STAT292
Isabelle (Izzy) Southon
300597453
Assignment 3

Question 1

a) Explain why this is a random effects design, rather than a fixed effects design.

This is a random effects design because we assume there is some treatment effect being the four personality types, and we ask how much of the variability in the response Y being subject's scores is due to the treatment of different personality types as opposed to residual unexplained variability.

- In psychology an experiment runs that observes a highly variable score between people with different personality types.
- How much of the variability is accounted for by varying the type of the subject's personality, as against the unexplained residual variability from other causes.
- Subject's scores are investigated with 4 different types of personality in which are selected at random.

This is a random effects design rather than a fixed effects design because it is not only the treatments (personality types) included that are of interest, as we are interested in the subject's scores on the certain test as well.

b) Write up the results of the analysis

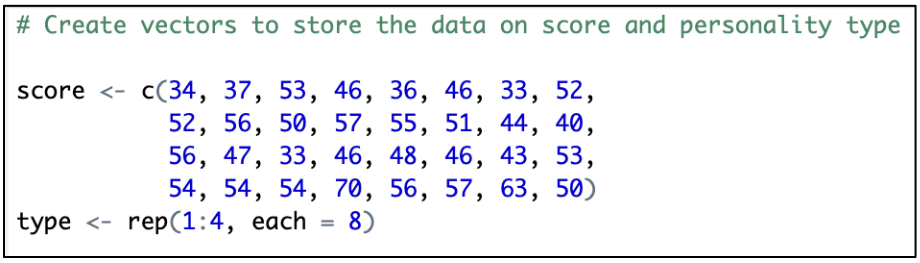
Model equation:
$Y_{ij} = \mu + A_i + E_{ij}$

Y_{ij} = random variable for j^{th} score, personality type i
$i = 1 – 4$
j = each individual observation (score) in each group (personality type)
$j = 1 – 8$
μ = overall mean
A_i = the random effect of personality type i
A_i distribution of random effects personality type i = $A_i \sim N(0, \sigma^2_A)$
E_{ij} = error terms for j at i
$E_{ij} \sim N(0, \sigma^2)$

Assumptions:

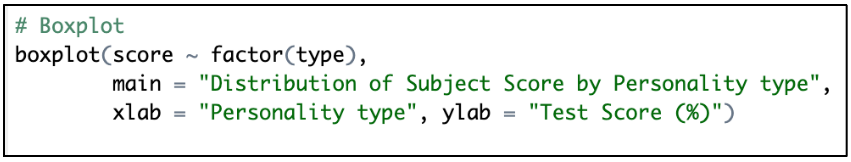
- The errors E come independently from a $N(0, \sigma^2)$ distribution
- The random effects A come independently from a $N(0, \sigma^2_A)$ distribution
- The random effects are independent of the errors

To begin the analysis, the data must be stored in vectors.

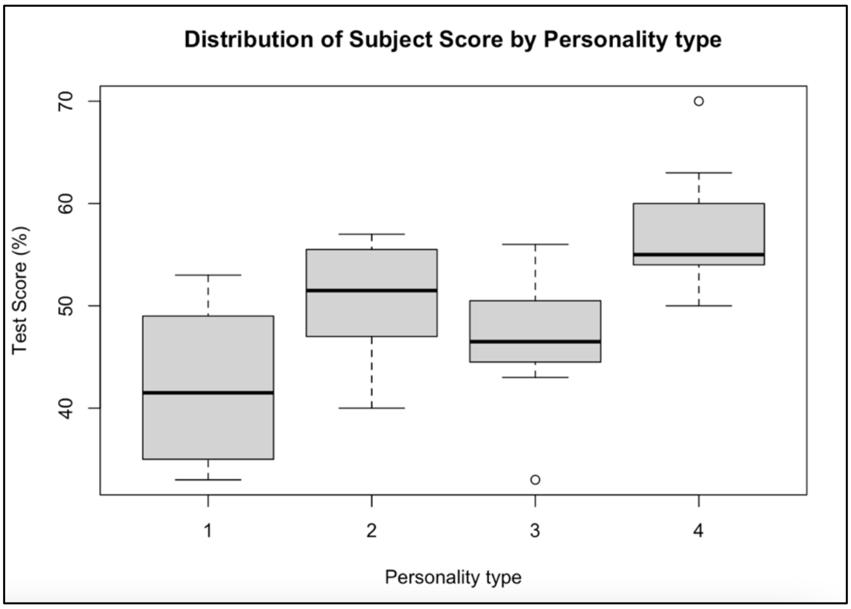


```
# Create vectors to store the data on score and personality type

score <- c(34, 37, 53, 46, 36, 46, 33, 52,
           52, 56, 50, 57, 55, 51, 44, 40,
           56, 47, 33, 46, 48, 46, 43, 53,
           54, 54, 54, 70, 56, 57, 63, 50)
type <- rep(1:4, each = 8)
```

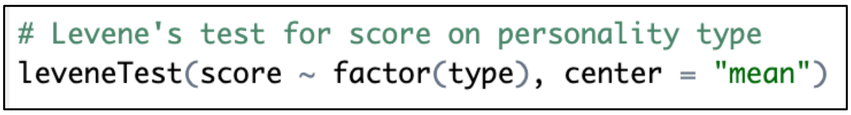
Boxplot



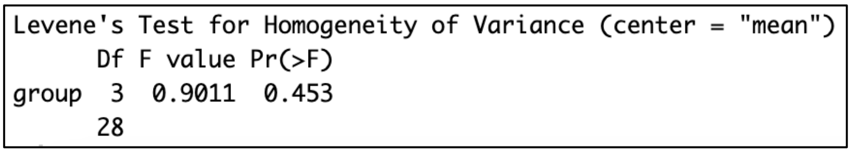
```
# Boxplot
boxplot(score ~ factor(type),
        main = "Distribution of Subject Score by Personality type",
        xlab = "Personality type", ylab = "Test Score (%)")
```



Levene’s test

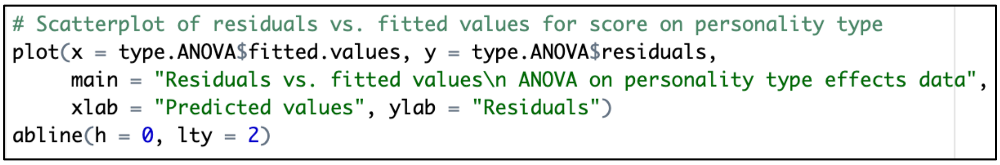


```
# Levene's test for score on personality type
leveneTest(score ~ factor(type), center = "mean")
```

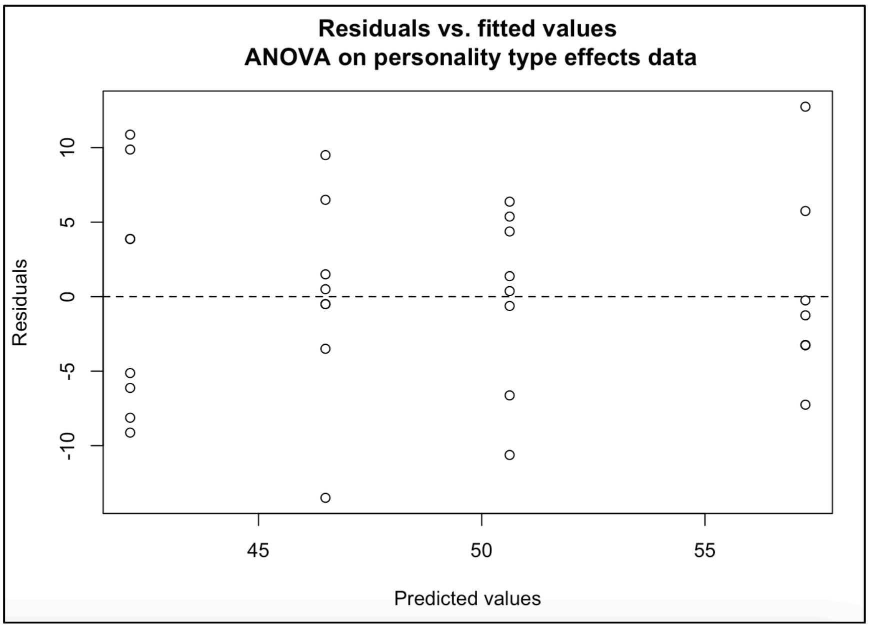


```
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  3  0.9011  0.453
      28
```

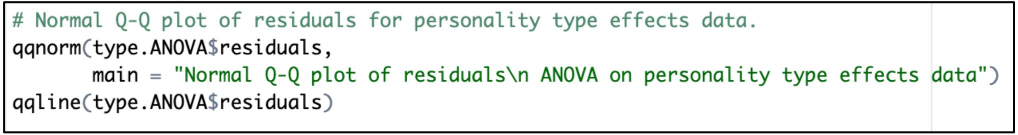
Scatterplot of residuals vs. fitted values



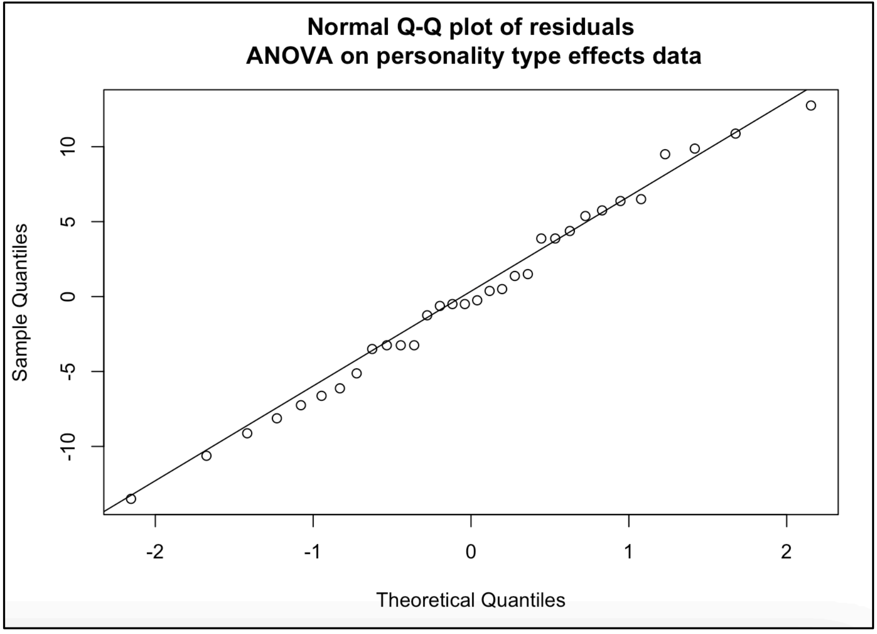
```
# Scatterplot of residuals vs. fitted values for score on personality type
plot(x = type.ANOVA$fitted.values, y = type.ANOVA$residuals,
     main = "Residuals vs. fitted values\n ANOVA on personality type effects data",
     xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```



Normal Q-Q plot



```
# Normal Q-Q plot of residuals for personality type effects data.
qqnorm(type.ANOVA$residuals,
       main = "Normal Q-Q plot of residuals\n ANOVA on personality type effects data")
qqline(type.ANOVA$residuals)
```



The boxplots indicate normality and constant variance. The general spread on all boxplots is relatively the same. Personality types 1 and 2 seem to have a higher interquartile range. The medians vary for each personality type.

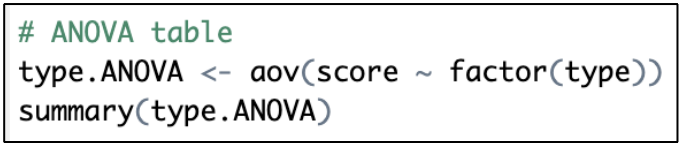
The levene's test is measuring equal variance and is confirmed for these data as the *p*-value of 0.453 is more than the significance level of 0.05. This means that we do not reject the null hypothesis of equal variance.

The residuals vs. fitted values scatterplot shows a relatively even spread along the zero line. The scatterplot shows equal variance. There doesn't look like there's any funnelling.

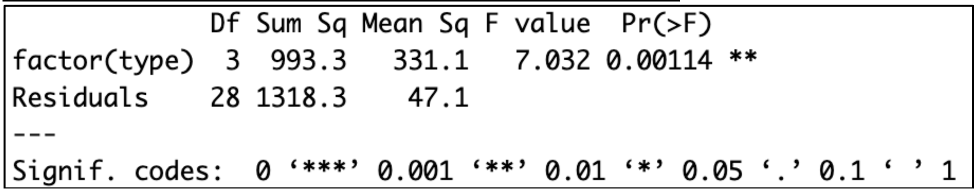
The normal Q-Q plot shows normality of the data. There is slight deviation from the Q-Q line, yet most data are relatively close to the line so normality can be confirmed.

All assumptions seem satisfied so this ANOVA will be valid.

ANOVA table:

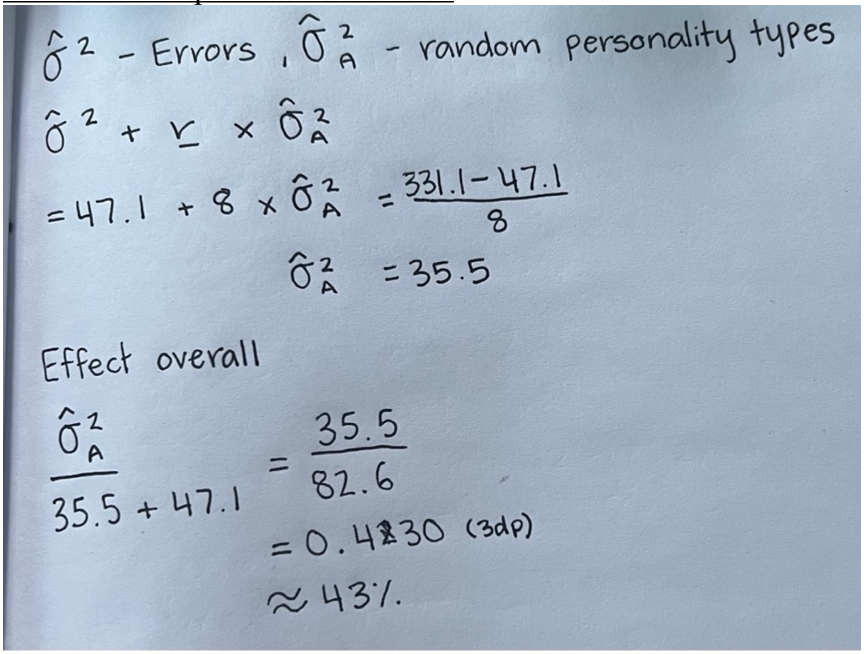


