ADS2 Mock Coding Challenge 2

Semester 2, 2023-24

1. Vitamin C and tooth growth

Lack of vitamin C leads to severe health issues. It is not produced in the human body and must be supplied with food. At the same time, personnel that have limited access to fresh vegetables (sailors, spacemen, travelers, etc) may suffer from the insufficiency of this compound in their food. Thus, a vitamin C formulation that can preserve its properties for a long time is of great need.

Researchers developed such a formulation. *In vitro* tests showed its efficiency. Now, they performed an *in vivo* trial. Guinea pigs received the newly developed formulation of Vitamin C or fresh orange juice (normalized according to the concentration of vitamin C) in addition to their standard diet (`supp`). Each type of additives included three concentrations (`dose`) of vitamin C: 0.5, 1, and 2 mg/ml. The measured outcome is the tooth length (`len`) in mm (stem cells that become teeth are sensitive to vitamin C).

Import, check, and organize the data appropriately. Reformat columns if needed.

```
teeth <- read.csv("teeth.csv")
head(teeth)
```

```
##   X  len supp dose
## 1 1 3.00   VC  0.5
## 2 2 5.95   VC  0.5
## 3 3 4.25   VC  0.5
## 4 4 3.65   VC  0.5
## 5 5 3.89   VC  0.5
## 6 6 5.34   VC  0.5
```

```
anyNA(teeth)
```

```
## [1] FALSE
```

```
anyDuplicated(teeth)
```

```
## [1] 0
```

Apparently, supp and dose must be grouping variables. It is reasonable to recode them and change their order. You can also check which values are in each column to make sure that there are no weird values like " " or similar.

```
teeth <- teeth %>%
mutate(dose = factor(dose, levels = c(0.5, 1, 2), ordered = T),
supp = as.factor(supp)) %>%
relocate(supp, dose)
str(teeth)
```

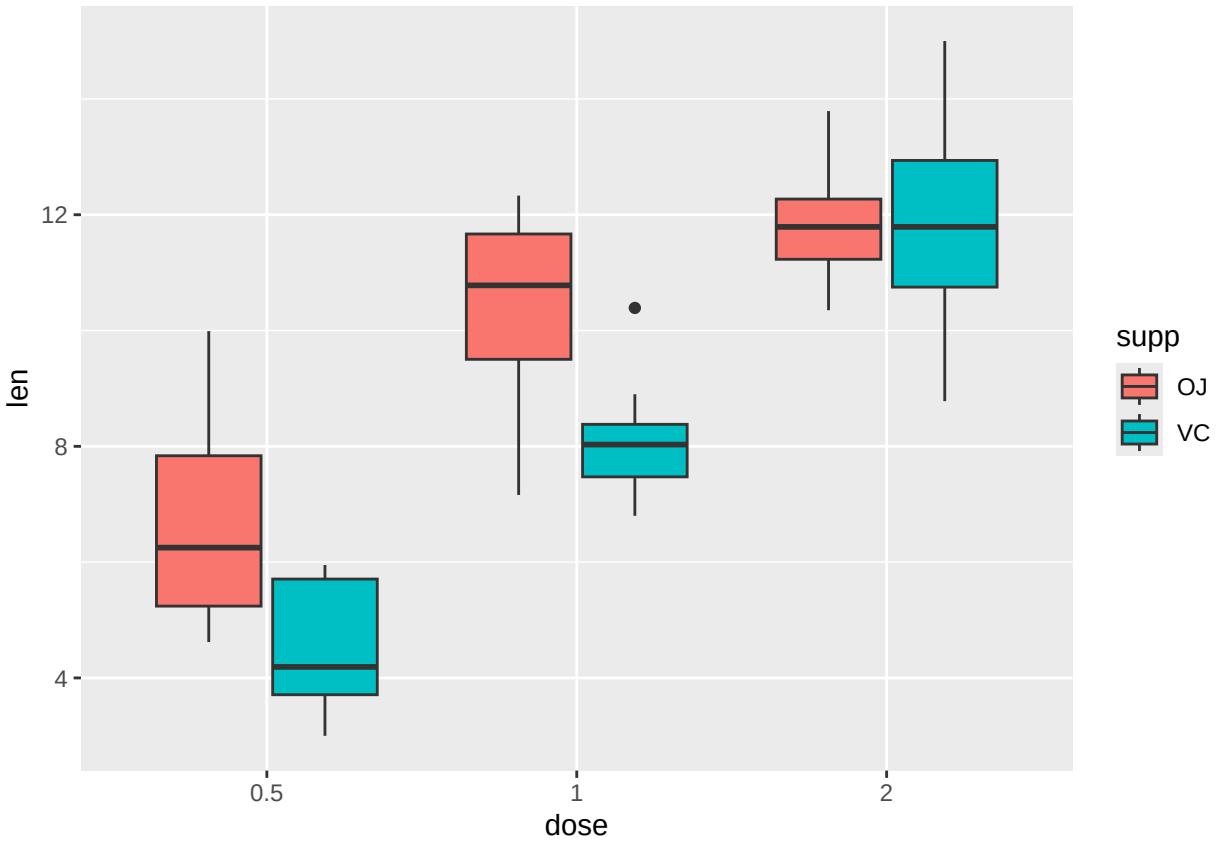
```
## 'data.frame':    60 obs. of  4 variables:
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Ord.factor w/ 3 levels "0.5"<"1"<"2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ X   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ len : num  3 5.95 4.25 3.65 3.89 5.34 5.83 5.91 3.4 4.13 ...
```

```
head(teeth)
```

```
##   supp dose X  len
## 1   VC  0.5 1 3.00
## 2   VC  0.5 2 5.95
## 3   VC  0.5 3 4.25
## 4   VC  0.5 4 3.65
## 5   VC  0.5 5 3.89
## 6   VC  0.5 6 5.34
```

