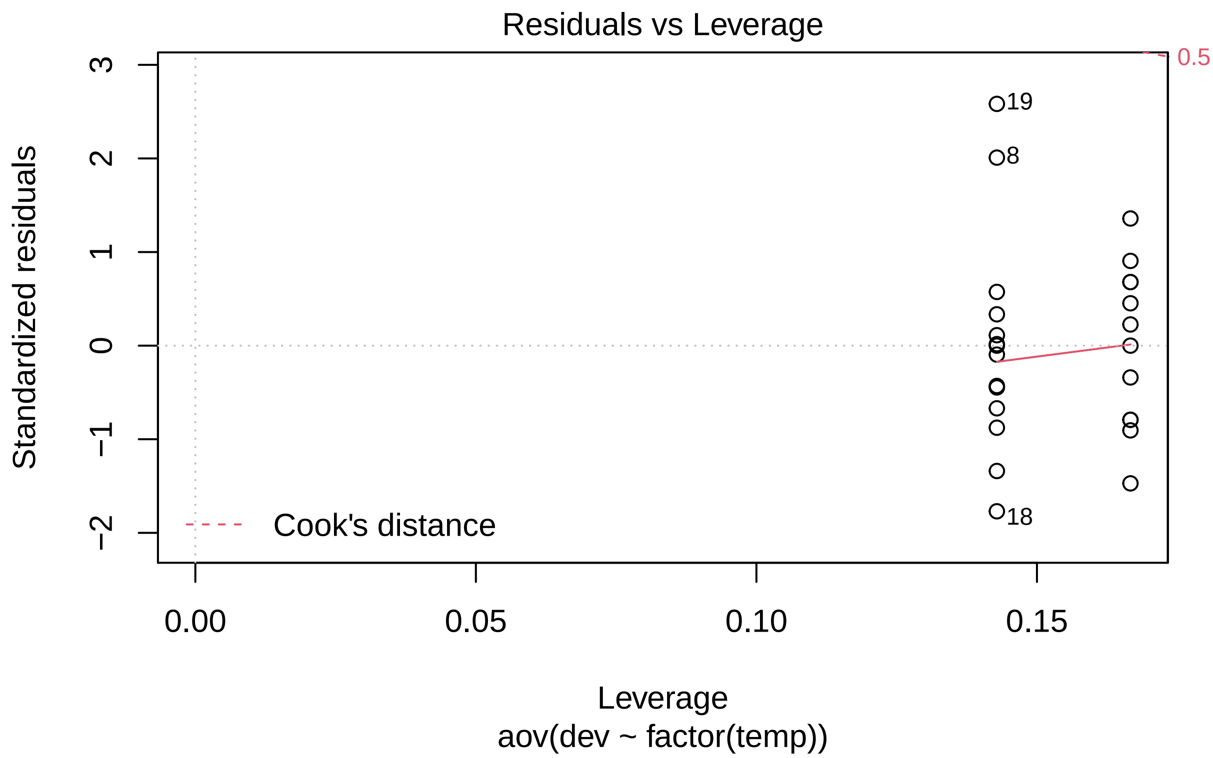
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(test_ANOVA)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(dev ~ factor(temp))
−1
0
1
2
3
−2
−1
0
1
2
18
8
19


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
19
8
18
65
70
75
80
85
Fitted values
aov(dev ~ factor(temp))




```r
#install.packages(car)
library(car)

# Checking for assumptions using Levene's test
leveneTest(dev~factor(temp))
```