```
# ANOVA table
type.ANOVA <- aov(score ~ factor(type))
summary(type.ANOVA)
```



```
             Df Sum Sq Mean Sq F value  Pr(>F)   
factor(type)  3  993.3   331.1   7.032 0.00114 **
Residuals    28 1318.3    47.1                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Estimated components of variance:



$\hat{\sigma}^2$ - Errors , $\hat{\sigma}^2_A$ - random personality types

$\hat{\sigma}^2 + \underline{r} \times \hat{\sigma}^2_A$

$= 47.1 + 8 \times \hat{\sigma}^2_A = \frac{331.1 - 47.1}{8}$

$\hat{\sigma}^2_A = 35.5$

Effect overall

$$\frac{\hat{\sigma}^2_A}{35.5 + 47.1} = \frac{35.5}{82.6}$$

$= 0.4\cancel{2}30$ (3dp)

$\approx 43\%$

The percentage of the total variance in the test scores due to personality is 43%. The total variance that is unexplained is (100 – 43%) is 57%.

Interpretation:

We measured the variation of random personality types over the total variation and got 43%. In the model 43% (A_i) is due to personality type variation, and 57% (E_{ij}) is due to unexplained errors. This means that 43% of variation in the test scores is explained by random personality types in this model. The unexplained variation of 57% could be due to many other factors such as the time of the day subject's took the test or the material in the test, but we don't know for sure what these factors are. Therefore, in psychology, this may be helpful as they know that 43% of variation in test scores is related to personality types.

STAT292
Isabelle (Izzy) Southon
300597453
Assignment 3

Question 2

a) Specify what kind of design this is and give the relevant model equation, including an interaction term

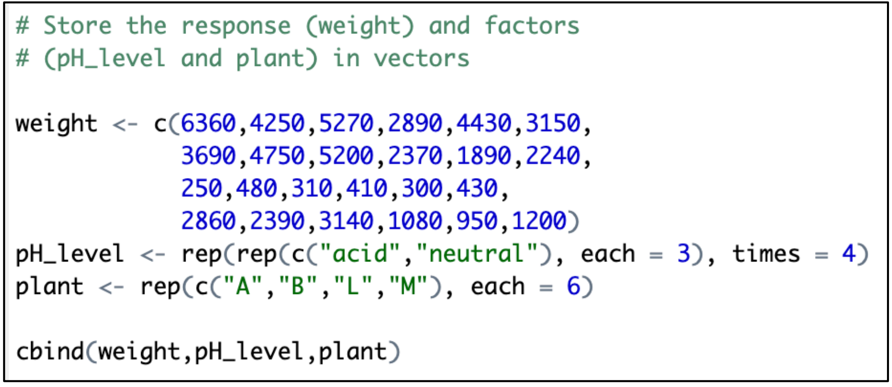
This is a balanced design two-way ANOVA as there are two factors. It is a balanced design because there are an equal number of observations in each cell. Factor A is plant name (four levels), and Factor B is pH level (two levels). This makes it a 4 x 2 design. The response variable Y is the plants uptake of zinc measured by weight (ppm).

The model equation for a two-way ANOVA is: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$

The interaction term is $(\alpha\beta)_{ij}$ which means if all its parameters are zero for each i (dry plant weight in ppm) for each j (pH level), there is no interaction between factors A (plant name) and B (pH level).

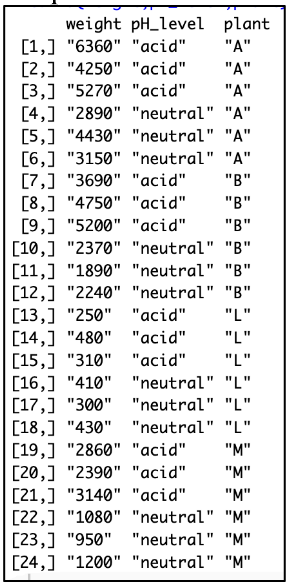
b) Analyse the data in R using the model from part (a), but try both the untransformed response variable and the log-transformed response variable. Choose one of these models for the presentation of your results, explaining the reason for your choice

Putting the data into R:



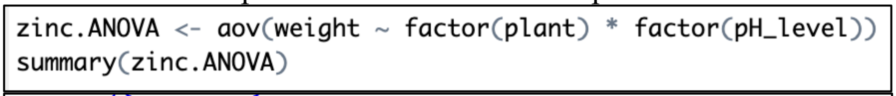
```
# Store the response (weight) and factors
# (pH_level and plant) in vectors

weight <- c(6360,4250,5270,2890,4430,3150,
            3690,4750,5200,2370,1890,2240,
            250,480,310,410,300,430,
            2860,2390,3140,1080,950,1200)
pH_level <- rep(rep(c("acid","neutral"), each = 3), times = 4)
plant <- rep(c("A","B","L","M"), each = 6)

cbind(weight,pH_level,plant)
```

Output:

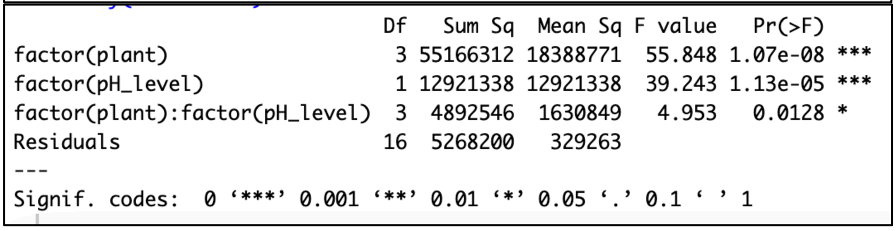


```
      weight pH_level  plant
 [1,] "6360" "acid"    "A"
 [2,] "4250" "acid"    "A"
 [3,] "5270" "acid"    "A"
 [4,] "2890" "neutral" "A"
 [5,] "4430" "neutral" "A"
 [6,] "3150" "neutral" "A"
 [7,] "3690" "acid"    "B"
 [8,] "4750" "acid"    "B"
 [9,] "5200" "acid"    "B"
[10,] "2370" "neutral" "B"
[11,] "1890" "neutral" "B"
[12,] "2240" "neutral" "B"
[13,] "250"  "acid"    "L"
[14,] "480"  "acid"    "L"
[15,] "310"  "acid"    "L"
[16,] "410"  "neutral" "L"
[17,] "300"  "neutral" "L"
[18,] "430"  "neutral" "L"
[19,] "2860" "acid"    "M"
[20,] "2390" "acid"    "M"
[21,] "3140" "acid"    "M"
[22,] "1080" "neutral" "M"
[23,] "950"  "neutral" "M"
[24,] "1200" "neutral" "M"
```

Untransformed response variable ANOVA output:

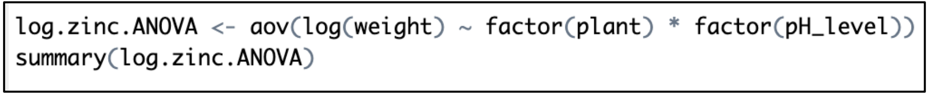


```
zinc.ANOVA <- aov(weight ~ factor(plant) * factor(pH_level))
summary(zinc.ANOVA)
```

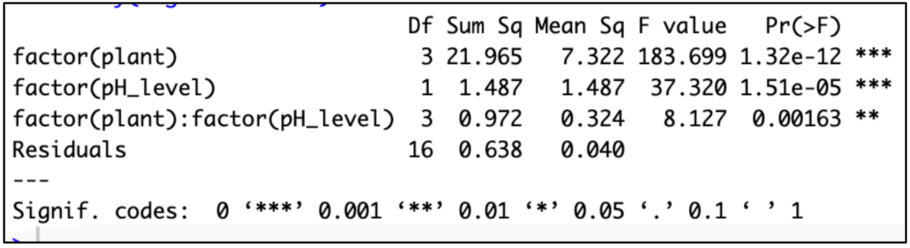


```
                              Df   Sum Sq  Mean Sq F value   Pr(>F)    
factor(plant)                  3 55166312 18388771  55.848 1.07e-08 ***
factor(pH_level)               1 12921338 12921338  39.243 1.13e-05 ***
factor(plant):factor(pH_level) 3  4892546  1630849   4.953   0.0128 *  
Residuals                     16  5268200   329263                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Log-transformed response variable ANOVA output:

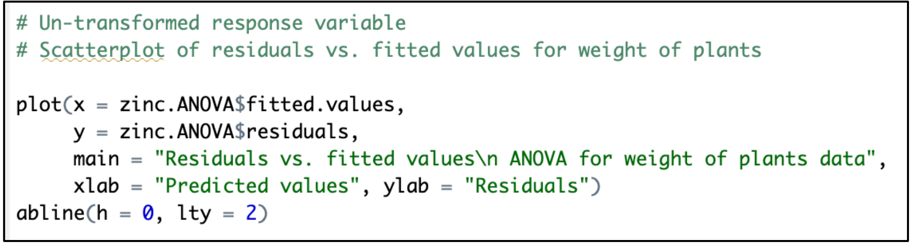


```
log.zinc.ANOVA <- aov(log(weight) ~ factor(plant) * factor(pH_level))
summary(log.zinc.ANOVA)
```

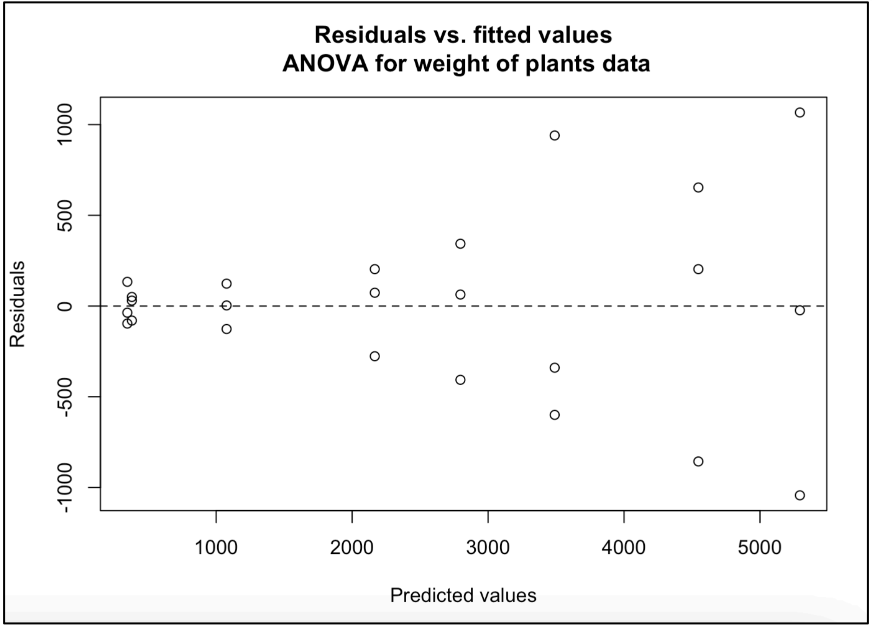