Plot the data in a useful way.

```
p.box <- ggplot(teeth) + geom_boxplot(aes(x=dose, y=len, fill = supp))
p.box
```



Choose, justify, state the statistical hypotheses, and carry out an appropriate test to answer whether the vitamin C formula is useful.

Choose the appropriate test

One way annova can be simulated We have 2 levels of supp × 3 levels of dose = 6 groups. It means we need to use ANOVA. As we have 2 factors, we should try to use a 2-way ANOVA if the data fit the requirements. Thus, the method of choice is a 2-way ANOVA or (if cannot run it) the Kruskal-Wallis test.

- is for model with interaction;
- is for model without interaction;

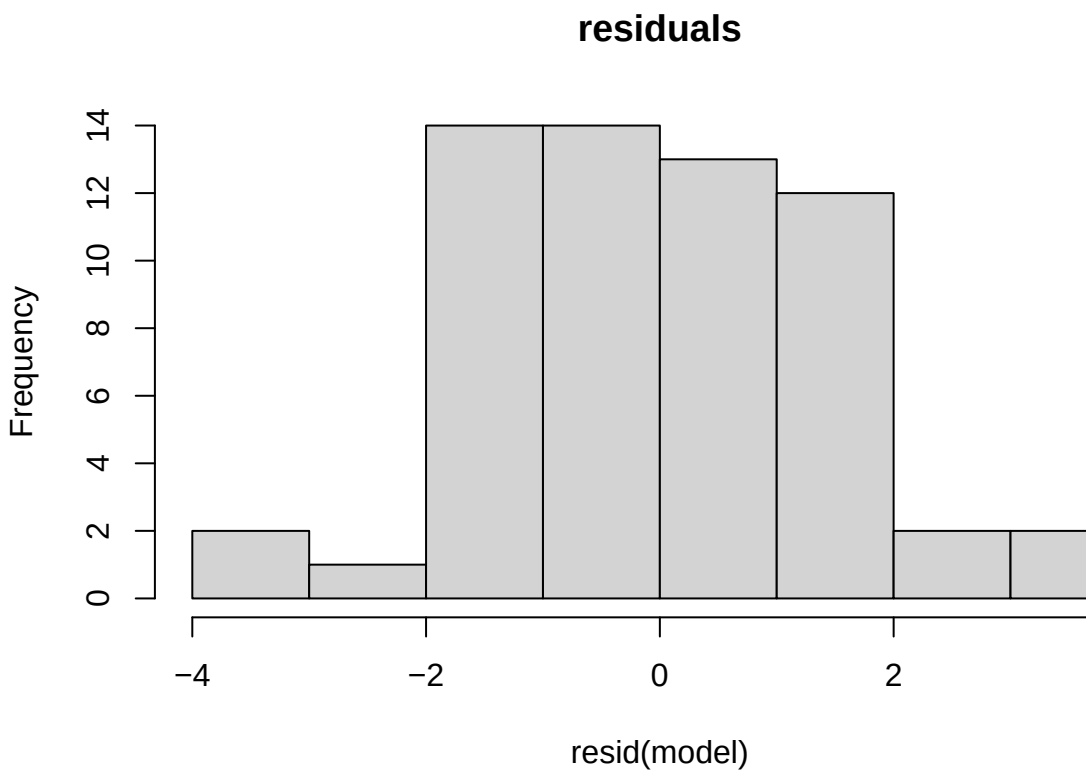
```
model <- aov(len ~ supp + dose + supp:dose, data = teeth)
```

Justify your test choice

Assumptions for a 2-way ANOVA test are:

1. **Independence of observations.** • We can assume it at once.

```r
hist(resid(model), main = "residuals")
```