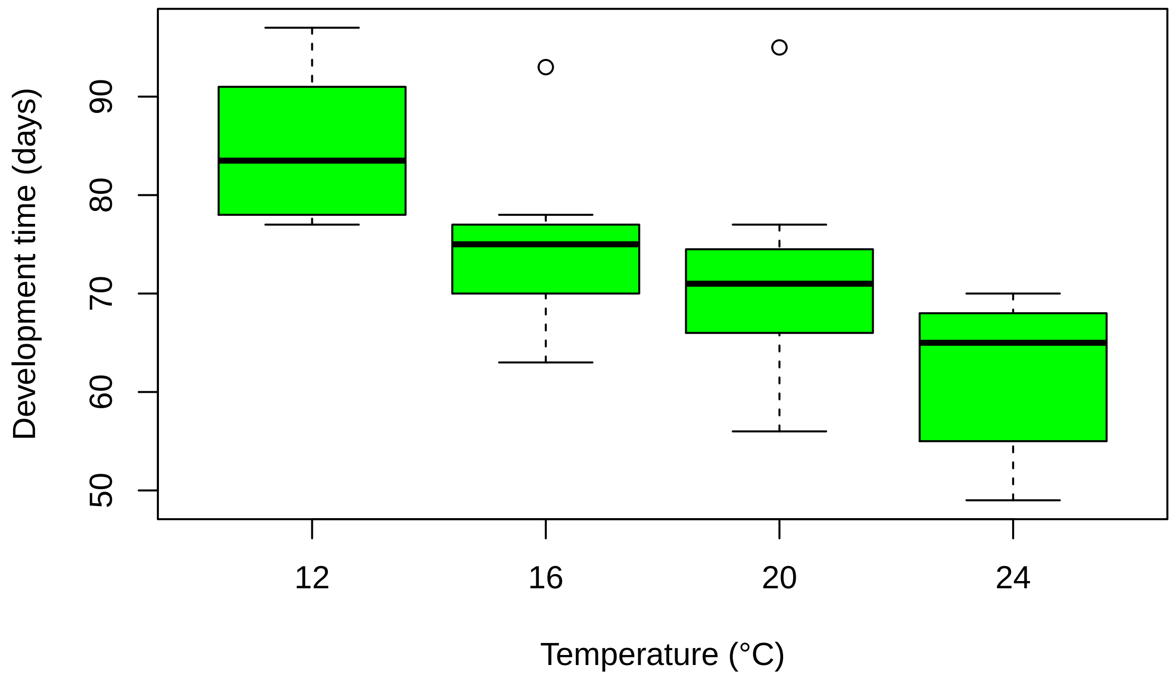
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.1406 0.9346
##       22
```

```r
# Checking for assumptions using Shapiro's test
shapiro.test(dev)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dev
## W = 0.97214, p-value = 0.6793
```

results show the factor temp is less so we reject null hypothesis. This shows a difference between mean development time.

```r
boxplot(dev~factor(temp), #factor category
        xlab = "Temperature (°C)",
        ylab = "Development time (days)",
        col = "green")
```



b

```r
#Tukey test
TukeyHSD(test_ANOVA)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = dev ~ factor(temp))
##
## $`factor(temp)`
##             diff       lwr       upr     p adj
## 16-12 -10.000000 -24.95236  4.952363 0.2747883
## 20-12 -13.142857 -28.09522  1.809505 0.0984799
## 24-12 -23.000000 -38.51680 -7.483200 0.0023828
## 20-16  -3.142857 -17.50862 11.222908 0.9286069
## 24-16 -13.000000 -27.95236  1.952363 0.1036498
## 24-20  -9.857143 -24.80951  5.095220 0.2862932
```