```
                               Df Sum Sq Mean Sq F value   Pr(>F)    
factor(plant)                   3 21.965   7.322 183.699 1.32e-12 ***
factor(pH_level)                1  1.487   1.487  37.320 1.51e-05 ***
factor(plant):factor(pH_level)  3  0.972   0.324   8.127  0.00163 ** 
Residuals                      16  0.638   0.040                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

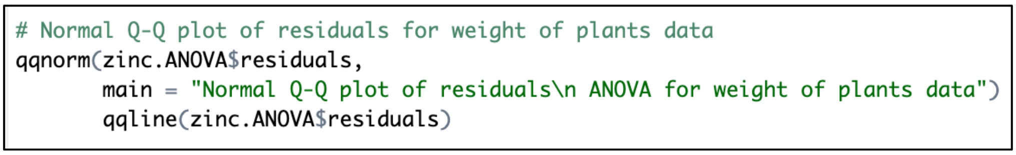
Stabilising variances



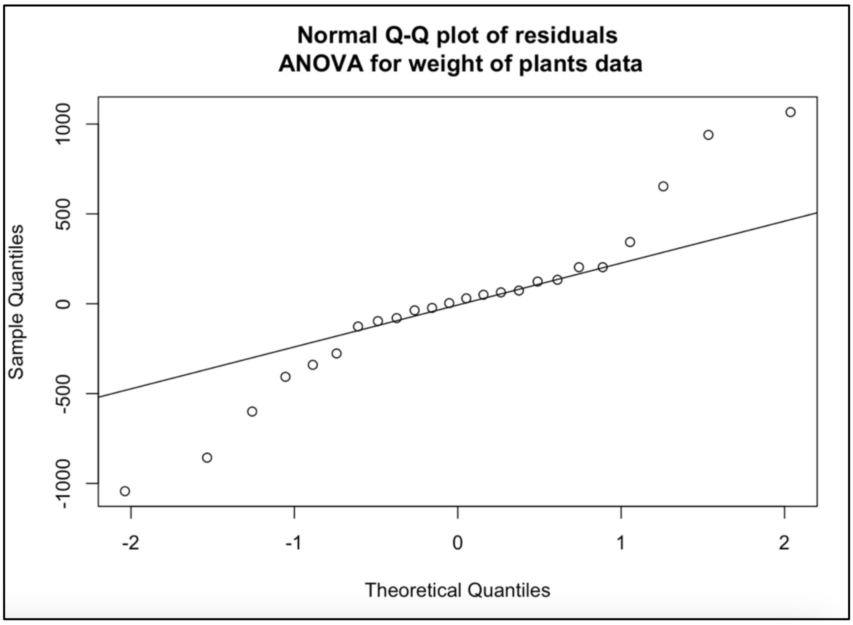
```
# Un-transformed response variable
# Scatterplot of residuals vs. fitted values for weight of plants

plot(x = zinc.ANOVA$fitted.values,
     y = zinc.ANOVA$residuals,
     main = "Residuals vs. fitted values\n ANOVA for weight of plants data",
     xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```



There is quite a bit of funnelling seen on the scatterplot above which is using the untransformed response variable of dry plant weight measuring plants uptake of zinc measured in parts per million (ppm). The plot of residuals vs predicted values shows this clear funnel shape, indicating that the groups with higher data means also have higher variances.

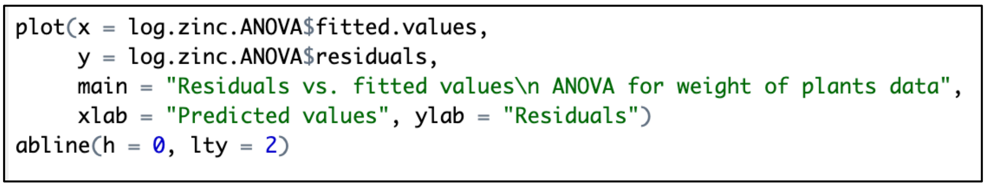


```
# Normal Q-Q plot of residuals for weight of plants data
qqnorm(zinc.ANOVA$residuals,
       main = "Normal Q-Q plot of residuals\n ANOVA for weight of plants data")
       qqline(zinc.ANOVA$residuals)
```

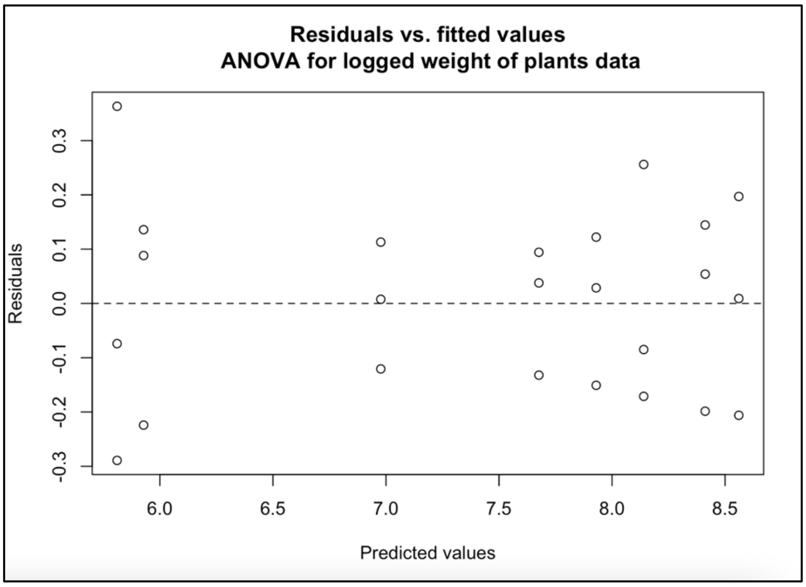


Normality of residuals does not look satisfied as low and high theoretical quantiles are far from the Q-Q line.

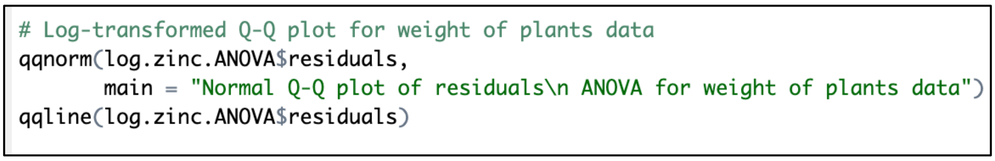
Because of this, the log-transformed response variable will be checked, to see if funnelling reduces and if normality increases. Using the raw data for this two way ANOVA would produce an inaccurate *p*-value because of unequal variance and having normality. This would mean that the null hypothesis would fail to be rejected when it is false. For example, at a 1% significance level for the raw data, the null hypothesis would fail to be rejected as *(p* = 0.0128 > 0.01), whereas for the log-transformed data where the assumptions are satisfied (*p* = 0.00163 < 0.01) the null hypothesis would be rejected and a significant interaction would be identified.



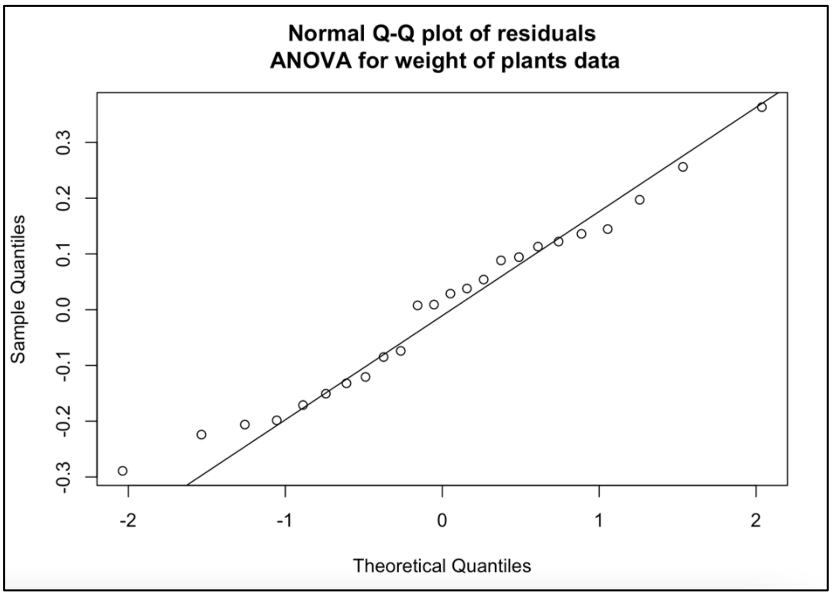
```
plot(x = log.zinc.ANOVA$fitted.values,
     y = log.zinc.ANOVA$residuals,
     main = "Residuals vs. fitted values\n ANOVA for weight of plants data",
     xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```



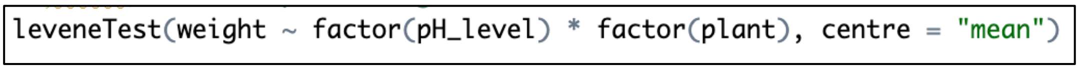
Funnelling is reduced on the log-transformed response variable and the assumption of constant variance is better satisfied.



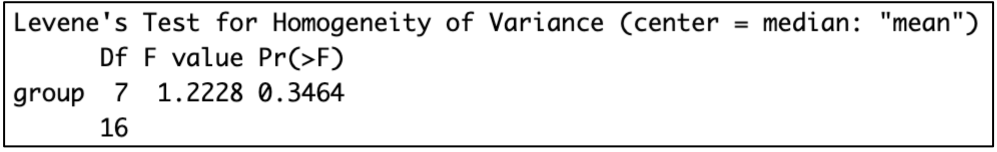
```
# Log-transformed Q-Q plot for weight of plants data
qqnorm(log.zinc.ANOVA$residuals,
       main = "Normal Q-Q plot of residuals\n ANOVA for weight of plants data")
qqline(log.zinc.ANOVA$residuals)
```



Normality is better satisfied with the log-transformed response variable being plant weight. For the report, logged data will be used as it satisfies constant variance and normality of residuals better than the raw data does.

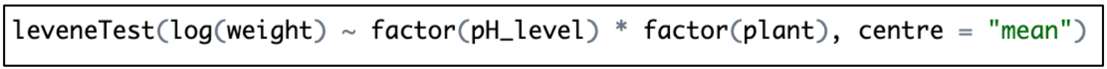


```
leveneTest(weight ~ factor(pH_level) * factor(plant), centre = "mean")
```

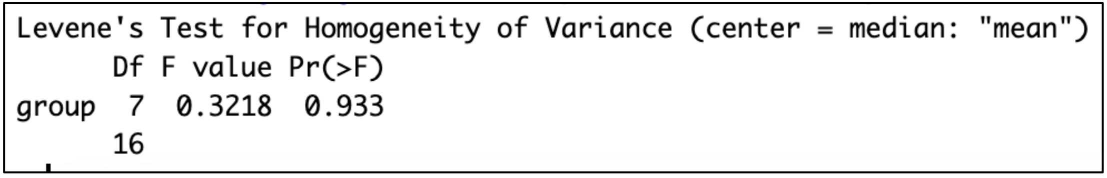


```
Levene's Test for Homogeneity of Variance (center = median: "mean")
      Df F value Pr(>F)
group  7  1.2228 0.3464
      16
```

The levene’s test for the raw data (the untransformed variable Y) produces a value of 0.3464, which is larger than the 0.05 significance level.



```
leveneTest(log(weight) ~ factor(pH_level) * factor(plant), centre = "mean")
```



```
Levene's Test for Homogeneity of Variance (center = median: "mean")
      Df F value Pr(>F)
group  7  0.3218  0.933
      16
```

For the log-transformed response variable Y, the Levene’s test produces a value of 0.933, which is much larger than the 0.05 significance level. This confirms that using the log-transformed data for variable Y (weight) is more appropriate to satisfy assumptions.

c) **Present the report in the usual way, using a 5% significance level. Whether or not you found a significant interaction, include an interaction graph in your report (using your choice of raw or logged data), and refer to it in the interpretation section, to help illustrate your results.**

Model equation:
$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$

Y = The response variable of plants uptake of zinc, measured in parts per million (ppm) of dry plant weight.
i = pH level (acid or neutral)
j = plant name
k = is the observations within i & j groups
Y_{ijk} = response variable Y (weight) at level i (pH level)
μ = overall mean
α_i = the random effect of plant i
β_j = the random effect of pH level j
$(\alpha\beta)_{ij}$ = the random effects of plant i and pH level j multiplied
E_{ijk} = error for j (pH level) at i (plant)

Assumptions:
Using the log transformed response variable Y (plants uptake of zinc, measured in parts per million (ppm) of dry plant weight), all assumptions seem satisfied. Normality of residuals was satisfied in the Q-Q plot, and constant variance was confirmed through the Levene’s test and plot of residuals vs. fitted values. Independence is assumed as we have to assume that the plants were randomly selected for this test. We are unable to check for independence as it depends on how the experiment was run. For example, it depends on if the plants were randomly selected, if the soils pH levels were actually what they were said to be (5.5 and 7), and that the plants uptake of zinc didn’t impact one another. It also depends on how accurately the dry plant weights were collected and whether it was at the same time. We also assume that the error term E_{ijk} follows a normal distribution of $E_{ijk} \sim N(0, \sigma^2)$