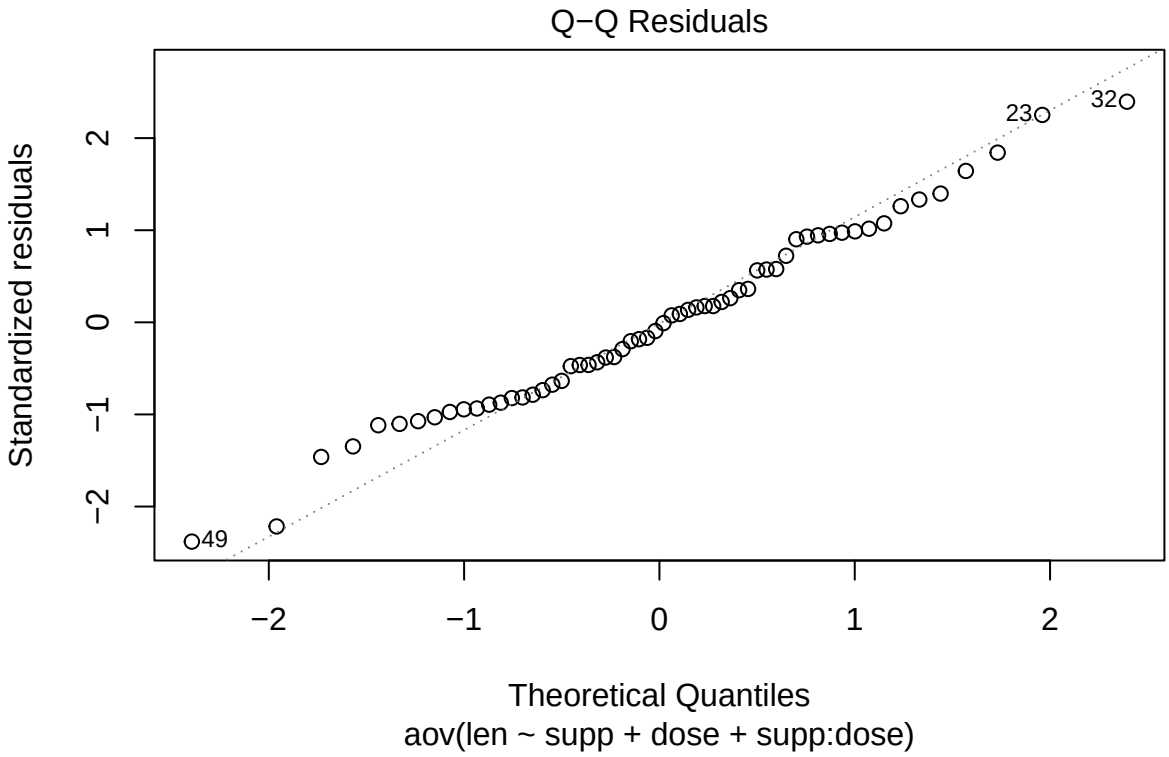


2. Normality of residuals:

```r
shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.98522, p-value = 0.6817
```

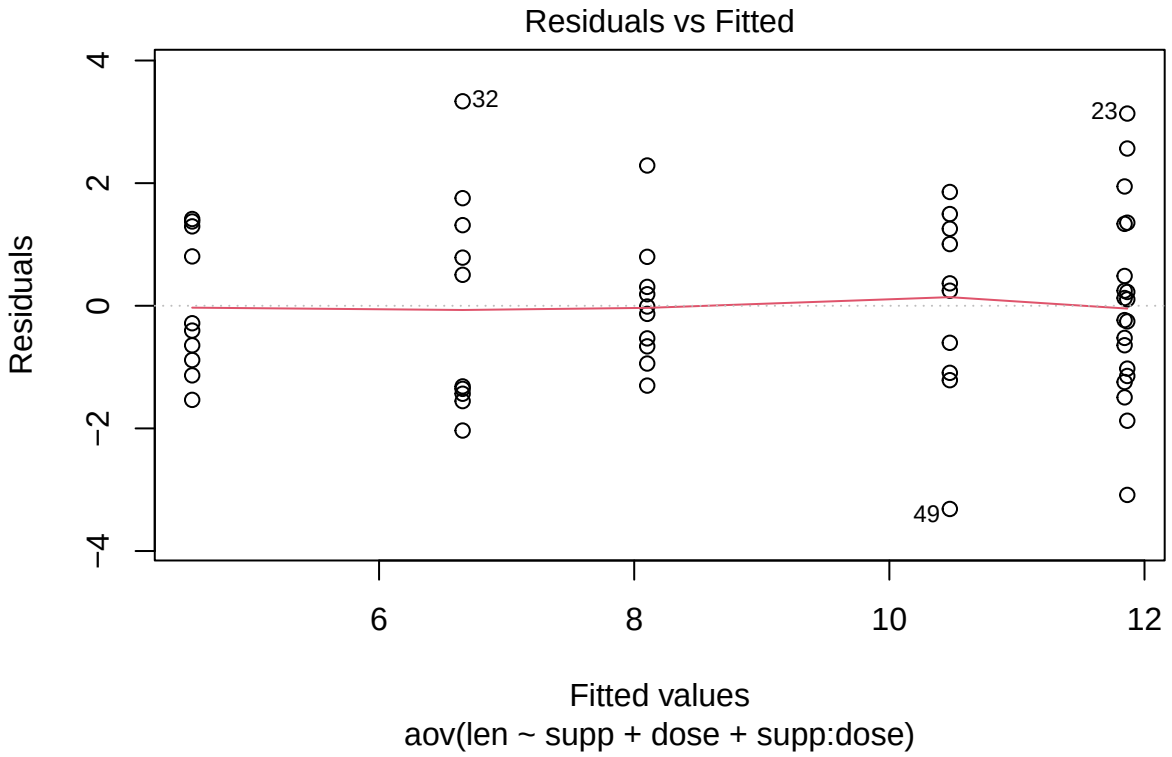
```r
plot(model, 2)
```



Based on the histogram, we cannot deny that the data do not obey a normal distribution. However, based on the p-value of the Shapiro test, we can believe that the residuals obey a normal distribution.