```
#plot(TukeyHSD(test_ANOVA))
```

c

Model Equation

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Null and alternative Hypotheses

H0 : All group means are equal

H1 : At least one of the group means is not equal

Assumptions about the data, and comments about whether diagnostic graphs support those assumptions

There are three assumptions that must be met for an ANOVA which include equal variance, normality and sample independence.

Levene's test had a p-value of 0.9346 which is greater than the 0.05 significance level hence we can assume that equal variance is met.

Shapiro's test had a p-value of 0.6793 which is slightly greater than the 0.05 significance level hence we can assume that normality is met.

When we look at the Residuals vs Fitted graph we can see that equal variance is met due to the distribution of the dots being equally distributed. The Normal Q-Q graph shows the distribution of dots is linear hence we can say that normality is met. Both of these diagnostic graphs support the assumptions made by the tests above.

ANOVA Test The ANOVA Test p-value is 0.00455 which is less than the 0.05 significance level hence we reject the null hypothesis as we have evidence that at least one of the group means is not equal.

From the box plot graph above we can see that the two end box plots have different group means in comparison to the two middle box plots where the group means are fairly similar. The first box plot has the highest mean of above 80 days while the far right box plot has the lowest mean of above 60 days. This box plot graph supports the ANOVA results.

Hence, in conclusion we can see that insects at a lower temperature develop slower than insects at a higher temperature and insects at a higher temperature develop faster than insects at a lower temperature.

Question 5

a

The teeth have been sampled from eight different females and eight different males as it increases representation of both females and males. If they had received eight teeth from one female and male, it would be comparing one person to another rather then a fair representation of a whole population sample.

b

The advantages of choosing eight from each group include:

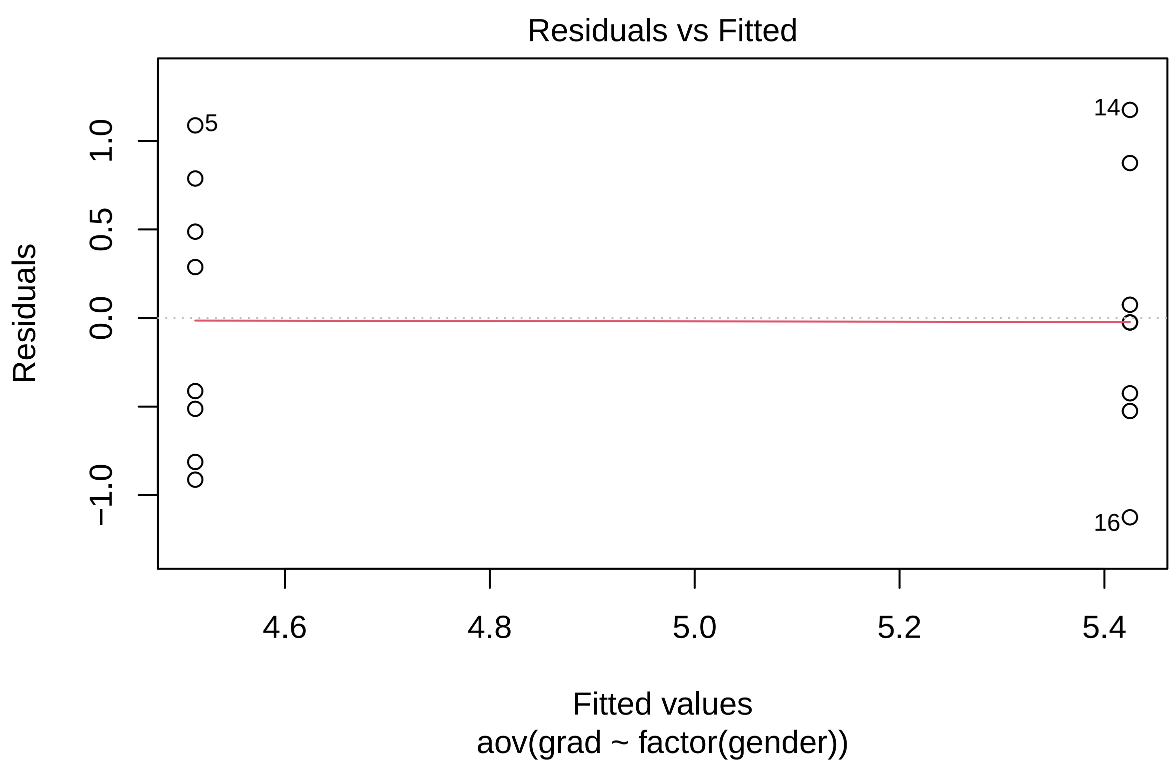
1) Fair and equal representation within a sample that is balanced.
2) Having more females then males would be more bias to females e.g. A population sample of 10 female and 6 males.

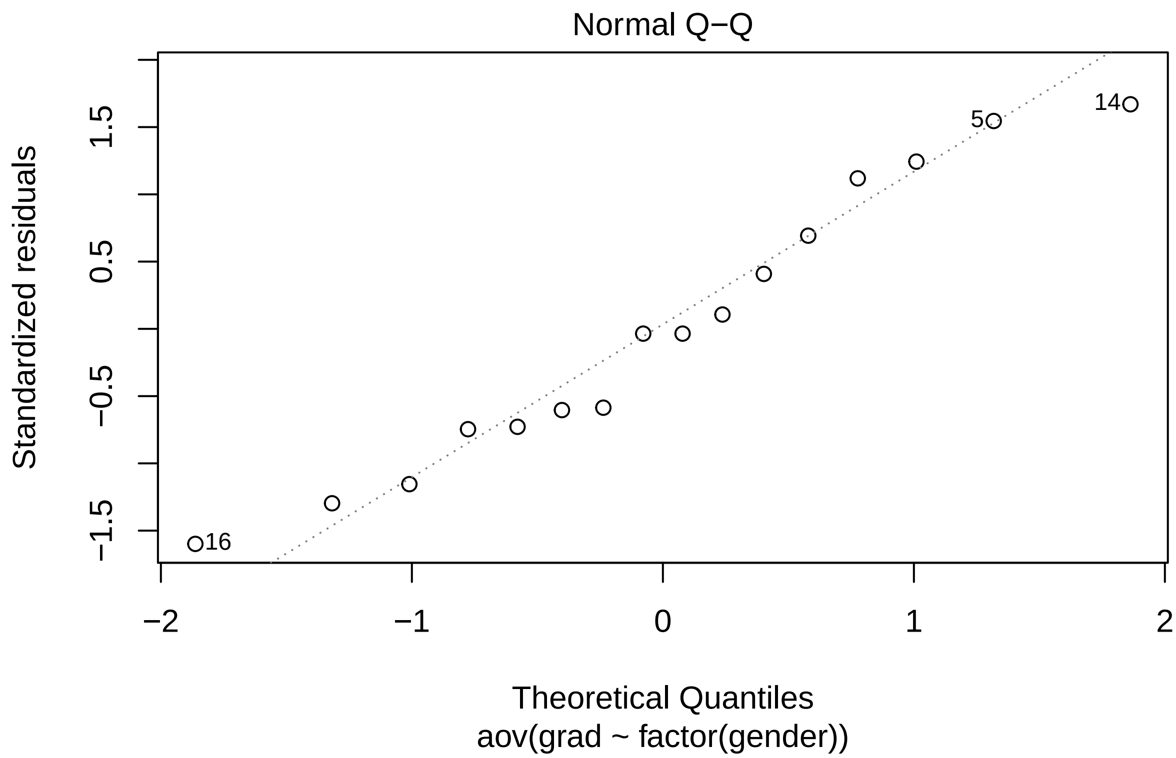
c

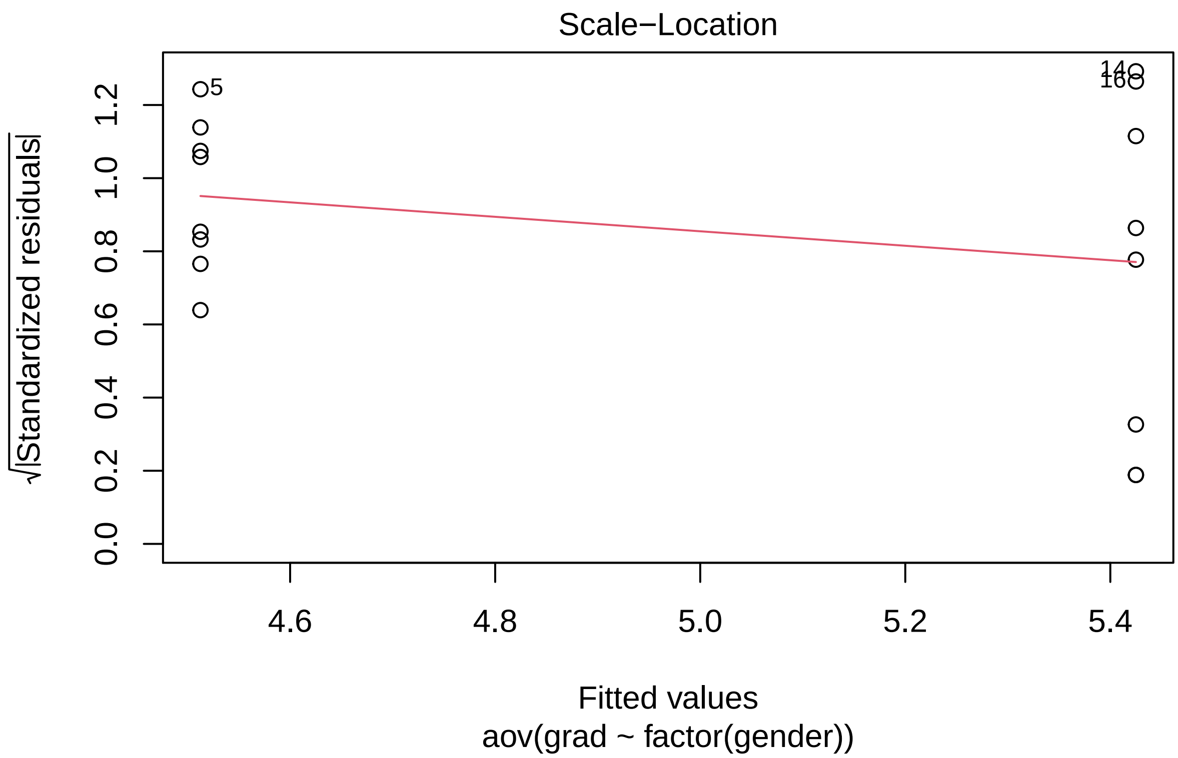
```
grad <- c(4.8,5.3,3.7,4.1,5.6,4.0,3.6,5.0,4.9,5.4,5.0,5.5,5.4,6.6,6.3,4.3)