STAT292
Isabelle (Izzy) Southon
300597453
Assignment 3

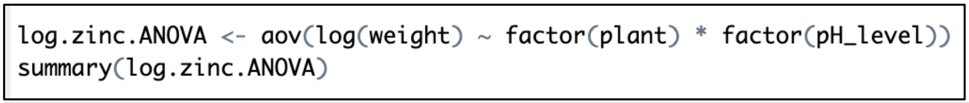
Hypotheses:

Interaction test

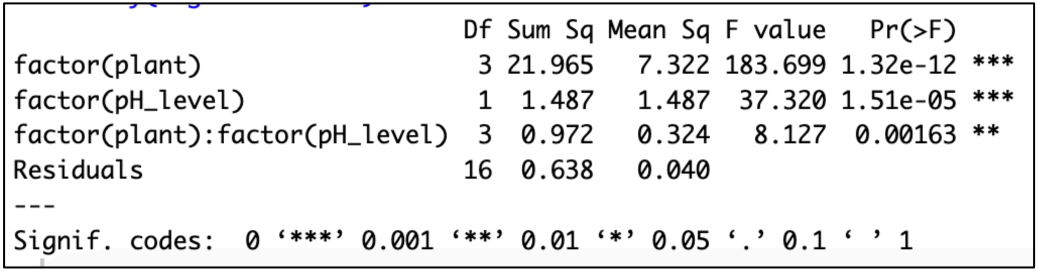
H_0: There is no interaction between plant and pH level, all $(\alpha\beta)_{ij} = 0$
The random effect of plant i and the random effect of pH level j are equal to zero

H_1: There is some interaction between plant and pH level, at least $(\alpha\beta)_{ij} \neq 0$
The random effect of plant i and the random effect of pH level j are not equal to zero

ANOVA table:



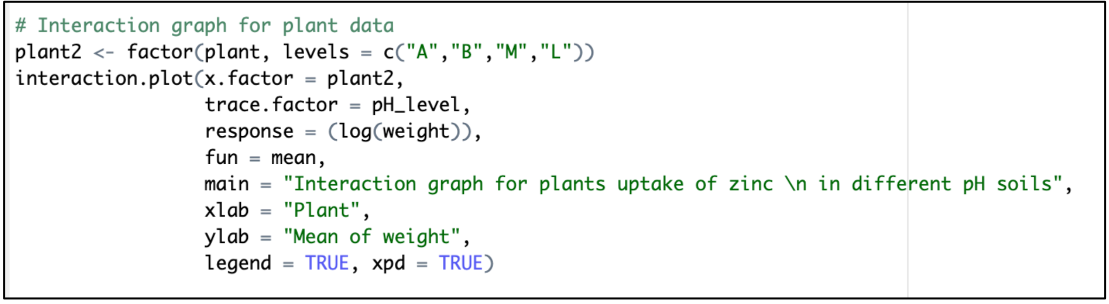
```
log.zinc.ANOVA <- aov(log(weight) ~ factor(plant) * factor(pH_level))
summary(log.zinc.ANOVA)
```



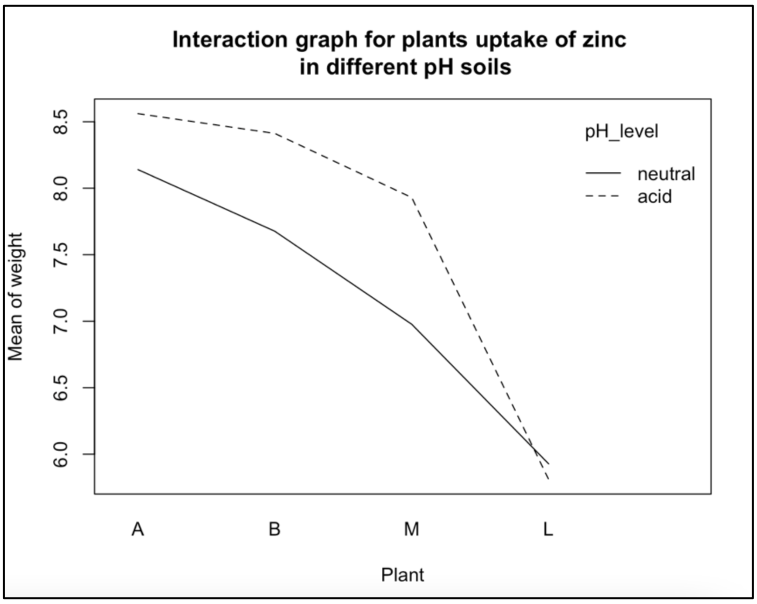
```
                              Df Sum Sq Mean Sq F value   Pr(>F)    
factor(plant)                  3 21.965   7.322 183.699 1.32e-12 ***
factor(pH_level)               1  1.487   1.487  37.320 1.51e-05 ***
factor(plant):factor(pH_level) 3  0.972   0.324   8.127  0.00163 ** 
Residuals                     16  0.638   0.040                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Statistical conclusion:

Since ($p = 0.00163 < 0.05$), we conclude that there is a significant interaction between plant and pH level at both the 1% and 5% significance level. We reject H_0 at the 5% significant level. We continue with simple effects. This is the complete model of two-way ANOVA and includes the interaction term that $(\alpha\beta)_{ij} \neq 0$.



```
# Interaction graph for plant data
plant2 <- factor(plant, levels = c("A","B","M","L"))
interaction.plot(x.factor = plant2,
                 trace.factor = pH_level,
                 response = (log(weight)),
                 fun = mean,
                 main = "Interaction graph for plants uptake of zinc \n in different pH soils",
                 xlab = "Plant",
                 ylab = "Mean of weight",
                 legend = TRUE, xpd = TRUE)
```



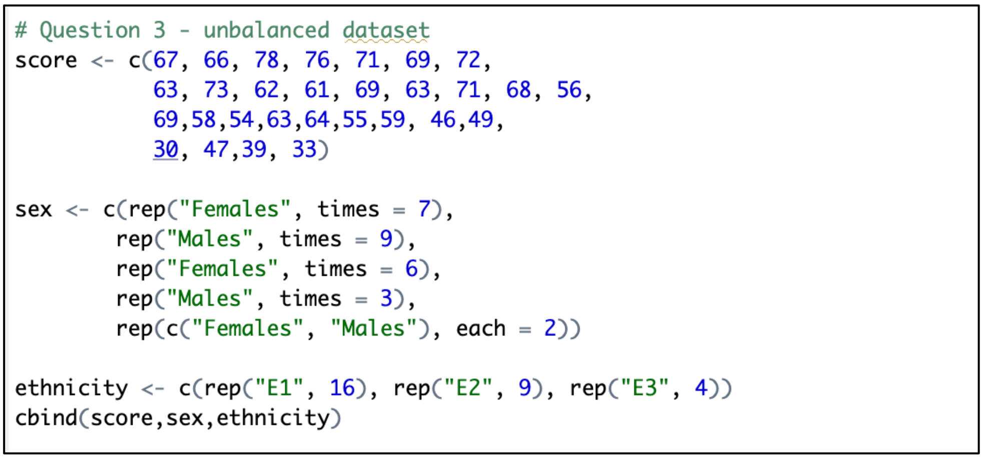
A simple effect is the effect of one factor on Y, when the other factor is fixed at certain levels. The non-parallel lines indicate an interaction.

Interpretation (simple effects):
The factors of plant and pH level interact in their effect on the plants uptake in zinc measured in dry plant weight. We cannot discuss the effect of plant on weight without specifying the pH level, and we cannot discuss the effect of the pH level on weight without specifying the plant. Since there is an interaction, we couldn't interpret the main effects terms in the ANOVA table, instead the interaction graph was produced. The same findings are evident in the interaction graph as the lines (showing pH level) are not parallel to one another, showing the interaction between plant and pH level effecting plants zinc measured in dry plant weight.

Question 3

a) Do a two-way ANOVA on the data

The dataset is unbalanced and is therefore incomplete. Below shows putting the dataset into R:



```
# Question 3 - unbalanced dataset
score <- c(67, 66, 78, 76, 71, 69, 72,
           63, 73, 62, 61, 69, 63, 71, 68, 56,
           69,58,54,63,64,55,59, 46,49,
           30, 47,39, 33)

sex <- c(rep("Females", times = 7),
        rep("Males", times = 9),
        rep("Females", times = 6),
        rep("Males", times = 3),
        rep(c("Females", "Males"), each = 2))

ethnicity <- c(rep("E1", 16), rep("E2", 9), rep("E3", 4))
cbind(score,sex,ethnicity)
```

Description: Children in a class are given a test, they are from 3 different ethnic groups. The teacher is interested in whether there are sex differences after allowing for ethnicity. Ethnicity the most important factor in this test, therefore it will be entered first into the dataset.

Model equation:
$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$

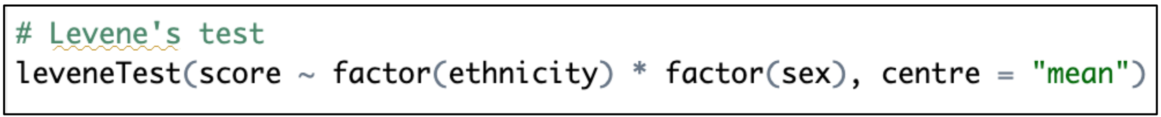
Y = the response variable being the score children get on the English test
i = sex (female or male)
j = ethnic group (E1, E2, or, E3)
k = the observations within i & j groups
μ = the overall mean
Y_{ijk} = response variable Y (weight) at level i (sex)
α_i = the random effect of sex i
β_j = the random effect of ethnic group j
$(\alpha\beta)_{ij}$ = the random effects of sex i and ethnic group j multiplied

E_{ijk} = error for j (ethnic group) at i (sex)

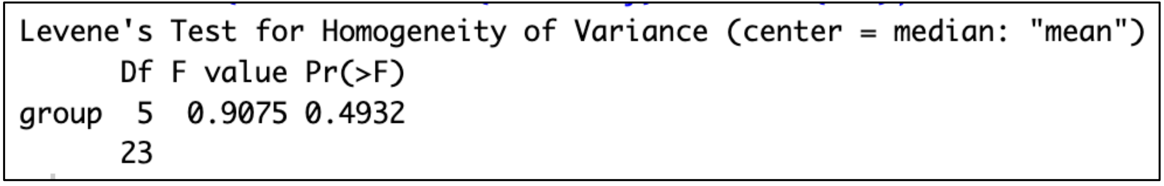
The interaction term is $(\alpha\beta)_{ij}$, corresponds to the interaction between i (sex) and j (ethnic group).

Assumptions:

Levene's test -



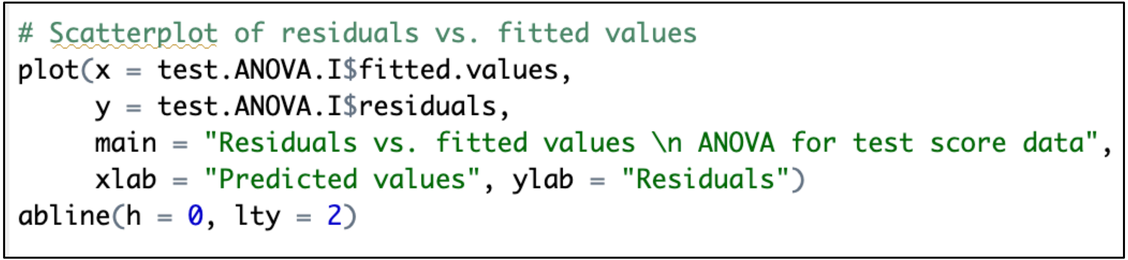
```
# Levene's test
leveneTest(score ~ factor(ethnicity) * factor(sex), centre = "mean")
```



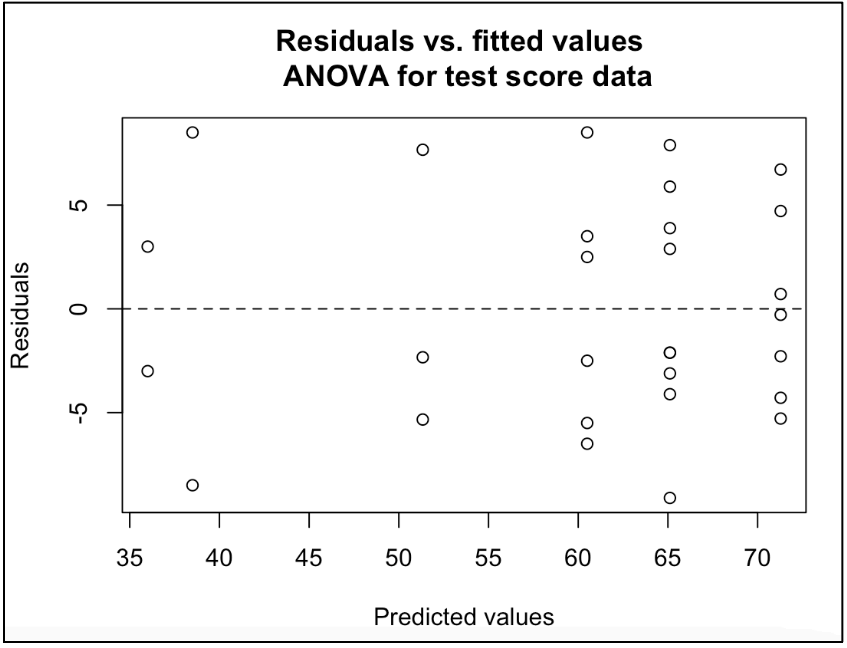
```
Levene's Test for Homogeneity of Variance (center = median: "mean")
      Df F value Pr(>F)
group  5  0.9075 0.4932
      23
```

The Levene's test produces a p-value of $p = 0.4932$, which is above the 5% significance level, suggesting that there are no issues with the assumption of constant variance ($p = 0.4932 > 0.05$).

Scatterplot of residuals vs. fitted values

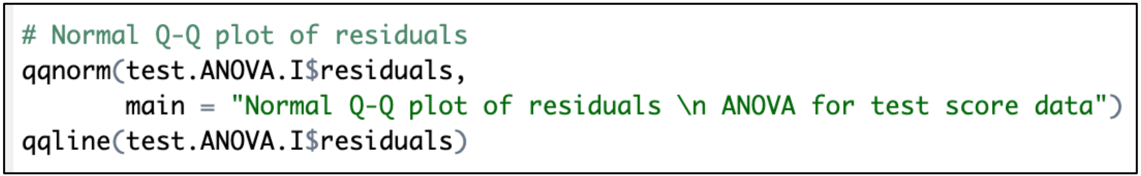


