Chapter 9 One-way Analysis of Variance

Angelo LaCommare-Soto

2025-04-05

Toxic Substances and Gene Regulation

Research Question 1: Does triclosan suppress gene expression in hypothalamic cells?

Section 1 - Importing Data

```
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```
library(tidyverse)
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
#commands header=TRUE in order to treat the first row of the data frame as a header and stringsAsFactor
gene <- read.csv("gene.csv",header=TRUE,stringsAsFactors = TRUE)
str(gene)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ replicate: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ treatment: Factor w/ 3 levels "1 micromolar",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ ct       : num  -0.00614 0.02071 -0.01457 -0.3457 -0.07838 ...
```

```
head(gene)
```

```
##   replicate treatment           ct
## 1         1      dmso -0.006138166
## 2         2      dmso  0.020711581
## 3         3      dmso -0.014573415
## 4         4      dmso -0.345704079
## 5         5      dmso -0.078384399
## 6         6      dmso  0.424088478
```

Question Answers

a. The dataset “gene.csv” contains 36 total observations of 3 different variables.
b. Given the question ‘Does triclosan suppress gene expression in hypothalamic cells?’, we are interested in the ‘treatment’ and ‘ct’ variables.
c. Given that the variable ‘treatment’ has three levels, I will compare three treatment groups.

Section 2 - Carrying out a One-way Analysis of Variance

```
# ANOVA
gene_aov <- aov(ct~treatment,data=gene)
gene_aov
```

```
## Call:
##    aov(formula = ct ~ treatment, data = gene)
##
## Terms:
##                 treatment Residuals
## Sum of Squares   10.32313  27.41426
## Deg. of Freedom         2        33
##
## Residual standard error: 0.9114468
## Estimated effects may be unbalanced
```

```
# ANOVA table
summary(gene_aov)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## treatment    2  10.32   5.162   6.213 0.00513 **
## Residuals   33  27.41   0.831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 3 - Estimating the R^2

```
# R squared
gene_R2 <- 10.32/(10.32+27.41)
gene_R2
```

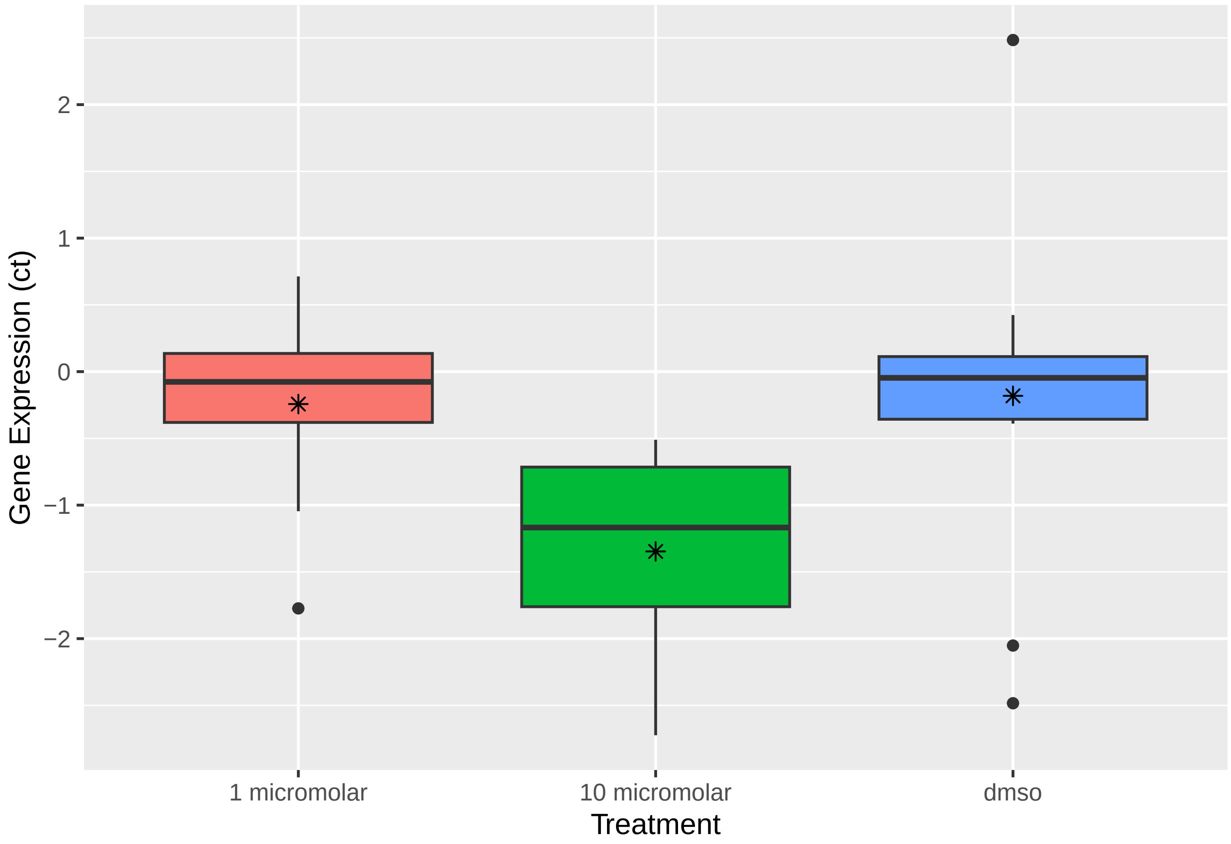
```
## [1] 0.2735224
```

Question Answer

a. About 27.35% of the variation in the data is explained by differences among treatment groups.

Challenge 1 - Plotting the ANOVA Data