3. Equality of variance:

```
plot(model, 1)
```



The plot indicates that the variances are approximately equal across groups, which meets the assumption for using the ANOVA test.

4. Equal group size (have to use different types of SS calculation for the ANOVA table if this requirement is violated): • The group size can be noticed in the data diagnosis step

The first plot shows the distance of residuals from the mean (is comparable across groups) and the second one shows residuals according to the normal distribution (close to the expected values)

We can use parametric ANOVA.

State the statistical hypotheses • H0: means of different supp groups are the same
• H1: means of different supp groups are NOT the same

Carry out an appropriate test to answer whether the vitamin C formula is useful.

```
summary(model)
```

Annova test

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1   33.3   33.32  15.473 0.000241 ***
## dose         2  396.0  198.02  91.965  < 2e-16 ***
## supp:dose    2   17.3    8.64   4.015 0.023683 *
## Residuals   54  116.3    2.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(model)
```

Post-hoc tests

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp + dose + supp:dose, data = teeth)
##
## $supp
##            diff       lwr        upr     p adj
## VC-OJ -1.490333 -2.249938 -0.7307291 0.0002408
##
## $dose
##         diff      lwr      upr   p adj
## 1-0.5 3.6930 2.574698 4.811302 0.0e+00
## 2-0.5 6.2595 5.141198 7.377802 0.0e+00
## 2-1   2.5665 1.448198 3.684802 2.8e-06
##
## $`supp:dose`
##                  diff        lwr        upr     p adj
## VC:0.5-OJ:0.5 -2.120 -4.0588347 -0.1811653 0.0243952
## OJ:1-OJ:0.5    3.819  1.8801653  5.7578347 0.0000048
## VC:1-OJ:0.5    1.447 -0.4918347  3.3858347 0.2524625
## OJ:2-OJ:0.5    5.189  3.2501653  7.1278347 0.0000000
## VC:2-OJ:0.5    5.210  3.2711653  7.1488347 0.0000000
## OJ:1-VC:0.5    5.939  4.0001653  7.8778347 0.0000000
## VC:1-VC:0.5    3.567  1.6281653  5.5058347 0.0000193
## OJ:2-VC:0.5    7.309  5.3701653  9.2478347 0.0000000
## VC:2-VC:0.5    7.330  5.3911653  9.2688347 0.0000000
## VC:1-OJ:1     -2.372 -4.3108347 -0.4331653 0.0082439
## OJ:2-OJ:1      1.370 -0.5688347  3.3088347 0.3090572
## VC:2-OJ:1      1.391 -0.5478347  3.3298347 0.2929176
## OJ:2-VC:1      3.742  1.8031653  5.6808347 0.0000074
## VC:2-VC:1      3.763  1.8241653  5.7018347 0.0000066
## VC:2-OJ:2      0.021 -1.9178347  1.9598347 1.0000000
```

Present and discuss your results. Is this novel formula useful? What would you suggest doing next?

The ANOVA results indicate significant effects of both the 'supp' factor ($F = 15.473$, $p < 0.001$) and the 'dose' factor ($F = 91.965$, $p < 0.001$) on the 'len' variable. Additionally, there is a significant interaction

between 'supp' and 'dose' ($F = 4.015$, $p = 0.024$), suggesting that the effect of one factor depends on the levels of the other factor. The residuals represent unexplained variance in the model, which is relatively low (Residuals: Mean Sq = 2.15).
The Post-hoc test result shows: • the new formulation is generally inferior to the fresh orange juice (~ 3.7 mm, $p < 0.001$);
• lower doses are worse than higher doses (~ 15.5 mm difference between the highest and the lowest concentration of the supplement, $p < 0.001$);
• at the highest dose, the new formulation is as good as the fresh orange juice ($p = 1$).
Altogether, the formula can be used to substitute natural dietary vitamin C, but only at a high dose.

What would you suggest doing next?

Still, the effect of each supplementation is a bit unclear due to the lack of the non-treated group.
The addition of the one would be cumbersome as we would have to give up the factorial design. It would be possible to either add a non-treated group and normalize all the other values to it or add a dose group with very low concentration.
As the formula is functional, it is possible to run some toxicological tests, test the long-term efficiency, and go for clinical trials. It may be possible to optimize the formula further, as the response is lower than that of the fresh juice at lower doses.

Possibly, more work can be spent on the bioavailability of vitamin C in the formula. But do not boldly write: "Let's increase the sample size". What will you see if the sample size is higher? What will you do with the even lower p-value? Think about it. If you still decide to ask for the one, you need to explain why you would like to do that, what you would like to see, and how much you want to increase the sample size.

