Assignment 5 - Robert Ivill 46012819

2023-05-11

Question 1: An experiment was performed to investigate the effect of three fertilizers on the size of tomatoes. Six plants were treated with one of the three fertilizers. At the end of the experiment three randomly chosen tomatoes were collected from each plant and weighed.

a. Read in the data and use str function to see the structure of the data. The variables should be: Weight, Fertilizer, Plant.

Using the code below, we can read in our data and confirm that the structure of the data is correct.

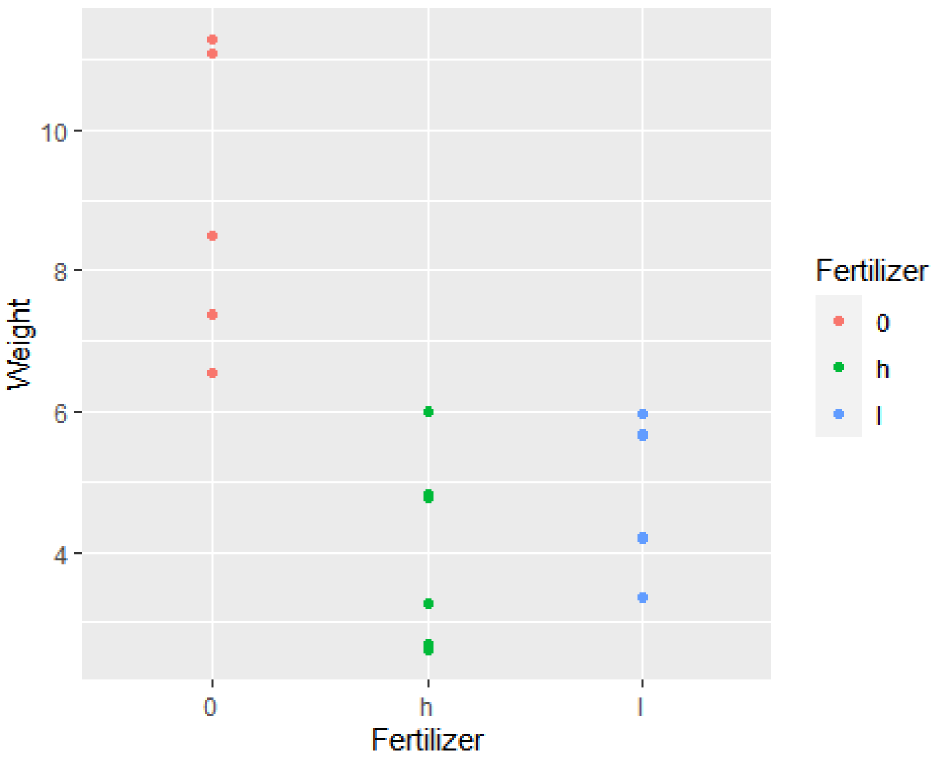
```
tomatoes_df <- read.csv("/Users/robbi/Downloads/tomatoes.csv")
str(tomatoes_df)
```

```
## 'data.frame':    18 obs. of  4 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Weight    : num  11.29 11.08 11.1 7.37 6.55 ...
##  $ Fertilizer: chr  "0" "0" "0" "0" ...
##  $ Plant     : int  1 1 1 2 2 2 3 3 3 4 ...
```