```
# creating a boxplot with means
p1 <- ggplot(gene,aes(x=treatment,y=ct, fill=treatment)) +
  geom_boxplot() +
  stat_summary(fun=mean, geom = 'point', shape = 8, size = 2)+
  ylab("Gene Expression (ct)") +
  xlab("Treatment") +
  theme(legend.position='none')
p1
```

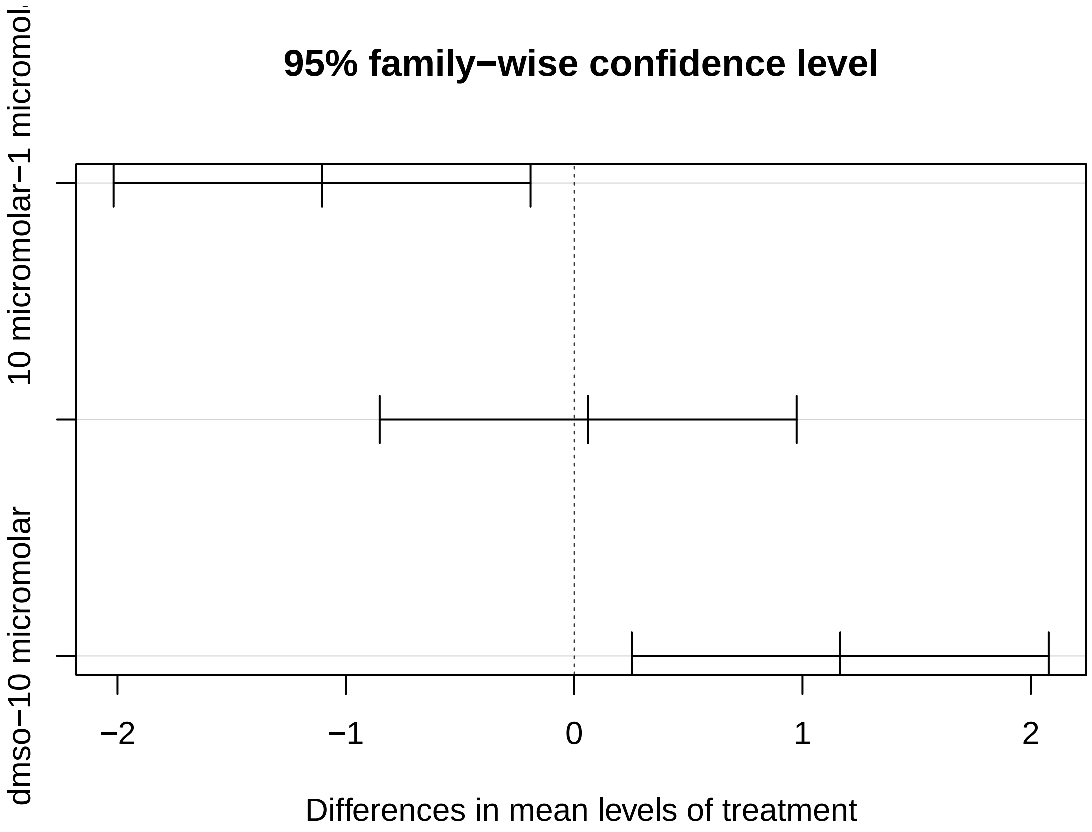


Section 4 - Carrying Out a Multiple Comparison Using the Tukey Test