```
# Scatterplot of residuals vs. fitted values
plot(x = test.ANOVA.I$fitted.values,
     y = test.ANOVA.I$residuals,
     main = "Residuals vs. fitted values \n ANOVA for test score data",
     xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```

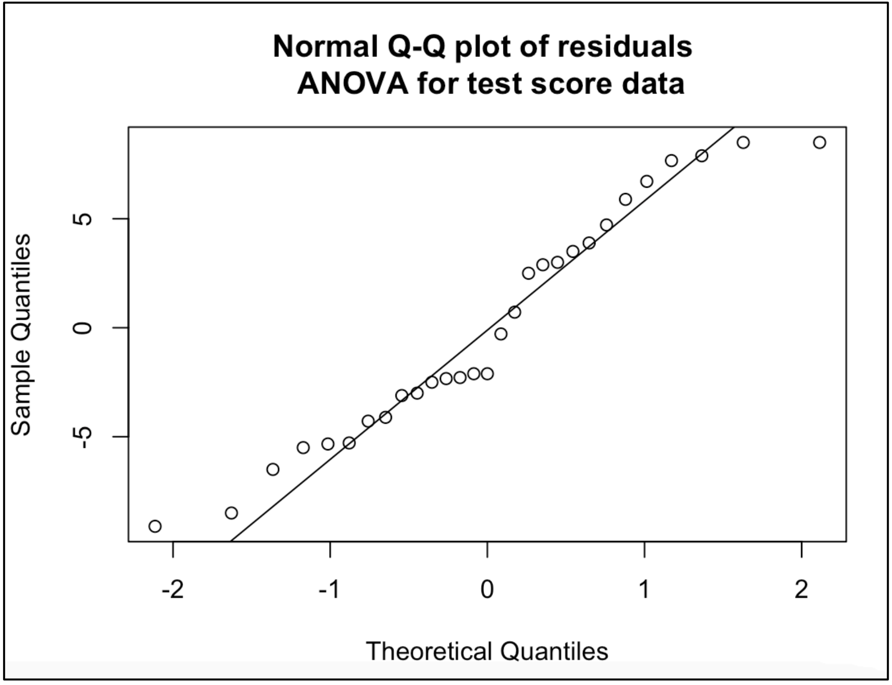


The plot of residuals vs. predicted values confirm constant variance like the Levene's test does. The variance looks relatively fine so constant variance for this data is satisfied.

Normal Q-Q plot



```
# Normal Q-Q plot of residuals
qqnorm(test.ANOVA.I$residuals,
       main = "Normal Q-Q plot of residuals \n ANOVA for test score data")
qqline(test.ANOVA.I$residuals)
```

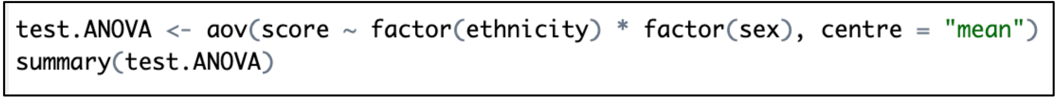


All data points are relatively close to the normal Q-Q line, so normality for this data can be assumed satisfied.

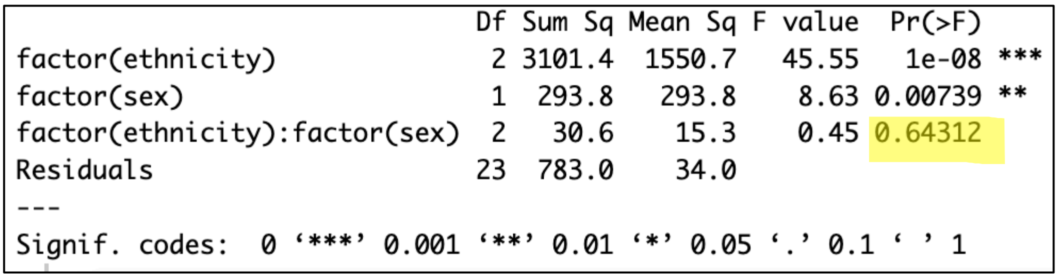
The assumptions of constant variance and normality are satisfied for this data in the Levene's test, scatterplot, and normal Q-Q plot. Independence cannot be confirmed through any tests, so we have to assume that independence was controlled when collecting the data. An ANOVA on this data is valid because the diagnostic graphs are acceptable, thus all assumptions seem satisfied.

Hypotheses:
H_0: There is no interaction between sex and ethnic group, all $(\alpha\beta)_{ij} = 0$
H_1: There is some interaction between sex and ethnic group, at least $(\alpha\beta)_{ij} \neq 0$

ANOVA table:
Type I ANOVA table

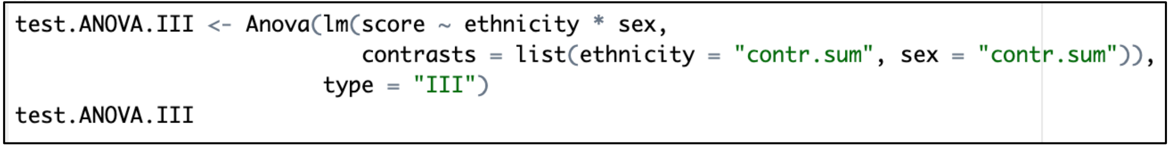


```
test.ANOVA <- aov(score ~ factor(ethnicity) * factor(sex), centre = "mean")
summary(test.ANOVA)
```

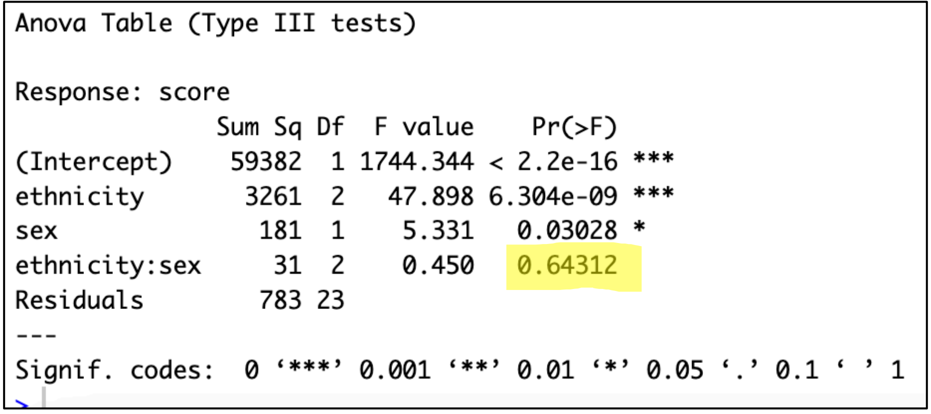


```
                               Df Sum Sq Mean Sq F value  Pr(>F)    
factor(ethnicity)               2 3101.4  1550.7   45.55   1e-08 ***
factor(sex)                     1  293.8   293.8    8.63 0.00739 ** 
factor(ethnicity):factor(sex)   2   30.6    15.3    0.45 0.64312    
Residuals                      23  783.0    34.0                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Type III ANOVA table



```
test.ANOVA.III <- Anova(lm(score ~ ethnicity * sex,
                           contrasts = list(ethnicity = "contr.sum", sex = "contr.sum")),
                        type = "III")
test.ANOVA.III
```

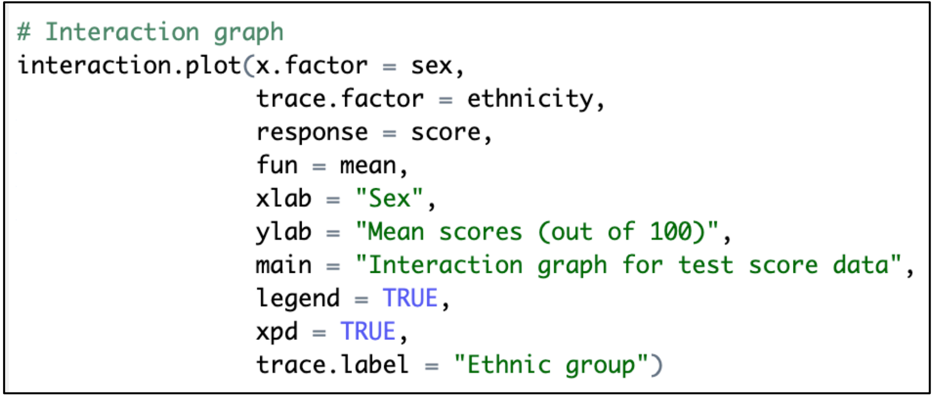


```
Anova Table (Type III tests)

Response: score
            Sum Sq Df  F value    Pr(>F)    
(Intercept)  59382  1 1744.344 < 2.2e-16 ***
ethnicity     3261  2   47.898 6.304e-09 ***
sex            181  1    5.331   0.03028 *  
ethnicity:sex   31  2    0.450   0.64312    
Residuals      783 23                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

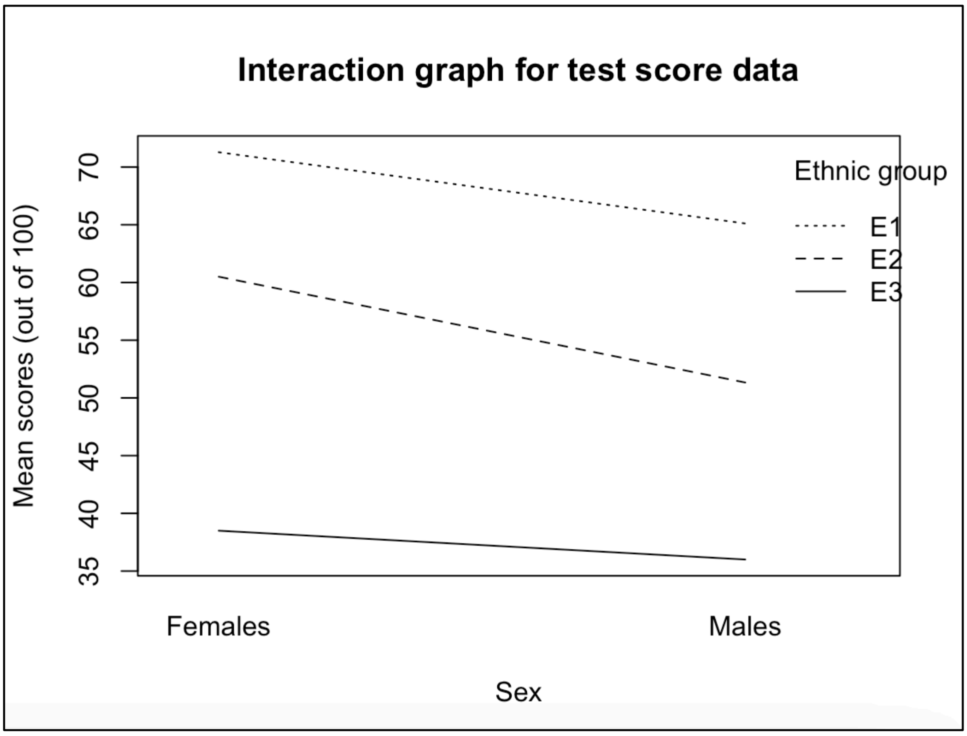
Statistical conclusion:

In both the Type I and Type III ANOVA tests, the A x B interaction is not significant at a 5% significance level because ($p = 0.643 > 0.05$). This means that we fail to reject H_0. This means that we proceed with main effects. Both effects, ethnicity and sex, are significant at the 5% and the 1% level, ($p = 1 \times 10^{-8}$ and $p = 0.00739$) using the Type I ANOVA table output.