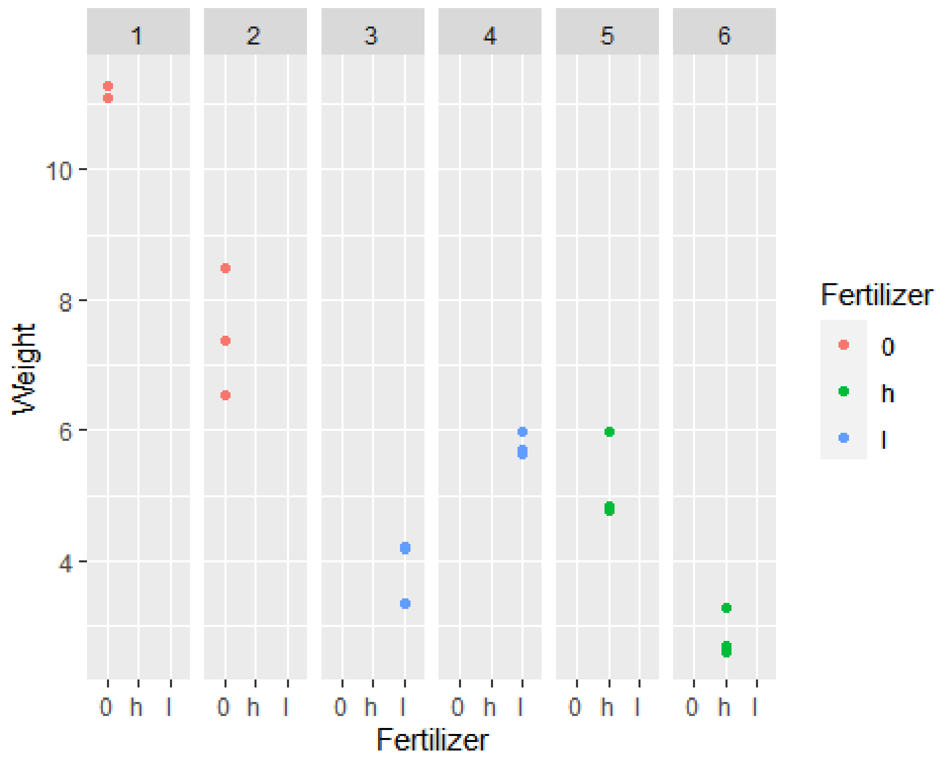
We can see that the structure of the data is correct, as the variables ‘Weight’, ‘Fertilizer’ and ‘Plant’ are all present in our data frame.

b. Explore the data using the graphs below. What do these graphs tell you?

```
library(ggplot2)
ggplot(tomatoes_df, aes(x=Fertilizer, y=Weight, color=Fertilizer)) +
geom_point()
```



```
ggplot(tomatoes_df, aes(x=Fertilizer, y=Weight, color=Fertilizer)) +
geom_point() +
facet_grid(~Plant)
```



The first graph is a scatterplot that shows the distribution of weights of tomatoes grown in different fertilizers. The second graph adds the facet_grid function for the plant variable so we can see the distribution of weights with each of the plants being separated to their own scatterplot

c. Formulate the research question.

The research question for this dataset could be 'Does the type of fertilizer that a tomato plant grows in affect the weight of the tomatoes?'

d. A linear mixed-effects model is to be used to answer the research question. Which of the variables are response variable, explanatory component (fixed effects), and structural component (random effects)?

The response variable is the weight of the tomato, as it is what is being measured to answer our question. The explanatory component is the type of fertilizer that the tomatoes are being grown in as it what we are observing the weights of the tomatoes in. The structural component is each tomato plant as they represent a group of tomatoes within the data.

e. Convert all categorical variables to factors, fit two linear mixed models using lmer function from lme4 package. Use anova function to test for the significance of the fixed effects. Are the fixed effects significant? Use the following code replacing V, response, fixed_effect, and random_effect with appropriate variables:

In the code below, V and fixed_effect was replaced with Fertilizer, response was replaced with Weight and random_effect was replaced with Plant.

```
#install.packages("lme4")
library(lme4)

## Warning: package 'lme4' was built under R version 4.2.3

## Loading required package: Matrix

tomatoes_df$Fertilizer <- as.factor(tomatoes_df$Fertilizer)
tomatoes_lmm0 <- lmer(Weight ~ (1|Plant), data=tomatoes_df)
tomatoes_lmm1 <- lmer(Weight ~ Fertilizer + (1|Plant), data=tomatoes_df)
anova(tomatoes_lmm0,tomatoes_lmm1)

## refitting model(s) with ML (instead of REML)

## Data: tomatoes_df
## Models:
## tomatoes_lmm0: Weight ~ (1 | Plant)
## tomatoes_lmm1: Weight ~ Fertilizer + (1 | Plant)
##                npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tomatoes_lmm0    3 61.537 64.208 -27.769   55.537
## tomatoes_lmm1    5 57.394 61.846 -23.697   47.394 8.1434  2    0.01705 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output for the anova function can be used to test the significance of the fixed effects. The p-value for the fixed effect of fertilizer in the second model is 0.01705, which is less than 0.05. This shows that the type of fertilizer used most likely have a significant effect on the weight of the tomatoes.

f. Use lmerTest and ranova function from lmerTest library to test for the significance of the random effects. Are the random effects significant?

Using the code below we can test the signifance of the random effects (the plant variable).

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

## Warning: package 'lmerTest' was built under R version 4.2.3

##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step

ranova(tomatoes_lmm1)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## Weight ~ Fertilizer + (1 | Plant)
##               npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>           5 -20.579 51.158
## (1 | Plant)      4 -30.827 69.654 20.496  1  5.975e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value for plant is much less than 0.05 (5.975e-6). This indicates that the random effect of plant is significant to our research question.

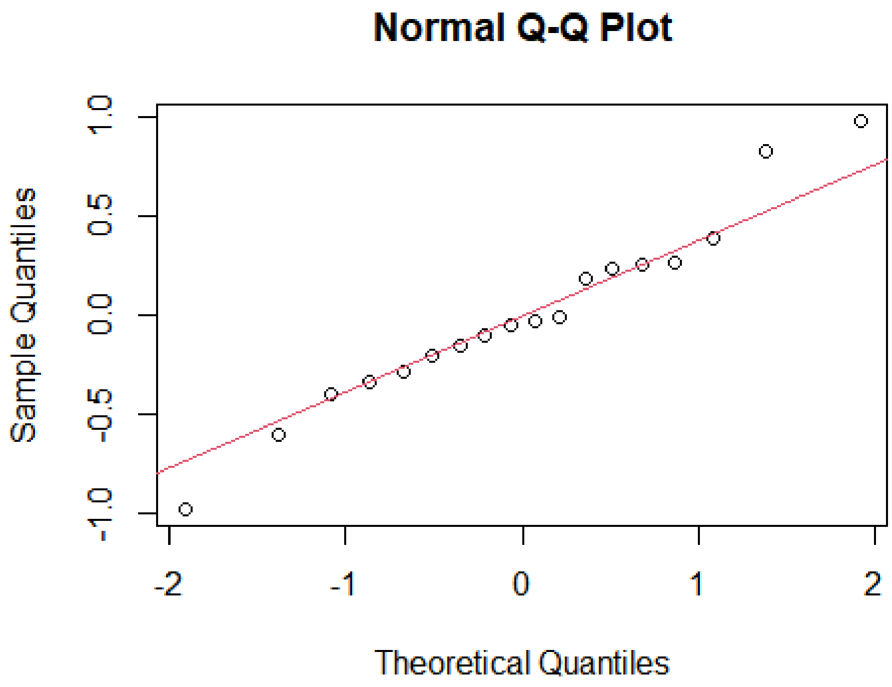
g. Based on e and f, answer the research question.

From our answers to e and f, we can answer the research question with the following statement: Both the explanatory component of the type of fertilizer and the random effect of the plant the tomato is apart of have an effect on the weights of tomatoes.

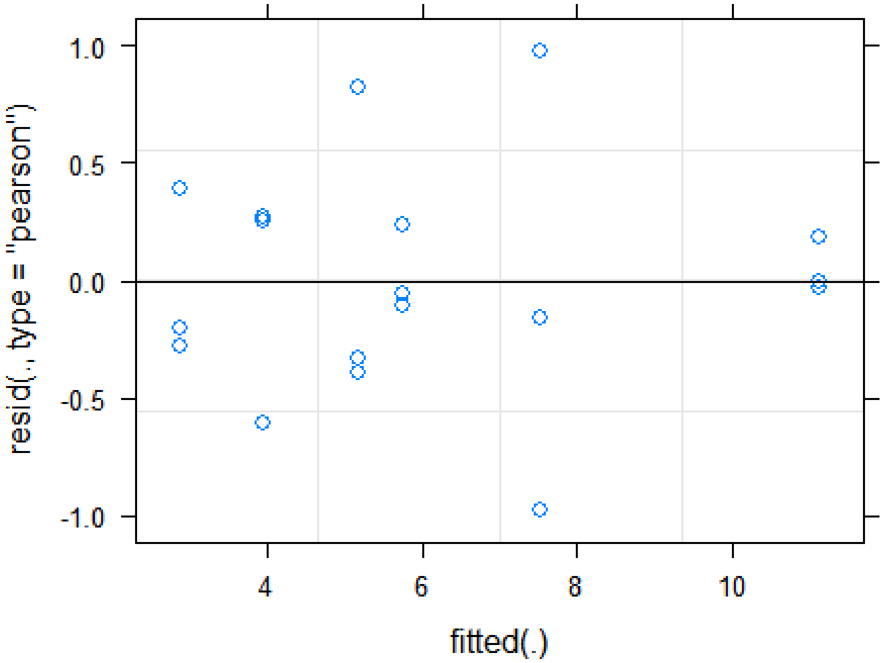
h. Check the assumptions of the model using the code below. Are the assumptions met?

To check the assumptions of the model, we create a Q-Q plot of the residuals and a fitted versus residuals plot.