Apart from the two way annova

Simulation and one way annova

```
one_way <- data.frame(
  group = rep(c("A", "B", "C"), each = 20),
  value = c(rnorm(20, mean = 10, sd = 2),
            rnorm(20, mean = 12, sd = 2),
            rnorm(20, mean = 15, sd = 2))
)
```

```
anova_result <- aov(value ~ group, data = one_way)
summary(anova_result)
```

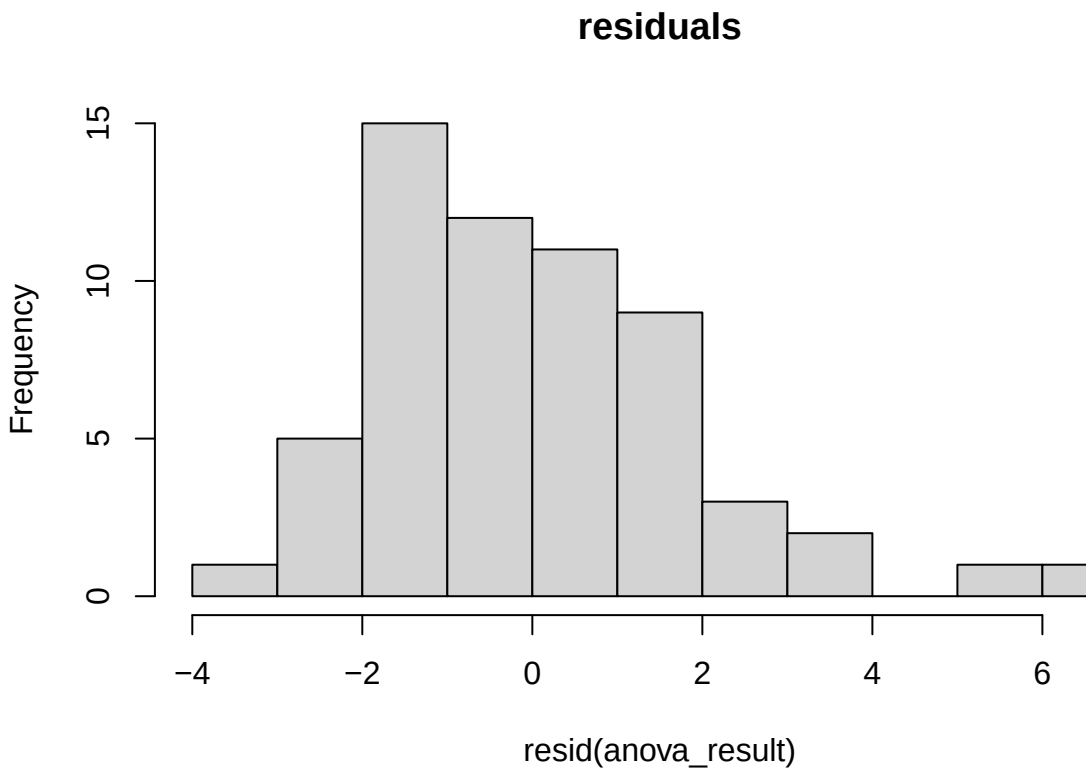
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        2  326.7  163.33   45.69 1.44e-12 ***
## Residuals   57  203.8    3.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Justify your test choice

Assumptions for a 2-way ANOVA test are:

1. Independence of observations. • We can assume it at once.

```
hist(resid(anova_result), main = "residuals")
```

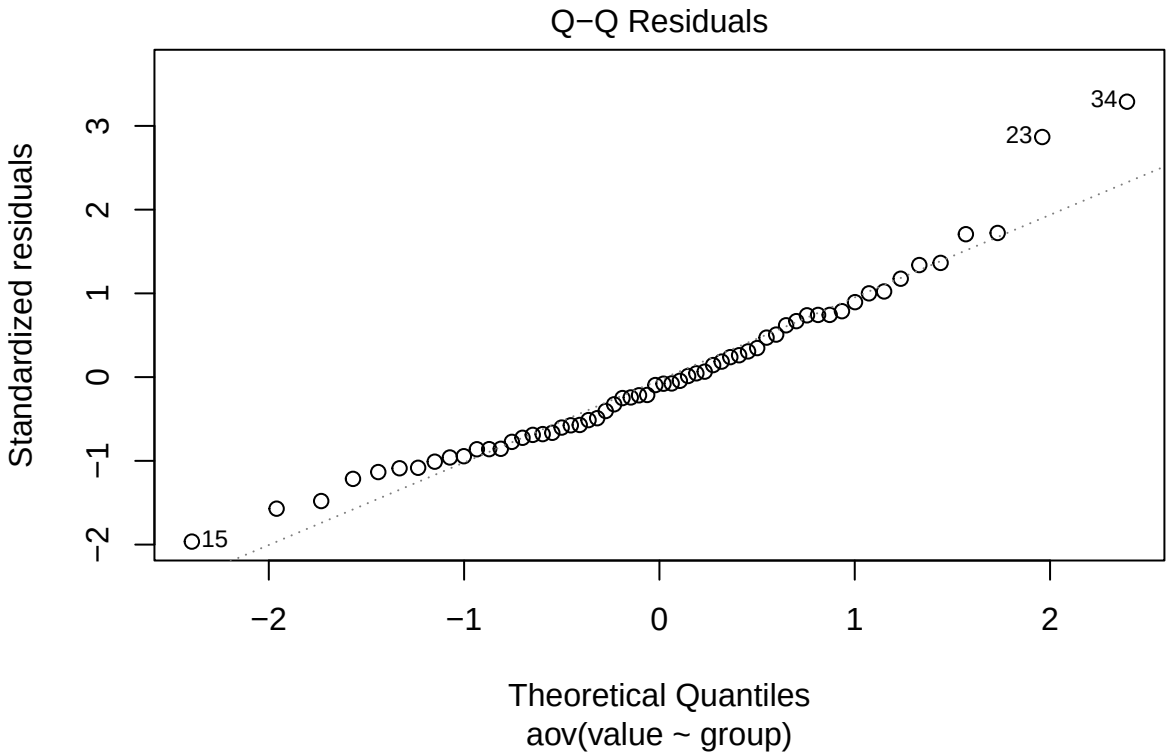


2. Normality of residuals:

```
shapiro.test(resid(anova_result))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(anova_result)
## W = 0.95373, p-value = 0.0234
```

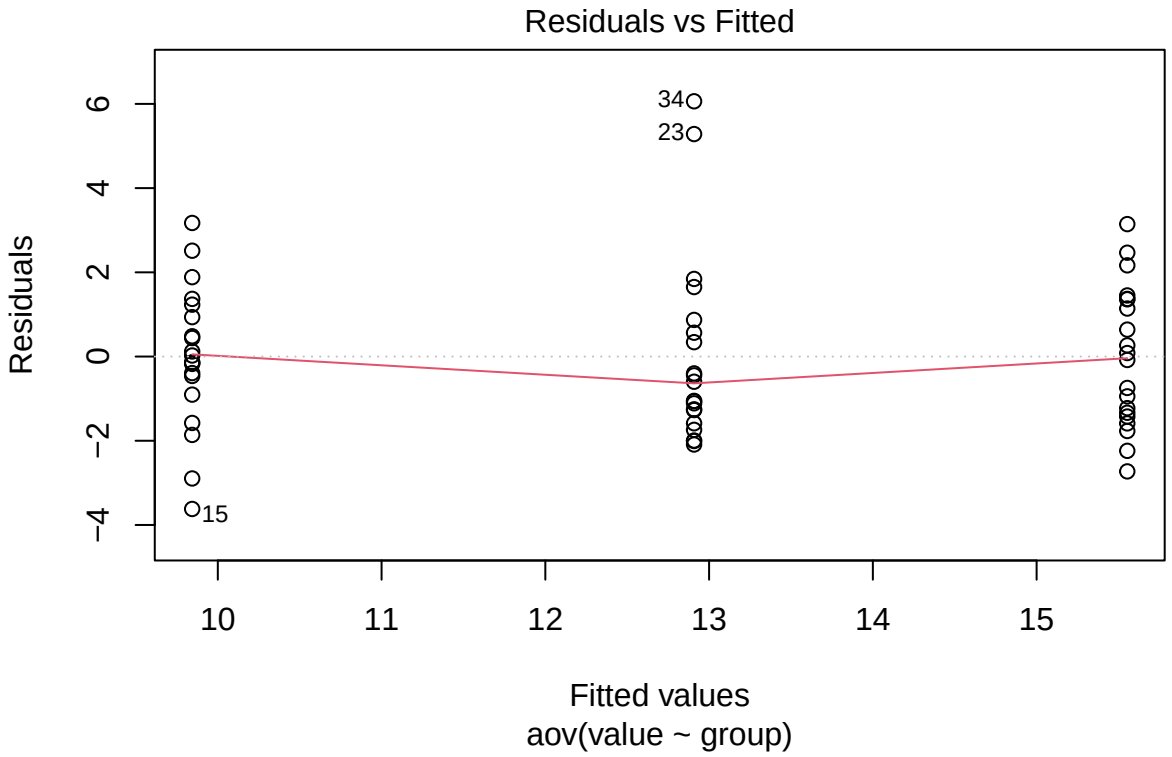
```
plot(anova_result, 2)
```



Based on the histogram, we cannot deny that the data do not obey a normal distribution. However, based on the p-value of the Shapiro test, we can believe that the residuals obey a normal distribution.

3. Equality of variance:

```
plot(anova_result, 1)
```



The plot indicates that the variances are approximately equal across groups, which meets the assumption for using the ANOVA test.