```
# Tukey test
t <- TukeyHSD(gene_aov)
t
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ct ~ treatment, data = gene)
##
## $treatment
##                                   diff        lwr        upr     p adj
## 10 micromolar-1 micromolar -1.10397225 -2.0170208 -0.1909237 0.0149593
## dmso-1 micromolar           0.06146762 -0.8515810  0.9745162 0.9850753
## dmso-10 micromolar          1.16543986  0.2523913  2.0784884 0.0098789
```

```
# Visualization of the Tukey test
plot(t)
```



Question Answers

a. Not all treatment groups differ from one another. The mean response of the 1 micromolar treatment of triclosan is not very different from the mean response of the negative control treatment. On the other hand, the mean response of the 10 micromolar treatment of triclosan is different from both the mean responses of the negative control 1 micromolar treatments of triclosan.
b. These results indicate that a higher concentration of triclosan is associated with a decrease in the gene expression of GnRH.
c. The assumptions of the ANOVA are 1) independence of observations, 2) normality within groups, and 3) and homoscedasticity (variance equality between groups).

Section 5 - Checking Model Assumptions

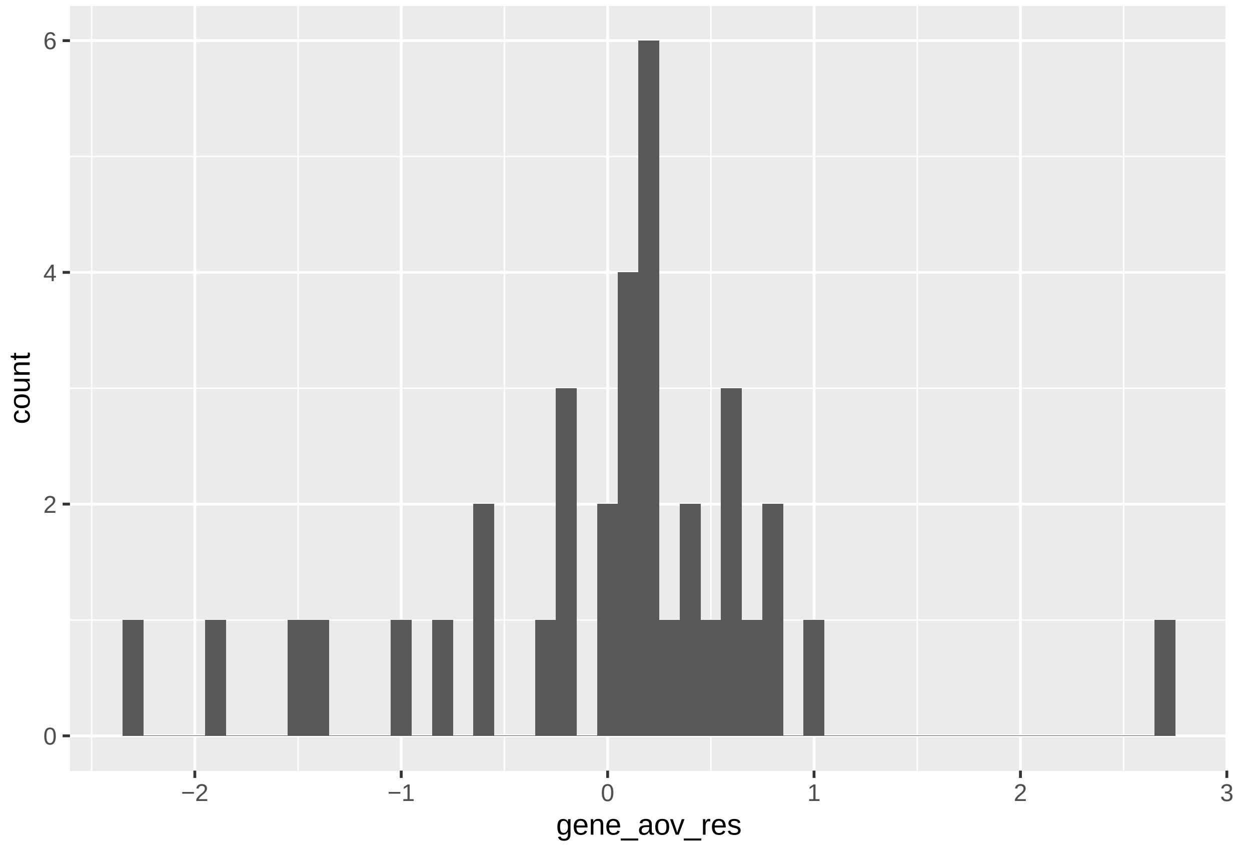
```
# In an ANOVA, the model residuals (the difference between each observation and the mean value) should