```
qqnorm(resid(tomatoes_lmm1))
qqline(resid(tomatoes_lmm1), col=2)
```



```
plot(tomatoes_lmm1)
```



The Q-Q plot shows that the residuals approximately follow a normal distribution. The residuals versus fitted plot has no clear pattern or trend across it. This shows that our assumptions of homoscedasticity, linearity and normality are correct.

Question 2

To investigate whether darker frogs have lower skin microbial diversity, frogs of five different species were captured, their darkness was measured and skin microbiome tests were performed. Based on the tests the diversity score was recorded.

a. Read in the data and see the structure of the data using str function. The variable names are: Diversity, Darkness, and Species.

Using the code below, we can read in the data and see its structure to verify that it is correct.

```
frogs_df <- read.csv("/Users/robbi/Downloads/frogs.csv")
str(frogs_df)

## 'data.frame':    50 obs. of  4 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Diversity: num  2.53 2.75 3.16 2.17 3.07 ...
##  $ Darkness : num  0.987 0.995 1.586 1.879 1.03 ...
##  $ Species  : int  1 1 1 1 1 1 1 1 1 1 ...
```

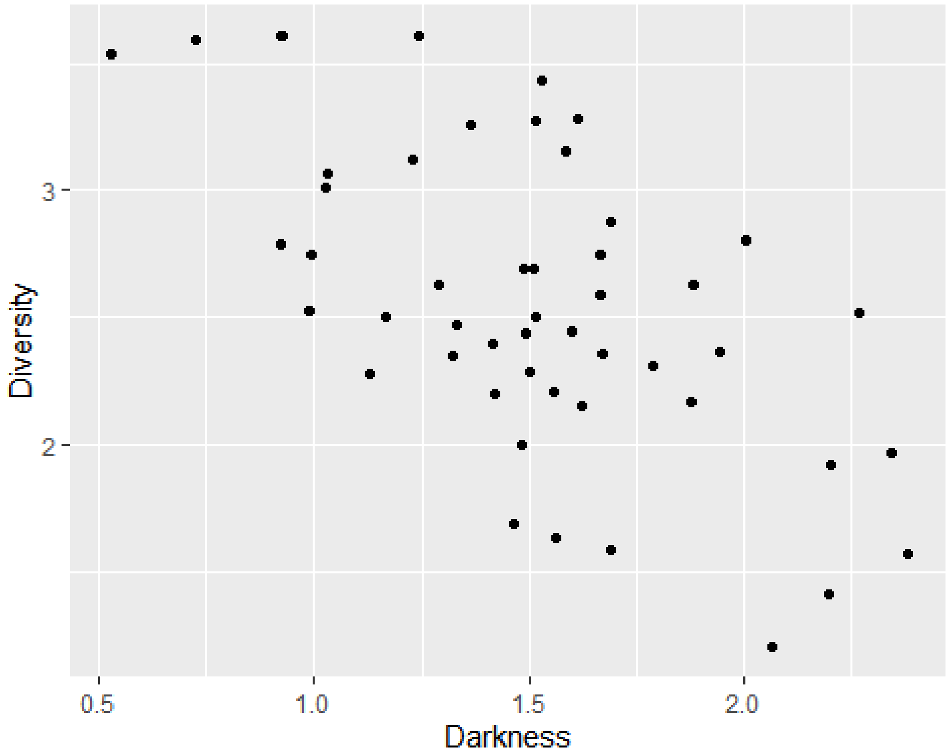
We see that the correct variables of Diversity, Darkness, Species are apart of the data frame.

b. Formulate the research question.

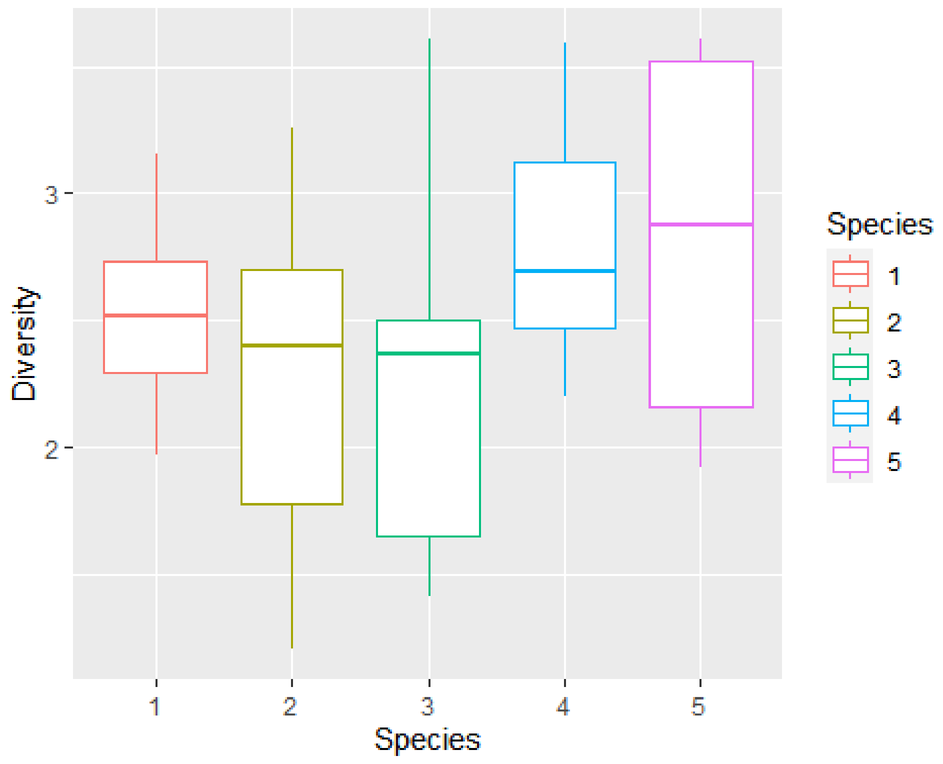
The research question for this can be: Are there any relationships between frog darkness and skin microbial diversity?

c. Convert any categorical variables to factors. Plot the data using the code below. Based on the plots, explain why you would use a linear mixed model for this analysis.