4. Equal group size (have to use different types of SS calculation for the ANOVA table if this requirement is violated): • The group size can be noticed in the data diagnosis step

The first plot shows the distance of residuals from the mean (is comparable across groups) and the second one shows residuals according to the normal distribution (close to the expected values)

```
TukeyHSD(anova_result)
```

We can use parametric ANOVA.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = value ~ group, data = one_way)
##
## $group
##         diff      lwr      upr     p adj
## B-A 3.064934 1.626138 4.503730 0.0000109
## C-A 5.710358 4.271562 7.149154 0.0000000
## C-B 2.645424 1.206628 4.084220 0.0001292
```

Not suitable for the annova

```
non_normal_data <- data.frame(
  group = rep(c("A", "B", "C"), each = 20),
  value = c(runif(20, min = 5, max = 15),
            runif(20, min = 8, max = 18),
            runif(20, min = 10, max = 20))
)
```

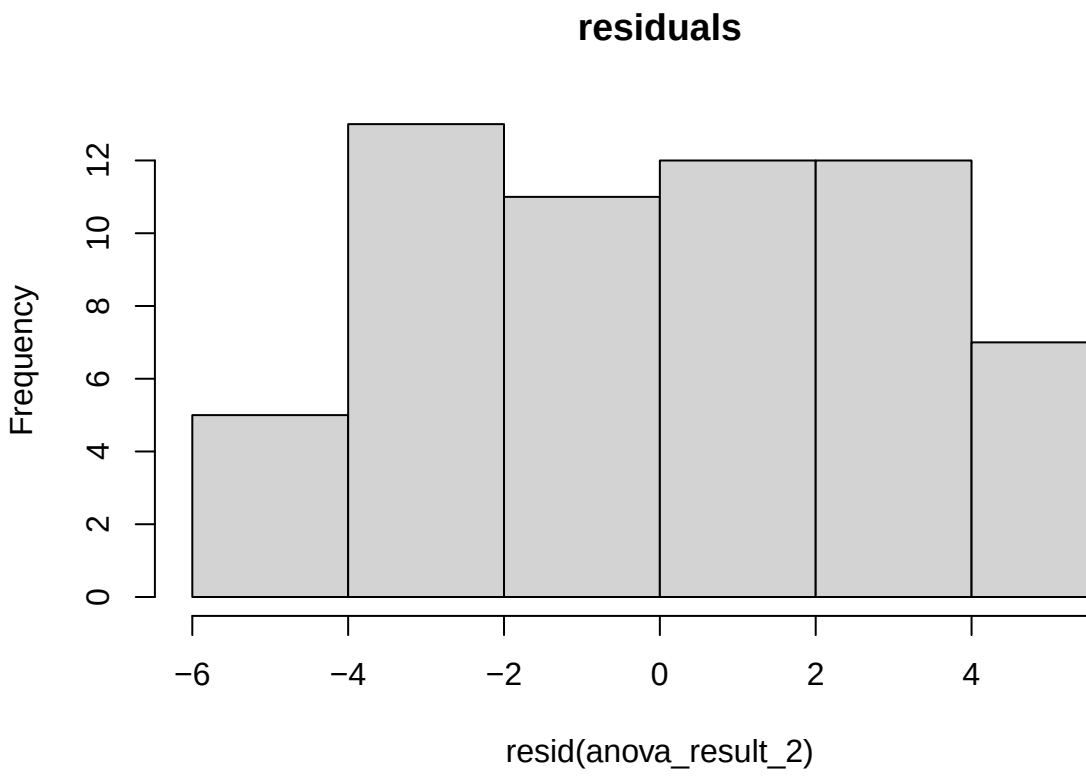
```
anova_result_2 <- aov(value ~ group, data = non_normal_data)
```

Justify your test choice

Assumptions for a 2-way ANOVA test are:

1. **Independence of observations.** • We can assume it at once.

```
hist(resid(anova_result_2), main = "residuals")
```

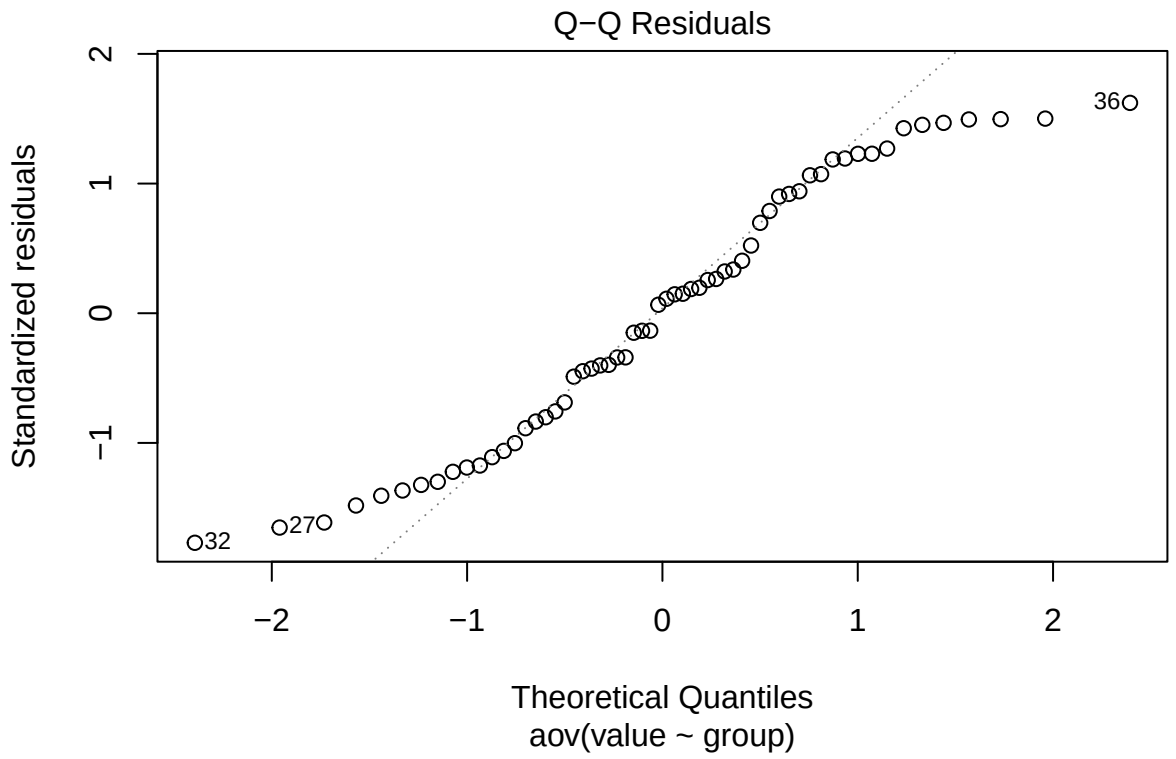


2. **Normality of residuals:**