gender <- rep(c("female", "male"), c(8,8))

test_ANOVA <- aov(grad~factor(gender))
summary(test_ANOVA)
```

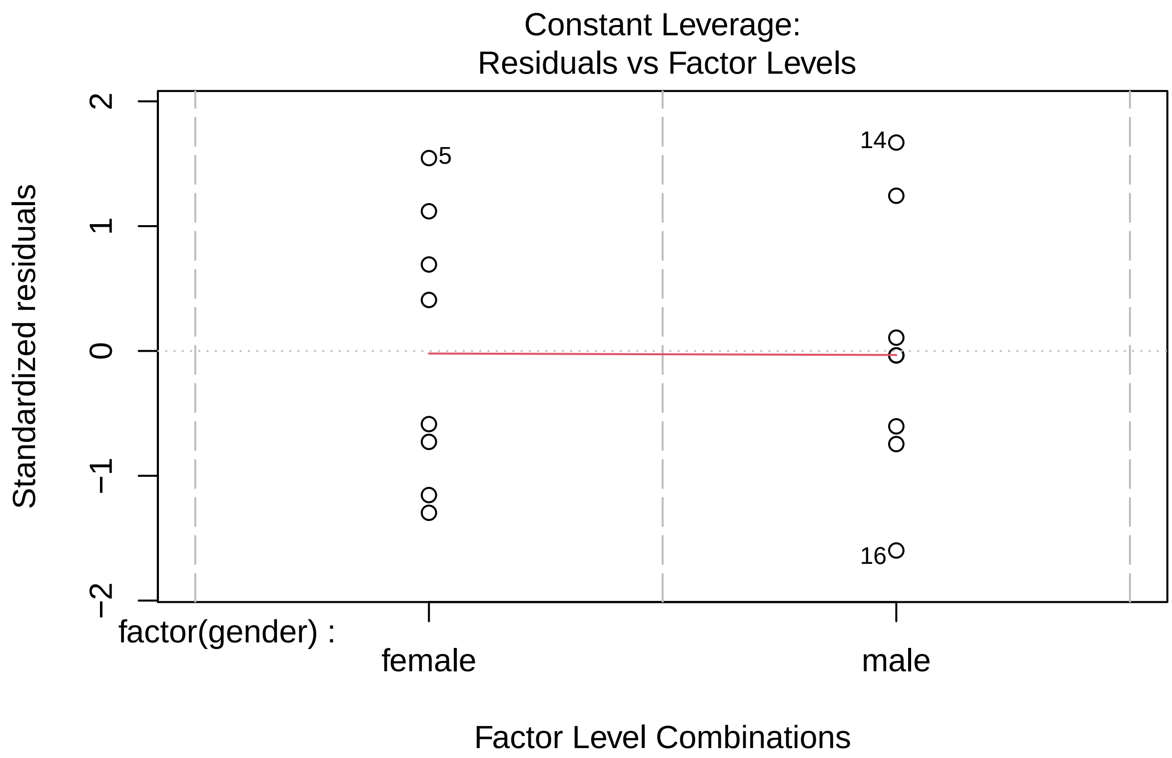
```
##                Df Sum Sq Mean Sq F value Pr(>F)
## factor(gender)  1  3.331   3.331   5.885 0.0294 *
## Residuals      14  7.924   0.566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(test_ANOVA)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(grad ~ factor(gender))
−1.5
−0.5
0.5
1.5
−2
−1
0
1
2
16
5
14