Interaction graph



```
# Interaction graph
interaction.plot(x.factor = sex,
                 trace.factor = ethnicity,
                 response = score,
                 fun = mean,
                 xlab = "Sex",
                 ylab = "Mean scores (out of 100)",
                 main = "Interaction graph for test score data",
                 legend = TRUE,
                 xpd = TRUE,
                 trace.label = "Ethnic group")
```



The lack of interaction between ethnic group and sex is evident on the interaction graph, as the trace lines (ethnic group lines) are all near parallel with one another. The main effect seen is the vertical separation of the 3 lines. In addition, mean test scores are higher for females than males.

Interpretation:
There is no interaction between ethnic group and sex, which means these factors together do not have an influence on children's test scores out of 100.

b) If a one-way ANOVA is done with factor Sex, the resulting ANOVA table is given below. Explain (relatively briefly) the discrepancy between the outcome of this test and the test for Sex in part (a).

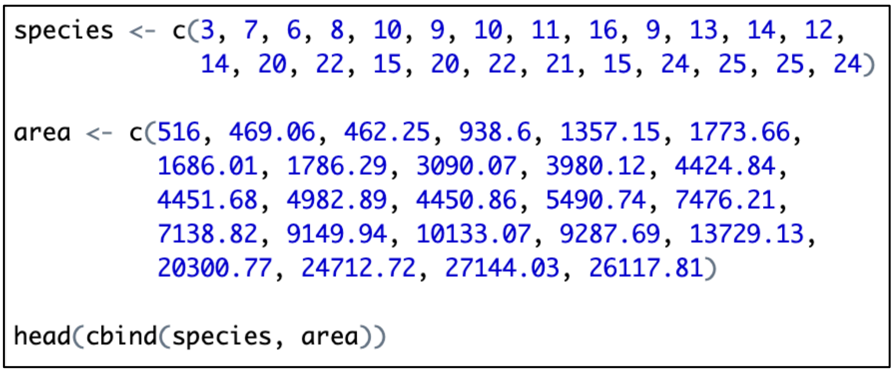
The one-way ANOVA shows that there is not a difference in means for sex between males and females test score. This is because (p = 0.321 > 0.05), so there isn't enough to reject the null hypothesis being that there is no difference in means over the two levels (male and female). In part a), sex had a significant effect at the 5% and 1% level, where (p = 0.00739 < 0.05 & 0.01). Whereas this one-way ANOVA table shows no significance between sex on test scores. The discrepancy between the outcomes of each of these tests can be explained as a one-way ANOVA is comparing means between male and females, whereas the two-way ANOVA was identifying whether there is an interaction between sex and ethnic group.

Question 4

Simple linear regression

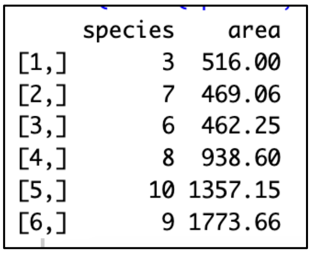
Putting the data into R:

STAT292
Isabelle (Izzy) Southon
300597453
Assignment 3



```
species <- c(3, 7, 6, 8, 10, 9, 10, 11, 16, 9, 13, 14, 12,
             14, 20, 22, 15, 20, 22, 21, 15, 24, 25, 25, 24)

area <- c(516, 469.06, 462.25, 938.6, 1357.15, 1773.66,
          1686.01, 1786.29, 3090.07, 3980.12, 4424.84,
          4451.68, 4982.89, 4450.86, 5490.74, 7476.21,
          7138.82, 9149.94, 10133.07, 9287.69, 13729.13,
          20300.77, 24712.72, 27144.03, 26117.81)

head(cbind(species, area))
```

Output:



```
     species    area
[1,]       3  516.00
[2,]       7  469.06
[3,]       6  462.25
[4,]       8  938.60
[5,]      10 1357.15
[6,]       9 1773.66
```

Description:

This simple linear regression will investigate whether it is an appropriate model in seeing if there is a relationship between the number of different species of macroinvertebrates found in mussel clumps (response variable Y) and the area (dm^2) of the mussel clumps (explanatory variable x).

Model equation:

The theoretical model equation is $Y = \beta 0 + \beta 1\text{x} + E$, where Y = the number of different species of macroinvertebrate founds in mussel clumps and x = the area of the mussel clumps. Y (number of species) and E (error term) are both random variables, and x (area of mussel clumps) is a quantity assumed without error. $\beta 0$ and $\beta 1$ are unknown parameters. $\beta 0$ is the intercept and $\beta 1$ is the slope. E is the deviations in expected values from the mean.

The fitted model is the line $\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1}x$

The fitted model line for untransformed area is: $\widehat{Y}$ = 9.856 + 0.0006593x

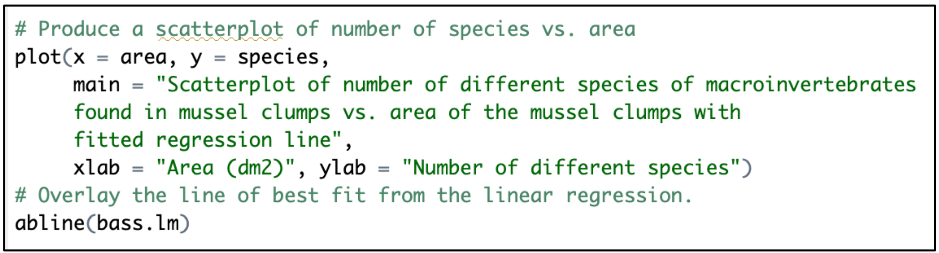
The fitted model line for the transformed area is: $\widehat{Y}$ = -25.7308 + 4.8756x

Assumptions:

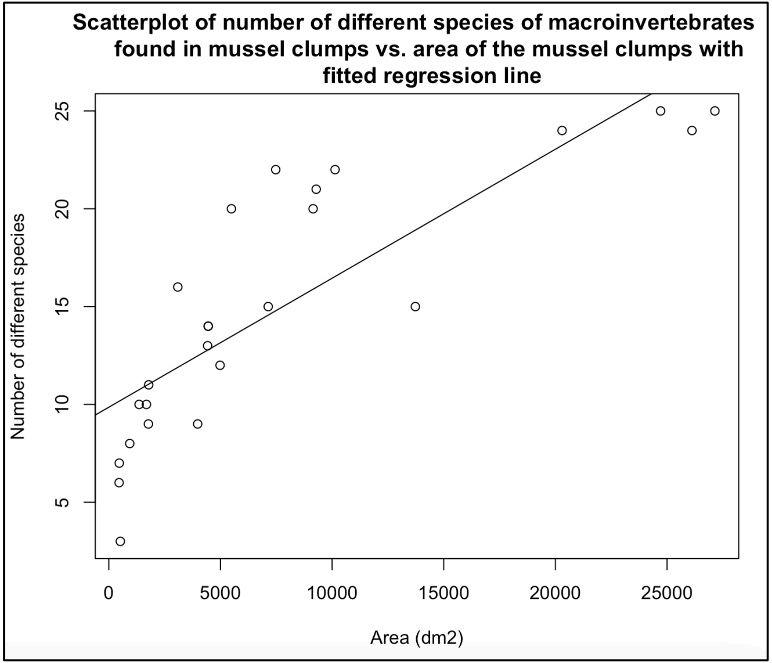
- Normality, the errors (E) come independently from a N(0,σ2) distribution, and they are independent of the area of mussel clumps (x).
- Constant variance
- Independence – the data was randomly sampled, and the area of mussel clumps (x) are independent from each other
- Linearity

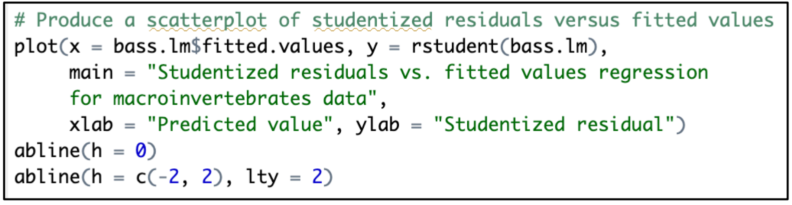
The assumptions of constant variance and normality will be checked below using Area and log(Area).

Area

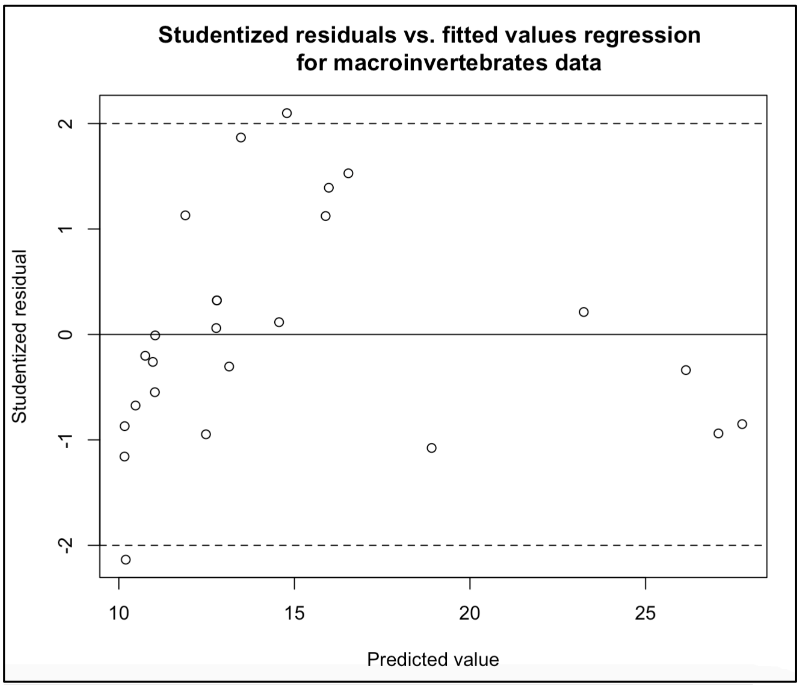


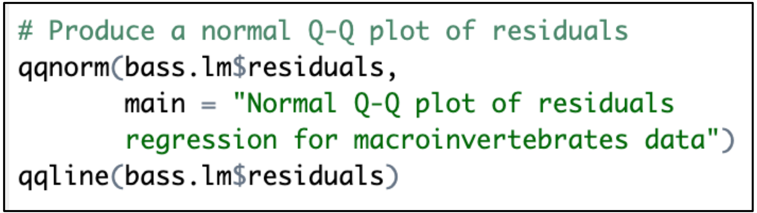
```
# Produce a scatterplot of number of species vs. area
plot(x = area, y = species,
     main = "Scatterplot of number of different species of macroinvertebrates
     found in mussel clumps vs. area of the mussel clumps with
     fitted regression line",
     xlab = "Area (dm2)", ylab = "Number of different species")
# Overlay the line of best fit from the linear regression.
abline(bass.lm)
```