```
frogs_df$Species <- as.factor(frogs_df$Species)
library(ggplot2)
ggplot(frogs_df, aes(x = Darkness, y = Diversity)) +
geom_point()
```



```
ggplot(frogs_df, aes(x=Species, y=Diversity, color=Species))+geom_boxplot()
```



From these plots, we can conclude that a linear mixed model for the analysis is the best. This is because the boxplot shows differences in diversity for each of the species and the scatterplot shows there is variation in diversity based on the level of darkness. This indicates that there is possibly other factors that effect the diversity than just darkness.

d. Fit a linear mixed model. Answer the research question stated above using anova function to compare models with and without fixed effects. Use the code below replacing response, fixed_effect, and random_effect with appropriate variables:

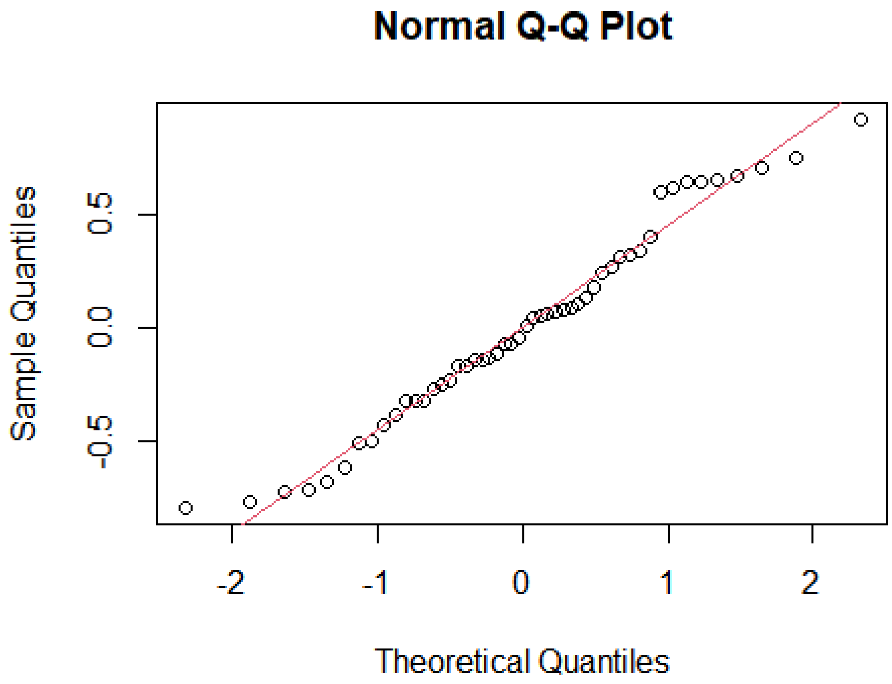
In the code below, we replaced the response variable with Diversity, the fixed_effect variable with Darkness, and the random_effect variable with Species.