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
5
14
16
4.6
4.8
5.0
5.2
5.4
Fitted values
aov(grad ~ factor(gender))



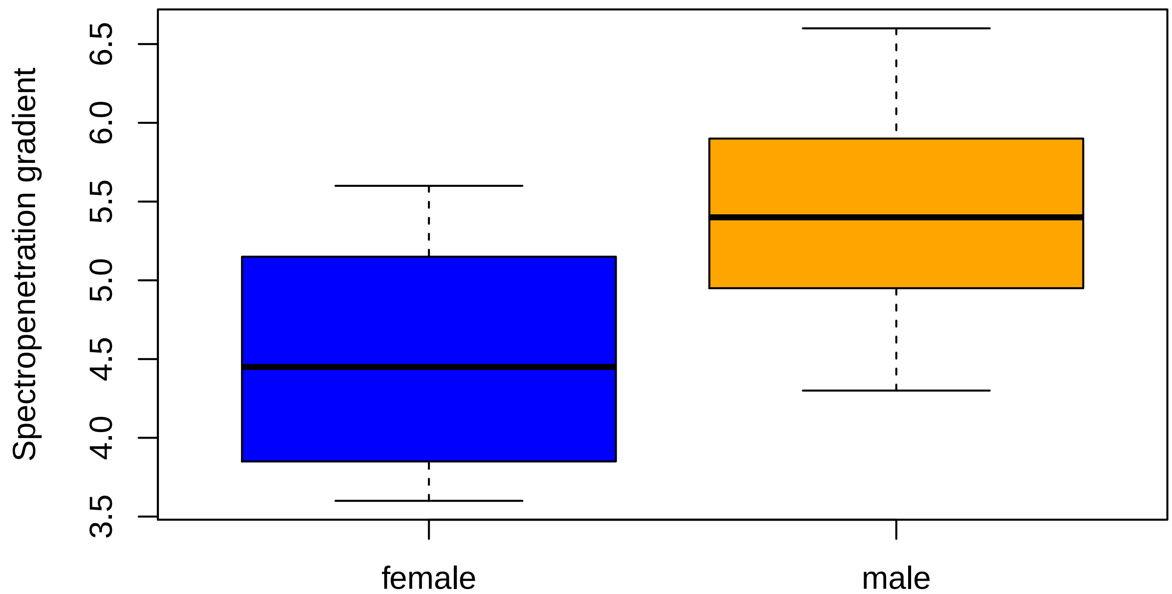

```
#install.packages(car)
library(car)

# Checking for assumptions using Levene's test
leveneTest(grad~factor(gender))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.4719 0.5034
##       14
```

```
# Checking for assumptions using Shapiro's test
shapiro.test(grad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grad
## W = 0.9621, p-value = 0.7001
```

```
boxplot(grad~factor(gender), #factor category
        xlab = "",
        ylab = "Spectropenetration gradient",
        col = c("blue", "orange"))
```



Model Equation

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Null and alternative Hypotheses

H0 : All group means are equal

H1 : At least one of the group means is not equal

Assumptions about the data, and comments about whether diagnostic graphs support those assumptions

There are three assumptions that must be met for an ANOVA which include equal variance, normality and sample independence.

Levene's test had a p-value of 0.5034 which is greater than the 0.05 significance level hence we can assume that equal variance is met.

Shapiro's test had a p-value of 0.7001 which is slightly greater than the 0.05 significance level hence we can assume that normality is met.

When we look at the Residuals vs Fitted graph we can see that equal variance is met due to the distribution of the dots being equally distributed. The Normal Q-Q graph shows the distribution of dots is linear hence we can say that normality is met. Both of these diagnostic graphs support the assumptions made by the tests above.

ANOVA Test The ANOVA Test p-value is 0.0294 which is less than the 0.05 significance level hence we reject the null hypothesis as we have evidence that at least one of the group means is not equal.

From the box plot graph above we can see that both the box plots have different group means in comparison to one another and the male mean is statistically more significant at a mean of 5.5 spectropenetration gradient. The female mean is at a mean of 4.5 spectropenetration gradient. This box plot supports the ANOVA results hence rejecting the null hypothesis.

Hence, in conclusion we can see that the extent to which X-rays can penetrate tooth enamel can be used to differentiate between females and males.

d

A Tukey test compares multiple groups of means where as there is only one comparison in this set of data that is looking at two groups; female and male.