```r
shapiro.test(resid(anova_result_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(anova_result_2)
## W = 0.94644, p-value = 0.01063
```

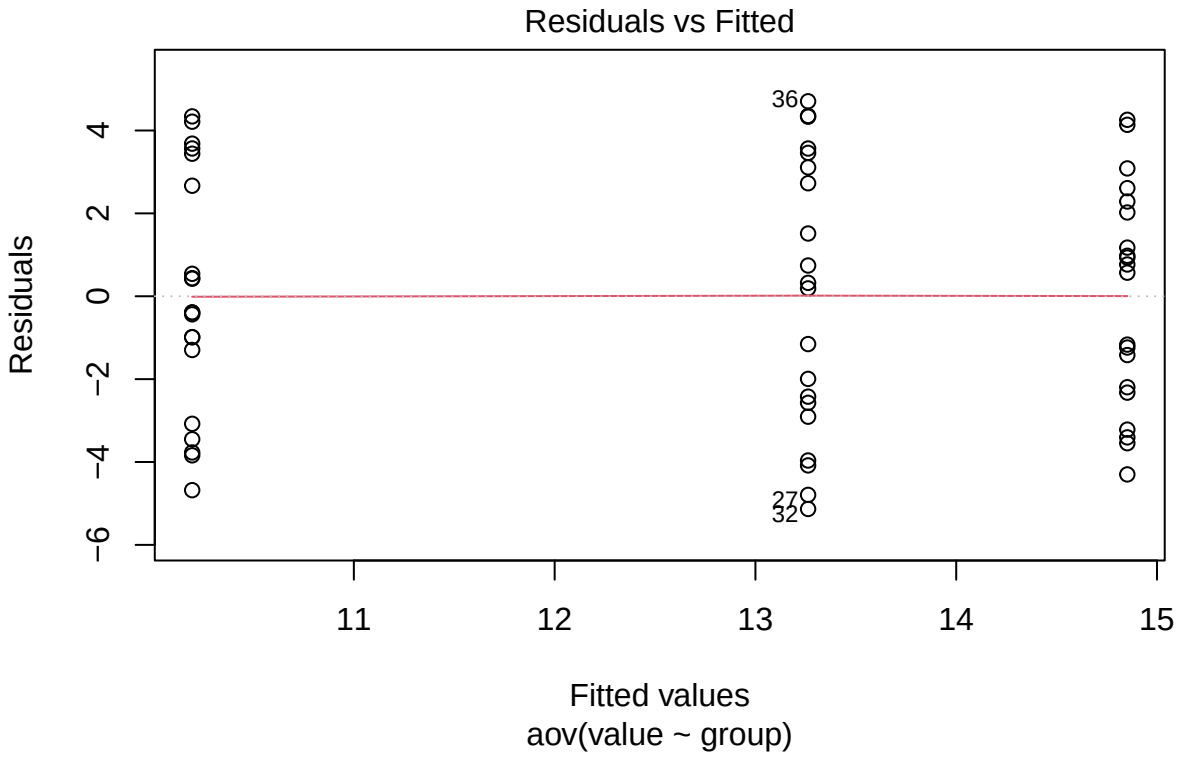
```r
plot(anova_result_2, 2)
```



Based on the histogram, we cannot deny that the data do not obey a normal distribution. However, based on the p-value of the Shapiro test, we can believe that the residuals obey a normal distribution.

3. Equality of variance:

```r
plot(anova_result_2, 1)
```




```
kruskal_result <- kruskal.test(value ~ group, data = non_normal_data)
print(kruskal_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by group
## Kruskal-Wallis chi-squared = 16.781, df = 2, p-value = 0.0002271
```

The dependent variable should be on an ordinal scale, a ratio scale, or an interval scale.
The observations should be independent. In other words, there should be no correlation between members of each group or between members of a group.
All groups should have the same shape distribution.

```
kruskal.test(value ~ group, data = non_normal_data)$posthoc
```

```
## NULL
```