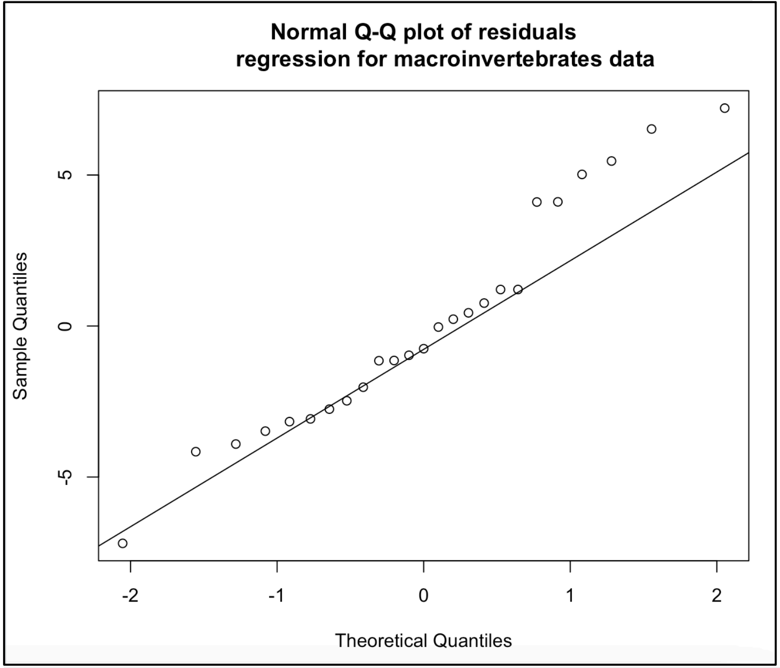


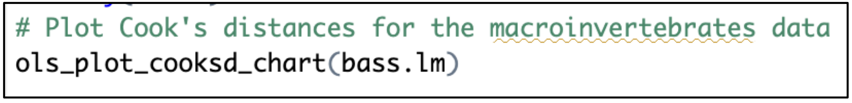
```
# Produce a scatterplot of studentized residuals versus fitted values
plot(x = bass.lm$fitted.values, y = rstudent(bass.lm),
     main = "Studentized residuals vs. fitted values regression
     for macroinvertebrates data",
     xlab = "Predicted value", ylab = "Studentized residual")
abline(h = 0)
abline(h = c(-2, 2), lty = 2)
```



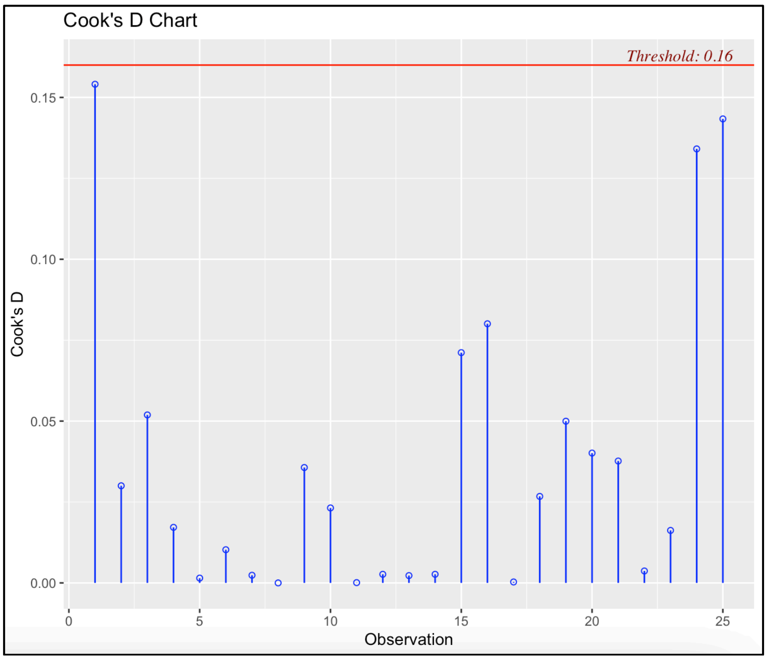


```
# Produce a normal Q-Q plot of residuals
qqnorm(bass.lm$residuals,
       main = "Normal Q-Q plot of residuals
       regression for macroinvertebrates data")
qqline(bass.lm$residuals)
```

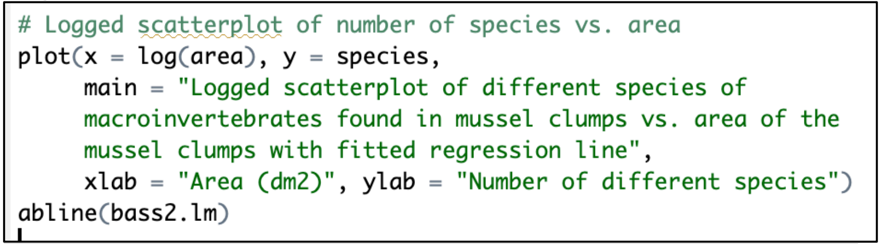




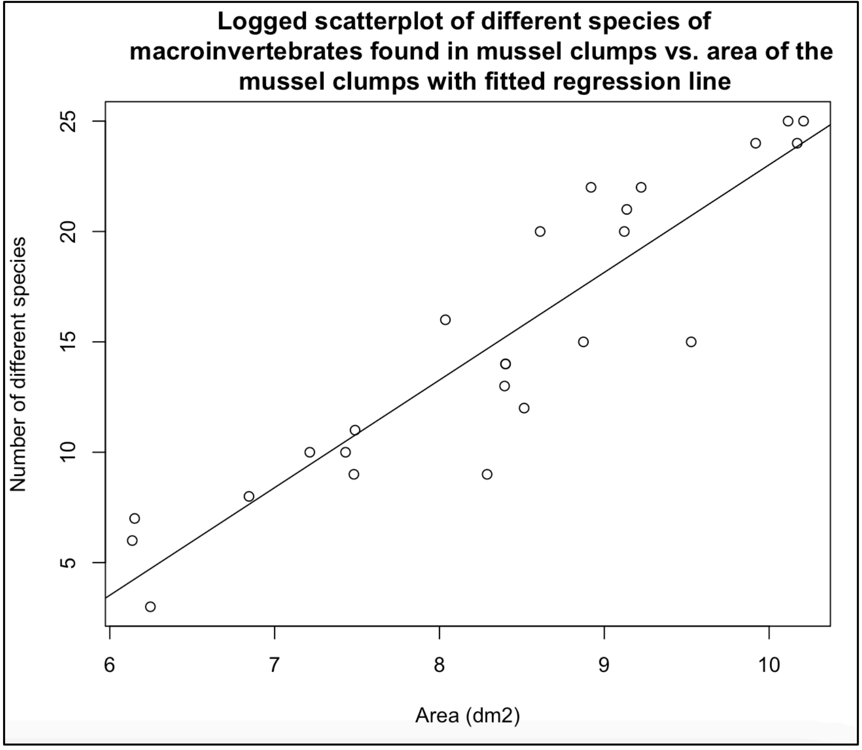
```
# Plot Cook's distances for the macroinvertebrates data
ols_plot_cooksd_chart(bass.lm)
```

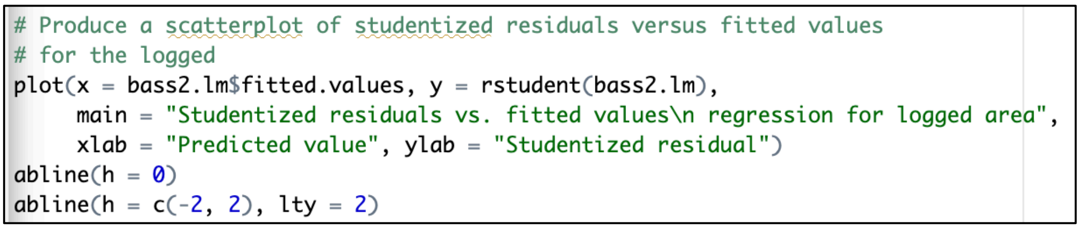


Log(Area)

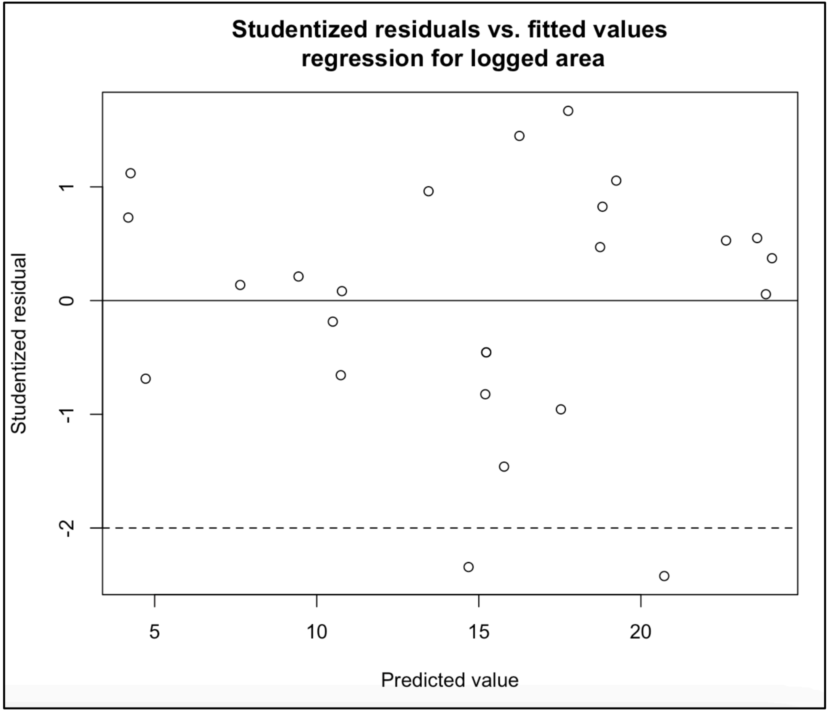


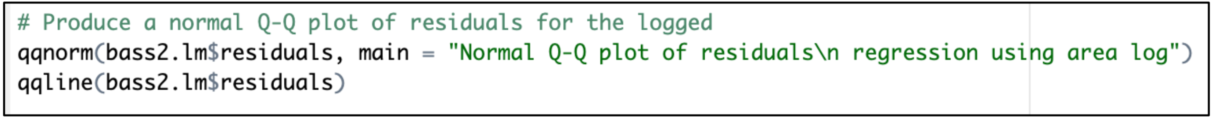
```
# Logged scatterplot of number of species vs. area
plot(x = log(area), y = species,
     main = "Logged scatterplot of different species of
     macroinvertebrates found in mussel clumps vs. area of the
     mussel clumps with fitted regression line",
     xlab = "Area (dm2)", ylab = "Number of different species")
abline(bass2.lm)
```



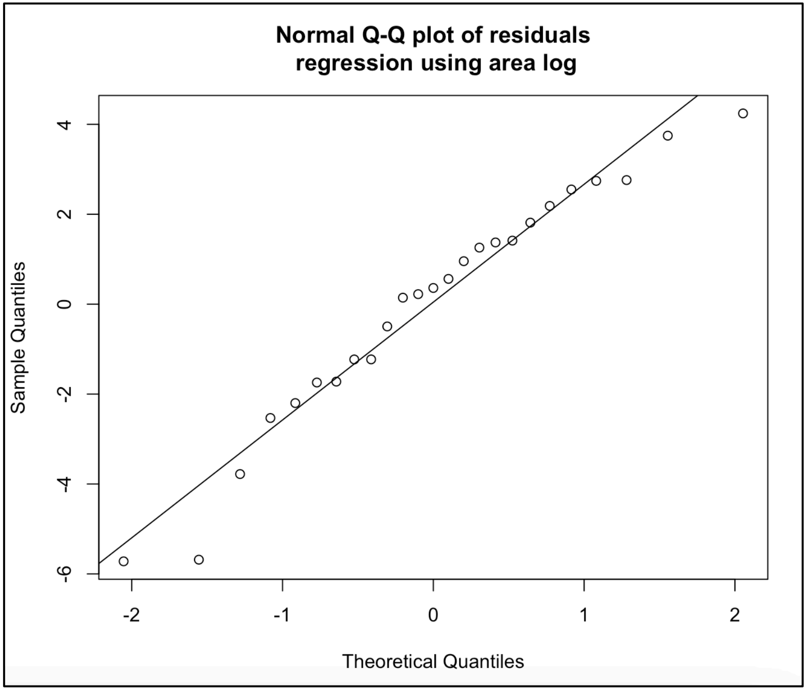


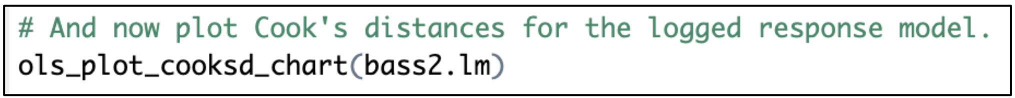
```
# Produce a scatterplot of studentized residuals versus fitted values
# for the logged
plot(x = bass2.lm$fitted.values, y = rstudent(bass2.lm),
     main = "Studentized residuals vs. fitted values\n regression for logged area",
     xlab = "Predicted value", ylab = "Studentized residual")
abline(h = 0)
abline(h = c(-2, 2), lty = 2)
```