```
mlm.frogs_df <- lmer(Diversity ~ Darkness + (1|Species), data = frogs_df)
mlm.frogs_df0 <- lmer(Diversity ~ 1 + (1|Species), data = frogs_df)
anova(mlm.frogs_df0, mlm.frogs_df)

## refitting model(s) with ML (instead of REML)

## Data: frogs_df
## Models:
## mlm.frogs_df0: Diversity ~ 1 + (1 | Species)
## mlm.frogs_df: Diversity ~ Darkness + (1 | Species)
##               npar    AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## mlm.frogs_df0    3 95.919 101.655 -44.96   89.919
## mlm.frogs_df     4 73.920  81.568 -32.96   65.920 23.999  1  9.638e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

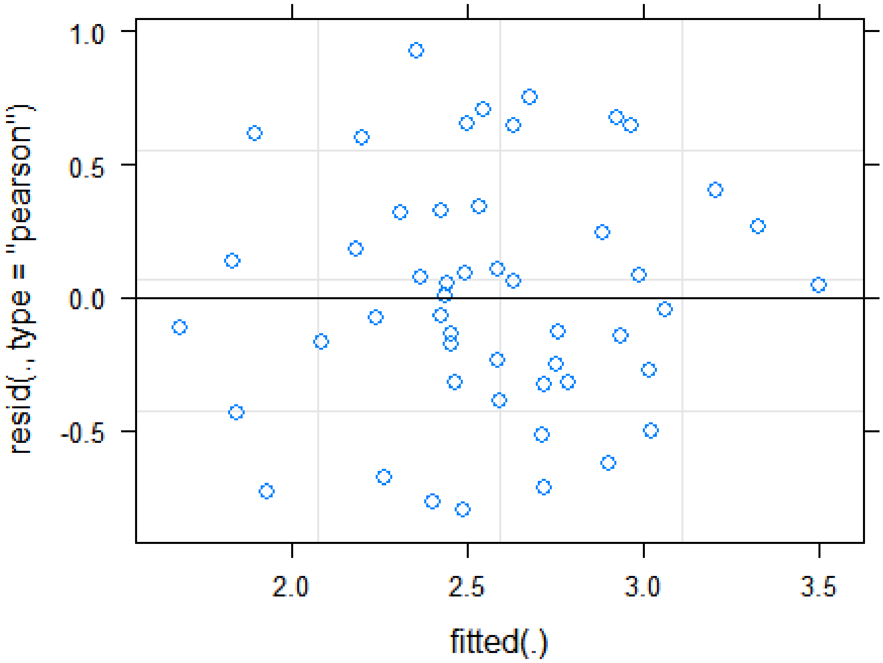
The linear mixed model shows that the p-value for the fixed effect of darkness is 9.639e-7, which is much less than 0.05. This indicates that the variable of darkness is highly significant to our model. This allows us to answer the research question: There is evidence to suggest that darker frogs do have lower skin microbiome diversity than lighter ones.

e. Check the assumptions of the model. Use the code below:

```
qqnorm(resid(mlm.frogs_df))
qqline(resid(mlm.frogs_df), col=2)
```



```
plot(mlm.frogs_df)
```



We can see from the Normal Q-Q plot that the residuals are roughly normally distributed. The residuals versus fitted plot shows no clear trends or patterns across the values. This shows the assumptions of normality, linearity and homoscedasticity are reasonable assumptions.