# extracting model residuals
gene_aov_res <- residuals(gene_aov)
gene_aov_res
```

```
##           1           2           3           4           5           6
##  0.17524933  0.20209908  0.16681408 -0.16431658  0.10300310  0.60547598
##           7           8           9          10          11          12
## -2.30260794 -1.87061250  2.66538294  0.57072361  0.05673750 -0.20794861
##          13          14          15          16          17          18
##  0.60463764  0.76315452  0.83638717 -0.33702006  0.21571431  0.71305356
##          19          20          21          22          23          24
## -0.64531150  0.23545823  0.14224736 -0.18409028 -0.96721807 -1.37701287
##          25          26          27          28          29          30
##  0.48183128  0.37349006  0.14824458  0.34210210  0.01497082  0.18426327
##          31          32          33          34          35          36
## -0.80274488 -0.59566112 -1.53070541  0.03198867  0.39604621  0.95617442
```

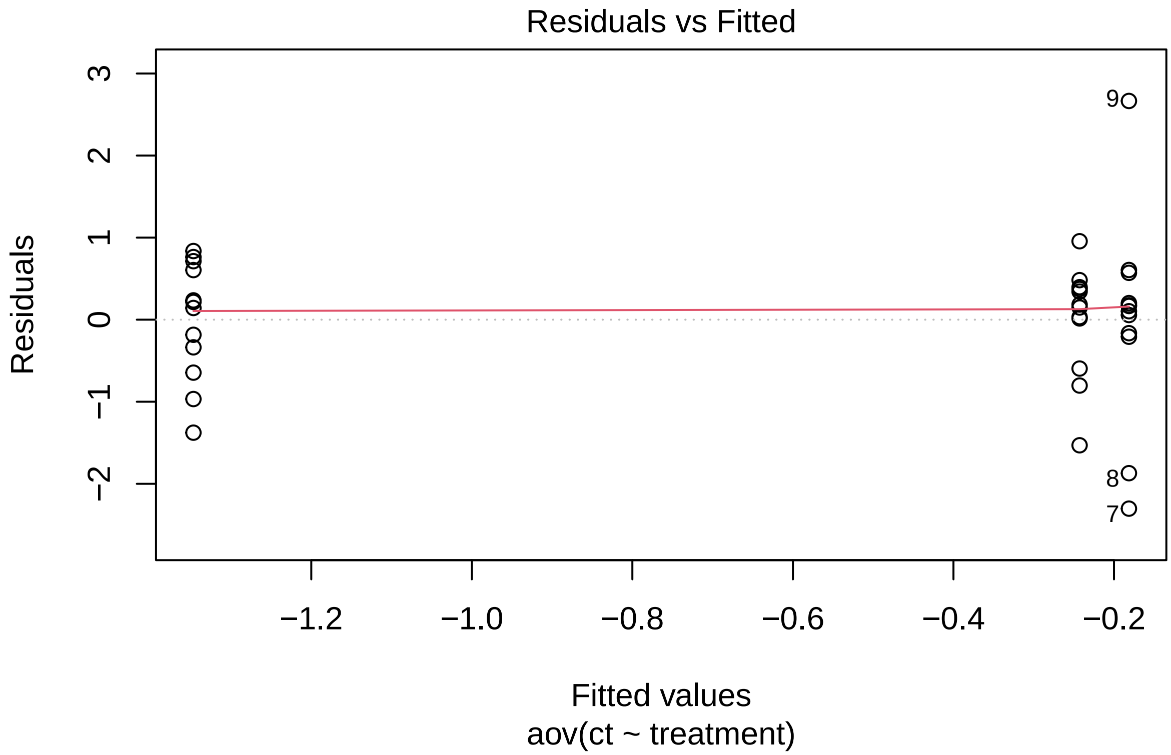
```
# converting aov_res into a dataframe for ggplot
gene_aov_res_df <- as.data.frame(gene_aov_res)
gene_aov_res_df
```

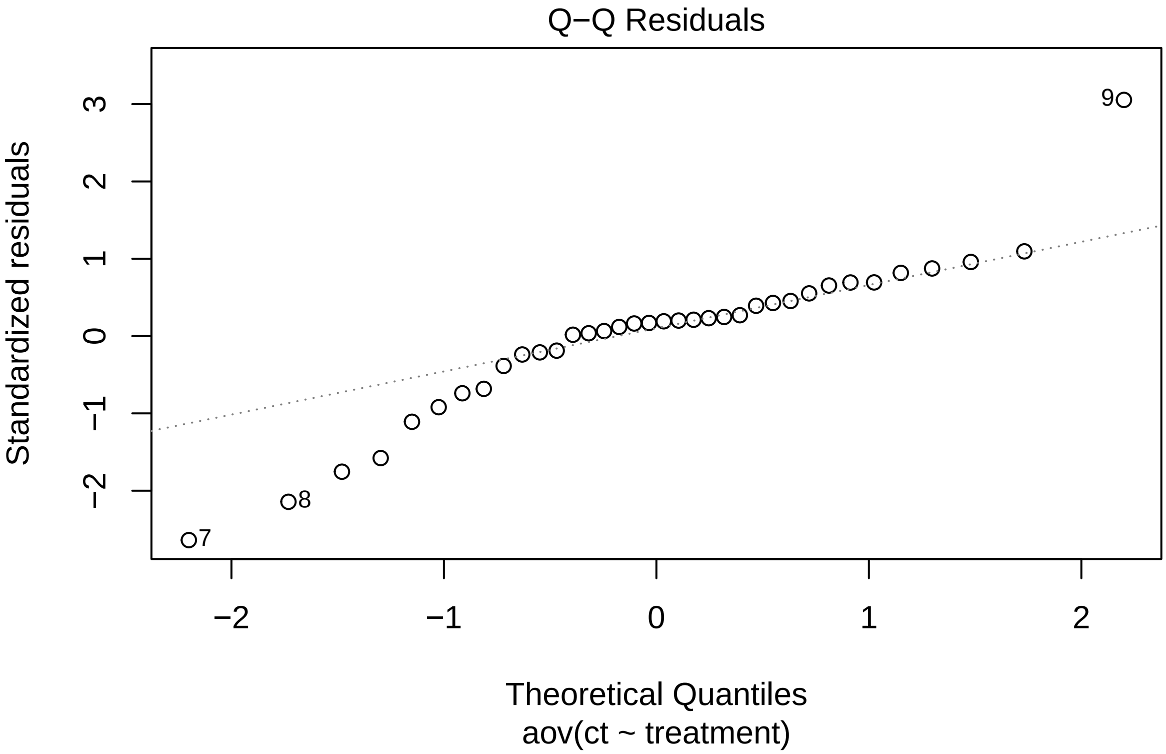
```
##    gene_aov_res
## 1    0.17524933
## 2    0.20209908
## 3    0.16681408
## 4   -0.16431658
## 5    0.10300310
## 6    0.60547598
## 7   -2.30260794
## 8   -1.87061250
## 9    2.66538294
## 10   0.57072361
## 11   