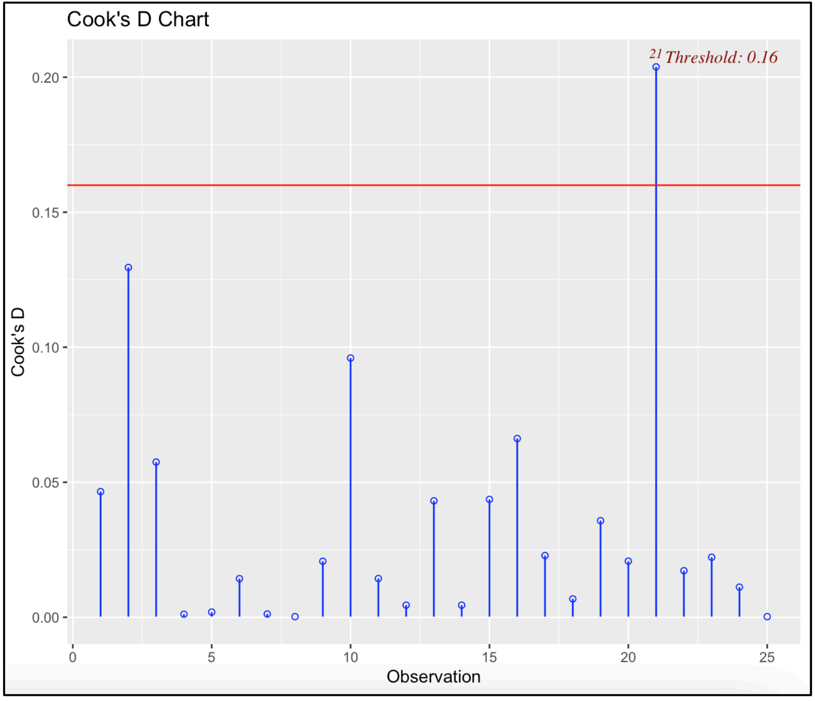


```
# Produce a normal Q-Q plot of residuals for the logged
qqnorm(bass2.lm$residuals, main = "Normal Q-Q plot of residuals\n regression using area log")
qqline(bass2.lm$residuals)
```





```
# And now plot Cook's distances for the logged response model.
ols_plot_cooksd_chart(bass2.lm)
```



Comments on whether the analysis and assumptions are valid

1) For the scatterplots of number of species and area of mussel clumps, the data with the logged area fits the regression line better than untransformed area does. For the scatterplot using the untransformed area, the data points are quite far away from the regression line and the line of best fit is quite high. Typically, the closer the data points are to the regression line, the stronger the relationship there is between Y and x (number of species and area of mussel clumps). Therefore, out of these two models the log-transformed area is more appropriate, as the points are closer to the line of best fit when compared to the scatterplot with the untransformed area. This means that linearity is satisfied when using the logged area.

2) The log transformed area produced a better satisfied studentized residuals vs. fitted values regression plot than the untransformed area did. Constant variance is violated for data using untransformed area because the data points aren't plotted evenly around the zero line and the data is not symmetrically distributed around the fitted line. The variance of residuals seems to decrease with the area of mussel clumps (x). This means that constant variance has been violated with the untransformed area. However, the for the logged area, the studentized residuals vs. fitted values plot looks valid because the points are evenly distributed around the fitted line, therefore constant variance is satisfied.

3) The logged area for the Q-Q plots is better than the untransformed area. This is because the points are closer to the Q-Q line and therefore normality is satisfised using the logged area.

4) For both the logged area and untransformed area, the Cook's distances are well below 1. This means there are no outliers or points of high leverage that have been detected. This illustrates how there are no influential points that have been detected, and therefore the line of best fit has not been affected.

Using the logged area, the assumptions of constant variance, normality and linearity are satisfied, and therefore the simple linear regression model is appropriate for this data.

Hypotheses:

$H_0 : \beta 1 = 0$

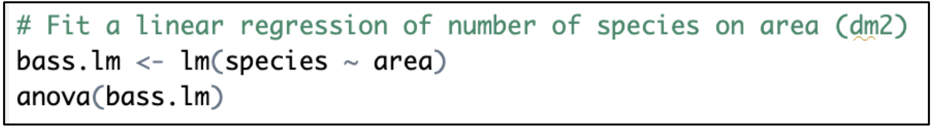
There is no linear relationship between the number of different species of macroinvertebrates and the area of the mussel clumps.

$H_1 : \beta 1 \neq 0$

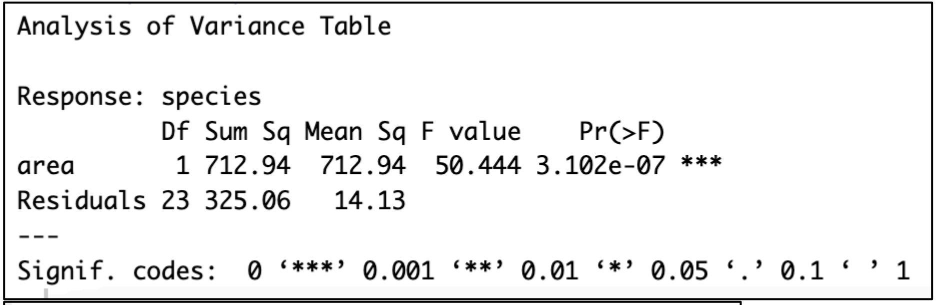
There is a linear relationship between the number of different species of macroinvertebrates and the area of the mussel clumps.

ANOVA tables:

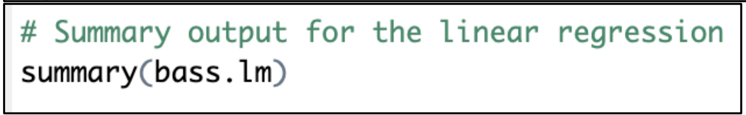
Linear regression (untransformed area)



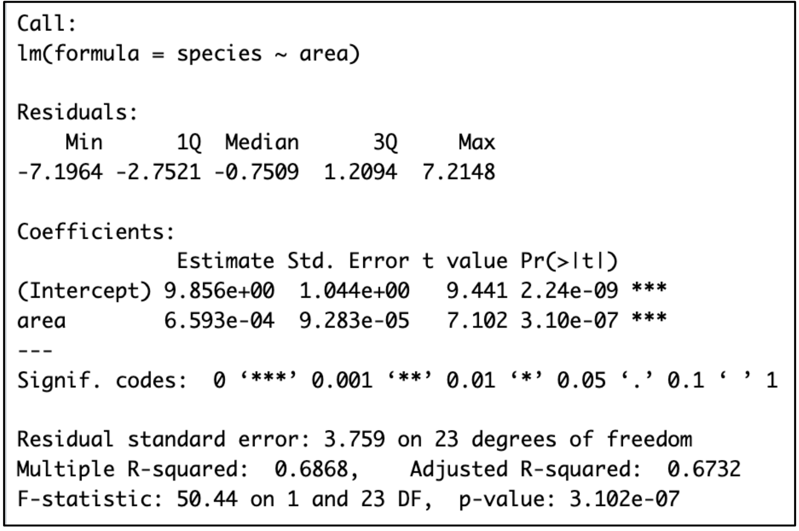
```
# Fit a linear regression of number of species on area (dm2)
bass.lm <- lm(species ~ area)
anova(bass.lm)
```



```
Analysis of Variance Table

Response: species
          Df Sum Sq Mean Sq F value    Pr(>F)    
area       1 712.94  712.94  50.444 3.102e-07 ***
Residuals 23 325.06   14.13                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

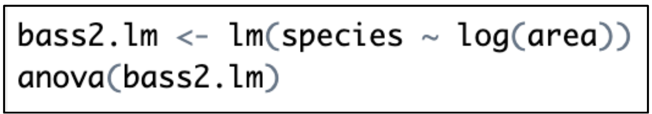


```
# Summary output for the linear regression
summary(bass.lm)
```

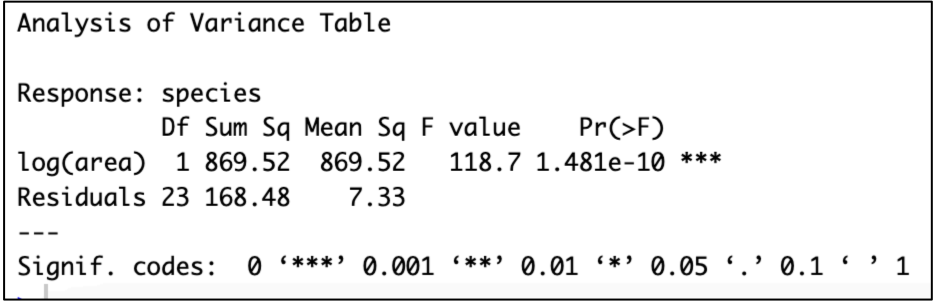


```
Call:
lm(formula = species ~ area)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1964 -2.7521 -0.7509  1.2094  7.2148 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.856e+00  1.044e+00   9.441 2.24e-09 ***
area        6.593e-04  9.283e-05   7.102 3.10e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.759 on 23 degrees of freedom
Multiple R-squared:  0.6868,	Adjusted R-squared:  0.6732 
F-statistic: 50.44 on 1 and 23 DF,  p-value: 3.102e-07
```

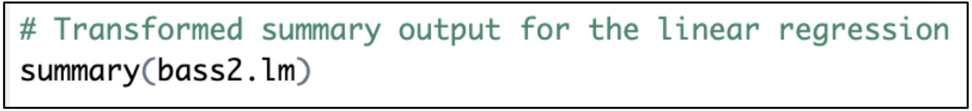
Linear regression (log area)



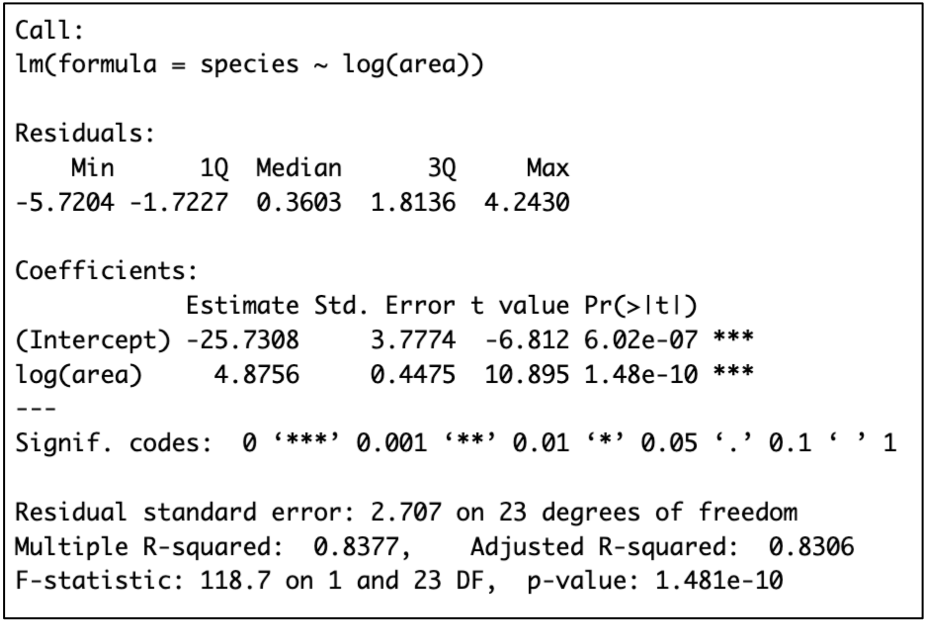
```
bass2.lm <- lm(species ~ log(area))
anova(bass2.lm)
```



```
Analysis of Variance Table

Response: species
          Df Sum Sq Mean Sq F value    Pr(>F)    
log(area)  1 869.52  869.52   118.7 1.481e-10 ***
Residuals 23 168.48    7.33                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
# Transformed summary output for the linear regression
summary(bass2.lm)
```



```
Call:
lm(formula = species ~ log(area))

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7204 -1.7227  0.3603  1.8136  4.2430 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -25.7308     3.7774  -6.812 6.02e-07 ***
log(area)     4.8756     0.4475  10.895 1.48e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.707 on 23 degrees of freedom
Multiple R-squared:  0.8377,	Adjusted R-squared:  0.8306 
F-statistic: 118.7 on 1 and 23 DF,  p-value: 1.481e-10
```

Statistical conclusions:

Using the logged area, the *F* statistic is 118.7 on (1, 23) df, while the *t* statistic is 10.895 on 23 df. The df for *t* are determined by the estimate of σ^2, from the MSE in the ANOVA Table, which has 23 df. The *F* statistic is equal to the t statistic squared, since there is only one numerator df; any minor difference is due to rounding. The *p*-values are the same from either statistic, $p = 1.481 \text{ x } 10^{-10}$, so we reject H_0 at the 5% and 1% significance levels as ($p = 1.481 \text{ x } 10^{-10} < 0.05$) and ($p = 1.481 \text{ x } 10^{-10} < 0.01$). These findings show that there is a linear relationship between the number of different species of macroinvertebrates and the area of the mussel clumps.

Interpretation:

From our statistical findings we can say that the area (in dm^2) of the mussel clumps is a useful predictor for the number of different species of macroinvertebrates found in mussel clumps. This investigation also found that it is appropriate to use simple linear regression to model the relationship between the number of different species of macroinvertebrates found in mussel clumps and the area (in dm2) of those mussel clumps.