0.05673750
## 12  -0.20794861
## 13   0.60463764
## 14   0.76315452
## 15   0.83638717
## 16  -0.33702006
## 17   0.21571431
## 18   0.71305356
## 19  -0.64531150
## 20   0.23545823
## 21   0.14224736
## 22  -0.18409028
## 23  -0.96721807
## 24  -1.37701287
## 25   0.48183128
## 26   0.37349006
## 27   0.14824458
## 28   0.34210210
## 29   0.01497082
## 30   0.18426327
## 31  -0.80274488
## 32  -0.59566112
## 33  -1.53070541
## 34   0.03198867
## 35   0.39604621
## 36   0.95617442
```

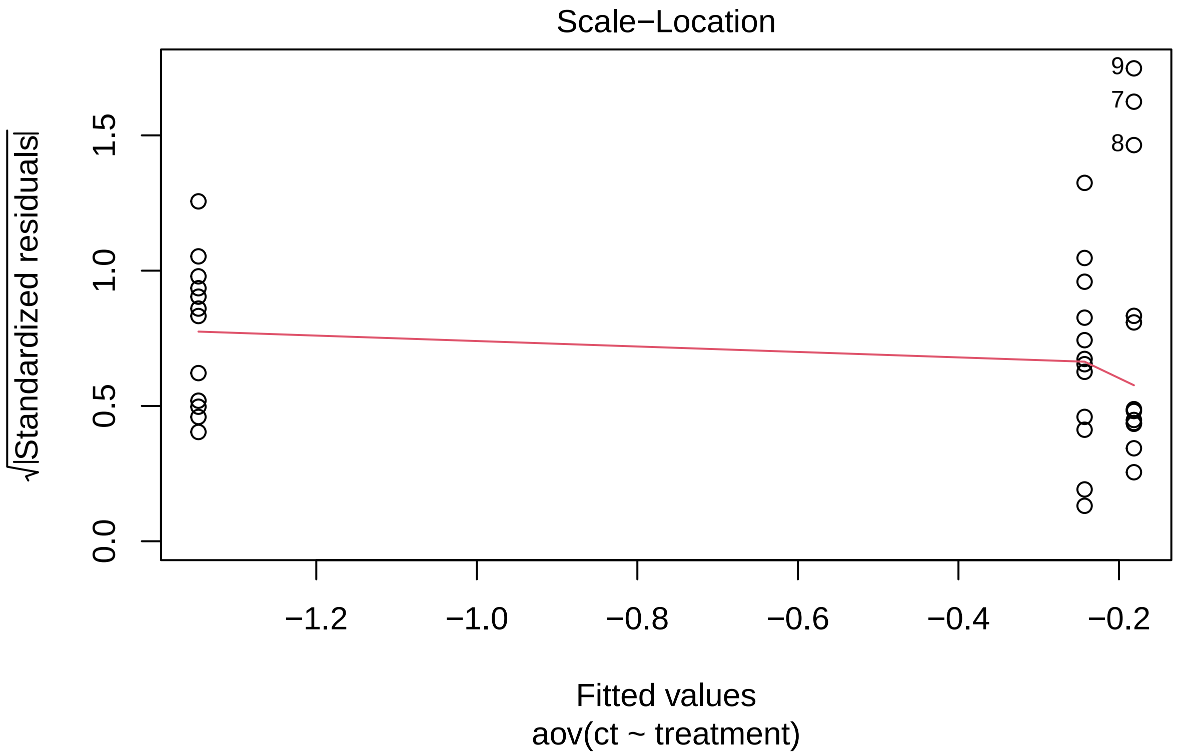
```
# histogram of model residuals
p2 <- ggplot(gene_aov_res_df, aes(x = gene_aov_res)) +
  geom_histogram(binwidth = 0.1)
p2
```

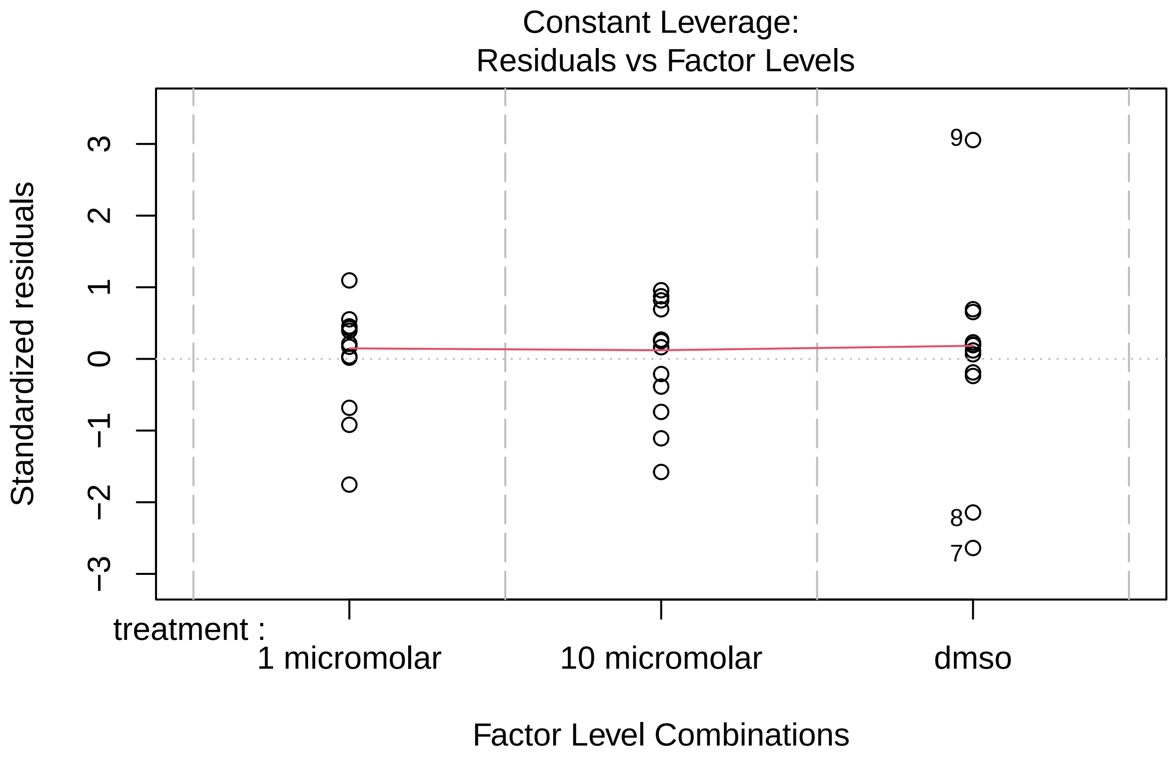


```
# The aov() function (done originally in the 3rd chunk) prepares the data for model checking plots: a p
# model checking plots
plot(gene_aov)
```


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
9
8
7
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
Fitted values
aov(ct ~ treatment)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−2
−1
0
1
2
9
8
7
Theoretical Quantiles
aov(ct ~ treatment)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
Fitted values
aov(ct ~ treatment)
9
7
8



```
# Another way we can test for normality in our data is by employing a Shapiro test. In this test, the n

shapiro.test(gene_aov_res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gene_aov_res
## W = 0.91122, p-value = 0.007014
```

Question Answer

a. Given the W statistic of 0.911 and its associated p-value of 0.007, there is sufficient evidence to reject the null hypothesis of the Schapiro test, which states that our model residuals comes from a normal distribution. Therefore, it is unlikely that the data is normally distributed.

Section 6 - Testing for Equal Variances

```
# Finally, we can use the Barlett's test to test for equal variances. For this, we will use the functio

bartlett.test(ct~treatment, data=gene)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ct by treatment
## Bartlett's K-squared = 5.0748, df = 2, p-value = 0.07907
```

```
# If the test results in a p-value > 0.05, then the data has equal variance (no difference in variance
```

```
# Hypothetically, if the variances are significantly different across factors, then the we have to tran

# transforming the data using the natural log
# transformed <- log(response_variable)
```

Discussion Question Answers

a. The estimated F statistic of 6.213 indicates that the variation among triclosan treatment group means was about 6 times larger than that of the variation within triclosan treatment groups.
b. The estimated coefficient of determination, R^2, of 0.2735 signifies that about 27.35% of the variation in the data is explained by differences among triclosan treatment groups, according to our ANOVA statistical model.
c. We test for model assumptions to ensure that our models are good fits for our data when attempting to derive statistical significance. On the other hand, model assumption testing can be good practice for fine tuning experimental set-ups to result in data that can be interpreted by statistical models.